ΕΘΝΙΚΌ ΜΕΤΣΌΒΙΟ ΠΟΛΥΤΕΧΝΕΊΟ
ΣΧΟΛΉ ΗΛΕΚΤΡΟΛΌΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΏΝ
ΤΟΜΈΑΣ ΤΕΧΝΟΛΟΓΊΑΣ ΠΛΗΡΟΦΟΡΙΚΉΣ &
ΥΠΟΛΟΓΙΣΤΏΝ

Δημιουργία Πλατφόρμας Αξιολόγησης και Ανάλυσης Αλγοριθμικών Συστημάτων ΤΝ που είναι Συμβατά με το Ευρωπαϊκό και Ελληνικό Δίκαιο

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ευαγγελία Γιαννακοπούλου

Επιβλέπων : Παναγιώτης Τσανάκας
Καθηγητής Ε.Μ.Π

Αθήνα, Μάρτιος 2025

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής &
Υπολογιστών

Δημιουργία Πλατφόρμας Αξιολόγησης και Ανάλυσης Αλγοριθμικών Συστημάτων ΤΝ που είναι Συμβατά με το Ευρωπαϊκό και Ελληνικό Δίκαιο

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ευαγγελία Γιαννακοπούλου

Επιβλέπων : Παναγιώτης Τσανάκας
Καθηγητής Ε.Μ.Π

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 7η Μαρτίου 2025

.....	.....	.....
Παναγιώτης Τσανάκας	Ανδρέας-Γεώργιος Σταφυλοπάτης	Γεώργιος Αλεξανδρίδης
Καθηγητής Ε.Μ.Π	Ομότιμος Καθηγητής Ε.Μ.Π.	Διδάκτωρ Ε.Μ.Π,

Αθήνα, Μάρτιος 2025

.....
Ευαγγελία Γιαννακοπούλου

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Η τεχνητή νοημοσύνη (ΤΝ) αποτελεί πλέον αναπόσπαστο μέρος της σύγχρονης πραγματικότητας, επηρεάζοντας κρίσιμους τομείς όπως η εργασία, η άμυνα, η υγεία και η εκπαίδευση. Με την εισαγωγή του AI Act, το οποίο θέτει νομικές και ηθικές απαιτήσεις για την υπεύθυνη ανάπτυξη και χρήση των συστημάτων ΤΝ, δημιουργήθηκε η ανάγκη για εργαλεία αξιολόγησης που να εξασφαλίζουν τη συμμόρφωση με αυτές τις προδιαγραφές. Στόχος της παρούσας διπλωματικής εργασίας είναι η δημιουργία ενός δυναμικού ερωτηματολογίου για την εκτίμηση των επιπτώσεων της ΤΝ, το οποίο εναρμονίζεται με τις απαιτήσεις του AI Act. Για τον σκοπό αυτό, βασίζεται στη δομή του capAI ενώ αξιοποιεί την τεχνική προσέγγιση της καναδικής πλατφόρμας εκτίμησης αντικτύπου για την ανάλυση και την αντιμετώπιση των επιπτώσεων που σχετίζονται με την ανάπτυξη αυτοματοποιημένων συστημάτων λήψης αποφάσεων.

Η εργασία ξεκινά με την παρουσίαση της ΤΝ, των διαφορετικών μεθόδων ανάπτυξής της και των ηθικών προκλήσεων που προκύπτουν από την χρήση της , όπως η μεροληψία, η αδιαφάνεια, η αυτοματοποίηση της εργασίας και η απώλεια της ιδιωτικότητας. Στη συνέχεια, εξετάζεται το νομικό πλαίσιο που διέπει την ΤΝ, με έμφαση στη νομοθεσία της Ε.Ε., των ΗΠΑ της Ελλάδας και της Κίνας. Ύστερα αναλύονται μερικά εργαλεια αλγοριθμικής εκτίμησης αντικτύπου (ΑΕΑ) που χρησιμοποιούνται για την ανάλυση της συμμόρφωσης των συστημάτων ΤΝ.

Ακολούθως, συνδυάζοντας περιεχόμενο και μεθοδολογίες από διάφορα άλλα ΑΕΑ, περιγράφεται η ανάπτυξη του ερωτηματολογίου και η υλοποίηση του. Το ερωτηματολόγιο αξιολογεί συστήματα ΤΝ με βάση πέντε βασικές μετρικές: ευημερία, μη κακοήθεια, δικαιοσύνη, αυτονομία και εξηγησιμότητα ενώ χωρίζεται με βάση τον κύκλο ανάπτυξης των ΤΝ σε πεντε βασικά στάδια : Σχεδιασμός, Ανάπτυξη, Αξιολόγηση, Λειτουργία και Απόσυρση.

Στη συνέχεια, πραγματοποιείται συγκριτική μελέτη με άλλα αντίστοιχα συστήματα, προκειμένου να βελτιστοποιηθεί η λειτουργικότητα και η ακρίβεια του εργαλείου.

Η εργασία ολοκληρώνεται με τα συμπεράσματα και τις προτάσεις για μελλοντικές βελτιώσεις, οι οποίες περιλαμβάνουν την ενσωμάτωση αυτόματων ελέγχων, δυναμικότητας και την περαιτέρω βελτίωση της αξιολόγησης ηθικών και νομικών πτυχών των συστημάτων ΤΝ.

Λέξεις Κλειδιά

Τεχνητή Νοημοσύνη, Αλγοριθμική Εκτίμηση Αντικτύπου, capAI, AI Act, Δικαιοσύνη, Εξηγησιμότητα, Μεροληψία, Ηθική , Νομοθεσία, Διαφάνεια, Ανθρώπινα Δικαιώματα, Αυτόνομα Συστήματα, Explainability, AI, Traceability, Trustworthy AI, Justice, Fairness, Robustness, Testing

Abstract

Artificial intelligence (AI) is now an integral part of modern reality, affecting critical areas such as employment, national defence, healthcare and education. The introduction of the AI Act, which sets legal and ethical requirements for the responsible development and use of AI systems, has created the need for assessment tools to ensure compliance with these standards. The objective of this thesis is to create a dynamic AI impact assessment questionnaire that is aligned with the requirements of the AI Act. To this end, it builds on the capAI structure while leveraging the technical approach of the Canadian Impact Assessment Platform to analyze and address the impacts associated with the development of automated decision-making systems.

The thesis starts with an introduction to AI, the different methods of its development and the ethical challenges that arise from its use , such as bias, opacity, automation of work and loss of privacy. The legal framework governing AI is then examined, with a focus on the legislation of the EU, the US, Greece and China. Then some algorithmic impact assessment (AIA) tools used to analyse the compliance of AI systems are analysed.

Subsequently, by combining content and methodologies from various other AIAs, the development of the questionnaire and its implementation is described. The questionnaire evaluates AI systems based on five key metrics: beneficence, non-maleficence, justice, autonomy and explicability and is divided based on the AI development cycle into five key stages : Design, Development, Evaluation, Operation and Retirement.

A comparative study with other similar systems is then carried out in order to challenge the functionality and accuracy of the tool.

The paper concludes with conclusions and recommendations for future improvements, which include the integration of automatic tests, dynamic questions and metrics and further improvement of the evaluation of ethical and legal aspects of AI systems.

Keywords

AI, Algorithmic Artificial Intelligence, Algorithmic Artificial Intelligence, capAI, AI Act, Justice, Explainability, Bias, Ethics , Legislation, Transparency, Human Rights, Autonomous Systems, Traceability, Trustworthy AI, Fairness, Robustness, Testing

Ευχαριστίες

Θα ήθελα να εκφράσω την ειλικρινή μου ευγνωμοσύνη, πρώτα απ’ όλα, στον καθηγητή Παναγιώτη Τσανάκα για την ευκαιρία που μου παρείχε να ασχοληθώ με ένα θέμα που με ενδιέφερε ιδιαίτερα.

Επίσης, θέλω να ευχαριστήσω θερμά τον διδάκτορα Βρεττό Μουλό για την άψογη συνεργασία μας, την υπομονή του και τις πολύτιμες συμβουλές του, οι οποίες με βοήθησαν να ολοκληρώσω αυτή τη διπλωματική εργασία.

Τέλος, δε θα μπορούσα να παραλείψω να ευχαριστήσω την οικογένεια και τους φίλους μου, που στάθηκαν δίπλα μου σε κάθε βήμα αυτής της προσπάθειας. Χωρίς αυτούς, αυτή η διαδρομή δε θα ήταν δυνατή.

Περιεχόμενα

Εισαγωγή

Σύγχρονη πραγματικότητα

Είναι γνωστό πως τα τελευταία 4 χρόνια η παρουσία της τεχνητής νοημοσύνης γίνεται όλο και πιο έντονη στη ζωή του μέσου ανθρώπου. Αυτό το φαινόμενο δεν περιορίζεται μόνο στις εφαρμογές γεννητικής ΤΝ όπως το ChatGPT και το DALL-E, αλλά παρατηρείται πως η ΤΝ εισβάλλει σταδιακά σε διάφορους τομείς, όπως εργασία, άμυνα, εκπαίδευση, και γενικότερα, επηρεάζοντας τις αποφάσεις που αφορούν την καθημερινή ζωή των ανθρώπων. Είναι αναμενόμενο λοιπόν, όπως γίνεται με την εισαγωγή κάθε νέας τεχνολογίας, να έχει προκύψει η ανάγκη για διαφάνεια και τεχνικούς τρόπους ελέγχου και ανάλυσης των συστημάτων αυτών. Πιο συγκεκριμένα, και η Ευρωπαϊκή Ένωση αλλά και διάφορα κράτη ανά τον κόσμο πρόσφατα έχουν μπει στη διαδικασία ανάπτυξης σχετικής νομοθεσίας για τη διασφάλιση της καλής χρήσης των συστημάτων αυτών με σεβασμό στις ήδη υπάρχουσες νομικές αρχές, αλλά και με βασικό γνώμονα τη βελτίωση της ανθρώπινης ποιότητας ζωής και την εφαρμογή των ανθρωπίνων δικαιωμάτων.

Μια από τις πιο σημαντικές εξελίξεις σε αυτόν τον τομέα είναι η εισαγωγή του AI Act από την Ευρωπαϊκή Ένωση, το οποίο πρακτικά επιβάλλει την ανάπτυξη εργαλείων αξιολόγησης συστημάτων ΤΝ. Το AI Act έχει ως στόχο τη ρύθμιση και την προώθηση της υπεύθυνης χρήσης της τεχνητής νοημοσύνης, με έμφαση στην προστασία των δικαιωμάτων των πολιτών, την αποφυγή προκαταλήψεων και την εξασφάλιση ότι τα συστήματα ΤΝ λειτουργούν με διαφάνεια και υπευθυνότητα χωρίς επιβλαβείς συνέπειες στο περιβάλλον και στην κοινωνία. Για την πρακτική εφαρμογή αυτού του νομικού πλαισίου, είναι απαραίτητη η συμβολή του τεχνικού τομέα, προκειμένου να αναπτυχθούν εργαλεία ελέγχου και ανάλυσης που θα εξασφαλίσουν ότι τα συστήματα ΤΝ συμμορφώνονται με τις νομικές και ηθικές απαιτήσεις.

Η ανάπτυξη τέτοιων εργαλείων αξιολόγησης είναι μονόδρομος προς ένα μέλλον όπου οι αποφάσεις των συστημάτων ΤΝ είναι κατανοητές από τους χρήστες, έγκυρες, χωρίς προκαταλήψεις και συμμορφώνονται με τα ηθικά και νομικά πρότυπα. Η εφαρμογή του AI Act καθιστά αυτή την ανάπτυξη ιδιαίτερα επιτακτική, ειδικά σε τομείς όπως η υγεία, η εργασία, η εθνική άμυνα και η δικαιοσύνη, όπου οι αποφάσεις μπορεί να επηρεάσουν ζωές και να έχουν σοβαρές κοινωνικές και οικονομικές συνέπειες. Έτσι, η δημιουργία εργαλείων αξιολόγησης δεν είναι μόνο τεχνική ανάγκη, αλλά και κοινωνική υποχρέωση, με στόχο την προστασία και τη βελτίωση της ποιότητας ζωής των ανθρώπων.

Αντικείμενο της διπλωματικής

Κύρια έμπνευση για την εκπόνηση της εν λόγω διπλωματικής εργασίας είναι η δημοσιευμένη εργασία “capAI: A procedure for conducting conformity assessment of AI systems in line with the EU Artificial Intelligence Act” [1]. Σε αυτή περιγράφεται το capAI που είναι μια δομημένη διαδικασία που στοχεύει στη διασφάλιση και την επίδειξη της συμμόρφωσης ενός συστήματος ΤΝ με τις απαιτήσεις του AI Act, καθώς και με τις ηθικές αρχές και τις οργανωτικές αξίες που ορίζει κάθε εταιρεία ή οργανισμός που αναπτύσσει τέτοια συστήματα. Επίσης, κομβική συμβολή σε αυτή την εργασίας είχε και η ανοιχτή πλατφόρμα της καναδικής κυβέρνησης η οποία με τη χρήση ενός ερωτηματολογίου που συμπληρώνει ο δημιουργός ενός συστήματος ΤΝ τον βοηθάει να αξιολογήσει και να μετριάσει τις επιπτώσεις που συνδέονται με την ανάπτυξη ενός αυτοματοποιημένου συστήματος λήψης αποφάσεων [2].

Ο βασικός στόχος αυτής της εργασίας είναι η δημιουργία ενός νέου, δυναμικού ερωτηματολογίου, το οποίο θα εναρμονίζεται με τις προδιαγραφές του AI Act βασιζόμενο στη δομή που προτείνει το capAI και χρησιμοποιώντας την τεχνική προσέγγιση της καναδικής πλατφόρμας, προσφέροντας έτσι μια ολοκληρωμένη μέθοδο αξιολόγησης. Το ερωτηματολόγιο αυτό θα λειτουργεί ως ένας οδηγός, ο οποίος θα βοηθά τις οργανώσεις να διασφαλίσουν ότι όλα τα στάδια ανάπτυξης των συστημάτων ΤΝ συμμορφώνονται με τις πρακτικές διασφάλισης ποιότητας και ηθικής, όπως αυτές ορίζονται στο capAI. Ταυτόχρονα, θα αποτελεί και ένα εργαλείο αξιολόγησης του τελικού προϊόντος, τα αποτελέσματα του οποίου θα είναι διαθέσιμα σε όλους τους ενδιαφερόμενους, προκειμένου να ενισχυθεί η διαφάνεια και να δομηθεί η απαραίτητη εμπιστοσύνη για μια παραγωγική συνεργασία.

Δομή της Διπλωματικής

Η διπλωματική εργασία αποτελείται από επτά κεφάλαια, καθένα από τα οποία επικεντρώνεται σε διαφορετικές πτυχές της τεχνητής νοημοσύνης και της αξιολόγησης των επιπτώσεών της. Στο Κεφάλαιο 1, γίνεται μια προσπάθεια ορισμού της ΤΝ και προσδιορισμού των τεχνικών ανάπτυξής της. Παράλληλα, αναλύονται τα ηθικά ζητήματα που προκύπτουν από τη χρήση τέτοιων τεχνολογιών από τον άνθρωπο. Αυτό το κεφάλαιο θέτει τις βάσεις για την κατανόηση της σημασίας της ΤΝ και των προκλήσεων που συνοδεύουν την εφαρμογή της.

Στο Κεφάλαιο 2, εξετάζονται τα νομικά πλαίσια που ρυθμίζουν την ΤΝ. Αναλύεται η ευρωπαϊκή νομοθεσία, όπως το AI Act, το Data Act, και το Data Governance Act, καθώς και νομοθεσίες από άλλες χώρες, όπως οι ΗΠΑ, η Κίνα και η Ελλάδα. Σε αυτό το κεφάλαιο γίνονται συγκρίσεις μεταξύ των διαφορετικών νομοθεσιών και σχολιάζεται η σημασία τους στη διαμόρφωση πολιτικών για την ΤΝ. Η κατανόηση αυτών των νομικών πλαισίων είναι απαραίτητη για την ανάπτυξη αποτελεσματικών και συμμορφούμενων συστημάτων ΤΝ.

Το Κεφάλαιο 3 παρουσιάζει διάφορες προσεγγίσεις στην Αλγοριθμική Εκτίμηση Αντικτύπου (ΑΕΑ). Εδώ, αναφέρονται τα εργαλεία και οι πλατφόρμες που έχουν αναπτυχθεί για την αξιολόγηση των συστημάτων ΤΝ, όπως το capAI και το ερωτηματολόγιο της κυβέρνησης του Καναδά. Επίσης, εξετάζονται άλλες διαφορετικές προσεγγίσεις ΑΕΑ από διάφορους οργανισμούς και ινστιτούτα. Το κεφάλαιο τονίζει την αναγκαιότητα της ύπαρξης τέτοιων εργαλείων για την αξιολόγηση των επιπτώσεων της ΤΝ.

Στο Κεφάλαιο 4, περιγράφεται η διαδικασία ανάπτυξης του ερωτηματολογίου. Εξηγείται πώς επιλέχθηκαν και διαμορφώθηκαν οι ερωτήσεις, καθώς και ο τρόπος ενσωμάτωσης των πέντε βασικών μετρικών: beneficence (ευημερία), non-maleficence (μη κακοήθεια), justice (δικαιοσύνη), autonomy (αυτονομία) και explicability (εξηγησιμότητα). Αυτές οι μετρικές αποτελούν τη βάση για την αξιολόγηση των ηθικών και κοινωνικών επιπτώσεων των συστημάτων ΤΝ.

Το Κεφάλαιο 5 αφορά την τεχνική υλοποίηση του συστήματος. Αναλύεται η ανάπτυξη του ερωτηματολογίου μέσω του React framework, η διαχείριση των απαντήσεων, ο υπολογισμός των μετρικών και η εξαγωγή των αποτελεσμάτων σε μορφή PDF. Επίσης, περιγράφεται η ανάπτυξη του συστήματος σε έναν Nginx server, ώστε να είναι δημόσια προσβάσιμο. Αυτό το κεφάλαιο παρέχει μια λεπτομερή περιγραφή των τεχνικών πτυχών της εργασίας.

Στο Κεφάλαιο 6, γίνεται μια συγκριτική μελέτη και βελτιστοποίηση. Αναλύονται άλλα υφιστάμενα συστήματα αξιολόγησης της ΤΝ που χρησιμοποιήθηκαν ως πηγές για την ανάπτυξη του ερωτηματολογίου. Η ανάλυση επικεντρώνεται στην πρωτοπορία και τη χρηστικότητα αυτών των συστημάτων, καθώς και στο κατά πόσο διευκολύνουν την πρακτική εφαρμογή των νομοθεσιών. Επίσης γίνεται και ανάλυση νοηματικής πληρότητας του ερωτηματολογίου με γνώμονα τις απαιτήσεις του AI Act.

Τέλος, στο Κεφάλαιο 7, συνοψίζονται τα βασικά ευρήματα της εργασίας. Επισημαίνεται η αναγκαιότητα τέτοιων εργαλείων στην εποχή της αυξημένης υιοθέτησης της ΤΝ και προτείνονται μελλοντικές βελτιώσεις. Μεταξύ αυτών περιλαμβάνονται η ενσωμάτωση αυτόματων ελέγχων κώδικα, οι δυναμικές αλλαγές και προσθήκες στα βάρη και τις ερωτήσεις, και ο περαιτέρω εμπλουτισμός του ερωτηματολογίου. Το κεφάλαιο κλείνει με προτάσεις για μελλοντικές επεκτάσεις της έρευνας και της ανάπτυξης του συστήματος.

Κεφάλαιο 1: Η Άνοδος της Τεχνητής Νοημοσύνης

1.1. Τεχνητή Νοημοσύνη

1.1.1. Ορισμος

Το 1955 ο John McCarthy συντάσσοντας την πρότασή του για ένα θερινό ερευνητικό συνέδριο με θέμα τη «τεχνητή νοημοσύνη» έγραψε: «Η μελέτη πρόκειται να προχωρήσει με βάση την εικασία ότι κάθε πτυχή της μάθησης ή οποιοδήποτε άλλο χαρακτηριστικό της νοημοσύνης μπορεί κατ' αρχήν να περιγραφεί με τόση ακρίβεια ώστε να μπορεί να δημιουργηθεί μια μηχανή που να την προσομοιώνει» [3]. Αυτή ήταν και η πρώτη φορά, που ο όρος «τεχνητή νοημοσύνη» χρησιμοποιήθηκε σε δημοσίευση [4].

Παρατηρούμε λοιπόν ότι από την αρχή της χρήσης του συγκεκριμένου όρου υπάρχει μια τάση ετεροκαθορισμού του. Για να μπορέσει δηλαδή να υπάρξει ένας αυστηρός και κλειστός ορισμός της ΤΝ θα πρέπει πρώτα να ορίσουμε αποτελεσματικά τον όρο “νοημοσύνη" ή “φυσική νοημοσύνη” ή “ανθρώπινη νοημοσύνη" μιας και η ΤΝ μπορεί να θεωρηθεί ως μια προσπάθεια προσομοίωσης των προηγούμενων.

Μια πιο πρακτική αν και ίσως εφήμερη προσέγγιση ακολουθείται από την Elaine Rich που γράφει “Η τεχνητή νοημοσύνη είναι η μελέτη του πώς να κάνουμε τους υπολογιστές να κάνουν πράγματα στα οποία προς το παρόν οι άνθρωποι είναι καλύτεροι” [5]. Αυτός ο ορισμός αποφεύγει την ασάφεια του όρου “νοημοσύνη” και για την ώρα μπορεί να θεωρηθεί ικανοποιητικός για τον ορισμό του πεδίου.

Βέβαια, μπορεί μικροί και γενικοί ορισμοί να είναι βοηθητικοί στην κατανόηση των ερευνητών για τον διαχωρισμό ενός πεδίου σε νομικά πλαίσια όμως είναι απαραίτητη μια παραπάνω συγκεκριμενοποίηση του όρου για την αποφυγή ανακριβειών. Για παράδειγμα, η NASA ακολουθώντας τον ορισμό της ΤΝ που περιέχεται στον ΕΟ 13960, ο οποίος παραπέμπει στο άρθρο 238(g) του National Defense Authorization Act του 2019, ορίζει την τεχνητή νοημοσύνη ως εξής [6]: (1) Οποιοδήποτε τεχνητό σύστημα που εκτελεί εργασίες υπό ποικίλες και απρόβλεπτες συνθήκες χωρίς σημαντική ανθρώπινη επίβλεψη ή που μπορεί να μαθαίνει από την εμπειρία και να βελτιώνει τις επιδόσεις του όταν εκτίθεται σε σύνολα δεδομένων. (2) Ένα τεχνητό σύστημα που αναπτύσσεται σε λογισμικό υπολογιστή, φυσικό υλικό ή άλλο πλαίσιο και επιλύει εργασίες που απαιτούν αντίληψη, νόηση, σχεδιασμό, μάθηση, επικοινωνία ή φυσική δράση που προσομοιάζει με εκείνη του ανθρώπου.(3) Ένα τεχνητό σύστημα σχεδιασμένο να σκέφτεται ή να ενεργεί όπως ένας άνθρωπος, συμπεριλαμβανομένων των γνωστικών αρχιτεκτονικών και των νευρωνικών δικτύων. (4) Ένα σύνολο τεχνικών, συμπεριλαμβανομένης της μηχανικής μάθησης που έχει σχεδιαστεί για να προσεγγίζει μια γνωστική εργασία. (5) Ένα τεχνητό σύστημα σχεδιασμένο να ενεργεί ορθολογικά, συμπεριλαμβανομένου ενός ευφυούς πράκτορα λογισμικού ή ενσώματου ρομπότ που επιτυγχάνει στόχους χρησιμοποιώντας αντίληψη, σχεδιασμό, συλλογισμό, μάθηση, επικοινωνία, λήψη αποφάσεων και δράση.

1.1.2. Είδη και μέθοδοι τεχνητής νοημοσύνης

Προς καλύτερη κατανόηση του πως λειτουργεί κάτι που χαρακτηρίζεται ως "τεχνητή νοημοσύνη" θα παραθέσουμε μερικά είδη της. Σύμφωνα με τους Stuart Russell και Peter Norvig στο "Τεχνητή Νοημοσύνη: Μια σύγχρονη προσέγγιση" μπορούμε να χωρίσουμε τα είδη της τεχνητής νοημοσύνης με βάση τη μέθοδο που χρησιμοποιούν οι ευφυείς πράκτορες για να καταλήξουν σε λύση στα εξής [7]:

- **Μέθοδος Αναζήτησης** : Εδώ το πρόβλημα αντιπροσωπεύεται από ένα σύνολο καταστάσεων πάνω στο οποίο ο ευφυής πράκτορας κάνει αναζήτηση. Πιο συγκεκριμένα έχουμε:
    - Αναζήτηση σε Απλά Περιβάλλοντα: Ο πράκτορας εξετάζει μια ακολουθία ενεργειών που σχηματίζουν μια διαδρομή προς μια κατάσταση στόχου. Τα προβλήματα εδώ βρίσκονται σε παρατηρήσιμα, αιτιοκρατικά, γνωστά και στατικά περιβάλλοντα
    - Αναζήτηση σε Πολύπλοκα Περιβάλλοντα: Εδώ ο πράκτορας αναζητά μια καλή κατάσταση χωρίς να τον νοιάζει η ακολουθία ενεργειών που τον οδήγησε σε αυτή καθώς έχει να κάνει με προβλήματα που δεν εμπίπτουν στους περιορισμούς των Απλών Περιβαλλόντων.
    - Αναζήτηση σε Ανταγωνιστικά Περιβάλλοντα: Εδώ δύο ή περισσότεροι πράκτορες συνυπάρχουν στο ίδιο περιβάλλον έχοντας αντικρουόμενους στόχους.
    - Ικανοποίηση Περιορισμών**:** Εδώ η κάθε κατάσταση ορίζεται πολυπαραγοντικά και ο στόχος των αλγορίθμων είναι μέσω αναζήτησης να ανακαλύψουν καταστάσεις που ικανοποιούν τους περιορισμούς.
- **Μέθοδος Λογικής:** Σε αυτή τη μέθοδο σχεδιάζονται πράκτορες που μπορούν να σχηματίσουν μια γνωσιακή αναπαράσταση ενός πολύπλοκου κόσμου, να χρησιμοποιήσουν μια διαδικασία συμπερασμού για να αντλήσουν νέες αναπαραστάσεις για τον κόσμο αυτό και να χρησιμοποιήσουν αυτές τις νέες αναπαραστάσεις για να αποφασίσουν ποια θα είναι η επόμενη κίνησή τους.
- **Πιθανοτική Μέθοδος:** Εδώ, σε αντίθεση με τις δύο προηγούμενες μεθόδους που οι αποφάσεις βασίζονται είτε σε απόλυτη γνώση των συνθηκών είτε σε εξέταση όλων των πιθανών σεναρίων σε περίπτωση αβεβαιότητας, χρησιμοποιούνται πιθανοτικές θεωρίες και νόμοι έτσι ώστε οι πράκτορες να μετατρέψουν την αβεβαιότητα σε βαθμούς πίστης με βάση τους οποίους θα πάρουν αποφάσεις.
- **Μηχανική Μάθηση:** Εδώ μελετάμε πως οι πράκτορες μπορούν να βελτιώσουν τη συμπεριφορά τους μέσω της επιμελούς μελέτης δεδομένων και παραδειγμάτων χωρίς να υπάρχει ρητός προγραμματισμός.
    - Μάθηση με επίβλεψη: Εδώ ο πράκτορας εκπαιδεύεται σε ένα σύνολο δεδομένων με ετικέτες, δηλαδή κάθε δεδομένο εισόδου συνοδεύονται από αντίστοιχη ετικέτα εξόδου. Ο στόχος είναι ο πράκτορας να μάθει την αντιστοίχιση μεταξύ εισόδων και εξόδων, ώστε να μπορεί να προβλέψει την έξοδο για νέα, αθέατα δεδομένα.
    - Μάθηση χωρίς επίβλεψη: Εδώ τα δεδομένα δεν έχουν ετικέτες και είναι στόχος του πράκτορα να εξάγει κάποια δομή ή μοτίβο από τα δεδομένα και να βγάλει συμπεράσματα.
    - Ενισχυτική μάθηση: Εδώ η μάθηση βασίζεται στην αλληλεπίδραση του πράκτορα με το περιβάλλον του. Πιο συγκεκριμένα, ο πράκτορας λαμβάνει ανατροφοδότηση με τη μορφή ανταμοιβών ή ποινών με βάση τις ενέργειές του και στόχος του είναι

να ανακαλύψει μια στρατηγική που μεγιστοποιεί τη σωρευτική ανταμοιβή με την πάροδο του χρόνου.

- **Μέθοδος αλληλεπίδρασης με φυσικό περιβάλλον:** Εδώ εξετάζουμε πως ένας πράκτορας μπορεί να βγάλει συμπεράσματα και με βάση αυτά να δράσει στον αφιλτράριστο πραγματικό κόσμο.
    - Επεξεργασία φυσικής γλώσσας: Εδώ οι πράκτορες προσπαθούν να επικοινωνήσουν με φυσική γλώσσα με τους ανθρώπους αλλά και να μάθουν διαβάζοντας κείμενα γραμμένα σε αυτή.
    - Μηχανική όραση: Εδώ οι πράκτορες εξοπλίζονται με κάμερες και προσπαθούν να ερμηνεύουν οπτικές πληροφορίες από το περιβάλλον τους με σκοπό να αναγνωρίσουν αντικείμενα, να πάρουν κάποια απόφαση ή να αλληλεπιδράσουν με το περιβάλλον τους.
    - Ρομποτική: Εδώ οι πράκτορες εφοδιάζονται και με αισθητήρες αλλά και με φυσικές συσκευές δράσης έτσι ώστε να μπορούν να μετακινούνται και να αλληλεπιδρούν με τον πραγματικό κόσμο παίρνοντας αποφάσεις με βάση τις αναλύσεις των ερεθισμάτων τους.

Είναι σημαντικό να επισημάνουμε ότι η παραπάνω κατηγοριοποίηση δεν είναι ούτε απόλυτα εξοντωτική και ούτε μπορεί κάθε μοντέλο τεχνητής νοημοσύνης να ενταχθεί σε μόνο μία από τις κατηγορίες καθώς υπάρχουν επικαλύψεις μεταξύ τους αλλά και πολλά μοντέλα που συνδυάζουν δύο ή περισσότερες από αυτές τις μεθόδους.

1.2 Τεχνητή νοημοσύνη και Ηθική

Όπως με όλες τις νέες τεχνολογίες έτσι και με την ΤΝ για να μπορέσει να υπάρξει ασφαλής ένταξή της στην ανθρώπινη ζωή πρέπει να διερευνηθούν οι πιθανές αρνητικές συνέπειες αλλά και ηθικά διλήμματα της αλλά και πως αυτά μπορούν να αντιμετωπιστούν ή τουλάχιστον μετριαστούν. Με βάση τους Stuart Russell και Peter Norvig στο “Τεχνητή Νοημοσύνη: Μια σύγχρονη προσέγγιση” αλλά και το άρθρο “Ethics of Artificial Intelligence and Robotics” της εγκυκλοπαίδειας “Stanford Encyclopedia of Philosophy” έχουμε εντοπίσει μερικούς βασικούς άξονες πάνω στους οποίους μπορεί να βασιστεί μία συζήτηση για το ηθικό πρόσημο της ΤΝ [7][8].

1.2.1. Απόρρητο και Ιδιωτικότητα

Το δικαίωμα στην ιδιωτική ζωή αναφέρεται στο δικαίωμα ενός ατόμου να διατηρεί τις προσωπικές του πληροφορίες, δραστηριότητες και επικοινωνίες απαλλαγμένες από την παρακολούθηση άλλων οντοτήτων, είτε αυτές είναι κυβερνήσεις και οι μυστικές τους υπηρεσίες, είτε είναι εταιρείες είτε και άλλοι πολίτες. Στο Άρθρο 12 της Οικουμενικής Διακήρυξης για τα Ανθρώπινα Δικαιώματα γίνεται νύξη στο εν λόγω δικαίωμα ως εξής : “Κανείς δεν επιτρέπεται να υποστεί αυθαίρετες επεμβάσεις στην ιδιωτική του ζωή, την οικογένεια, την κατοικία ή την αλληλογραφία του, ούτε προσβολές της τιμής και της υπόληψης του. Καθένας έχει το δικαίωμα να τον προστατεύουν οι νόμοι από επεμβάσεις και προσβολές αυτού του είδους” [9]. Είναι προφανές, λοιπόν, πως το δικαίωμα αυτό θεωρείται θεμελιώδες για την προσωπική ελευθερία και αυτονομία των ατόμων επιτρέποντας τους να επιλέγουν ποιες προσωπικές πληροφορίες μοιράζονται, με ποια άτομα ή οντότητες και υπό ποιες συνθήκες.

Σήμερα, όμως αυτό το δικαίωμα απειλείται και παραβιάζεται συνεχώς. Η πληθώρα των καμερών παρακολούθησης και γενικά η χρήση αισθητήρων που παράγουν δεδομένα για τη ζωή μας,

η συνεχόμενη τάση ψηφιοποίησης και αποθήκευσης σε μέρη συνδεδεμένα στον παγκόσμιο ιστό προσωπικών πληροφοριών των ατόμων (κοινωνικών, ιατρικών κλπ), η ενθάρρυνση εκούσιου μοιράσματος πληροφοριών μέσω των μέσων κοινωνικής δικτύωσης αλλά και η στοχευμένη συγκομιδή δεδομένων από εταιρείες με σκοπό το κέρδος προς αντάλλαγμα "δωρεάν" υπηρεσιών είναι μόνο μερικοί από τους λόγους που συμβάλλουν στην πρωτοφανής διαθεσιμότητα προσωπικών δεδομένων που υπάρχει στις μέρες μας. Και αν συνδυάσουμε αυτή τη διαθεσιμότητα με την εξέλιξη των τεχνολογιών ΤΝ, που βασικό τους χαρακτηριστικό είναι η επεξεργασία και εξαγωγή συμπερασμάτων χρησιμοποιώντας μεγάλα σύνολα δεδομένων (εφαρμογές μηχανικής μάθησης) καταλήγουμε με έναν πολύ ισχυρό και άρα δυνητικά επικίνδυνο συνδυασμό. Είναι σημαντικό επίσης να αναφέρουμε ότι ενώ παλιά κύριοι υπόλογοι για τέτοιες παραβιάσεις ήταν οι μυστικές υπηρεσίες και οι κυβερνήσεις τους, τώρα, λόγω των λόγω που προαναφέρθηκαν υπόλογοι είναι επίσης και εταιρείες στον τομέα της τεχνολογίας αλλά και μεμονωμένοι πολίτες ή αυτόνομες οργανώσεις μέσω τεχνικών παραβίασης βάσεων δεδομένων.

Ένας τρόπος λοιπόν για να μπορέσουν να εξισορροπηθούν τα οφέλη της χρήσης μεγάλων ποσοτήτων δεδομένων σε εφαρμογές ΤΝ αλλά και να εξασφαλιστεί η ιδιωτικότητα των πολιτών υιοθετούνται οι εξής τεχνικές :

- **Αποταυτοποίηση:** Η απαλοιφή προσωπικών πληροφοριών ταυτοποίησης (π.χ. όνομα, αριθμός μητρώου κοινωνικής ασφάλισης) έτσι ώστε οι πληροφορίες να μπορούν να χρησιμοποιηθούν για τη στατιστική τους αξία χωρίς να υπάρχει παραβίαση απορρήτου πολιτών. Βέβαια, αυτό αρκετές φορές δεν είναι αρκετό καθώς η επαναταυτοποίηση είναι πολλές φορές εύκολη μέσω της διασταύρωσης με άλλα σύνολα δεδομένων αλλά και της ύπαρξης μοναδικών συνδυασμών χαρακτηριστικών ανά άτομο στο αρχικό σύνολο.
- **Γενίκευση Πεδίων:** Η ενοποίηση πολλών συγκεκριμένων τιμών χαρακτηριστικών σε μια γενικότερη κατηγορία. (π.χ. αντί να αποθηκεύουμε ακριβή ηλικία ατόμων μπορούμε να αποθηκεύσουμε ηλικιακή ομάδα όπως "20 - 30 χρονών").
- **Ερωτήματα Συνάθροισης:** Μπορούμε εναλλακτικά να κρατάμε όλες τις συγκεκριμένες εγγραφές κρυφές και να επιτρέπουμε ερωτήματα στη βάση δεδομένων που λαμβάνουν μια απόκριση και επιστρέφουν μια συνοψισμένη μετρική πάνω στα δεδομένα (π.χ. έναν μέσο όρο ή μια μέτρηση) μόνο εάν η αρχική απόκριση αφορούσε ικανοποιητικό αριθμό εγγραφών και η ερώτηση δεν ήταν πολύ συγκεκριμένη. Πάλι εδώ εάν δεν υπάρχει η κατάλληλη μέριμνα η επαναταυτοποίηση ίσως είναι πιθανή μέσω πολλών ερωτημάτων.
- **Διαφορικό Απόρρητο:** Είναι ένα μαθηματικό πλαίσιο που διασφαλίζει ότι η συμπερίληψη ή ο αποκλεισμός των δεδομένων ενός μεμονωμένου ατόμου σε ένα σύνολο δεδομένων δεν μπορεί να επηρεάσει σημαντικά το αποτέλεσμα οποιασδήποτε ανάλυσης που πραγματοποιείται σε αυτό το σύνολο δεδομένων. Αυτό παρέχει την εγγύηση ότι η ιδιωτικότητα των ατόμων προστατεύεται, ακόμη και αν κάποιος πιθανός επιτιθέμενος έχει πρόσβαση και μπορεί να διασταυρώσει τα δεδομένα με εξωτερικές πληροφορίες.
- **Ομόσπονδη μάθηση:** Αντί, όπως στις προηγούμενες προσεγγίσεις να έχουμε ένα κεντρικό μοντέλο που κρατάει και εκπαιδεύεται πάνω σε όλα τα δεδομένα, μπορούμε να ακολουθήσουμε το μοντέλο της ομόσπονδης μάθησης όπου το σύνολο των αποκεντρωμένων συσκευών των χρηστών εκπαιδεύουν συνεργατικά ένα κοινό μοντέλο μηχανικής μάθησης, διατηρώντας τα δεδομένα τους τοπικά αποθηκευμένα και άρα κρυφά ενώ μοιράζονται με το κεντρικό μοντέλο μόνο ενημερώσεις του μοντέλου (όπως οι κλίσεις ή ενημερώσεις σε παραμέτρους)

- **Ασφαλή Συνάθροιση:** Η ασφαλής συνάθροιση επιτρέπει σε πολλούς συμμετέχοντες μιας διαδικασίας ομόσπονδης μάθησης να συνεισφέρουν τις ενημερώσεις τους (όπως τα βάρη ή παραμέτρους ή κλίσεις του μοντέλου) με τρόπο που αποκρύπτει τις επιμέρους συνεισφορές, αποκαλύπτοντας μόνο το τελικό συγκεντρωτικό αποτέλεσμα.

1.2.2. Μεροληψία στα συστήματα λήψης αποφάσεων

Στόχος των σχεδιαστών λογισμικού ΤΝ είναι τα συστήματα τους να μπορούν να χαρακτηριστούν ως "δίκαια" καθώς μια λάθος κρίση τους μπορεί να έχει μεγάλες συνέπειες στα εμπλεκόμενα άτομα. Βέβαια, υπάρχουν διαφορετικοί ορισμοί για τη δικαιοσύνη οπότε παραθέτουμε συνοπτικά μερικούς από τους πιο ευρέως χρησιμοποιημένους:

- **Ατομική δικαιοσύνη (individual fairness):** Κάθε άτομο πρέπει να αντιμετωπίζεται όπως τα όμοιά του ανεξαρτήτως της ταυτότητάς του η των κατηγοριών στις οποίες ανήκει.
- **Ομαδική δικαιοσύνη (group fairness):** Το σύστημα θα πρέπει να βγάζει ίσα ή ισοδύναμα αποτελέσματα για δυο διαφορετικές κατηγορίες εάν αυτά συνοψιστούν σε κάποιο στατιστικό στοιχείο.
- **Δικαιοσύνη μέσω αγνοίας (fairness through unawareness):** Με βάση αυτή την προσέγγιση θεωρείται θεμιτό ο αλγόριθμος να αγνοεί ευαίσθητα χαρακτηριστικά (π.χ. φυλή, φύλο, εθνικότητα) κατά τη λήψη αποφάσεων, θεωρώντας έτσι ότι η δικαιοσύνη επιτυγχάνεται εάν τα χαρακτηριστικά αυτά δε χρησιμοποιούνται ρητά στη διαδικασία λήψης αποφάσεων. Βέβαια, αυτή η προσέγγιση έχει και πολλούς κριτικούς καθώς οι αλγόριθμοι τεχνητής μπορούν να μαντεύουν λανθάνοντα χαρακτηριστικά με βάση τα υπόλοιπα συνεχίζοντας έτσι τη μεροληψία (π.χ. εάν ο αλγόριθμός καταλήξει να είναι μεροληπτικός με βάση τον ταχυδρομικό κώδικα μπορεί αυτό το αποτέλεσμα να κρύβει φυλετική μεροληψία από πίσω καθώς άτομα ίδιας φυλής/εθνότητας συνηθίζουν να μένουν σε κοντινές γειτονιές)
- **Ίσα αποτελέσματα (equal outcome):** Η δικαιοσύνη ίσων αποτελεσμάτων ή δημογραφική ισοτιμία είναι μια προσέγγιση ομαδικής δικαιοσύνης σε συστήματα λήψης αποφάσεων που εξασφαλίζει ότι τα αποτελέσματα κατανέμονται ισότιμα σε διαφορετικές δημογραφικές ομάδες που χωρίζονται από ευαίσθητα χαρακτηριστικά όπως το φύλο ή η εθνικότητα.
- **Ίσες ευκαιρίες (equal opportunity):** Εξασφαλίζει ίσες ευκαιρίες επιτυχίας για άτομα από διαφορετικές ομάδες, αλλά μόνο μεταξύ ατόμων που έχουν παρόμοια προσόντα.
- **Ίσος αντίκτυπος (equal impact):** Η αλγοριθμική λήψη αποφάσεων γίνεται με τρόπο που προσπαθεί να διασφαλίσει ότι οι επιπτώσεις μιας απόφασης κατανέμονται ισότιμα σε διάφορες δημογραφικές ομάδες.

Δυστυχώς, όμως παρά τις προσπάθειες των προγραμματιστών για αλγοριθμική δικαιοσύνη στη λήψη αποφάσεων πολλά συστήματα έχουν βρεθεί να αναπαράγουν ήδη υπάρχουσες κοινωνικές προκαταλήψεις. Από πιο μικρού αντίκτυπου συστήματα όπως οι διαφημίσεις που παραδίδονται από το Google AdSense που είναι πιο πιθανό να σχετίζονται με εγκληματική δραστηριότητα, εάν το όνομα που αναζητείται είναι στερεοτυπικά μαύρο μέχρι πολύ μεγαλύτερου αντίκτυπου αποφάσεις όπως αυτές του αλγορίθμου COMPAS ενός συμβουλευτικού συστήματος πρόβλεψης για το αν ένας κατηγορούμενος θα υποπέσει εκ νέου σε αξιόποινη πράξη που, παρόλο που βρέθηκε να είναι εξίσου επιτυχημένο σε μια ομάδα τυχαίων ανθρώπων, βρέθηκε επίσης να παράγει περισσότερα ψευδώς θετικά αποτελέσματα και λιγότερα ψευδώς αρνητικά αποτελέσματα για τους μαύρους κατηγορούμενους [10][11].

Ένα βοηθητικό πλαίσιο για την πιο γενική εξέταση της μεροληψίας στα υπολογιστικά συστήματα έχει προταθεί από τους Friedman και Nissenbaum [12]. Αυτό το πλαίσιο περιλαμβάνει

τρεις κύριες κατηγορίες μεροληψιών: την προϋπάρχουσα προκατάληψη, την τεχνική προκατάληψη και την αναδυόμενη προκατάληψη.Η **προϋπάρχουσα προκατάληψη** αναφέρεται σε προκαταλήψεις που βασίζονται σε κοινωνικές δυναμικές που έχουν εγκαθιδρυθεί για ιστορικούς και πολιτικούς λόγους και προϋπάρχουν της δημιουργίας του συστήματος. Αυτές οι προκαταλήψεις μπορούν να εισέλθουν στο σύστημα είτε μέσω των συνειδητών προσπαθειών ατόμων ή θεσμών, είτε ασυνείδητα λόγω της κανονικοποίησης και ευρείας αποδοχής που έχουν στην κοινωνία. Για παράδειγμα, αν μια κοινωνία έχει εγκαθιδρύσει συγκεκριμένες ανισότητες ή στερεότυπα, αυτά μπορούν να μεταφερθούν στο σύστημα μέσω των δεδομένων ή των αλγορίθμων που το διέπουν.Η **τεχνική προκατάληψη** αφορά ανακριβή συμπεράσματα στα οποία μπορεί να καταλήξει ένα υπολογιστικό σύστημα λόγω των περιορισμών που εμφανίζονται όταν προσπαθούμε να αυτοματοποιήσουμε πραγματικές συνθήκες. Αυτοί οι περιορισμοί μπορεί να προκύπτουν από την πολυπλοκότητα των δεδομένων, την ανεπάρκεια των αλγορίθμων ή την αδυναμία του συστήματος να κατανοήσει πλήρως την ανθρώπινη συμπεριφορά. Η **αναδυόμενη προκατάληψη** αναφέρεται στις προκαταλήψεις που εμφανίζονται λόγω των μεταβαλλόμενων κοινωνικών και περιβαλλοντικών συνθηκών που δεν μπορούν να αντανακλαστούν ή αφομοιωθούν από το σύστημα. Αυτό οδηγεί το σύστημα να καταλήγει σε πεπερασμένα ή λανθασμένα συμπεράσματα, καθώς δεν είναι σε θέση να προσαρμοστεί στις νέες συνθήκες.

Για να μετριαστεί η παρουσία αυτών των μεροληψιών στα συστήματα, μπορεί να γίνει εστίαση σε τρεις άξονες: την **είσοδο, την ανάπτυξη και την επίβλεψη**. Στην **είσοδο**, πρέπει να γίνει προσπάθεια να μην εισαχθούν κατά τη διάρκεια του σχεδιασμού και της ανάπτυξης της ΤΝ ήδη υπάρχουσες κοινωνικές προκαταλήψεις. Για αυτόν τον λόγο, λαμβάνονται μέτρα όπως η πρόσληψη προγραμματιστών από διαφορετικά κοινωνικά υπόβαθρα, η αναλυτική εξέταση των δεδομένων εκπαίδευσης για τυχόν μεροληψίες και η υπεύθυνη λήψη αποφάσεων, όπως η επιλογή του μοντέλου, με αναλογισμό των πιθανών αρνητικών επιπτώσεων.Κατά τη διάρκεια της **ανάπτυξης**, είναι απαραίτητο να γίνονται συνεχή τεστ και αξιολογήσεις των αποτελεσμάτων του συστήματος, ώστε να μπορούν να διορθωθούν έγκαιρα λανθασμένα συμπεράσματα. Αυτή η διαδικασία βοηθά στην εντοπισμό και διόρθωση των μεροληψιών που μπορεί να έχουν εισαχθεί κατά τη φάση της εισόδου.Τέλος, στην **επίβλεψη**, είναι σημαντικό να υπάρχουν μέθοδοι που ελέγχουν το μοντέλο μετά το πέρας της ανάπτυξής του. Ο στόχος είναι να εντοπιστούν και να αντιμετωπιστούν οι προκαταλήψεις πριν προκαλέσουν σημαντική ζημία. Αυτό μπορεί να περιλαμβάνει τη συνεχή παρακολούθηση της απόδοσης του συστήματος σε πραγματικές συνθήκες και την εφαρμογή διορθωτικών μέτρων όταν αυτό είναι απαραίτητο.

1.2.3. Αυτόνομα συστήματα

Τα Ηνωμένα Έθνη έχουν διεξάγει πληθώρα συζητήσεων με σκοπό την απαγόρευση της χρήσης των θανατηφόρων αυτόματων όπλων με βάση τη Σύμβαση της Γενεύης για ορισμένα συμβατικά όπλα. Παρόλο που αυτές δεν έχουν οδηγήσει κάπου λόγω της διαφωνίας των μεγάλων δυνάμεων είναι λογικό ότι το να είναι αυτόματα συστήματα υπεύθυνα για υποκειμενικές αποφάσεις που είναι πιθανό να καταλήξουν στην αφαίρεση ανθρώπινων ζωών βρίσκει πολλούς ηθικά αντίθετους. [13] Πιο συγκεκριμένα είναι δύσκολο να υπάρξει εγγύηση ότι τα όπλα αυτά θα τηρούν τις αρχές της διάκρισης μεταξύ μαχητών και αμάχων, της αναλογικότητας και της στρατιωτικής αναγκαιότητας της βίας σε συγκρούσεις καθώς αυτά αποτελούν πολύπλοκες ηθικές αποφάσεις που δύσκολα μπορούν να κωδικοποιηθούν [14].

Βέβαια, πέρα από τα αυτόματα όπλα, ηθικά διλήμματα προκύπτουν και από τη χρήση άλλων δυνητικά θανατηφόρων/καταστροφικών μηχανών όπως τα αυτόματα αυτοκίνητα. Ενώ, όπως και με τα όπλα, υπάρχει το επιχείρημα ότι ένα πλήρως αυτόματο αυτοκίνητο θα έχει κωδικοποιηθεί έτσι ώστε να υπακούει ακριβώς τους κώδικες οδικής κυκλοφορίας και άρα θα είναι λιγότερο

επικίνδυνο από τον μέσο άνθρωπο οδηγό γι' αυτό τον λόγο, ερωτήματα όπως "Που πέφτει η ευθύνη σε περίπτωση λάθους του αυτόματου συστήματος;" ή "Πως πρέπει αυτοί οι πράκτορες να δρουν σε περιπτώσεις κρίσης που η επιλογή είναι ανάμεσα σε δύο καταστροφικές συνθήκες;" εξακολουθούν να μην έχουν ικανοποιητικές και ευρέως αποδεκτές απαντήσεις.

1.2.4. Διαφάνεια, Επεξηγησιμότητα και Εμπιστοσύνη

Όπως με κάθε νέα τεχνολογία, έτσι και με την ΤΝ, για να μπορέσει να υιοθετηθεί ουσιαστικά από το δημογραφικό για το οποίο προορίζεται, είναι απαραίτητο να προηγηθεί μια περίοδος χτισίματος εμπιστοσύνης προς αυτή την τεχνολογία. Αυτό συνήθως γίνεται με επίσημες διαδικασίες επαλήθευσης και επικύρωσης που τελικά οδηγούν στην πιστοποίηση των νέων τεχνολογικών προϊόντων. Βέβαια, λόγω της καινοτομίας τους τα συστήματα ΤΝ χρειάζονται διαδικασίες επαλήθευσης και επικύρωσης που δεν έχουν αναπτυχθεί πλήρως ακόμα.

Είναι επίσης προφανές ότι ο τελικός χρήστης μιας τέτοιας τεχνολογίας θα χρειαστεί και ένα βαθμό δικής του εποπτείας του αποτελέσματος για να μπορέσει να το εμπιστευθεί πλήρως. Εδώ έρχεται η ανάγκη για εξηγησιμότητα των ΤΝ, καθώς οι περισσότερες λειτουργούν σαν μαύρα κουτιά που δέχονται δεδομένα και παράγουν συμπεράσματα χωρίς να είναι ξεκάθαρο, ακόμα και στους ίδιους τους ειδικούς πολλές φορές, πως κατέληξαν σε αυτά. Η δικαιολόγηση λοιπόν μιας απόφασης είναι απαραίτητη ενώ για να είναι ουσιαστική θα πρέπει να είναι αντιπροσωπευτική της λογικής που ακολούθησε το σύστημα, πλήρης και να μην παραλείπει πληροφορίες που μπορούν να μεταβάλλουν το νόημα και ανάλογη του γνωστικού επιπέδου που κατέχει ο αναγνώστης της και των νοηματικών αφαιρέσεων που μπορεί να επιβάλλει η κάθε συνθήκη.

Γι ακόμη πιο διεξοδικό έλεγχο μπορούν να χρησιμοποιηθούν και audits που καταγράφουν την ακολουθία αποφάσεων που παίρνει το σύστημα στη διάρκεια της ζωής του και από τα οποία μπορούν να εξαχθούν διάφορων ειδών στατιστικά στοιχεία που μπορούν να επιβεβαιώσουν ή να διαψεύσουν την ορθή χρήση του συστήματος (π.χ. στατιστικά που ελέγχουν τη μεροληψία που επιδεικνύει το σύστημα προς συγκεκριμένες κοινωνικές ομάδες).

1.2.5. Αυτοματοποίηση και εργασία

Είναι γνωστό πως με κάθε νέα τεχνολογική εξέλιξη θα υπάρξουν συγκεκριμένες θέσεις εργασίας που σταδιακά θα εξαλειφθούν και τα παραγωγικά τους αποτελέσματα θα καλυφθούν από τη νέα τεχνολογία. Αυτό ιστορικά δεν έχει προκαλέσει ιδιαίτερα μακροπρόθεσμους περιόδους ανεργίας καθώς οι νέες τεχνολογίες βοηθούν και άρα αυξάνουν την παραγωγικότητα η οποία με τη σειρά της αυξάνει τη ζήτηση για αγαθά και έτσι δημιουργούνται νέες θέσεις εργασίας στον κλάδο, οι οποίες μπορούν να καλυφθούν από τα άτομα που έχασαν την εργασίας τους λόγω της τεχνολογίας εξαρχής. Αυτό βέβαια δεν είναι εγγυημένο ότι δε θα έχει ιδιαίτερα αρνητικές συνέπειες για άτομα καθώς πράγματα όπως ο ρυθμός με τον οποίο γίνονται οι αλλαγές λόγω της τεχνολογίας στον κλάδο σε συνδυασμό με την προσβασιμότητα της επανεκπαίδευσης που παρέχεται από την κοινωνία παίζουν πολύ σημαντικό ρόλο στο πόσο καταστροφική θα είναι σε ατομικό επίπεδο αυτή η μετάβαση.

Θα μπορούσε κανείς να υποθέσει ότι με την τεχνολογική εξέλιξη έχουμε μια μετατόπιση προς απασχόληση που βασίζεται περισσότερο στην υψηλή εκπαίδευση, και η οποία θα διαρκεί λιγότερες ώρες και θα έχει πολύ ικανοποιητικές απολαβές αυξάνοντας καθολικά την ποιότητα ζωής. Όμως, αυτό που έχει παρατηρηθεί να συμβαίνει είναι μια "εργατική πόλωση". Έχει, δηλαδή, μετρηθεί αύξηση και των θέσεων που απαιτούν υψηλότερη εξειδίκευση αλλά και των θέσεων που απαιτούν χαμηλότερη εξειδίκευση ενώ οι μεσαίας ειδίκευσης μειώνονται και αντικαθίσταται από αυτόματα συστήματα σε επίπεδο πιθανής εξαφάνισης [15]. Αυτή η πόλωση, βέβαια, δε διευρύνει τις

ανισότητες μόνο στον εργασιακό τομέα αλλά και λόγω της σύνδεσής της με μια γενική μισθολογική ανισότητα χτίζει τα θεμέλια για μία βαθιά άνιση κοινωνία.

1.2.6. Ηθική των μηχανών

Μία βασική φιλοσοφική ερώτηση που ακολουθεί από πάντα τη μελέτη και ανάπτυξη ΤΝ είναι το πού ακριβώς είναι το όριο ανάμεσα στο "έξυπνο μηχάνημα" και στο "άτομο με συνείδηση που πρέπει να έχει ανθρώπινα δικαιώματα". Η απάντηση σε αυτό το ερώτημα δεν είναι εύκολη καθώς δεν έχει αναχθεί επιτυχώς σε κάποιο προηγούμενο ηθικό δίλημμα/κοινωνικό αγώνα. Δεν μπορούν, ας πούμε, οι ΤΝ να θεωρηθούν περιθωριοποιημένη κοινωνική ομάδα και με βάση αυτό να αποφασιστεί ότι πρέπει να έχουν δικαίωμα στην ψήφο καθώς είναι κάποια αβάσιμη προκατάληψη που τους στερεί το δικαίωμα αυτό, καθώς ο προγραμματισμός τους και τα υλικά τους θα ανήκουν πάντα σε κάποιο άλλο πρόσωπο και άρα υπάρχει περίπτωση να είναι απλά φερέφωνα των απόψεων αυτού προσώπου αυτού. Επίσης, δεν είναι απαραίτητο ότι όλο το φάσμα των ανθρώπινων δικαιωμάτων και υποχρεώσεων πρέπει να αποδοθεί στις ΤΝ μαζί. (π.χ. μπορεί να αποφασιστεί ότι κάθε ΤΝ είναι "νομικό πρόσωπο" όπως οι εταιρείες και άρα έχει δικαιώματα και υποχρεώσεις αλλά όχι τα ίδια με ένα ¨ φυσικό πρόσωπο"). Ακόμα, έξτρα ερωτήματα δημιουργούνται αν αναλογιστούμε τη ρεαλιστική πιθανότητα ΤΝ να λειτουργήσουν και σαν υποβοηθήματα και τελικά αναπόσπαστα κομμάτια φυσικών ατόμων (π.χ. προσθήκη υπολογιστικού τεχνητού μέλους σε άτομο με αναπηρία).

1.2.7. Αλληλεπίδραση ανθρώπου-ρομπότ

Άλλο ένα βασικό ηθικό δίλημμα που ακολουθεί τις ΤΝ είναι το πόσο χειριστικές μπορεί να είναι. Δεδομένου ότι είναι συστήματα με μεγάλη έμφαση στην συλλογή προσωπικών (και μη) πληροφοριών, ενώ βασικός στόχος τους είναι συνήθως το κέρδος, μπορούν να προβούν σε ιδιαίτερα στοχευμένη και περίτεχνη χειραγώγηση της συμπεριφοράς ευάλωτων χρηστών τους οδηγώντας του σε επιπόλαιες και γενικά επιζήμιες για αυτούς αποφάσεις. Μερικά παραδείγματα τέτοιας χειριστικής συμπεριφοράς είναι η πολιτική προπαγάνδα υποβοηθούμενη από μέσα αυτοματισμού που έχει αυξηθεί σε μερικά από τα πιο διάσημα μέσα κοινωνικής δικτύωσης τα τελευταία χρόνια αλλά και οποιαδήποτε μορφής "deep fake" περιεχομένου του οποίου η ενταμένη παρουσία στο διαδίκτυο διαβρώνει της αίσθηση της πραγματικότητας πολλών χρηστών ενώ κλονίζει την εμπιστοσύνη άλλων προς το διαδικτυακό περιεχόμενο γενικότερα [16].

Το πρόβλημα αυτό γίνεται πιο έντονο όταν οι χρήστες δεν αλληλεπιδρούν με κάτι που αναγνωρίζουν ξεκάθαρα ως μηχανή (π.χ. λάπτοπ) αλλά με ένα αυτόματο σύστημα που έχει σχεδιαστεί με σκοπό να μιμείται ανθρώπινα χαρακτηριστικά και αντιδράσεις (π.χ. care robots) και άρα οι χρήστες μπορούν να αρχίζουν να του αποδίδουν συνείδηση και συναισθήματα και να δημιουργούν διαπροσωπικούς δεσμούς μαζί του. Αυτό, ενώ μπορεί να είναι ευεργετικό για κάποιους χρήστες, αφήνει πολλούς χωρίς πραγματικό ανθρώπινο δίκτυο υποστήριξης και άρα έρμαια των στους στόχους της ΤΝ και των κατασκευαστών της.

1.2.8. Μοναδικότητα

Μοναδικότητα είναι το σημείο στην εξέλιξη των ΤΝ όπου έχει δημιουργηθεί μια ΤΝ που είναι και πιο "έξυπνη" από κάθε άνθρωπο αλλά και μπορεί να σχεδιάσει και η ίδια μια ΤΝ που είναι

ακόμα πιο έξυπνη από αυτή. Είναι δηλαδή το σημείο στο οποίο θα υπάρξει μια έκρηξη στην εξέλιξη της "εξυπνάδας" πέρα από κάθε εποπτεία που μπορεί να έχει ο άνθρωπος. Δεν είναι γνωστό ακόμα αν αυτό είναι μία ρεαλιστική εκτίμηση για το που μπορεί να καταλήξει το πεδίο ανάπτυξης της ΤΝ ή απλά ένα σενάριο επιστημονικής φαντασίας βασισμένο στον πυρηνικό φόβο των ανθρώπων για το άγνωστο. Δεν είναι καν ξεκάθαρο στην φάση που βρισκόμαστε αν το σενάριο της μοναδικότητας ταυτίζεται με κάτι πιθανά αποκαλυπτικό για το ανθρώπινο είδος όπως σε πολλές ιστορίες φαντασίας ή η αυξημένη νοημοσύνη είναι πάντα συνοδευόμενη με αυξημένη ηθική αντίληψη. Άλλη μια σημαντική παράμετρος σε αυτή την συζήτηση είναι και η πιθανότητα ανάμειξης/επέκτασης της ανθρώπινης νοημοσύνης με τεχνητά μέσα και άρα η δημιουργία υβριδικής νοημοσύνης που κάνει την έννοια της μοναδικότητας άστοχη με την κατάρριψη του διπόλου "άνθρωπος"-"μηχανή".

Κεφάλαιο 2: Νομικά Πλαίσια για ΤΝ

2.1 Ευρωπαϊκή νομοθεσία

2.1.1 Data Governance Act

Ο κανονισμός (ΕΕ) 2022/868, γνωστός και ως Data Governance Act (DGA), είναι μια νομοθεσία της Ευρωπαϊκής Ένωσης που αποσκοπεί στη δημιουργία ενός πλαισίου για τη διακυβέρνηση των δεδομένων εντός της ένωσης, προωθώντας την κοινή χρήση και επαναχρησιμοποίηση των δεδομένων και διασφαλίζοντας παράλληλα την προστασία της ιδιωτικής ζωής και την ασφάλεια [17][18].

Ο κανονισμός αυτός θεσπίζει ένα ολοκληρωμένο πλαίσιο για τη διακυβέρνηση δεδομένων εντός της Ευρωπαϊκής Ένωσης, εστιάζοντας σε τρεις κύριους άξονες: την επαναχρησιμοποίηση δεδομένων που κατέχουν φορείς του δημόσιου τομέα, την παροχή υπηρεσιών διαμεσολάβησης δεδομένων και την εθελοντική κοινοποίηση δεδομένων για αλτρουιστικούς σκοπούς. Παράλληλα, δημιουργεί το Ευρωπαϊκό Συμβούλιο Καινοτομίας Δεδομένων, το οποίο έχει ως σκοπό την υποστήριξη αυτών των στόχων. Ο κανονισμός δεν επιβάλλει στους φορείς του δημόσιου τομέα την υποχρέωση να επιτρέπουν την επαναχρησιμοποίηση δεδομένων, αλλά συμπληρώνει τους υφιστάμενους νόμους σχετικά με την πρόσβαση και την επαναχρησιμοποίηση δεδομένων, διασφαλίζοντας παράλληλα τη συμμόρφωση με τους κανόνες προστασίας δεδομένων και τη νομοθεσία περί ανταγωνισμού.

Στο πλαίσιο της επαναχρησιμοποίησης δεδομένων που κατέχουν φορείς του δημόσιου τομέα ο κανονισμός καθορίζει τους όρους για την περαιτέρω χρήση ορισμένων κατηγοριών δεδομένων, όπως εμπορικά εμπιστευτικά δεδομένα, στατιστικά δεδομένα και δεδομένα που προστατεύονται από δικαιώματα διανοητικής ιδιοκτησίας. Οι φορείς του δημόσιου τομέα απαγορεύεται γενικά να χορηγούν αποκλειστικά δικαιώματα επαναχρησιμοποίησης δεδομένων, εκτός από ειδικές περιπτώσεις όπου αυτό είναι απαραίτητο για υπηρεσίες δημοσίου συμφέροντος. Τα δεδομένα πρέπει να ανωνυμοποιούνται ή να τροποποιούνται για την προστασία της ιδιωτικής ζωής και της εμπιστευτικότητας πριν από την επαναχρησιμοποίηση, ενώ η πρόσβαση μπορεί να παρέχεται μέσα σε ασφαλή περιβάλλοντα επεξεργασίας. Τα τέλη για την περαιτέρω χρήση δεδομένων πρέπει να είναι διαφανή, αμερόληπτα και αναλογικά, με δυνατότητα εκπτώσεων ή δωρεάν πρόσβασης για μην εμπορικούς σκοπούς ή για χρήση από μικρομεσαίες επιχειρήσεις.

Οι υπηρεσίες διαμεσολάβησης δεδομένων, οι οποίες διευκολύνουν την ανταλλαγή δεδομένων μεταξύ κατόχων και χρηστών, πρέπει να λειτουργούν με ουδετερότητα και να μη χρησιμοποιούν τα δεδομένα για άλλους σκοπούς. Οι πάροχοι αυτών των υπηρεσιών υποχρεούνται να ενημερώνουν την αρμόδια αρχή του κράτους μέλους τους πριν από την έναρξη των δραστηριοτήτων τους και στη συνέχεια μπορούν να λειτουργούν σε ολόκληρη την ΕΕ χωρίς πρόσθετες κοινοποιήσεις. Επιπλέον, πρέπει να εξασφαλίζουν διαφάνεια, δικαιοσύνη και μη διάκριση, να λαμβάνουν μέτρα για την πρόληψη της απάτης και να διασφαλίζουν την ασφάλεια των δεδομένων. Ο κανονισμός θεσπίζει επίσης μια διαδικασία κοινοποίησης και προϋποθέσεις για τους παρόχους, ώστε να διασφαλίζεται η συμμόρφωση με τους κανόνες.

Στο πλαίσιο του αλτρουισμού των δεδομένων, ο κανονισμός προωθεί την εθελοντική ανταλλαγή δεδομένων για στόχους γενικού ενδιαφέροντος, όπως η υγειονομική περίθαλψη, η κλιματική αλλαγή ή η επιστημονική έρευνα. Οι οντότητες που ασχολούνται με τον αλτρουισμό δεδομένων μπορούν να εγγραφούν ως αναγνωρισμένοι οργανισμοί αλτρουισμού δεδομένων και να χρησιμοποιήσουν μια κοινή ετικέτα της ΕΕ. Αυτοί οι οργανισμοί πρέπει να λειτουργούν σε μην

κερδοσκοπική βάση, να συμμορφώνονται με τις απαιτήσεις διαφάνειας και προστασίας δεδομένων και να παρέχουν εργαλεία για την απόκτηση και ανάκληση της συγκατάθεσης. Ο κανονισμός εισάγει επίσης ένα ευρωπαϊκό έντυπο συγκατάθεσης για τον αλτρουισμό δεδομένων, το οποίο διευκολύνει τη συλλογή συγκατάθεσης σε ολόκληρη την ΕΕ.

Σχετικά με τις αρμόδιες αρχές και τις διαδικαστικές διατάξεις, κάθε κράτος μέλος οφείλει να ορίσει αρμόδιες αρχές για την εποπτεία των υπηρεσιών διαμεσολάβησης δεδομένων και των οργανισμών αλτρουισμού δεδομένων. Αυτές οι αρχές πρέπει να είναι ανεξάρτητες, αμερόληπτες και επαρκώς εξοπλισμένες. Τα άτομα και οι οργανισμοί έχουν το δικαίωμα να υποβάλλουν καταγγελίες στις αρμόδιες αρχές εάν πιστεύουν ότι τα δικαιώματά τους βάσει του κανονισμού έχουν παραβιαστεί. Οι αρμόδιες αρχές είναι υπεύθυνες για την παρακολούθηση της συμμόρφωσης και τη λήψη μέτρων επιβολής, όταν αυτό είναι απαραίτητο.

Το Ευρωπαϊκό Συμβούλιο Καινοτομίας Δεδομένων έχει συσταθεί με στόχο να συμβουλεύει την Ευρωπαϊκή Επιτροπή σε θέματα διακυβέρνησης δεδομένων, να προωθεί τη διατομεακή κοινή χρήση δεδομένων και να αναπτύσσει κατευθυντήριες γραμμές για κοινούς ευρωπαϊκούς χώρους δεδομένων. Το Συμβούλιο διευκολύνει τη συνεργασία μεταξύ των κρατών μελών, παρέχει συμβουλές σχετικά με τα τεχνικά πρότυπα και υποστηρίζει την ανάπτυξη μιας συνεπούς προσέγγισης για τη διακυβέρνηση δεδομένων σε ολόκληρη την ΕΕ. Επιπλέον, συνδράμει στην εφαρμογή του κανονισμού και στην αξιολόγηση των επιπτώσεών του.

Στο πλαίσιο της διεθνούς πρόσβασης και διαβίβασης δεδομένων ο κανονισμός καθορίζει τους όρους για τη διαβίβαση μην προσωπικών δεδομένων σε τρίτες χώρες, διασφαλίζοντας ότι αυτές οι διαβιβάσεις δεν έρχονται σε σύγκρουση με το δίκαιο της ΕΕ ή με την εθνική νομοθεσία. Οι αποφάσεις δικαστηρίων ή αρχών τρίτων χωρών που απαιτούν διαβίβαση δεδομένων πρέπει να βασίζονται σε διεθνείς συμφωνίες, όπως οι συνθήκες αμοιβαίας δικαστικής συνδρομής. Ελλείψει τέτοιων συμφωνιών, οι διαβιβάσεις μπορούν να πραγματοποιούνται μόνο εάν το νομικό σύστημα της τρίτης χώρας εξασφαλίζει επαρκή προστασία και δικαστική προσφυγή.

Η Ευρωπαϊκή Επιτροπή εξουσιοδοτείται να εκδίδει κατ' εξουσιοδότηση πράξεις για τη συμπλήρωση του κανονισμού, ιδίως σε θέματα που αφορούν τη διαβίβαση ιδιαίτερα ευαίσθητων δεδομένων και το εγχειρίδιο κανόνων για τους οργανισμούς αλτρουισμού δεδομένων. Αυτές οι πράξεις καταρτίζονται σε διαβούλευση με εμπειρογνώμονες και ενδιαφερόμενους φορείς και τίθενται σε ισχύ εκτός εάν το Ευρωπαϊκό Κοινοβούλιο ή το Συμβούλιο αντιταχθεί εντός συγκεκριμένης προθεσμίας.

Οι τελικές και μεταβατικές διατάξεις του κανονισμού προβλέπουν ότι τα κράτη μέλη πρέπει να θεσπίσουν κυρώσεις για τις παραβάσεις του κανονισμού, οι οποίες πρέπει να είναι αποτελεσματικές, αναλογικές και αποτρεπτικές. Η Επιτροπή θα αξιολογήσει τον κανονισμό έως τις 24 Σεπτεμβρίου 2025 και θα υποβάλει έκθεση στο Ευρωπαϊκό Κοινοβούλιο και το Συμβούλιο. Ο κανονισμός τέθηκε σε ισχύ στις 23 Ιουνίου 2022 και εφαρμόζεται από τις 24 Σεπτεμβρίου 2023, με μεταβατικές ρυθμίσεις για τις οντότητες που ήδη παρέχουν υπηρεσίες διαμεσολάβησης δεδομένων.

Συμπερασματικά, νόμος για τη διακυβέρνηση των δεδομένων αποσκοπεί στην προώθηση της κοινής χρήσης και επαναχρησιμοποίησης δεδομένων μεταξύ τομέων και συνόρων, ιδίως για δεδομένα που κατέχουν φορείς του δημόσιου τομέα, διασφαλίζοντας παράλληλα τη συμμόρφωση με τη νομοθεσία περί προστασίας δεδομένων. Επιδιώκει να οικοδομήσει εμπιστοσύνη στην κοινή χρήση δεδομένων με την καθιέρωση μηχανισμών που διασφαλίζουν τη διαφάνεια και τον έλεγχο της χρήσης των δεδομένων, ενισχύοντας την εμπιστοσύνη μεταξύ των ατόμων και των επιχειρήσεων. Ο κανονισμός αποσκοπεί επίσης στη στήριξη της καινοτομίας με γνώμονα τα δεδομένα, επιτρέποντας την πρόσβαση σε δεδομένα για έρευνα, ανάπτυξη και άλλους σκοπούς, με ιδιαίτερη έμφαση στις μικρομεσαίες και τις νεοσύστατες επιχειρήσεις. Επιπλέον, επιδιώκει να διασφαλίσει την κυριαρχία των δεδομένων με την ενίσχυση της στρατηγικής αυτονομίας της ΕΕ, την προώθηση της ελεύθερης

ροής δεδομένων εντός της ΕΕ και τη διασφάλιση ότι οι διαβιβάσεις δεδομένων σε τρίτες χώρες συμμορφώνονται με τα πρότυπα της ΕΕ.

2.1.2 Data Act

Ο κανονισμός (ΕΕ) 2023/2854, γνωστός και ως Data Act, είναι μια νομοθεσία της Ευρωπαϊκής Ένωσης που αποσκοπεί στην εναρμόνιση των κανόνων για τη δίκαιη πρόσβαση στα δεδομένα και τη χρήση τους εντός της. [19][20] Επιδιώκει να δημιουργήσει μια πιο δίκαιη και ανταγωνιστική οικονομία δεδομένων, πλαισιώνοντας την ασφαλή κοινή χρήση δεδομένων και διασφαλίζοντας ότι τα δεδομένα που παράγονται από συνδεδεμένα προϊόντα και υπηρεσίες είναι προσβάσιμα στους χρήστες και σε τρίτους υπό δίκαιους, εύλογους και αμερόληπτους όρους. Ο κανονισμός αυτός εισάγει ένα ολοκληρωμένο πλαίσιο για τη διαχείριση και την ανταλλαγή δεδομένων, με στόχο την προώθηση της καινοτομίας, του ανταγωνισμού και της προστασίας των δικαιωμάτων των χρηστών. Οι γενικές διατάξεις του κανονισμού καθορίζουν το πλαίσιο λειτουργίας και τους βασικούς στόχους του, με έμφαση στην ενίσχυση της πρόσβασης σε δεδομένα, τη διασφάλιση της διαφάνειας και την προστασία των δικαιωμάτων των χρηστών και των επιχειρήσεων.

Ένα από τα κεντρικά σημεία του κανονισμού είναι η παροχή στους χρήστες συνδεδεμένων προϊόντων (όπως συσκευές IoT) ή συναφών υπηρεσιών του δικαιώματος πρόσβασης σε δεδομένα που παράγονται από αυτά τα προϊόντα. Αυτό περιλαμβάνει τόσο δεδομένα προϊόντων (δεδομένα που παράγονται από την ίδια τη συσκευή) όσο και δεδομένα σχετικών υπηρεσιών (δεδομένα που παράγονται κατά την παροχή μιας υπηρεσίας). Οι κάτοχοι δεδομένων, όπως οι κατασκευαστές ή οι πάροχοι υπηρεσιών, υποχρεούνται να θέτουν τα δεδομένα αυτά στη διάθεση των χρηστών και, κατόπιν αιτήματος, τρίτων. Τα δεδομένα πρέπει να παρέχονται σε δομημένο, κοινώς χρησιμοποιούμενο και αναγνώσιμο από μηχανήματα μορφότυπο, δωρεάν. Ενώ οι κάτοχοι των δεδομένων μπορούν να προστατεύουν τα εμπορικά μυστικά, πρέπει να διασφαλίζουν την πρόσβαση στα δεδομένα, εφαρμόζοντας μέτρα εμπιστευτικότητας για την προστασία των ευαίσθητων πληροφοριών.

Οι χρήστες έχουν επίσης το δικαίωμα να ζητήσουν από τους κατόχους δεδομένων να μοιραστούν τα δεδομένα τους με τρίτους, όπως παρόχους υπηρεσιών aftermarket ή άλλες επιχειρήσεις. Η διάταξη αυτή αποσκοπεί στην προώθηση του ανταγωνισμού και της καινοτομίας, επιτρέποντας σε τρίτους να προσφέρουν νέες υπηρεσίες με βάση τα δεδομένα. Οι κάτοχοι δεδομένων δεν πρέπει να κάνουν διακρίσεις μεταξύ συγκρίσιμων κατηγοριών αποδεκτών δεδομένων και πρέπει να παρέχουν δεδομένα υπό δίκαιους, εύλογους και αμερόληπτους όρους. Οι κάτοχοι δεδομένων μπορούν να ζητούν εύλογη αποζημίωση για τη διάθεση των δεδομένων, αλλά η αποζημίωση αυτή πρέπει να είναι αναλογική και αμερόληπτη, διασφαλίζοντας ότι δε δημιουργεί εμπόδια στην πρόσβαση.

Ο κανονισμός απαγορεύει επίσης τις καταχρηστικές συμβατικές ρήτρες στις συμφωνίες ανταλλαγής δεδομένων μεταξύ επιχειρήσεων, ιδίως όταν το ένα μέρος έχει ισχυρότερη διαπραγματευτική θέση. Οι όροι που αποκλίνουν κατάφωρα από την ορθή εμπορική πρακτική θεωρούνται καταχρηστικοί και δεν μπορούν να εφαρμοστούν. Η διάταξη αυτή αποσκοπεί στην προστασία των μικρότερων επιχειρήσεων από το να εξαναγκάζονται σε δυσμενείς συμφωνίες και διασφαλίζει ότι οι συμβάσεις ανταλλαγής δεδομένων είναι δίκαιες και ισότιμες.

Σε περιπτώσεις εξαιρετικής ανάγκης, όπως η αντιμετώπιση δημόσιων εκτάκτων αναγκών (π.χ. φυσικές καταστροφές, κρίσεις υγείας) ή η εκπλήρωση συγκεκριμένων καθηκόντων δημόσιου συμφέροντος, οι φορείς του δημόσιου τομέα, η Ευρωπαϊκή Επιτροπή, η Ευρωπαϊκή Κεντρική Τράπεζα ή οι φορείς της Ένωσης μπορούν να ζητήσουν πρόσβαση σε δεδομένα που κατέχουν

ιδιωτικοί φορείς. Τα αιτήματα πρέπει να είναι συγκεκριμένα, διαφανή και αναλογικά, και οι κάτοχοι δεδομένων μπορούν να απορρίψουν τα αιτήματα εάν δεν έχουν τον έλεγχο των δεδομένων ή εάν έχουν ήδη υποβληθεί παρόμοια αιτήματα. Οι κάτοχοι δεδομένων μπορούν να δικαιούνται δίκαιη αποζημίωση για την παροχή δεδομένων, εκτός από τις περιπτώσεις δημόσιων έκτακτων αναγκών, όπου τα δεδομένα πρέπει να παρέχονται δωρεάν.

Ο κανονισμός αποσκοπεί επίσης στο να διευκολύνει τους πελάτες να εναλλάσσονται μεταξύ υπηρεσιών επεξεργασίας δεδομένων (π.χ. υπηρεσίες νέφους) με την άρση εμποδίων όπως το υψηλό κόστος αλλαγής και οι συμβατικές δεσμεύσεις. Οι πάροχοι υπηρεσιών επεξεργασίας δεδομένων πρέπει να διασφαλίζουν ότι οι πελάτες μπορούν να επιτύχουν λειτουργική ισοδυναμία κατά τη μετάβαση σε μια νέα υπηρεσία, δηλαδή η νέα υπηρεσία θα πρέπει να προσφέρει συγκρίσιμη λειτουργικότητα. Τα τέλη αλλαγής θα μειωθούν σταδιακά και τελικά θα καταργηθούν έως το 2027, διασφαλίζοντας ότι οι πελάτες δε θα τιμωρούνται οικονομικά για την αλλαγή παρόχου.

Για την προστασία από παράνομη πρόσβαση σε δεδομένα, ο κανονισμός περιλαμβάνει εγγυήσεις για την αποτροπή της παράνομης πρόσβασης σε μην προσωπικά δεδομένα από κυβερνήσεις τρίτων χωρών. Οι πάροχοι υπηρεσιών επεξεργασίας δεδομένων πρέπει να λαμβάνουν μέτρα για να διασφαλίζουν ότι τα δεδομένα που τηρούνται στην ΕΕ δε διαβιβάζονται ούτε αποκτούν πρόσβαση σε αυτά κατά παράβαση της νομοθεσίας της ΕΕ. Οι αποφάσεις δικαστηρίων ή αρχών τρίτων χωρών που απαιτούν πρόσβαση σε δεδομένα πρέπει να βασίζονται σε διεθνείς συμφωνίες ή να πληρούν συγκεκριμένους όρους, όπως η αναλογικότητα και η δυνατότητα δικαστικού ελέγχου.

Ο κανονισμός θέτει επίσης βασικές απαιτήσεις για τη διαλειτουργικότητα μεταξύ δεδομένων, μηχανισμών ανταλλαγής δεδομένων και υπηρεσιών. Αυτό περιλαμβάνει τη διασφάλιση ότι οι δομές δεδομένων, οι μορφότυποι και οι μέθοδοι πρόσβασης είναι τυποποιημένες και αναγνώσιμες από μηχανήματα. Η Ευρωπαϊκή Επιτροπή μπορεί να εγκρίνει κοινές προδιαγραφές για να διασφαλίσει τη διαλειτουργικότητα, ιδίως σε περιπτώσεις όπου δεν υπάρχουν διαθέσιμα ή δεν επαρκούν εναρμονισμένα πρότυπα.

Για την επίλυση διαφορών, ο κανονισμός θεσπίζει ένα πλαίσιο που επιτρέπει στους χρήστες, τους κατόχους και τους αποδέκτες δεδομένων να επιλύουν τις διαφορές τους μέσω πιστοποιημένων φορέων επίλυσης διαφορών. Οι φορείς αυτοί πρέπει να είναι αμερόληπτοι, ανεξάρτητοι και ικανοί να λαμβάνουν ταχείες και οικονομικά αποδοτικές αποφάσεις. Ο μηχανισμός αυτός διασφαλίζει ότι οι διαφορές που σχετίζονται με την πρόσβαση σε δεδομένα και την κοινοχρησία μπορούν να επιλύονται αποτελεσματικά, χωρίς να απαιτούνται χρονοβόρες και δαπανηρές νομικές διαδικασίες.

Ο κανονισμός συμπληρώνει τους υφιστάμενους νόμους της ΕΕ για την προστασία των δεδομένων, όπως ο Γενικός Κανονισμός για την Προστασία Δεδομένων (ΓΚΠΔ), και διασφαλίζει ότι τα δεδομένα προσωπικού χαρακτήρα προστατεύονται σύμφωνα με τους νόμους αυτούς. Επιπλέον, προβλέπει εξαιρέσεις για τις μικρομεσαίες και τις πολύ μικρές επιχειρήσεις, ιδίως όσον αφορά τις υποχρεώσεις κοινοχρησίας δεδομένων. Οι εξαιρέσεις αυτές ισχύουν εκτός εάν οι μικρομεσαίες είναι υπεργολάβοι για μεγαλύτερες επιχειρήσεις. Αυτό αποσκοπεί στη μείωση του κανονιστικού φόρτου για τις μικρότερες επιχειρήσεις, επιτρέποντάς τους να επικεντρωθούν στην καινοτομία και την ανάπτυξη χωρίς να κατακλύζονται από απαιτήσεις συμμόρφωσης.

Τέλος, ο κανονισμός θα εφαρμοστεί από τις 12 Σεπτεμβρίου 2025, δίνοντας στις επιχειρήσεις χρόνο για να προσαρμοστούν στους νέους κανόνες. Αυτή η μεταβατική περίοδος επιτρέπει στους κατασκευαστές, τους παρόχους υπηρεσιών και άλλους ενδιαφερόμενους φορείς να εφαρμόσουν τις απαραίτητες τεχνικές και οργανωτικές αλλαγές για να συμμορφωθούν με τον κανονισμό.

Συμπερασματικά, ο νόμος για τα δεδομένα αποσκοπεί στη δημιουργία μιας δικαιότερης και πιο ανταγωνιστικής οικονομίας δεδομένων, διασφαλίζοντας ότι τα δεδομένα που παράγονται από συνδεδεμένα προϊόντα και υπηρεσίες είναι προσβάσιμα στους χρήστες, σε τρίτους και σε φορείς του δημόσιου τομέα υπό σαφείς και δίκαιους όρους. Επιδιώκει να προωθήσει την καινοτομία, να

προστατεύσει τα εμπορικά απόρρητα και να διασφαλίσει ότι η κοινή χρήση δεδομένων δεν υπονομεύει τα δικαιώματα των ατόμων ή την ασφάλεια των δεδομένων. Με την εναρμόνιση των κανόνων για την πρόσβαση στα δεδομένα και τη χρήση τους, ο κανονισμός αποτελεί βασική συνιστώσα της ευρύτερης στρατηγικής της ΕΕ για τη δημιουργία μιας ενιαίας αγοράς δεδομένων, όπου τα δεδομένα θα μπορούν να διακινούνται ελεύθερα μεταξύ τομέων και συνόρων, διασφαλίζοντας παράλληλα την προστασία της ιδιωτικής ζωής, την ασφάλεια και τον θεμιτό ανταγωνισμό.

2.1.3 AI Act

Ο κανονισμός (ΕΕ) 2024/1689, γνωστός και ως AI Act, θεσπίζει ένα εναρμονισμένο νομικό πλαίσιο για την ανάπτυξη, την εμπορία και τη χρήση συστημάτων τεχνητής νοημοσύνης (ΤΝ) στην Ευρωπαϊκή Ένωση. [21] Οι πρωταρχικοί του στόχοι είναι να προωθήσει την ανθρωποκεντρική και αξιόπιστη τεχνητή νοημοσύνη, να διασφαλίσει υψηλό επίπεδο προστασίας της υγείας, της ασφάλειας, των θεμελιωδών δικαιωμάτων, της δημοκρατίας, του κράτους δικαίου και της προστασίας του περιβάλλοντος και να στηρίξει την καινοτομία, αποτρέποντας παράλληλα τις επιβλαβείς επιπτώσεις των συστημάτων τεχνητής νοημοσύνης. Ο κανονισμός εφαρμόζεται στα συστήματα ΤΝ που διατίθενται στην αγορά, τίθενται σε λειτουργία ή χρησιμοποιούνται στην ΕΕ, ανεξάρτητα από το αν ο πάροχος εδρεύει στην ΕΕ ή σε τρίτη χώρα. Καλύπτει επίσης τα συστήματα ΤΝ των οποίων η παραγωγή χρησιμοποιείται στην ΕΕ, ακόμη και αν το ίδιο το σύστημα βρίσκεται εκτός της ΕΕ. Ωστόσο, αποκλείει τα συστήματα ΤΝ που χρησιμοποιούνται αποκλειστικά για στρατιωτικούς σκοπούς, σκοπούς άμυνας ή εθνικής ασφάλειας, καθώς και εκείνα που αναπτύσσονται αποκλειστικά για επιστημονική έρευνα.

Ο κανονισμός υιοθετεί μια προσέγγιση με βάση τον κίνδυνο, κατηγοριοποιώντας τα συστήματα ΤΝ με βάση το επίπεδο κινδύνου που ενέχουν.Περιγράφονται 4 βασικές κατηγορίες κινδύνου: **απαράδεκτος κίνδυνος, υψηλός κίνδυνος, περιορισμένος κίνδυνος και ελάχιστος ή ανύπαρκτος κίνδυνος** [22].

1. **Απαράδεκτος κίνδυνος:** Αυτή η κατηγορία περιλαμβάνει συστήματα ΤΝ που θεωρούνται σαφής απειλή για την ασφάλεια, τις ζωές και τα δικαιώματα των ανθρώπων. Ο κανονισμός απαγορεύει οκτώ πρακτικές, όπως η επιβλαβής χειραγώγηση και εξαπάτηση με τη χρήση ΤΝ, η εκμετάλλευση τρωτών σημείων, η κοινωνική βαθμολόγηση, η αξιολόγηση ή πρόβλεψη ποινικών αδικημάτων σε επίπεδο ατόμου, η μη στοχευμένη συλλογή δεδομένων από το διαδίκτυο ή CCTV για τη δημιουργία ή επέκταση βάσεων δεδομένων αναγνώρισης προσώπων, η αναγνώριση συναισθημάτων σε χώρους εργασίας και εκπαιδευτικούς χώρους, η βιομετρική κατηγοριοποίηση για την εξαγωγή προστατευόμενων χαρακτηριστικών και η βιομετρική αναγνώριση σε πραγματικό χρόνο για σκοπούς επιβολής του νόμου σε δημόσιους χώρους.
2. **Υψηλός κίνδυνος:** Αυτή η κατηγορία περιλαμβάνει συστήματα ΤΝ που μπορούν να προκαλέσουν σοβαρούς κινδύνους για την υγεία, την ασφάλεια ή τα θεμελιώδη δικαιώματα. Παραδείγματα περιλαμβάνουν συστήματα ΤΝ που χρησιμοποιούνται σε κρίσιμες υποδομές (π.χ. μεταφορές), εκπαιδευτικά ιδρύματα, απασχόληση, βασικές υπηρεσίες (π.χ. υγειονομική περίθαλψη), επιβολή του νόμου, μετανάστευση και έλεγχο συνόρων. Τα συστήματα αυτά πρέπει να συμμορφώνονται με αυστηρές απαιτήσεις, όπως η διαχείριση κινδύνων, η χρήση υψηλής ποιότητας δεδομένων, η διαφάνεια και η ανθρώπινη εποπτεία.
3. **Περιορισμένος κίνδυνος:** Αυτή η κατηγορία αναφέρεται σε συστήματα ΤΝ με συγκεκριμένες υποχρεώσεις διαφάνειας. Για παράδειγμα, όταν χρησιμοποιούνται συστήματα όπως τα chatbots, οι χρήστες πρέπει να ενημερώνονται ότι αλληλεπιδρούν με

μια μηχανή. Επίσης, τα συστήματα που παράγουν περιεχόμενο (π.χ. deepfakes) πρέπει να επισημαίνονται σαφώς ως τεχνητά παραγόμενα.

4. **Ελάχιστος ή ανύπαρκτος κίνδυνος:** Αυτή η κατηγορία περιλαμβάνει τα περισσότερα συστήματα ΤΝ που χρησιμοποιούνται σήμερα στην ΕΕ, όπως τα βιντεοπαιχνίδια με ΤΝ ή τα φίλτρα ανεπιθύμητων μηνυμάτων. Για αυτά τα συστήματα, ο κανονισμός δεν εισάγει πρόσθετες απαιτήσεις.

Τα συστήματα τεχνητής νοημοσύνης υψηλού κινδύνου πρέπει να συμμορφώνονται με αυστηρές απαιτήσεις, συμπεριλαμβανομένης της εκτίμησης ρίσκου για τον εντοπισμό, την αξιολόγηση και τον μετριασμό των κινδύνων καθ' όλη τη διάρκεια του κύκλου ζωής του συστήματος, της διακυβέρνησης δεδομένων για να διασφαλίζεται ότι χρησιμοποιούνται σύνολα δεδομένων υψηλής ποιότητας, αντιπροσωπευτικά και αμερόληπτα για την εκπαίδευση, την επικύρωση και τις δοκιμές, και της διαφάνειας για την παροχή σαφών οδηγιών χρήσης και την ενημέρωση των χρηστών σχετικά με τις δυνατότητες, τους περιορισμούς και τους κινδύνους του συστήματος. Επιπλέον, τα συστήματα τεχνητής νοημοσύνης υψηλού κινδύνου πρέπει να επιτρέπουν την ανθρώπινη εποπτεία, επιτρέποντας την ανθρώπινη παρέμβαση, διόρθωση ή διακοπή, εάν είναι απαραίτητο, και να πληρούν υψηλά πρότυπα ακρίβειας, ευρωστίας και κυβερνοασφάλειας για να διασφαλίζουν ανθεκτικότητα έναντι σφαλμάτων, βλαβών και κυβερνοεπιθέσεων. Τα συστήματα αυτά πρέπει να υποβάλλονται σε αξιολόγηση συμμόρφωσης για να αποδεικνύεται η συμμόρφωση με τις απαιτήσεις του κανονισμού. Η αξιολόγηση αυτή μπορεί να διενεργηθεί από τον πάροχο (αυτοαξιολόγηση) ή από κοινοποιημένο οργανισμό (αξιολόγηση από τρίτο μέρος). Τα συστήματα που συμμορφώνονται πρέπει να φέρουν τη σήμανση CE, επιτρέποντάς τους να κυκλοφορούν ελεύθερα στην αγορά της ΕΕ.

Σύμφωνα πάλι με τον κανονισμό οι χρήστες πρέπει να ενημερώνονται όταν αλληλεπιδρούν με ένα σύστημα τεχνητής νοημοσύνης και το περιεχόμενο που παράγεται από τεχνητή νοημοσύνη (π.χ. deepfakes) πρέπει να επισημαίνεται σαφώς ως τεχνητά παραγόμενο ή παραποιημένο. Τα άτομα που επηρεάζονται από αποφάσεις που λαμβάνονται από συστήματα ΤΝ υψηλού κινδύνου έχουν το δικαίωμα να λαμβάνουν εξηγήσεις για την απόφαση, ιδίως όταν αυτή επηρεάζει σημαντικά την υγεία, την ασφάλεια ή τα θεμελιώδη δικαιώματά τους.

Τα μοντέλα τεχνητής νοημοσύνης γενικής χρήσης είναι εκείνα που μπορούν να εκτελέσουν ένα ευρύ φάσμα εργασιών και δεν περιορίζονται σε μια συγκεκριμένη εφαρμογή. Οι πάροχοι τέτοιων μοντέλων πρέπει να διατηρούν τεχνική τεκμηρίωση και να παρέχουν πληροφορίες σχετικά με τις δυνατότητες, τους περιορισμούς και τις προβλεπόμενες χρήσεις του μοντέλου. Ορισμένα μοντέλα ΤΝ γενικού σκοπού ενδέχεται να ενέχουν συστημικούς κινδύνους λόγω των υψηλών δυνατοτήτων τους και της ευρείας χρήσης τους. Οι πάροχοι τέτοιων μοντέλων πρέπει να λαμβάνουν πρόσθετα μέτρα για τον εντοπισμό και τον μετριασμό αυτών των κινδύνων, συμπεριλαμβανομένης της διενέργειας δοκιμών με αντιπάλους και της εφαρμογής ισχυρών μέτρων κυβερνοασφάλειας.

Τα κράτη μέλη ενθαρρύνονται να δημιουργήσουν ρυθμιστικά «sandboxes» τεχνητής νοημοσύνης για να παρέχουν ένα ελεγχόμενο περιβάλλον για τη δοκιμή καινοτόμων συστημάτων τεχνητής νοημοσύνης. Αυτά τα sandboxes είναι ιδιαίτερα ωφέλιμα για τις μικρομεσαίες επιχειρήσεις και τις νεοσύστατες επιχειρήσεις, καθώς παρέχουν πρόσβαση σε εμπειρογνωμοσύνη, πόρους και κανονιστική καθοδήγηση. Οι πάροχοι συστημάτων ΤΝ που δεν ταξινομούνται ως υψηλού κινδύνου ενθαρρύνονται να υιοθετήσουν εθελοντικούς κώδικες δεοντολογίας για την προώθηση ηθικών και αξιόπιστων πρακτικών ΤΝ.

Το Ευρωπαϊκό Συμβούλιο Τεχνητής Νοημοσύνης θα συντονίζει την εφαρμογή του κανονισμού σε όλα τα κράτη μέλη, θα παρέχει καθοδήγηση και θα διευκολύνει την ανταλλαγή βέλτιστων πρακτικών. Μια επιστημονική ομάδα ανεξάρτητων εμπειρογνωμόνων θα υποστηρίξει το Συμβούλιο και την Επιτροπή στην παρακολούθηση και αξιολόγηση των συστημάτων τεχνητής νοημοσύνης, ιδίως εκείνων που ενέχουν συστημικούς κινδύνους. Ένα συμβουλευτικό φόρουμ θα

παρέχει πληροφορίες από τα ενδιαφερόμενα μέρη, συμπεριλαμβανομένων της βιομηχανίας, της ακαδημαϊκής κοινότητας, της κοινωνίας των πολιτών και των οργανώσεων καταναλωτών.

Τα κράτη μέλη πρέπει να ορίσουν αρχές εποπτείας της αγοράς για την παρακολούθηση της συμμόρφωσης με τον κανονισμό. Οι αρχές αυτές έχουν την εξουσία να διερευνούν, να επιθεωρούν και να λαμβάνουν διορθωτικά μέτρα, συμπεριλαμβανομένης της ανάκλησης ή απόσυρσης μη συμμορφούμενων συστημάτων ΤΝ. Οι πάροχοι πρέπει να δημιουργήσουν συστήματα παρακολούθησης μετά τη διάθεση στην αγορά για την παρακολούθηση των επιδόσεων των συστημάτων ΤΝ υψηλού κινδύνου μετά τη διάθεσή τους στην αγορά. Οποιαδήποτε σοβαρά περιστατικά, όπως εκείνα που οδηγούν σε θάνατο, σοβαρή βλάβη ή σημαντικές διαταραχές, πρέπει να αναφέρονται στις αρμόδιες αρχές. Τα κράτη μέλη πρέπει επίσης να θεσπίσουν αποτελεσματικές, αναλογικές και αποτρεπτικές κυρώσεις για τη μη συμμόρφωση, συμπεριλαμβανομένων προστίμων ύψους έως και 6% του ετήσιου παγκόσμιου κύκλου εργασιών του παρόχου για σοβαρές παραβάσεις.

Ο κανονισμός θα εφαρμοστεί από τις 2 Αυγούστου 2026. Ωστόσο, ορισμένες διατάξεις, όπως οι απαγορεύσεις για απαράδεκτες πρακτικές ΤΝ, θα ισχύουν από τις 2 Φεβρουαρίου 2025. Τα συστήματα τεχνητής νοημοσύνης που κυκλοφορούν ήδη στην αγορά ή λειτουργούν πριν από την εφαρμογή του κανονισμού δε θα χρειαστεί να συμμορφωθούν με τις νέες απαιτήσεις, εκτός εάν υποστούν σημαντικές αλλαγές στον σχεδιασμό ή στον επιδιωκόμενο σκοπό τους. Ο κανονισμός δε θίγει άλλους νόμους της ΕΕ, συμπεριλαμβανομένων εκείνων για την προστασία των δεδομένων, την προστασία των καταναλωτών και την ασφάλεια των προϊόντων, διασφαλίζοντας μια συνεκτική προσέγγιση για τη ρύθμιση της ΤΝ σε ολόκληρη την ΕΕ.

Συμπερασματικά, ο κανονισμός αυτός αποτελεί σημαντική πρόοδο στη διακυβέρνηση της τεχνητής νοημοσύνης στην Ευρωπαϊκή Ένωση. Με τη θέσπιση ενός ολοκληρωμένου νομικού πλαισίου, ο κανονισμός αποσκοπεί στην προώθηση της καινοτομίας, διασφαλίζοντας παράλληλα τα δημόσια συμφέροντα και τα θεμελιώδη ανθρώπινα δικαιώματα. Μέσω της έμφασής του στους ηθικούς προβληματισμούς, την προστασία των δεδομένων και τη συνεργασία των ενδιαφερομένων μερών, ο κανονισμός επιδιώκει να καταστήσει την ΕΕ παγκόσμιο ηγέτη στην υπεύθυνη τεχνητή νοημοσύνη διασφαλίζοντας ότι οι τεχνολογίες ΤΝ αναπτύσσονται με τρόπο που ευθυγραμμίζεται με τις αξίες και τις αρχές της Ένωσης.

2.1.4 Στρατηγική της Ευρώπης

Η στρατηγική της Ευρωπαϊκής Ένωσης σχετικά με τη νομοθεσία γύρω από την τεχνολογία, τα δεδομένα και κυρίως την τεχνητή νοημοσύνη (ΤΝ) αποτελεί μια ολοκληρωμένη προσέγγιση που στοχεύει στην προώθηση της καινοτομίας, την προστασία των θεμελιωδών δικαιωμάτων και τη διασφάλιση της ανταγωνιστικότητας της ΕΕ στον παγκόσμιο χώρο. Όπως είδαμε, η ΕΕ έχει αναπτύξει μια σειρά από νομοθετικά πλαίσια, όπως το AI Act, το Data Governance Act (DGA) και το Data Act, τα οποία συνθέτουν μια ευρύτερη στρατηγική για τη δημιουργία μιας ενιαίας αγοράς δεδομένων και την ανάπτυξη αξιόπιστης τεχνητής νοημοσύνης.

Η Ευρωπαϊκή Ένωση θέτει στο επίκεντρο της στρατηγικής της την προστασία των ανθρωπίνων δικαιωμάτων, της δημοκρατίας και του κράτους δικαίου. Οι νομοθεσίες της, όπως το AI Act, έχουν σχεδιαστεί για να διασφαλίζουν ότι η τεχνητή νοημοσύνη αναπτύσσεται με τρόπο που σέβεται τις αξίες της ΕΕ, όπως η ισότητα, η διαφάνεια και η προστασία της ιδιωτικής ζωής. Αυτή η ανθρωποκεντρική προσέγγιση εξασφαλίζει ότι οι τεχνολογίες ΤΝ χρησιμοποιούνται για να βελτιώσουν την ποιότητα ζωής των πολιτών, χωρίς να θίγονται τα θεμελιώδη δικαιώματά τους.

Παράλληλα, η ΕΕ επιδιώκει να στηρίξει την έρευνα, την ανάπτυξη και την εφαρμογή νέων τεχνολογιών, ιδίως στον τομέα της ΤΝ και των δεδομένων. Μέσω νομοθεσιών όπως το Data Act και το DGA, η ΕΕ ενθαρρύνει την κοινή χρήση δεδομένων μεταξύ επιχειρήσεων, φορέων του δημόσιου

τομέα και ερευνητικών ιδρυμάτων, δημιουργώντας ένα ευνοϊκό περιβάλλον για την καινοτομία. Αυτή η προσέγγιση δεν προωθεί μόνο την τεχνολογική πρόοδο, αλλά και την οικονομική ανάπτυξη, καθώς οι επιχειρήσεις μπορούν να αξιοποιήσουν τα δεδομένα για τη δημιουργία νέων υπηρεσιών και προϊόντων.

Η διασφάλιση της ασφάλειας και της προστασίας των δεδομένων αποτελεί άλλο έναν βασικό στόχο της στρατηγικής της ΕΕ. Η Ένωση έχει θέσει υψηλά πρότυπα για την προστασία των δεδομένων, όπως φαίνεται στον Γενικό Κανονισμό για την Προστασία Δεδομένων (GDPR) και στις πρόσφατες νομοθεσίες όπως το DGA και το Data Act. Αυτοί οι κανονισμοί εξασφαλίζουν ότι τα δεδομένα χρησιμοποιούνται με τρόπο που προστατεύει την ιδιωτική ζωή και την ασφάλεια των πολιτών, ενώ ταυτόχρονα επιτρέπουν την ελεύθερη ροή των δεδομένων εντός της ΕΕ.

Επιπλέον, η ΕΕ επιδιώκει να δημιουργήσει μια ενιαία αγορά δεδομένων, όπου τα δεδομένα μπορούν να διακινούνται ελεύθερα μεταξύ των κρατών μελών, ενισχύοντας την οικονομική ανάπτυξη και την ανταγωνιστικότητα. Αυτό επιτυγχάνεται μέσω της εναρμόνισης των κανόνων για την πρόσβαση στα δεδομένα και τη χρήση τους, όπως προβλέπει το Data Act. Η ενιαία αγορά δεδομένων αποσκοπεί να δημιουργήσει ένα οικοσύστημα όπου τα δεδομένα μπορούν να χρησιμοποιηθούν για την ανάπτυξη νέων τεχνολογιών και υπηρεσιών, ενισχύοντας την καινοτομία και την ανταγωνιστικότητα της ΕΕ στον παγκόσμιο χώρο.

Τέλος, η ΕΕ αναγνωρίζει τη σημασία των μικρομεσαίων επιχειρήσεων (ΜΜΕ) και παρέχει εξαιρέσεις και υποστήριξη για να μην κατακλύζονται από κανονιστικούς φόρτους. Μέσω ρυθμιστικών sandboxes και εθελοντικών κωδίκων δεοντολογίας, οι ΜΜΕ μπορούν να αναπτύσσουν καινοτόμες λύσεις χωρίς να αντιμετωπίζουν υπερβολικές απαιτήσεις. Αυτή η προσέγγιση ενισχύει την ανταγωνιστικότητα των μικρότερων επιχειρήσεων και βοηθά στη δημιουργία ενός δυναμικού και καινοτόμου επιχειρηματικού περιβάλλοντος.

Η στρατηγική της ΕΕ για την τεχνολογία, τα δεδομένα και την ΤΝ βασίζεται σε μια ισορροπημένη προσέγγιση που συνδυάζει την προώθηση της καινοτομίας με την προστασία των δικαιωμάτων των πολιτών. Μέσω των νομοθετικών της πλαισίων, η ΕΕ επιδιώκει να δημιουργήσει ένα περιβάλλον όπου η τεχνολογία αναπτύσσεται με τρόπο που ευθυγραμμίζεται με τις αξίες της Ένωσης, όπως η δημοκρατία, η διαφάνεια και η ισότητα. Παράλληλα, η ΕΕ θέτει τις βάσεις για μια ανταγωνιστική οικονομία δεδομένων, η οποία μπορεί να οδηγήσει σε νέες ευκαιρίες για επιχειρήσεις, ερευνητές και πολίτες.

Η στρατηγική αυτή αποτελεί μια σημαντική πρόκληση, καθώς απαιτεί την ισορροπία μεταξύ της προστασίας των δικαιωμάτων και της προώθησης της καινοτομίας. Ωστόσο, αν εφαρμοστεί αποτελεσματικά, μπορεί να καταστήσει την ΕΕ παγκόσμιο ηγέτη στον τομέα της υπεύθυνης τεχνολογίας και της διακυβέρνησης των δεδομένων, διασφαλίζοντας ότι η τεχνολογική πρόοδος εξυπηρετεί το κοινό συμφέρον και τις ανθρώπινες αξίες.

2.2 Ελληνική Νομοθεσία

Το νομικό πλαίσιο της Ελλάδας σχετικά με τη χρήση της Τεχνητής Νοημοσύνης (ΤΝ) στον ιδιωτικό και δημόσιο τομέα καθορίζεται στα Κεφάλαια Α και Β του Μέρους Α του Νόμου 4961/2022 [23]. Ο νόμος αυτός έχει ως κύριο στόχο τη δημιουργία ενός νομικού πλαισίου που να ρυθμίζει τη χρήση της ΤΝ τόσο από τον δημόσιο όσο και από τον ιδιωτικό τομέα, με σεβασμό στα δικαιώματα των φυσικών και νομικών προσώπων και με βάση τις αρχές της λογοδοσίας και της διαφάνειας.

Στον δημόσιο τομέα, ο νόμος προβλέπει ότι πριν από την έναρξη λειτουργίας ενός συστήματος ΤΝ, η δημόσια υπηρεσία που είναι υπεύθυνη για αυτό πρέπει να διεξάγει μια

Αλγοριθμική Εκτίμηση Αντικτύπου (ΑΕΑ). Η ΑΕΑ καλύπτει διάφορες πτυχές, όπως τον στόχο της ΤΝ, τα τεχνικά χαρακτηριστικά της, τις δυνατότητές της, το είδος των αποφάσεων ή πράξεων που παράγει, τα δεδομένα που συλλέγει, καθώς και μια ανάλυση των ωφελειών και των κινδύνων της χρήσης της, τόσο σε επίπεδο ατομικών ελευθεριών όσο και σε επίπεδο κοινωνικού αντικτύπου. Επιπλέον, οι δημόσιες υπηρεσίες που χρησιμοποιούν ΤΝ έχουν την υποχρέωση να δημοσιοποιούν πληροφορίες όπως ο χρόνος έναρξης της χρήσης της ΤΝ, τα τεχνικά της χαρακτηριστικά, το είδος των αποφάσεων ή πράξεων που παράγει, καθώς και τη διενέργεια της ΑΕΑ. Κάθε απόφαση ή πράξη που προκύπτει από τη χρήση της ΤΝ πρέπει να συνοδεύεται από μια κατανοητή, ακριβή και προσβάσιμη εξήγηση προς όλα τα νομικά πρόσωπα που αφορά ενώ η Εθνική Αρχή Διαφάνειας είναι υπεύθυνη για τη διαχείριση περιπτώσεων παραβίασης αυτών των κανόνων. Επιπλέον, ο νόμος προβλέπει τη δημιουργία ενός μητρώου, όπου είναι υποχρεωτική η αναλυτική καταγραφή όλων των συστημάτων ΤΝ που χρησιμοποιούνται από τις δημόσιες υπηρεσίες, το οποίο θα ελέγχεται από την Εθνική Αρχή Διαφάνειας. Τέλος, ο νόμος ορίζει τους όρους και τις προϋποθέσεις που πρέπει να περιλαμβάνονται στις προκηρύξεις για δημόσιες συμβάσεις που αφορούν την ανάπτυξη και εφαρμογή ΤΝ.

Στον ιδιωτικό τομέα, ο νόμος προβλέπει ότι εάν μια ιδιωτική εταιρεία χρησιμοποιεί ΤΝ για λήψη αποφάσεων που επηρεάζουν την εργασιακή κατάσταση ή τις συνθήκες εργασίας των ατόμων, αυτές οι αποφάσεις πρέπει να συνοδεύονται από μια κατανοητή, ακριβή και προσβάσιμη εξήγηση. Επιπλέον, πρέπει να παρέχονται διαβεβαιώσεις ότι δεν υπάρχει μεροληψία λόγω κοινωνικών προκαταλήψεων. Όπως και στον δημόσιο τομέα, οι ιδιωτικές εταιρείες υποχρεούνται να καταγράφουν αναλυτικά τα συστήματα ΤΝ που χρησιμοποιούν σε ένα μητρώο.

Επιπλέον, ο νόμος τονίζει ότι καμία διαδικασία που προβλέπεται σε αυτόν δεν αντικαθιστά τις διαδικασίες που επιβάλλει ο Γενικός Κανονισμός για την Προστασία Δεδομένων (ΓΚΠΔ) ενώ προβλέπει και τη σύσταση τριών επιτροπών: της Συντονιστικής Επιτροπής, η οποία θα συντονίζει την ανάπτυξη της ΤΝ, της Επιτροπής Εποπτείας, η οποία θα λειτουργεί ως εκτελεστικό όργανο της Συντονιστικής Επιτροπής, και του Παρατηρητηρίου, το οποίο θα αναλαμβάνει την εποπτεία της κινητικότητας στον τομέα της ΤΝ.

Συνολικά, το ελληνικό νομικό πλαίσιο ευθυγραμμίζεται με τις αρχές της Ευρωπαϊκής Ένωσης, δίνοντας έμφαση στη λογοδοσία, τη διαφάνεια και την προστασία των ατομικών δικαιωμάτων. Ωστόσο, η εφαρμογή και η επιβολή του νόμου παραμένουν σε πρώιμα στάδια σε σύγκριση με πιο καθιερωμένα πλαίσια, όπως αυτά της ΕΕ.

2.3 Νομοθεσίας στις ΗΠΑ

Οι Ηνωμένες Πολιτείες εξελίσσουν ενεργά την προσέγγισή τους στη νομοθεσία για την τεχνητή νοημοσύνη, με ένα μείγμα πρωτοβουλιών σε ομοσπονδιακό και πολιτειακό επίπεδο. Ωστόσο, η έλλειψη ενός ολοκληρωμένου ομοσπονδιακού νόμου για την ΤΝ έχει οδηγήσει σε ένα κατακερματισμένο ρυθμιστικό τοπίο, το οποίο περιπλέκεται περαιτέρω από την αλλαγή στη διοίκηση που συνέβη το 2025 [24].

Στο ομοσπονδιακό επίπεδο, η νομοθεσία για την τεχνητή νοημοσύνη χαρακτηρίζεται από μια σειρά πρωτοβουλιών που στοχεύουν στην προώθηση της καινοτομίας, αλλά χωρίς να υπάρχει ένας ενιαίος νόμος που να ρυθμίζει τους κινδύνους. Ένα από τα πιο σημαντικά νομοθετικά κείμενα είναι ο **National Artificial Intelligence Initiative Act του 2020 (NAII)**. Αυτός ο νόμος επικεντρώνεται στην προώθηση της έρευνας, της ανάπτυξης και της καινοτομίας της τεχνητής

νοημοσύνης σε όλους τους βασικούς ομοσπονδιακούς οργανισμούς. Ο πρωταρχικός στόχος του είναι να εδραιώσει τη θέση των Ηνωμένων Πολιτειών ως παγκόσμιου ηγέτη στην καινοτομία της τεχνητής νοημοσύνης, και όχι να ρυθμίσει τους κινδύνους. Ο νόμος κατευθύνει οργανισμούς όπως το Υπουργείο Ενέργειας (DOE) και το Εθνικό Ίδρυμα Επιστημών (NSF) να προωθήσουν εργαλεία ΤΝ, να διεξάγουν μελέτες εργατικού δυναμικού και να αναπτύξουν εθελοντικά πρότυπα. Ωστόσο, αυτά τα πρότυπα αποσκοπούν κυρίως στην προώθηση της ηγετικής θέσης των ΗΠΑ στον τομέα της τεχνητής νοημοσύνης και όχι στη διασφάλιση της ασφάλειας ή της αξιοπιστίας, όπως συμβαίνει με την προσέγγιση της ΕΕ. Ο νόμος NAII υπογραμμίζει επίσης τη σημασία της αντανάκλασης των «αμερικανικών αξιών» στην ανάπτυξη της ΤΝ, αλλά δεν έχει δεσμευτικούς κανονισμούς για τις επιχειρήσεις.

Μια άλλη σημαντική εξέλιξη είναι η **Έκθεση της διακομματικής ομάδας εργασίας της Βουλής των Αντιπροσώπων για την τεχνητή νοημοσύνη (2024)**. Η έκθεση αυτή, που δημοσιεύθηκε τον Δεκέμβριο του 2024, χρησιμεύει ως οδηγός για το Κογκρέσο στην αντιμετώπιση των εξελίξεων στην τεχνητή νοημοσύνη. Η έκθεση τονίζει την ανάγκη να προωθηθεί η καινοτομία της τεχνητής νοημοσύνης με ταυτόχρονη προστασία από τους κινδύνους, την ενθάρρυνση πολιτικών για συγκεκριμένους τομείς και τη σταδιακή προσέγγιση της ρύθμισης. Υπογραμμίζει επίσης τη σημασία της διατήρησης του ανθρώπου στο επίκεντρο της πολιτικής για την ΤΝ, της διασφάλισης της ηθικής χρήσης και της αντιμετώπισης των επιπτώσεων στο εργατικό δυναμικό. Αν και δεν είναι νομικά δεσμευτική, η έκθεση παρέχει ένα πλαίσιο για τη μελλοντική νομοθεσία για την ΤΝ και υπογραμμίζει την ανάγκη για προσεκτική, βασισμένη στον κίνδυνο διακυβέρνηση που εξισορροπεί την καινοτομία με την ασφάλεια.

Στο πλαίσιο των **εκτελεστικών διαταγμάτων,** η κυβέρνηση Μπάιντεν (2021-2025) εξέδωσε διάφορα διατάγματα, μεταξύ των οποίων το εκτελεστικό διάταγμα για την «ασφαλή, ασφαλή και αξιόπιστη ανάπτυξη και χρήση της ΤΝ» (Executive Order on "Safe, Secure, and Trustworthy Development and Use of AI") και το νομοσχέδιο για την ΤΝ (AI Bill of Rights), το οποίο επικεντρώθηκε στη ρύθμιση των κινδύνων της ΤΝ και στην προώθηση της ηθικής χρήσης. Αυτές οι πολιτικές κατεύθυναν τις ομοσπονδιακές υπηρεσίες να εφαρμόσουν νέες κατευθυντήριες γραμμές, να διορίσουν αξιωματούχους ΤΝ και να συμμετέχουν σε διεθνείς συνεργασίες. Ωστόσο, με την ανάληψη των καθηκόντων του τον Ιανουάριο του 2025, ο Τραμπ ανακάλεσε τις περισσότερες από τις πολιτικές του Μπάιντεν για την ΤΝ, εκπληρώνοντας μια προεκλογική υπόσχεση για την άρση των εμποδίων στην καινοτομία της ΤΝ. Ειδικότερα, από τα ελάχιστα που έχουν ακόμα παραμείνει άθικτα είναι το εκτελεστικό διάταγμα 14141 (Υποδομές ΤΝ) και το εκτελεστικό διάταγμα 14144 (Κυβερνοασφάλεια). Στη θέση τους, ο Τραμπ υπέγραψε το εκτελεστικό διάταγμα: «Άρση των εμποδίων για την αμερικανική ηγεσία στην τεχνητή νοημοσύνη», το οποίο θέτει ως προτεραιότητα τη διατήρηση της παγκόσμιας κυριαρχίας των ΗΠΑ στον τομέα της τεχνητής νοημοσύνης, την προώθηση της οικονομικής ανταγωνιστικότητας και τη διασφάλιση ότι η ανάπτυξη της τεχνητής νοημοσύνης είναι απαλλαγμένη από "ιδεολογικές προκαταλήψεις ή σχεδιασμένες κοινωνικές ατζέντες". Η εν λόγω εντολή αναθέτει επίσης τη δημιουργία ενός σχεδίου δράσης για την τεχνητή νοημοσύνη εντός 180 ημερών για την καθοδήγηση της μελλοντικής πολιτικής για την τεχνητή νοημοσύνη.

Σε πολιτειακό επίπεδο, η νομοθεσία για την τεχνητή νοημοσύνη έχει αρχίσει να αναπτύσσεται με πιο συγκεκριμένα ρυθμιστικά πλαίσια. Ένα από τα πιο αξιοσημείωτα παραδείγματα είναι ο **Νόμος του Κολοράντο για την Τεχνητή Νοημοσύνη (2024)**. Αυτός ο νόμος αποτελεί την πρώτη ολοκληρωμένη νομοθεσία για την τεχνητή νοημοσύνη στις ΗΠΑ, υιοθετώντας μια προσέγγιση με βάση τον κίνδυνο, παρόμοια με την πράξη της ΕΕ για την τεχνητή νοημοσύνη. Στοχεύει κυρίως στους προγραμματιστές και τους φορείς ανάπτυξης συστημάτων ΤΝ υψηλού κινδύνου, απαιτώντας διαφάνεια, μετριασμό των κινδύνων και κατάλληλη τεκμηρίωση. Οι επιχειρήσεις πρέπει να γνωστοποιούν τον σκοπό, τις προβλεπόμενες χρήσεις, τα οφέλη και τους

περιορισμούς των συστημάτων τεχνητής νοημοσύνης τους, μαζί με υψηλού επιπέδου περιλήψεις των δεδομένων εκπαίδευσης και μέτρα για τον μετριασμό των προκαταλήψεων. Οι προγραμματιστές υποχρεούνται επίσης να θεσπίσουν πολιτικές και διαδικασίες διαχείρισης κινδύνου, συχνά ευθυγραμμισμένες με πρότυπα του κλάδου, όπως το πλαίσιο διαχείρισης κινδύνου NIST AI Risk Management Framework ή τα σχετικά πρότυπα ISO. Ο νόμος υπογραμμίζει τη σημασία των αξιολογήσεων των επιπτώσεων της ΤΝ και των δεικτών συμμόρφωσης για την απόδειξη της εύλογης μέριμνας για τον μετριασμό των αλγοριθμικών διακρίσεων. Θέτοντας αυτές τις απαιτήσεις, ο νόμος του Κολοράντο για την τεχνητή νοημοσύνη παρέχει ένα θεμελιώδες πλαίσιο για τη μελλοντική ρύθμιση της τεχνητής νοημοσύνης στις ΗΠΑ.

Επιπλέον, η **Πολιτική του Ανώτατου Δικαστηρίου του Ιλινόις για την τεχνητή νοημοσύνη (2025)** αποτελεί ένα άλλο σημαντικό βήμα στη ρύθμιση της ΤΝ σε πολιτειακό επίπεδο. Αυτή η πολιτική, που τέθηκε σε ισχύ από την 1η Ιανουαρίου 2025, αφορά την υπεύθυνη ενσωμάτωση της ΤΝ στα δικαστικά συστήματα. Αναγνωρίζοντας τις ραγδαίες εξελίξεις στις τεχνολογίες γενετικής τεχνητής νοημοσύνης, η πολιτική περιγράφει κατευθυντήριες γραμμές για τη δεοντολογική χρήση, τη λογοδοσία και τη διασφάλιση της δικαστικής ακεραιότητας. Τονίζει τη σημασία της εκπαίδευσης και της προσαρμογής των επαγγελματιών του νομικού κλάδου που χρησιμοποιούν ΤΝ, διασφαλίζοντας ότι τα εργαλεία ΤΝ χρησιμοποιούνται υπεύθυνα στις δικαστικές διαδικασίες. Η πολιτική υπογραμμίζει επίσης την ανάγκη για δικαστική υπευθυνότητα και επαγγελματική συμπεριφορά κατά την ενσωμάτωση της ΤΝ στις νομικές διαδικασίες. Με την προώθηση των δεοντολογικών προτύπων και της εξουσιοδοτημένης χρήσης της ΤΝ στο δικαστικό σώμα, το Ιλινόις τοποθετείται ως ηγέτης στη δικαστική διακυβέρνηση της ΤΝ.

Το νομοθετικό τοπίο των ΗΠΑ στον τομέα της τεχνητής νοημοσύνης παραμένει κατακερματισμένο, με έμφαση στην καινοτομία και όχι στη συνολική ρύθμιση. Ενώ οι ομοσπονδιακές προσπάθειες, όπως ο νόμος NAII και το εκτελεστικό διάταγμα του Τραμπ, δίνουν προτεραιότητα στην ηγεσία της ΤΝ, πρωτοβουλίες σε πολιτειακό επίπεδο, όπως ο νόμος του Κολοράντο για την ΤΝ και η πολιτική ΤΝ του Ανώτατου Δικαστηρίου του Ιλινόις, παρέχουν σαφέστερα ρυθμιστικά πλαίσια. Οι επιχειρήσεις πρέπει να περιηγηθούν σε αυτό το πολύπλοκο περιβάλλον, διασφαλίζοντας τη συμμόρφωση με τους υφιστάμενους νόμους και προετοιμάζοντας τις μελλοντικές εξελίξεις στη διακυβέρνηση της ΤΝ. Καθώς η τεχνητή νοημοσύνη συνεχίζει να μεταμορφώνει τους κλάδους, οι ΗΠΑ θα πρέπει να εξισορροπήσουν την καινοτομία με τη διαχείριση των κινδύνων για να διατηρήσουν την παγκόσμια ηγετική τους θέση στον τομέα της τεχνητής νοημοσύνης.

2.4 Κινέζικη Νομοθεσία

Η Κίνα έχει εισαγάγει διάφορους βασικούς κανονισμούς για τη ρύθμιση των τεχνολογιών τεχνητής νοημοσύνης (ΤΝ), εξισορροπώντας την καινοτομία με την ασφάλεια, τις ηθικές εκτιμήσεις και τη δημόσια ασφάλεια. Αυτοί οι κανονισμοί αποσκοπούν τόσο στην προώθηση της ανάπτυξης της τεχνολογίας όσο και στην αντιμετώπιση των κινδύνων που μπορεί να προκύψουν από τη χρήση της. Παρακάτω παρουσιάζεται μια αναλυτική περιγραφή των κύριων διατάξεων και των επιπτώσεών τους [25][26].

Με ισχύ από τις 15 Αυγούστου 2023, τα **Γενετικά Μέτρα Τεχνητής Νοημοσύνης** ρυθμίζουν τις γεννητικές υπηρεσίες τεχνητής νοημοσύνης που προσφέρονται στο κοινό στην Κίνα, συμπεριλαμβανομένων εκείνων που παρέχονται από ξένες εταιρείες. Αυτά τα μέτρα στοχεύουν στην εξισορρόπηση της καινοτομίας και της ασφάλειας, ενθαρρύνοντας τις επενδύσεις σε τεχνολογίες ΤΝ, όπως αλγόριθμοι, τσιπ και πλατφόρμες λογισμικού. Ένα από τα κύρια χαρακτηριστικά αυτών

των μέτρων είναι η έμφαση στην προστασία της ιδιωτικής ζωής των δεδομένων, καθώς και στον έλεγχο του περιεχομένου και στη διαφάνεια των υπηρεσιών τεχνητής νοημοσύνης. Επιπλέον, οι ξένες εταιρείες που επιθυμούν να εισέλθουν στην κινεζική αγορά οφείλουν να συμμορφώνονται και με τους νόμους περί ξένων επενδύσεων, γεγονός που επιβάλλει πρόσθετες απαιτήσεις στη διαδικασία εγκατάστασης και λειτουργίας τους στην Κίνα.

Οι **Διατάξεις Βαθιάς Σύνθεσης,** που τέθηκαν σε ισχύ από τις 10 Ιανουαρίου 2023, ρυθμίζουν τη χρήση τεχνολογιών βαθιάς σύνθεσης, όπως τα deepfakes, για τη δημιουργία ή την αλλοίωση διαδικτυακού περιεχομένου. Αυτές οι διατάξεις απαιτούν από τους παρόχους υπηρεσιών να επισημαίνουν το περιεχόμενο που παράγεται με τεχνητή νοημοσύνη (π.χ. με ετικέτες όπως «Παράγεται από τεχνητή νοημοσύνη»). Επιπλέον, οι πάροχοι οφείλουν να δημιουργήσουν συστήματα διαχείρισης για την εγγραφή των χρηστών, την ασφάλεια των δεδομένων και τις δεοντολογικές αναθεωρήσεις. Ο κύριος στόχος αυτών των διατάξεων είναι η πρόληψη της διάδοσης παράνομων πληροφοριών και η διατήρηση ενός υγιούς οικοσυστήματος στον κυβερνοχώρο, προστατεύοντας ταυτόχρονα το κοινό από παραπλανητικό ή επιβλαβές περιεχόμενο.

Τα **Μέτρα Δεοντολογικής Αναθεώρησης**, που τέθηκαν σε ισχύ από την 1η Δεκεμβρίου 2023, αντιμετωπίζουν τις δεοντολογικές προκλήσεις που προκύπτουν από τις επιστημονικές και τεχνολογικές δραστηριότητες, συμπεριλαμβανομένης της ανάπτυξης της τεχνητής νοημοσύνης. Αυτά τα μέτρα απαιτούν από τις εταιρείες να συστήνουν εσωτερικές επιτροπές επανεξέτασης για την ηθικά ευαίσθητη έρευνα. Επιπλέον, υποχρεώνουν τις εταιρείες να υποβάλλουν τις δραστηριότητές τους σε εξωτερική επαλήθευση από εμπειρογνώμονες, ειδικά σε περιπτώσεις όπου οι δραστηριότητες ΤΝ μπορεί να επηρεάσουν την κοινή γνώμη ή την κοινωνική συνείδηση. Αυτή η διαδικασία εξασφαλίζει ότι οι τεχνολογικές εξελίξεις γίνονται με τρόπο που σέβεται τις ηθικές αρχές και τις κοινωνικές αξίες.

Οι **Διατάξεις Σύστασης Αλγορίθμου**, που τέθηκαν σε ισχύ από την 1η Μαρτίου 2022, επιβάλλουν την κατάθεση αλγορίθμων στη Διοίκηση Κυβερνοχώρου της Κίνας για αλγορίθμους που επηρεάζουν την κοινή γνώμη ή την κοινωνική δέσμευση. Η κατάθεση αυτή πρέπει να ολοκληρωθεί εντός 10 εργάσιμων ημερών από την παροχή της υπηρεσίας, με αυστηρές κυρώσεις για τη μη συμμόρφωση. Μέχρι τις 30 Ιουνίου 2024, έχουν κατατεθεί πάνω από 1.400 αλγόριθμοι, γεγονός που αντικατοπτρίζει την ταχεία ανάπτυξη της τεχνητής νοημοσύνης στην Κίνα και την αυξανόμενη ευαισθητοποίηση των εταιρειών για την τήρηση των κανονισμών.

Το κανονιστικό πλαίσιο της Κίνας για την τεχνητή νοημοσύνη εξελίσσεται με ταχείς ρυθμούς, αντανακλώντας τη δέσμευση της χώρας να προωθήσει την καινοτομία, ενώ ταυτόχρονα αντιμετωπίζει ηθικές, ασφαλιστικές και κοινωνικές ανησυχίες. Οι επιχειρήσεις που δραστηριοποιούνται στην κινεζική αγορά ΤΝ πρέπει να περιηγηθούν προληπτικά σε αυτούς τους κανονισμούς, διασφαλίζοντας τη συμμόρφωση, μετριασμό των κινδύνων και αξιοποίηση των ευκαιριών. Με την αυστηρότερη επιβολή της νομοθεσίας και την αυξανόμενη ευαισθητοποίηση του κοινού για τους κινδύνους που σχετίζονται με την ΤΝ, οι εταιρείες θα πρέπει να δώσουν προτεραιότητα στη διαφάνεια, τις ηθικές πρακτικές και τις ισχυρές στρατηγικές συμμόρφωσης.

2.5 Συμπεράσματα και Συγκρίσεις

Η ανάπτυξη εμπεριστατωμένων και ολοκληρωμένων νόμων γύρω από την τεχνητή νοημοσύνη είναι επιτακτική ανάγκη, καθώς αποτελεί θεμελιώδες ζήτημα ανθρωπίνων δικαιωμάτων, ασφάλειας και κοινωνικής εμπιστοσύνης. Οι τεχνολογίες τεχνητής νοημοσύνης έχουν τη δυνατότητα να μετασχηματίσουν τις βιομηχανίες και να βελτιώσουν τις ζωές, αλλά ενέχουν επίσης σημαντικούς κινδύνους, συμπεριλαμβανομένων των προκαταλήψεων, των διακρίσεων και των

απειλών κατά της ιδιωτικής ζωής και της ασφάλειας. Η αποτελεσματική ρύθμιση είναι απαραίτητη για τον μετριασμό αυτών των κινδύνων και την ταυτόχρονη προώθηση της καινοτομίας.

Όπως παρατηρούμε από τις αντίστοιχες νομοθεσίες που μελετήσαμε προηγουμένως οι διάφορες χώρες βρίσκονται σε διαφορετικά στάδια επίτευξης αυτής της ισορροπίας. Η Ευρωπαϊκή Ένωση ξεχωρίζει ως παγκόσμιος ηγέτης στην ηθική διακυβέρνηση της τεχνητής νοημοσύνης, με την ολοκληρωμένη και εστιασμένη στα ανθρώπινα δικαιώματα προσέγγισή της. Η Κίνα, όντας ιδιαίτερα προηγμένη στο ρυθμιστικό της πλαίσιο και την επιβολή του, δίνει προτεραιότητα στον κρατικό έλεγχο και την ασφάλεια. Οι Ηνωμένες Πολιτείες, παρά την ηγετική τους θέση στην καινοτομία της ΤΝ, δεν διαθέτουν συνεκτικό ομοσπονδιακό πλαίσιο, με αποτέλεσμα μια κατακερματισμένη και λιγότερο ολοκληρωμένη προσέγγιση στη ρύθμιση. Η Ελλάδα, ευθυγραμμιζόμενη με τις αρχές της ΕΕ, βρίσκεται στα πρώτα στάδια της εφαρμογής του πλαισίου της για την ΤΝ, εστιάζοντας στη διαφάνεια και τη λογοδοσία.

Συμπερασματικά**,** η παγκόσμια κοινότητα πρέπει να δώσει προτεραιότητα στην ανάπτυξη ισχυρών, ηθικών και διαφανών κανονισμών για την τεχνητή νοημοσύνη, ώστε να διασφαλιστεί ότι οι τεχνολογικές εξελίξεις ευθυγραμμίζονται με τα ανθρώπινα δικαιώματα και τις κοινωνικές αξίες. Ενώ σε συγκεκριμένες περιπτώσεις, όπως αυτή της Κίνας και της ΕΕ, έχει σημειωθεί σημαντική πρόοδος, σε άλλες, όπως οι ΗΠΑ, υπάρχει ανάγκη για πιο ενοποιημένη εργασία και ολοκληρωμένα πλαίσια. Μόνο μέσω τέτοιων προσπαθειών μπορούμε να αξιοποιήσουμε τα οφέλη της τεχνητής νοημοσύνης και ταυτόχρονα να διασφαλίσουμε τα δικαιώματα και την ασφάλεια των ατόμων παγκοσμίως.

Κεφάλαιο 3: Προσεγγίσεις στην Αλγοριθμική Εκτίμηση Αντικτύπου

3.1 Εργαλείο Αλγοριθμικής Εκτίμησης Αντικτύπου του Καναδά

Το εργαλείο Αλγοριθμικής Εκτίμησης Αντικτύπου του Καναδά είναι ένα εργαλείο αξιολόγησης κινδύνων που αναπτύχθηκε για την κυβερνητική επίβλεψη συστημάτων αυτοματοποιημένης λήψης αποφάσεων [2]. Πρόκειται για ένα ερωτηματολόγιο που έχει σχεδιαστεί για την αξιολόγηση των κινδύνων και των επιπτώσεων των συστημάτων αυτών βοηθώντας τις κυβερνητικές υπηρεσίες και τους οργανισμούς να διαχειριστούν αποτελεσματικά αυτούς τους κινδύνους. Το εργαλείο αποτελείται από 51 ερωτήσεις κινδύνου και 34 ερωτήσεις μετριασμού, οι οποίες καλύπτουν τομείς όπως ο σχεδιασμός του έργου, οι δυνατότητες του συστήματος, οι αλγόριθμοι, οι τύποι αποφάσεων, οι επιπτώσεις και η χρήση δεδομένων.

Το εργαλείο υπολογίζει μια ακατέργαστη βαθμολογία επιπτώσεων (raw impact score) με βάση τους τομείς κινδύνου και μια βαθμολογία μετριασμού (mitigation score) με βάση τα μέτρα για τη μείωση των κινδύνων. Εάν η βαθμολογία μετριασμού είναι μικρότερη από το 80% της μέγιστης βαθμολογίας, χρησιμοποιείται σαν τελικό αποτέλεσμα η ακατέργαστη βαθμολογία επιπτώσεων. Εάν είναι 80% ή υψηλότερη, το 15% αφαιρείται από την ακατέργαστη βαθμολογία επιπτώσεων. Τα επίπεδα επιπτώσεων κυμαίνονται από το Επίπεδο I (μικρή επίπτωση) έως το Επίπεδο IV (πολύ μεγάλη επίπτωση), με βάση το ποσοστό τελικής βαθμολογίας (0%-100%). Οι τομείς κινδύνου που αξιολογούνται περιλαμβάνουν το έργο, το σύστημα, τον αλγόριθμο, την απόφαση, τον αντίκτυπο και τα δεδομένα, ενώ οι τομείς μετριασμού επικεντρώνονται στη διαβούλευση, την ποιότητα των δεδομένων, τη διαδικαστική δικαιοσύνη και την προστασία της ιδιωτικής ζωής.

Το εργαλείο πρέπει να ολοκληρώνεται κατά τη φάση σχεδιασμού ενός έργου και πριν από την παραγωγή για την επικύρωση του συστήματος. Πρέπει επίσης να επανεξετάζεται και να επικαιροποιείται τακτικά, ιδίως όταν αλλάζει η λειτουργικότητα ή το πεδίο εφαρμογής του συστήματος. Κατά την ολοκλήρωση της εκτίμησης αντικτύπου, είναι σημαντικό να συλλέγονται λεπτομερείς πληροφορίες σχετικά με το έργο, συμπεριλαμβανομένου του πλαισίου απόφασης, της ευπάθειας του πελάτη, των λεπτομερειών του αλγορίθμου, των πηγών δεδομένων και των διαβουλεύσεων με τους ενδιαφερόμενους φορείς.

Μόλις ολοκληρωθούν, τα τελικά αποτελέσματα της εκτίμησης αντικτύπου πρέπει να δημοσιεύονται στην Πύλη Ανοικτής Διακυβέρνησης σε προσβάσιμη μορφή.Έτσι διασφαλίζεται ότι τα αυτοματοποιημένα συστήματα λήψης αποφάσεων αναπτύσσονται και εφαρμόζονται με κατάλληλη διαχείριση κινδύνων, διαφάνεια και λογοδοσία, ευθυγραμμιζόμενα με τα πρότυπα δεοντολογίας, νομικής και πολιτικής.

3.2 Μέθοδος CapAI

Σκοπός του capAI είναι να παρέχει μια διαδικασία αξιολόγησης της συμμόρφωσης για τα συστήματα τεχνητής νοημοσύνης που να διασφαλίζει ότι είναι νομικά συμβατά, ηθικά ορθά και τεχνικά στιβαρά σύμφωνα με την πράξη της ΕΕ για την τεχνητή νοημοσύνη (AI Act) [1]. Χρησιμεύει ως εργαλείο διακυβέρνησης για να βοηθήσει τους οργανισμούς να επιδείξουν την τήρηση καθορισμένων ηθικών αξιών μεταφράζοντας αυτές τις αξίες από ασαφής ιδέες σε πρακτικά και μετρήσιμα κριτήρια.

Το capAI θα πρέπει να χρησιμοποιείται για συστήματα ΤΝ υψηλής επικινδυνότητας που απαιτούν αξιολογήσεις συμμόρφωσης σύμφωνα με το AI Act, καθώς και για συστήματα ΤΝ χαμηλής επικινδυνότητας που δεν εμπίπτουν στο ρυθμιστικό πεδίο εφαρμογής αλλά μπορούν να επωφεληθούν από εθελοντικούς κώδικες δεοντολογίας. Επιπλέον, το capAI μπορεί να χρησιμοποιηθεί για ενσωματωμένα συστήματα ΤΝ εντός προϊόντων ή υπηρεσιών που ρυθμίζονται βάσει τομεακών κανονισμών. Η διαδικασία αξιολόγησής του περιλαμβάνει τρία βασικά έγγραφα: Το πρωτόκολλο εσωτερικής αναθεώρησης (Internal Review Protocol) , το συνοπτικό δελτίο δεδομένων (Summary Datasheet) και τον εξωτερικό πίνακα αποτελεσμάτων. (External Scorecard).

Το **Πρωτόκολλο Εσωτερικής Επισκόπησης (IRP)** είναι ένα δομημένο έγγραφο που καθοδηγεί τους οργανισμούς στην αξιολόγηση της ευαισθητοποίησης, των επιδόσεων και των πόρων που απαιτούνται για την πρόληψη και την αντιμετώπιση πιθανών αστοχιών στα συστήματα ΤΝ καθ' όλη τη διάρκεια του κύκλου ζωής τους. Δημιουργείται ακολουθώντας έναν κατάλογο ελέγχου που εξετάζει βασικά ερωτήματα σχετικά με την οργανωτική διακυβέρνηση και τη συγκεκριμένη περίπτωση χρήσης του συστήματος ΤΝ, με τη συμμετοχή βασικών παραγόντων, όπως ανώτατα στελέχη, ιδιοκτήτες προϊόντων, διαχειριστές έργων και επιστήμονες δεδομένων σε κάθε στάδιο ανάπτυξης. Το IRP έχει σχεδιαστεί για να ολοκληρώνεται χρονολογικά, από το στάδιο του σχεδιασμού έως το στάδιο της απόσυρσης του συστήματος ΤΝ.

Το **Συνοπτικό Δελτίο Δεδομένων (SDS)** είναι ένα συνοπτικό έγγραφο υψηλού επιπέδου που περιγράφει τον σκοπό, τη λειτουργικότητα και τις επιδόσεις ενός συστήματος ΤΝ, το οποίο πληροί τις απαιτήσεις δημόσιας καταχώρισης σύμφωνα με το AI Act. Δημιουργείται με τη συγκέντρωση συγκεκριμένων πληροφοριών που απαιτούνται σύμφωνα με το Παράρτημα VIII του AI Act, συμπεριλαμβανομένων λεπτομερειών όπως τα στοιχεία επικοινωνίας του παρόχου, η εμπορική ονομασία του συστήματος ΤΝ, ο σκοπός για τον οποίο προορίζεται, η κατάσταση και τυχόν σχετικές λεπτομέρειες πιστοποίησης. Το SDS χρησιμεύει ως συνοπτική αναφορά για τα ενδιαφερόμενα μέρη και υποβάλλεται στη δημόσια βάση δεδομένων της ΕΕ για τα συστήματα ΤΝ υψηλού κινδύνου.

Ο **Εξωτερικός πίνακας αποτελεσμάτων (ESC)** είναι ένα δημόσιο έγγραφο αναφοράς που συνοψίζει βασικές πληροφορίες σχετικά με ένα σύστημα ΤΝ, εστιάζοντας στον σκοπό, τις αξίες, τα δεδομένα και τη διακυβέρνησή του, για να καταδείξει τη δεοντολογική συμμόρφωση και τις ορθές πρακτικές. Δημιουργείται με τη δημιουργία του μέσω του IRP, το οποίο πρέπει να ολοκληρωθεί πρώτα, καθώς το ESC βασίζεται σε ελεγμένα συνοπτικά δεδομένα από το IRP. Το ESC έχει σχεδιαστεί για να είναι διαθέσιμο στους πελάτες και τα ενδιαφερόμενα μέρη, παρέχοντας διαφάνεια χωρίς να αποκαλύπτει ευαίσθητες ή ανταγωνιστικές πληροφορίες.

3.3 Άλλα εργαλεία Αλγοριθμική Εκτίμησης Αντικτύπου

Παρακάτω θα περιγράψουμε μερικά άλλα διάσημα εργαλεία αλγοριθμικής εκτίμησης αντικτύπου για να συλλέξουμε πληροφορίες για τους διάφορους τρόπους εφαρμογής της ιδέας πίσω από ένα ΑΕΑ. Για να μπορέσει η ανάλυση να είναι πιο δομημένη θα επικεντρωθούμε σε 8 βασικούς άξονες βασισμένους στην μεθοδολογία του “A systematic review of artificial intelligence impact assessments”[27]. Οι άξονες αυτοί έχουν ως εξής :

- **Σκοπός** : Γιατί δημιουργήθηκε αυτή η ΑΕΑ? Ποια θετική επίδραση θα έχει στην κοινωνία αν χρησιμοποιηθει σωστα;
- **Πεδίο εφαρμογής** : Για ποιά συστήματα συστήνεται/είναι σχεδιασμένο αυτή η ΑΕΑ;
- **Θεματικές** : Ποιές θεματικές καλύπτει αυτή η ΑΕΑ; (π.χ. ανθρώπινα δικαιώματα, ηθική, προστασία δεδομένων, ιδιωτικότητα, ασφάλεια,προστασία , περιβάλλον, εργασιακά δικαιώματα, δημοκρατία)

- **Οργάνωση** : Είναι η διαδικασία εκτίμησης αντικτύπου ενσωματωμένη σε άλλες οργανωτικές δομές της εταιρείας ή είναι κάτι ξεχωριστό και ανεξάρτητο;
- **Χρονοδιάγραμμα** : Πότε ολοκληρώνεται η ΑΕΑ; Πότε πρέπει να ανανεωθεί; Αφορά μακροπρόθεσμους ή βραχυπρόθεσμους κινδύνους ;
- **Διαδικασία και μέθοδοι :** Ποιά είναι η διαδικασία ολοκλήρωσης της ΑΕΑ;
- **Διαφάνεια :** Περιλαμβάνει ερωτήσεις που σχετίζονται με την επεξηγησιμότητα και την διαφάνεια του συστήματος ;
- **Προκλήσεις :** Έγιναν συμβιβασμοί μεταξύ καινοτομίας και ρυθμιστικής πολιτικής;

3.3.1 Ada Lovelace Institute

Η αξιολόγηση αλγοριθμικών επιπτώσεων του Ινστιτούτου Ada Lovelace είναι ένα πλαίσιο ΑΕΑ που αναπτύχθηκε το 2022 και είναι ειδικά προσαρμοσμένο για την Εθνική Πλατφόρμα Ιατρικής Απεικόνισης (NMIP) [28]. Η αξιολόγηση αυτή αποσκοπεί στην αξιολόγηση των πιθανών κοινωνικών επιπτώσεων των αλγοριθμικών συστημάτων στην υγειονομική περίθαλψη πριν από την ανάπτυξή τους, προωθώντας τη λογοδοσία και τη δημόσια εμπιστοσύνη. Η μοναδική εστίασή της έγκειται στη γεφύρωση του χάσματος μεταξύ των προγραμματιστών του ιδιωτικού τομέα και των αναγκών του δημόσιου τομέα, γεγονός που την καθιστά μια νέα μελέτη περίπτωσης στον τομέα των αξιολογήσεων αλγοριθμικών επιπτώσεων.

Σκοπός της παρούσας εκτίμησης επιπτώσεων είναι η αξιολόγηση των πιθανών κοινωνικών επιπτώσεων των αλγοριθμικών συστημάτων που χρησιμοποιούνται στην υγειονομική περίθαλψη πριν από την ανάπτυξή τους, διασφαλίζοντας τη λογοδοσία, τη διαφάνεια και την εμπιστοσύνη του κοινού. Με τον εντοπισμό των πιθανών βλαβών και οφελών, η αξιολόγηση βοηθά τις ομάδες έργου να μετριάσουν τους κινδύνους και να μεγιστοποιήσουν τα θετικά αποτελέσματα των συστημάτων τους. Η διαδικασία αυτή ωφελεί την κοινωνία, καθώς προωθεί την υπεύθυνη ανάπτυξη της ΤΝ, διασφαλίζοντας ότι οι τεχνολογίες υγειονομικής περίθαλψης είναι ασφαλείς, δίκαιες και ευθυγραμμισμένες με τα ηθικά πρότυπα, βελτιώνοντας τελικά τη φροντίδα των ασθενών και την εμπιστοσύνη του κοινού στις λύσεις υγειονομικής περίθαλψης που βασίζονται στην ΤΝ.

Η εν λόγω εκτίμηση επιπτώσεων έχει σχεδιαστεί για αλγοριθμικά συστήματα που χρησιμοποιούν δεδομένα από την Εθνική Πλατφόρμα Ιατρικής Απεικόνισης (NMIP). Εφαρμόζεται σε έργα που περιλαμβάνουν έρευνα, εκπαίδευση ή δοκιμή ιατρικών προϊόντων που χρησιμοποιούν δεδομένα απεικόνισης NMIP. Η αξιολόγηση είναι προσαρμοσμένη για συστήματα που θα μπορούσαν να επηρεάσουν σημαντικά τη φροντίδα των ασθενών, τις ροές εργασίας στον τομέα της υγειονομικής περίθαλψης και τα ευρύτερα κοινωνικά αποτελέσματα, διασφαλίζοντας ότι τα συστήματα αυτά εξετάζονται λεπτομερώς για τις πιθανές επιπτώσεις τους στα άτομα και τις κοινότητες .

Η αξιολόγηση καλύπτει ένα ευρύ φάσμα θεμάτων, συμπεριλαμβανομένων ηθικών προβληματισμών όπως η προστασία της ιδιωτικής ζωής των δεδομένων, η συναίνεση, η αυτονομία και η επιτήρηση. Αντιμετωπίζει επίσης πιθανές προκαταλήψεις, διακρίσεις και ανισότητες που θα μπορούσαν να προκύψουν από το σύστημα. Επιπλέον, η αξιολόγηση εξετάζει τις περιβαλλοντικές επιπτώσεις, όπως η κατανάλωση ενέργειας που απαιτείται για την εκπαίδευση και την εκτέλεση μοντέλων τεχνητής νοημοσύνης, και τις πιθανές επιπτώσεις στις σχέσεις μεταξύ ασθενών και επαγγελματιών υγείας.

Η αξιολόγηση είναι ενσωματωμένη στην οργανωτική δομή του εργαστηρίου ΤΝ του NHS και αποτελεί μέρος της διαδικασίας πρόσβασης στα δεδομένα του NMIP. Δεν είναι αυτόνομη, αλλά εμπλέκει πολλαπλά ενδιαφερόμενα μέρη, συμπεριλαμβανομένων των ομάδων έργου, της επιτροπής πρόσβασης στα δεδομένα NMIP (DAC) και του κοινού σε συμμετοχικά εργαστήρια. Ενώ η αρχική

αναστοχαστική άσκηση ολοκληρώνεται από την ομάδα έργου, στο συμμετοχικό εργαστήριο συμμετέχουν εξωτερικοί ενδιαφερόμενοι και η τελική αξιολόγηση επανεξετάζεται από την DAC, εξασφαλίζοντας μια συνεργατική και πολυπρισματική προσέγγιση.

Η εκτίμηση επιπτώσεων πρέπει να ολοκληρώνεται κατά τη φάση ανάπτυξης του κύκλου ζωής της ΤΝ, πριν από την εγκατάσταση του συστήματος. Θα πρέπει να επικαιροποιείται τακτικά, ιδίως όταν υπάρχουν σημαντικές αλλαγές στο σύστημα, όπως αλλαγές στο πεδίο εφαρμογής, στην εφαρμογή ή στη βάση χρηστών. Η αξιολόγηση επανεξετάζεται επίσης μετά από μια σταθερή περίοδο δύο ετών ή όπως συνιστάται από την DAC, διασφαλίζοντας ότι η αξιολόγηση παραμένει σχετική και αντικατοπτρίζει την τρέχουσα κατάσταση του συστήματος.

Η διαδικασία αυτής της αξιολόγησης περιλαμβάνει επτά βήματα: 1) Η ομάδα έργου ολοκληρώνει μια άσκηση αναστοχασμού για τον εντοπισμό πιθανών επιπτώσεων. 2) Η ομάδα υποβάλλει αίτηση για πρόσβαση σε δεδομένα NMIP και η αίτηση φιλτράρεται από την DAC. 3) Το NHS AI Lab συντονίζει ένα συμμετοχικό εργαστήριο με εκπροσώπους του κοινού και των ασθενών για να συζητήσει τις πιθανές επιπτώσεις του έργου. 4) Η ομάδα συνθέτει τα ευρήματα του εργαστηρίου και επικαιροποιεί την άσκηση αναστοχασμού. 5) Η DAC εξετάζει την επικαιροποιημένη αξιολόγηση και λαμβάνει απόφαση για την πρόσβαση στα δεδομένα. 6) Η ολοκληρωμένη αξιολόγηση δημοσιεύεται στον δικτυακό τόπο του NMIP. 7) Η αξιολόγηση επανεξετάζεται επαναληπτικά καθώς εξελίσσεται το έργο.

Η αξιολόγηση αυτή περιλαμβάνει ερωτήσεις σχετικά με τη διαφάνεια, την ανιχνευσιμότητα και την επεξηγηματικότητα. Ρωτά πώς τα αποτελέσματα του συστήματος θα είναι εξηγήσιμα και ερμηνεύσιμα για τους χρήστες, πώς τα άτομα μπορούν να αμφισβητήσουν ή να προσβάλουν τα πορίσματα του συστήματος και πώς το σύστημα διασφαλίζει τη λογοδοσία. Οι ερωτήσεις αυτές διασφαλίζουν ότι το σύστημα είναι διαφανές στις λειτουργίες του και ότι οι αποφάσεις του μπορούν να γίνουν κατανοητές και να προσβληθούν από τους επηρεαζόμενους.

Η αξιολόγηση εξετάζει τις αναγκαίες αντισταθμίσεις, όπως αυτές μεταξύ καινοτομίας και ρύθμισης και μεταξύ διαφορετικών μεροληψιών. Προτρέπει τις ομάδες να αξιολογήσουν πώς το σύστημά τους θα μπορούσε να προκαλέσει βλάβη ακούσια ή εκούσια, μεταξύ άλλων μέσω προκαταλήψεων ή κακής χρήσης. Η αξιολόγηση ενθαρρύνει επίσης τις ομάδες να εξετάσουν πώς θα μετριάσουν αυτούς τους κινδύνους, διασφαλίζοντας ότι το σύστημα έχει σχεδιαστεί για να ελαχιστοποιήσει τη ζημία και ταυτόχρονα να μεγιστοποιήσει τα οφέλη. Αυτό περιλαμβάνει την αντιμετώπιση πιθανών μεροληψιών, τη διασφάλιση της δικαιοσύνης και την εξέταση των περιβαλλοντικών και κοινωνικών επιπτώσεων του συστήματος.

3.3.2 European Law Institute

Τώρα θα μελετήσουμε την έκθεση για την αξιολόγηση των επιπτώσεων των αλγοριθμικών συστημάτων λήψης αποφάσεων (Algorithmic Decision-Making Systems) που χρησιμοποιούνται από τη δημόσια διοίκηση αναπτύχθηκε από το Ευρωπαϊκό Ινστιτούτο Δικαίου (ELI) [29]. Η έκθεση αυτή περιγράφει μια τυποποιημένη διαδικασία για την αξιολόγηση των επιπτώσεων των συστημάτων ΤΝ στη δημόσια διοίκηση, με στόχο την ενίσχυση της διαφάνειας και της λογοδοσίας. Οι πρότυποι κανόνες δημιουργήθηκαν ώστε να μπορούν να προσαρμοστούν σε διάφορα νομικά πλαίσια εντός και εκτός της ΕΕ, αντικατοπτρίζοντας το εξελισσόμενο τοπίο της ρύθμισης της ΤΝ.

Ο σκοπός της εκτίμησης επιπτώσεων που περιγράφεται στο έγγραφο είναι να εκτιμηθούν οι πιθανές επιπτώσεις των αλγοριθμικών συστημάτων λήψης αποφάσεων (ADMS) που χρησιμοποιούνται από τη δημόσια διοίκηση στο κοινό. Η εν λόγω αξιολόγηση αποσκοπεί στην ευαισθητοποίηση σχετικά με τους κινδύνους που συνδέονται με τα ADMS, να επιτρέψει στις διοικητικές αρχές να λαμβάνουν τεκμηριωμένες αποφάσεις σχετικά με τη χρήση τους, να επιτρέψει

στους εμπειρογνώμονες και το κοινό να συμμετέχουν στη διαδικασία λήψης αποφάσεων, να ενισχύσει τη διαφάνεια και να καταστήσει τη δημόσια διοίκηση υπεύθυνη για τη χρήση αυτών των συστημάτων. Με τον τρόπο αυτό, επιδιώκει να διασφαλίσει ότι η ανάπτυξη των ADMS ευθυγραμμίζεται με τις δημοκρατικές αρχές, το κράτος δικαίου και το δικαίωμα στη χρηστή διοίκηση, ενισχύοντας τελικά την εμπιστοσύνη του κοινού στη χρήση τέτοιων τεχνολογιών.

Η αξιολόγηση έχει σχεδιαστεί για αλγοριθμικά συστήματα λήψης αποφάσεων (ADMS) που χρησιμοποιούνται από δημόσιες αρχές, τα οποία είναι πιθανό να έχουν σημαντικές επιπτώσεις στο κοινό, ενώ είναι ιδιαίτερα σημαντική για τα συστήματα υψηλού κινδύνου, όπως αυτά που χρησιμοποιούνται σε κρίσιμες υποδομές, στην επιλεξιμότητα κοινωνικής ασφάλειας ή στην προληπτική αστυνόμευση.

Η αξιολόγηση καλύπτει μια σειρά κρίσιμων θεμάτων, συμπεριλαμβανομένων των ανθρωπίνων δικαιωμάτων, τα οποία επικεντρώνονται στην προστασία των θεμελιωδών δικαιωμάτων, όπως η προστασία της ιδιωτικής ζωής και η απαγόρευση των διακρίσεων. Εξετάζει επίσης ηθικά ζητήματα που εκτείνονται πέρα από την απλή νομική συμμόρφωση, διασφαλίζοντας ότι τα συστήματα λειτουργούν εντός αποδεκτών ηθικών ορίων. Η προστασία των δεδομένων αποτελεί βασικό μέλημα, τονίζοντας την ανάγκη για εγγυήσεις για τα προσωπικά δεδομένα και την τήρηση των σχετικών νόμων, ενώ τα δικαιώματα προστασίας της ιδιωτικής ζωής πρέπει να διατηρούνται καθ' όλη τη διάρκεια του κύκλου ζωής του συστήματος. Επιπλέον, η αξιολόγηση αξιολογεί τους κινδύνους ασφάλειας που σχετίζονται με τα τρωτά σημεία του συστήματος και τις επιπτώσεις στην ασφάλεια, ιδίως σε περιβάλλοντα κρίσιμων υποδομών. Εξετάζονται επίσης οι περιβαλλοντικές επιπτώσεις, παράλληλα με τις επιπτώσεις στα εργασιακά δικαιώματα και τη δυναμική του εργατικού δυναμικού. Τέλος, η αξιολόγηση εξετάζει τη δυνητική επιρροή αυτών των συστημάτων στις δημοκρατικές διαδικασίες και τη δημόσια εμπιστοσύνη, διασφαλίζοντας ότι συμβάλλουν θετικά στην κοινωνία.

Η αξιολόγηση εντάσσεται σε μια ευρύτερη οργανωτική δομή, καθώς περιλαμβάνει συντονισμό με άλλες διαδικασίες, όπως οι κανονισμοί για την προστασία των δεδομένων ή οι κανονισμοί για την ασφάλεια των προϊόντων. Διεξάγεται κατά κύριο λόγο από την αρχή εφαρμογής, δηλαδή τη δημόσια αρχή που χρησιμοποιεί ή προτίθεται να χρησιμοποιήσει το ADMS. Ωστόσο, περιλαμβάνει επίσης τη συνεργασία με τους παρόχους συστημάτων και τους παρόχους δεδομένων, ενώ για τα συστήματα υψηλού κινδύνου, περιλαμβάνει έλεγχο από εμπειρογνώμονα από ανεξάρτητο συμβούλιο εμπειρογνωμόνων. Η εποπτική αρχή επιβλέπει την όλη διαδικασία, διασφαλίζοντας τη συμμόρφωση με τους κανόνες και παρέχοντας καθοδήγηση.

Η αξιολόγηση των επιπτώσεων πρέπει να ολοκληρώνεται κατά τη διάρκεια των φάσεων ανάπτυξης και ανάπτυξης του κύκλου ζωής της ΤΝ. Θα πρέπει να συνοδεύει την ανάπτυξη του συστήματος, ξεκινώντας αρκετά νωρίς ώστε να επιτρέπει ουσιαστικές αλλαγές αν εντοπιστούν προβλήματα. Η αξιολόγηση θα πρέπει να επικαιροποιείται κάθε φορά που υπάρχουν ενδείξεις για ουσιαστικές αρνητικές επιπτώσεις που δεν είχαν προβλεφθεί προηγουμένως, μετά από έξι μήνες χρήσης, εάν δεν έχει γίνει επανεξέταση τους τελευταίους τρεις μήνες, και στη συνέχεια κάθε δύο χρόνια. Με τον τρόπο αυτό διασφαλίζεται ότι η αξιολόγηση παραμένει επίκαιρη και αντικατοπτρίζει τυχόν αλλαγές στο σύστημα ή στο πλαίσιο χρήσης του.

Η διαδικασία της αξιολόγησης περιλαμβάνει διάφορα στάδια: διαλογή για να διαπιστωθεί αν είναι απαραίτητη η αξιολόγηση αντικτύπου,εκτίμηση των βασικών ζητημάτων που πρέπει να εξεταστούν, προετοιμασία έκθεσης αξιολόγησης των επιπτώσεων που περιλαμβάνει λεπτομερή περιγραφή του συστήματος και των επιπτώσεών του, και για τα συστήματα υψηλού κινδύνου, διενέργεια ελέγχου από εμπειρογνώμονες και συμμετοχή του κοινού. Η έκθεση αξιολόγησης των επιπτώσεων πρέπει να αξιολογεί τις επιδόσεις, την αποτελεσματικότητα και τις πιθανές επιπτώσεις του συστήματος στα θεμελιώδη δικαιώματα, τη δημοκρατία και την κοινωνική ευημερία. Το τελικό βήμα περιλαμβάνει τη δημοσίευση της έκθεσης και, εάν είναι απαραίτητο, την επικαιροποίησή της

με βάση τις συνεχείς αναθεωρήσεις και την ανατροφοδότηση από τους εμπειρογνώμονες και το κοινό.

Η αξιολόγηση περιλαμβάνει ερωτήσεις σχετικά με τη διαφάνεια, την ανιχνευσιμότητα και την επεξηγηματικότητα. Απαιτεί από την αρχή εφαρμογής να περιγράψει πώς οι αποφάσεις του συστήματος μπορούν να εξηγηθούν στο κοινό και να διασφαλίσει ότι οι λειτουργίες του συστήματος είναι ανιχνεύσιμες καθ' όλη τη διάρκεια του κύκλου ζωής του. Η αξιολόγηση επιβάλλει επίσης μέτρα για να διασφαλιστεί ότι η λογική του συστήματος είναι διαφανής και ότι οι αποφάσεις του μπορούν να γίνουν κατανοητές από όσους επηρεάζονται από αυτές. Αυτό είναι ζωτικής σημασίας για τη διατήρηση της εμπιστοσύνης του κοινού και τη διασφάλιση της λογοδοσίας.

Η αξιολόγηση εξετάζει τις αναγκαίες αντισταθμίσεις, όπως αυτές μεταξύ καινοτομίας και ρύθμισης και μεταξύ διαφορετικών μεροληψιών. Απαιτεί από την αρχή εφαρμογής να αξιολογήσει την αναγκαιότητα και την αναλογικότητα της χρήσης του συστήματος, λαμβάνοντας υπόψη τις αντισταθμίσεις μεταξύ διαφόρων παραγόντων, όπως η αποτελεσματικότητα, η ακρίβεια και οι δυνητικοί κίνδυνοι. Η αξιολόγηση ενθαρρύνει επίσης την αρχή να διερευνήσει εναλλακτικές λύσεις για το προτεινόμενο σύστημα και να αιτιολογήσει γιατί το επιλεγμένο σύστημα είναι αποδεκτό παρά τους συναφείς κινδύνους. Με τον τρόπο αυτό διασφαλίζεται ότι τα οφέλη του συστήματος μεγιστοποιούνται, ενώ ελαχιστοποιούνται τα πιθανά αρνητικά αποτελέσματα.

3.3.3 Moje Państwo Foundation

Η παρούσα αξιολόγηση αλγοριθμικών επιπτώσεων είναι ένα προτεινόμενο εργαλείο που έχει σχεδιαστεί για την αξιολόγηση των επιπτώσεων της τεχνητής νοημοσύνης (AI) και των συστημάτων αυτοματοποιημένης λήψης αποφάσεων (ADM) στον δημόσιο τομέα, με στόχο τη διασφάλιση της διαφάνειας, της λογοδοσίας και της διαχείρισης κινδύνων[30]. Η πρωτοβουλία αναπτύχθηκε από το Moje Państwo Foundation με την υποστήριξη του Ευρωπαϊκού Ταμείου Τεχνητής Νοημοσύνης, μια κοινή πρωτοβουλία του Δικτύου Ευρωπαϊκών Ιδρυμάτων (NEF). Το εργαλείο αντλεί έμπνευση από τα υφιστάμενα πλαίσια σε χώρες όπως ο Καναδάς και η Ολλανδία προωθώντας την υπεύθυνη χρήση της ΤΝ, διασφαλίζοντας παράλληλα τα ανθρώπινα δικαιώματα και τις ατομικές ελευθερίες.

Σκοπός της παρούσας αξιολόγησης αλγοριθμικών επιπτώσεων είναι να αξιολογήσει τους κινδύνους που συνδέονται με την εφαρμογή συστημάτων τεχνητής νοημοσύνης (AI) και αυτοματοποιημένης λήψης αποφάσεων (ADM) στο δημόσιο τομέα, διασφαλίζοντας ότι τα συστήματα αυτά χρησιμοποιούνται με τρόπο διαφανή, υπεύθυνο και με σεβασμό των θεμελιωδών ανθρωπίνων δικαιωμάτων και ελευθεριών. Η αξιολόγηση αποσκοπεί στο να βοηθήσει τις δημόσιες αρχές να κατανοήσουν και να μετριάσουν καλύτερα τους κινδύνους της αυτοματοποιημένης λήψης αποφάσεων, διευκολύνοντας τη διαχείριση των εντοπισμένων κινδύνων καθ' όλη τη διάρκεια του κύκλου ζωής του συστήματος. Με τον τρόπο αυτό, ωφελεί την κοινωνία προωθώντας την ασφαλέστερη και αποτελεσματικότερη χρήση των συστημάτων AI/ADM, μειώνοντας το ενδεχόμενο βλάβης και διασφαλίζοντας ότι οι τεχνολογίες αυτές χρησιμοποιούνται με τρόπο που ευθυγραμμίζεται με τις δημόσιες αξίες και τα ανθρώπινα δικαιώματα.

Η παρούσα εκτίμηση επιπτώσεων έχει σχεδιαστεί για συστήματα AI και ADM που χρησιμοποιούνται στο δημόσιο τομέα, ιδίως για εκείνα που εφαρμόζονται από δημόσιες αρχές τόσο σε επίπεδο κεντρικής όσο και τοπικής αυτοδιοίκησης. Αφορά συστήματα που διεκπεραιώνουν καθήκοντα όπως η διαχείριση υποθέσεων πολιτών, η ανίχνευση απάτης και άλλες διοικητικές λειτουργίες, με έμφαση στη διασφάλιση ότι τα συστήματα αυτά χρησιμοποιούνται υπεύθυνα και δεοντολογικά .

Η αξιολόγηση καλύπτει ένα ευρύ φάσμα θεμάτων, συμπεριλαμβανομένων των ανθρωπίνων δικαιωμάτων, της δεοντολογίας, της προστασίας δεδομένων, της ιδιωτικής ζωής, της ασφάλειας, της προστασίας και των περιβαλλοντικών ζητημάτων. Δίνει έμφαση στην ανάγκη αξιολόγησης του αντίκτυπου των συστημάτων AI/ADM στα θεμελιώδη δικαιώματα, όπως η προστασία της ιδιωτικής ζωής και η απαγόρευση των διακρίσεων, και εξετάζει τις ηθικές επιπτώσεις της χρήσης αυτών των τεχνολογιών. Η αξιολόγηση ασχολείται επίσης με τους πιθανούς κινδύνους της αλγοριθμικής μεροληψίας, της διαφάνειας και της λογοδοσίας, διασφαλίζοντας ότι τα συστήματα δεν θέτουν σε κίνδυνο τις δημόσιες αξίες ή δεν βλάπτουν τους πολίτες .

Η αξιολόγηση δεν εντάσσεται σε συγκεκριμένη οργανωτική δομή, αλλά προτείνεται ως εργαλείο για τους δημόσιους φορείς που μπορούν να χρησιμοποιήσουν κατά την εφαρμογή των συστημάτων AI/ADM. Μπορεί να διενεργηθεί από τους ίδιους τους δημόσιους φορείς, αν και το ιδανικό μοντέλο προτείνει ότι θα μπορούσε να διενεργηθεί από έναν ανεξάρτητο δημόσιο φορέα. Επί του παρόντος, ελλείψει τέτοιων εξειδικευμένων φορέων, η αξιολόγηση πρέπει να διεξάγεται από τους φορείς που ενδιαφέρονται να εφαρμόσουν το σύστημα.

Η εκτίμηση προορίζεται να ολοκληρωθεί κατά τα στάδια σχεδιασμού και ανάπτυξης του κύκλου ζωής του συστήματος AI/ADM, πριν το σύστημα τεθεί σε λειτουργία. Θα πρέπει να επικαιροποιείται εάν αλλάζει η λειτουργικότητα του συστήματος ή το πεδίο εφαρμογής της αυτοματοποιημένης λήψης αποφάσεων, διασφαλίζοντας ότι η αξιολόγηση παραμένει επίκαιρη και ότι οι κίνδυνοι διαχειρίζονται συνεχώς.

Η διαδικασία της αλγοριθμικής εκτίμησης επιπτώσεων (AEA) περιλαμβάνει δύο βασικά μέρη:

- Χαρακτηρισμός: Καθορίζονται τα χαρακτηριστικά του συστήματος AI/ADM απαντώντας σε συγκεκριμένα ερωτήματα που σχετίζονται με τις λειτουργίες, την προβλεπόμενη χρήση και τις πιθανές επιπτώσεις του.
- Αξιολόγηση των επιπτώσεων: Αποδίδεται ένα επίπεδο επιπτώσεων στο σύστημα, προσδιορίζονται οι σχετικοί κίνδυνοι και εφαρμόζονται συστάσεις για τη διαχείριση των εν λόγω κινδύνων, εξασφαλίζοντας τη συμμετοχή των ενδιαφερόμενων μερών καθ' όλη τη διάρκεια της αξιολόγησης.

Η αξιολόγηση περιλαμβάνει ερωτήσεις σχετικά με τη διαφάνεια, την ανιχνευσιμότητα και την επεξηγηματικότητα. Ρωτά εάν τεκμηριώνεται η διαδικασία λήψης αποφάσεων του συστήματος, συμπεριλαμβανομένων των πηγών δεδομένων, των πιθανών προκαταλήψεων και της ακρίβειας του αλγορίθμου. Επίσης, εξετάζεται η διαθεσιμότητα του πηγαίου κώδικα του συστήματος και άλλων πληροφοριών, όπως οι στόχοι και οι κανόνες λειτουργίας του, στο κοινό. Επιπλέον, μελετάται κατά πόσον το σύστημα επιτρέπει την ανθρώπινη εποπτεία και κατά πόσον οι αποφάσεις μπορούν να αμφισβητηθούν διασφαλίζοντας ότι οι λειτουργίες του συστήματος είναι διαφανείς και υπόλογες.

Η αξιολόγηση λαμβάνει υπόψη τις αναγκαίες αντισταθμίσεις, ιδίως μεταξύ καινοτομίας και κανονιστικού πλαισίου, καθώς και μεταξύ διαφορετικών μεροληψιών. Τονίζει τη σημασία της εξισορρόπησης των πλεονεκτημάτων των συστημάτων AI/ADM, όπως η αυξημένη αποτελεσματικότητα και η μείωση του κόστους, με τους δυνητικούς κινδύνους για τα ανθρώπινα δικαιώματα και τις δημόσιες αξίες. Η αξιολόγηση υπογραμμίζει επίσης την ανάγκη να αντιμετωπιστεί η αλγοριθμική μεροληψία και να διασφαλιστεί ότι τα συστήματα δεν επηρεάζουν δυσανάλογα ορισμένες ομάδες, όπως οι εθνοτικές μειονότητες ή τα οικονομικά μειονεκτούντα άτομα. Με τον εντοπισμό και τον μετριασμό αυτών των κινδύνων, η αξιολόγηση αποσκοπεί στην επίτευξη ισορροπίας μεταξύ της αξιοποίησης της τεχνολογικής καινοτομίας και της προστασίας των θεμελιωδών δικαιωμάτων.

3.3.4 Commission Nationale de l'Informatique et des Libertés

Το ερωτηματολόγιο συμμόρφωσης συστημάτων ΤΝ είναι ένα πλαίσιο που αναπτύχθηκε από την CNIL (Commission Nationale de l'Informatique et des Libertés) για να καθοδηγήσει τους οργανισμούς στην αξιολόγηση της συμμόρφωσης των συστημάτων ΤΝ τους με τους κανονισμούς προστασίας δεδομένων, ιδίως με τον ΓΚΠΔ [31].

Σκοπός της παρούσας εκτίμησης επιπτώσεων είναι να παρέχει έναν ολοκληρωμένο οδηγό αυτοαξιολόγησης για τους οργανισμούς που εφαρμόζουν ή χρησιμοποιούν συστήματα ΤΝ, διασφαλίζοντας ότι τα συστήματα αυτά συμμορφώνονται με τους κανονισμούς προστασίας δεδομένων και τις δεοντολογικές εκτιμήσεις. Στόχος του οδηγού είναι να βοηθήσει τους οργανισμούς να εντοπίσουν και να μετριάσουν τους κινδύνους που συνδέονται με τα συστήματα ΤΝ, ιδίως όσον αφορά τα προσωπικά δεδομένα και τα θεμελιώδη δικαιώματα. Με τον τρόπο αυτό, ωφελεί την κοινωνία προωθώντας την υπεύθυνη χρήση της ΤΝ, προστατεύοντας την ιδιωτική ζωή των ατόμων και διασφαλίζοντας ότι τα συστήματα ΤΝ δεν διαιωνίζουν προκαταλήψεις ή διακρίσεις. Ο οδηγός υπογραμμίζει επίσης τη σημασία της διαφάνειας και της ανθρώπινης εποπτείας, οι οποίες μπορούν να οικοδομήσουν την εμπιστοσύνη του κοινού στις τεχνολογίες ΤΝ.

Η παρούσα εκτίμηση επιπτώσεων έχει σχεδιαστεί για συστήματα ΤΝ που χρησιμοποιούνται σε διάφορους τομείς και σενάρια. Αφορά τόσο τους παρόχους όσο και τους χρήστες συστημάτων ΤΝ, συμπεριλαμβανομένων φυσικών ή νομικών προσώπων, δημόσιων αρχών, οργανισμών ή άλλων οργανισμών που αναπτύσσουν ή χρησιμοποιούν συστήματα ΤΝ. Ο οδηγός προορίζεται να είναι εφαρμόσιμος σε όλους τους τύπους συστημάτων ΤΝ, ανεξάρτητα από τις συγκεκριμένες τεχνικές που χρησιμοποιούνται, όπως η συνεχής μάθηση ή ο αυτόματος σχολιασμός.

Η αξιολόγηση καλύπτει ένα ευρύ φάσμα θεμάτων, συμπεριλαμβανομένων των ανθρωπίνων δικαιωμάτων, της δεοντολογίας, της προστασίας δεδομένων, της ιδιωτικής ζωής, της ασφάλειας και της προστασίας. Εξετάζει τον πιθανό αντίκτυπο των συστημάτων τεχνητής νοημοσύνης στα θεμελιώδη δικαιώματα, όπως η ελευθερία της έκφρασης, η ελευθερία της σκέψης και ο σεβασμός της ιδιωτικής ζωής. Ο οδηγός εξετάζει επίσης τους κινδύνους μεροληψίας και διακρίσεων, την ανάγκη για διαφάνεια και επεξηγηματικότητα, καθώς και τη σημασία της εξασφάλισης των συστημάτων τεχνητής νοημοσύνης έναντι επιθέσεων. Επιπλέον, θίγει τις κοινωνικές και δημοκρατικές επιπτώσεις της ΤΝ, όπως το ενδεχόμενο η ΤΝ να επηρεάσει τις απόψεις ή να επιδεινώσει τις ανισότητες .

Η αξιολόγηση αυτή εντάσσεται στο πλαίσιο της γαλλικής αρχής προστασίας δεδομένων. Δεν συμπληρώνεται από τρίτο μέρος, αλλά προορίζεται ως εργαλείο αυτοαξιολόγησης για εσωτερική χρήση από τους οργανισμούς. Ωστόσο, η CNIL ενθαρρύνει τους οργανισμούς να ζητούν επίσημες συμβουλές ή να διεξάγουν, εάν είναι απαραίτητο, εκτίμηση αντικτύπου σχετικά με την προστασία των δεδομένων (DPIA), η οποία μπορεί να περιλαμβάνει αξιολογήσεις τρίτων .

Η αξιολόγηση των επιπτώσεων πρέπει να ολοκληρώνεται καθ' όλη τη διάρκεια του κύκλου ζωής της ΤΝ, από τις αρχικές φάσεις σχεδιασμού και ανάπτυξης έως την ανάπτυξη και τη συνεχή χρήση του συστήματος ΤΝ. Θα πρέπει να επικαιροποιείται τακτικά, ιδίως όταν υπάρχουν σημαντικές αλλαγές στο σύστημα ΤΝ, όπως ενημερώσεις στον αλγόριθμο, αλλαγές στις πηγές δεδομένων ή αλλαγές στο επιχειρησιακό πλαίσιο του συστήματος. Συνιστάται η συνεχής παρακολούθηση και η περιοδική επαναξιολόγηση για τη διασφάλιση της διαρκούς συμμόρφωσης και της διαχείρισης κινδύνων.

Η διαδικασία αυτής της αξιολόγησης περιλαμβάνει διάφορα στάδια. Πρώτα, οι οργανισμοί πρέπει να προσδιορίσουν τον στόχο και την αναλογικότητα του συστήματος ΤΝ, διασφαλίζοντας ότι οι επιλεγμένες τεχνικές είναι απαραίτητες και δικαιολογημένες. Στη συνέχεια, πρέπει να αξιολογήσουν την ποιότητα και τη νομιμότητα των δεδομένων εκπαίδευσης, εξετάζοντας ζητήματα όπως η προέλευση των δεδομένων, η μεροληψία και η συμμόρφωση με τον ΓΚΠΔ. Η ανάπτυξη και

η εκπαίδευση του αλγορίθμου πρέπει στη συνέχεια να ελεγχθεί ως προς την αξιοπιστία, τη δικαιοσύνη και τις επιδόσεις. Κατά τη φάση της παραγωγής, πρέπει να εφαρμόζονται μέτρα ανθρώπινης εποπτείας και διαφάνειας. Τέλος, οι οργανισμοί πρέπει να τεκμηριώνουν την όλη διαδικασία, να διενεργούν τακτικούς ελέγχους και να διασφαλίζουν ότι τα άτομα μπορούν να ασκήσουν τα δικαιώματά τους όσον αφορά το σύστημα ΤΝ .

Η αξιολόγηση περιλαμβάνει ερωτήσεις σχετικά με τη διαφάνεια, την ιχνηλασιμότητα και την επεξηγηματικότητα. Οι οργανισμοί καλούνται να διασφαλίσουν ότι η διαδικασία λήψης αποφάσεων του συστήματος ΤΝ καταγράφεται και μπορεί να εξηγηθεί εκ των υστέρων. Πρέπει επίσης να παρέχουν σαφείς και κατανοητές πληροφορίες στα άτομα που αλληλεπιδρούν με το σύστημα, συμπεριλαμβανομένων εξηγήσεων σχετικά με τον τρόπο λειτουργίας του συστήματος και τυχόν περιορισμούς ή παραδοχές που μπορεί να έχει. Ο οδηγός υπογραμμίζει τη σημασία της χρήσης τεχνικών και μεθοδολογικών εργαλείων για να καταστεί δυνατή η επεξηγηματικότητα και προτείνει ότι οι προσεγγίσεις ανοικτού κώδικα μπορούν να ενισχύσουν τη διαφάνεια.

Η αξιολόγηση αναγνωρίζει την ανάγκη για συμβιβασμούς, ιδίως μεταξύ καινοτομίας και ρύθμισης, καθώς και μεταξύ διαφορετικών τύπων μεροληψίας. Για παράδειγμα, διερωτάται κατά πόσον τα σημαντικά πλεονεκτήματα της χρήσης ενός συστήματος ΤΝ (όπως η τεχνική αποτελεσματικότητα ή η μείωση του κόστους) αντισταθμίζουν τους δυνητικούς κινδύνους, συμπεριλαμβανομένων των επιπτώσεων στα θεμελιώδη δικαιώματα. Ο οδηγός ενθαρρύνει επίσης τους οργανισμούς να εξισορροπήσουν την πολυπλοκότητα των λύσεων ΤΝ με την ανάγκη για επεξηγηματικότητα και να εξετάσουν τα αντισταθμιστικά οφέλη που συνεπάγεται η συνεχής μάθηση σε σενάρια, όπου η ποιότητα των δεδομένων και η απόδοση του συστήματος πρέπει να παρακολουθούνται προσεκτικά για την αποφυγή επιδείνωσης ή μεροληψίας .

3.3.5 INTERPOL και UNICRI

Το ερωτηματολόγιο αξιολόγησης κινδύνων για υπεύθυνη καινοτομία ΤΝ αναπτύχθηκε από την INTERPOL και το Διαπεριφερειακό Ινστιτούτο Ερευνών για το Έγκλημα και τη Δικαιοσύνη των Ηνωμένων Εθνών (UNICRI) για να βοηθήσει τις υπηρεσίες επιβολής του νόμου να αξιολογήσουν τους κινδύνους που συνδέονται με τα συστήματα ΤΝ [32].

Σκοπός του παρόντος ερωτηματολογίου αξιολόγησης κινδύνων είναι να υποστηρίξει τις υπηρεσίες επιβολής του νόμου στην αξιολόγηση των κινδύνων που συνδέονται με τα συστήματα ΤΝ από την άποψη της υπεύθυνης καινοτομίας ΤΝ. Στόχος του είναι να προσδιορίσει τις πιθανές δυσμενείς επιπτώσεις σε άτομα, ομάδες και την κοινωνία, καθώς και την πιθανότητα εμφάνισης τέτοιων επιπτώσεων. Αξιολογώντας οργανωτικά, τεχνικά μέτρα και μέτρα ασφαλείας, το ερωτηματολόγιο βοηθά τις υπηρεσίες να κατανοήσουν και να μετριάσουν τους κινδύνους, διασφαλίζοντας ότι τα συστήματα ΤΝ χρησιμοποιούνται με τρόπο που σέβεται τα ανθρώπινα δικαιώματα και τις ηθικές αρχές. Αυτό ωφελεί την κοινωνία προωθώντας την ασφαλή και υπεύθυνη χρήση της ΤΝ στην επιβολή του νόμου, μειώνοντας τη βλάβη και ενισχύοντας την εμπιστοσύνη στις τεχνολογίες ΤΝ .

Η παρούσα εκτίμηση επιπτώσεων έχει σχεδιαστεί για συστήματα ΤΝ που χρησιμοποιούνται ή προορίζονται για χρήση από υπηρεσίες επιβολής του νόμου. Επικεντρώνεται σε συστήματα που ενδέχεται να επηρεάσουν άτομα και κοινότητες, ιδίως όσους εμπλέκονται σε διαδικασίες λήψης αποφάσεων, σε έρευνες και στην παροχή δημόσιων υπηρεσιών.

Η αξιολόγηση καλύπτει ένα ευρύ φάσμα θεμάτων, συμπεριλαμβανομένων των ανθρωπίνων δικαιωμάτων, της δεοντολογίας, της προστασίας δεδομένων, της ιδιωτικής ζωής, της ασφάλειας, της προστασίας και των περιβαλλοντικών ζητημάτων. Αντιμετωπίζει συγκεκριμένα τους κινδύνους που σχετίζονται με τη μη συμμόρφωση με τους νόμους, την ελαχιστοποίηση της βλάβης, την ανθρώπινη

αυτονομία, τη δικαιοσύνη και το ενδεχόμενο διακρίσεων ή προκαταλήψεων στα συστήματα τεχνητής νοημοσύνης. Το ερωτηματολόγιο εξετάζει επίσης τον αντίκτυπο των συστημάτων ΤΝ σε ευάλωτες ομάδες και το ενδεχόμενο κακής χρήσης ή μη εξουσιοδοτημένης πρόσβασης, εξασφαλίζοντας μια ολοκληρωμένη αξιολόγηση των κινδύνων.

Η αξιολόγηση δεν εντάσσεται σε μια συγκεκριμένη οργανωτική δομή, αλλά έχει ως στόχο να συμπληρώσει τις υφιστάμενες πρακτικές διαχείρισης κινδύνων στις υπηρεσίες επιβολής του νόμου. Έχει σχεδιαστεί για να ολοκληρώνεται από την ομάδα ή το μέλος ή τα μέλη του προσωπικού που είναι υπεύθυνα για το έργο ΤΝ ή για ένα συγκεκριμένο στάδιο εντός του κύκλου ζωής της ΤΝ. Αν και μπορεί να περιλαμβάνει διαβουλεύσεις με εσωτερικούς ή εξωτερικούς εμπειρογνώμονες, η κύρια ευθύνη ανήκει στην υπηρεσία που εφαρμόζει το σύστημα ΤΝ.

Το ερωτηματολόγιο αξιολόγησης κινδύνων συνιστάται να χρησιμοποιείται νωρίς στις φάσεις σχεδιασμού και ανάπτυξης του κύκλου ζωής της ΤΝ, καθώς και κατά τη διάρκεια των δοκιμών και της αξιολόγησης. Θα πρέπει επίσης να διεξάγεται περιοδικά καθ' όλη τη διάρκεια του κύκλου ζωής του συστήματος ΤΝ, ώστε να λαμβάνονται υπόψη οι αλλαγές στο σύστημα, το περιβάλλον του ή οι νέοι κίνδυνοι που ενδέχεται να προκύψουν. Οι επικαιροποιήσεις είναι ιδιαίτερα αναγκαίες μετά την εφαρμογή μέτρων μετριασμού ή όταν συμβαίνουν σημαντικές αλλαγές στη λειτουργία ή το πλαίσιο του συστήματος .

Η διαδικασία της αξιολόγησης περιλαμβάνει πέντε αλληλένδετα βήματα: προετοιμασία για την αξιολόγηση κινδύνων, συμπλήρωση του ερωτηματολογίου, ερμηνεία των αποτελεσμάτων, κοινοποίηση των αποτελεσμάτων στον αρμόδιο ιδιοκτήτη κινδύνων και αντιμετώπιση των κινδύνων. Το ερωτηματολόγιο είναι δομημένο γύρω από δύο κύριες κατηγορίες -επιπτώσεις και πιθανότητες- με κάθε ερώτηση να βαθμολογείται σε κλίμακα από το 1 έως το 5. Η συνολική βαθμολογία κινδύνου υπολογίζεται πολλαπλασιάζοντας τη βαθμολογία επιπτώσεων επί τη βαθμολογία πιθανότητας, επιτρέποντας στους οργανισμούς να εντοπίζουν και να ιεραρχούν τους κινδύνους. Τα αποτελέσματα χρησιμοποιούνται στη συνέχεια για την ανάπτυξη στρατηγικών μετριασμού και την ενημέρωση για τη λήψη αποφάσεων.

Η αξιολόγηση περιλαμβάνει ερωτήσεις σχετικά με τη διαφάνεια, την ανιχνευσιμότητα και την επεξηγηματικότητα. Εξετάζονται ο αντίκτυπος των περιορισμών στην ανθρώπινη εποπτεία, η πιθανότητα μεροληψίας στη λήψη αποφάσεων, ενώ δίνεται και έμφαση στην ικανότητα των χρηστών να κατανοούν τα αποτελέσματα του συστήματος ΤΝ. Οι ερωτήσεις αυτές διασφαλίζουν ότι οι λειτουργίες του συστήματος ΤΝ είναι διαφανείς, υπόλογες και κατανοητές τόσο για τους χρήστες όσο και για τα άτομα που επηρεάζονται.

Η αξιολόγηση εξετάζει τις αναγκαίες αντισταθμίσεις, ιδίως μεταξύ καινοτομίας και ρύθμισης, καθώς και μεταξύ διαφορετικών προκαταλήψεων. Τονίζει τη σημασία της εξισορρόπησης των πλεονεκτημάτων των συστημάτων τεχνητής νοημοσύνης με τους δυνητικούς κινδύνους για τα ανθρώπινα δικαιώματα και τις δημόσιες αξίες. Το ερωτηματολόγιο εξετάζει την πιθανότητα και τον αντίκτυπο των προκαταλήψεων στα συστήματα ΤΝ, το ενδεχόμενο διακρίσεων και την ανάγκη για δικαιοσύνη στη λήψη αποφάσεων ΤΝ. Με τον εντοπισμό και τον μετριασμό αυτών των κινδύνων, η αξιολόγηση αποσκοπεί στην επίτευξη ισορροπίας μεταξύ της αξιοποίησης της τεχνολογικής καινοτομίας και της προστασίας των θεμελιωδών δικαιωμάτων.

3.4 Αναγκαιότητα των ΑΕΑ

Όπως είδαμε στο προηγούμενο υποκεφάλαιο τα εργαλεία αλγοριθμικής εκτίμησης αντικτύπου (ΑΕΑ) αποτελούν βασικά μέσα για την αξιολόγηση των επιπτώσεων των συστημάτων τεχνητής νοημοσύνης (ΤΝ) σε διάφορους τομείς, όπως η υγειονομική περίθαλψη, η δημόσια

διοίκηση και η επιβολή του νόμου. Αυτά τα εργαλεία έχουν σχεδιαστεί για να εξισορροπούν την καινοτομία με την ανάγκη για ασφάλεια, ηθική και διαφάνεια, ενώ παράλληλα στοχεύουν στη διασφάλιση της συμμόρφωσης με τους κανονισμούς και την προστασία των ανθρωπίνων δικαιωμάτων. Οι ΑΕΑ καλύπτουν ένα ευρύ φάσμα θεμάτων, όπως η προστασία της ιδιωτικής ζωής, η αντιμετώπιση των προκαταλήψεων, η περιβαλλοντική βιωσιμότητα και η διαφάνεια στις λειτουργίες των συστημάτων ΤΝ. Επιπλέον, ενθαρρύνουν τη συμμετοχή των ενδιαφερόμενων μερών και την ανθρώπινη εποπτεία, διασφαλίζοντας ότι οι τεχνολογίες ΤΝ χρησιμοποιούνται υπεύθυνα και με σεβασμό στις δημόσιες αξίες. Η συστηματική εφαρμογή των ΑΕΑ βοηθά στη μείωση των κινδύνων και στην ενίσχυση της εμπιστοσύνης του κοινού στις τεχνολογίες ΤΝ, ενώ ταυτόχρονα υποστηρίζει την υπεύθυνη ανάπτυξη και εφαρμογή τους. Πρακτικά οι εκτιμήσεις αλγοριθμικών επιπτώσεων (ΑΕΑ) έχουν αναδειχθεί ως προτεινόμενο ρυθμιστικό εργαλείο για την αντιμετώπιση των πιθανών βλαβών των συστημάτων τεχνητής νοημοσύνης ,όμως δεδομένης της πολυπλοκότητας και της αδιαφάνειας των τεχνολογιών αυτών είναι εύλογο να αμφισβητηθεί κατά πόσον οι ΑΕΑ αποτελούν την καλύτερη λύση, ιδίως δεδομένου ότι οι εναλλακτικές ρυθμιστικές προσεγγίσεις δεν έχουν διερευνηθεί διεξοδικά σε αυτή την εργασία. Βασιζόμενοι, λοιπόν, στις γνώσεις του «An Institutional View of Algorithmic Impact Assessments», θα επιδιώξουμε την εξέταση της λογική πίσω από τις ΑΕΑ και την πιθανή αποτελεσματικότητά τους στον μετριασμό των κινδύνων που συνδέονται με τα συστήματα ΤΝ [33].

Όπως έχει αναφερθεί και στο κεφάλαιο 1 τα συστήματα τεχνητής νοημοσύνης, ενώ είναι αναμφισβήτητα ισχυρά και αποτελεσματικά, μπορούν να προκαλέσουν δυσμενείς επιπτώσεις σε διάφορους τομείς, συμπεριλαμβανομένων των διακρίσεων, της διαδικαστικής αδικίας, της σωματικής βλάβης και γενικά της παραβίασης ανθρωπίνων δικαιωμάτων. Μία από τις σημαντικότερες προκλήσεις για τη ρύθμιση αυτών των συστημάτων είναι η έλλειψη τεχνικών γνώσεων μεταξύ του κοινού και των δημόσιων αρχών. Αυτό το κενό γνώσεων καθιστά αναγκαία τη συμμετοχή του ιδιωτικού τομέα, καθώς οι δημιουργοί ΤΝ διαθέτουν την απαιτούμενη τεχνογνωσία για την αξιολόγηση και τον μετριασμό των πιθανών βλαβών. Ωστόσο, οι ιδιωτικοί οργανισμοί καθοδηγούνται κυρίως από κίνητρα κέρδους και αποφυγής ευθύνης, γεγονός που περιπλέκει την προθυμία τους να συνεργαστούν καλόπιστα. Παρά τις προκλήσεις αυτές, οι ΑΕΑ προσφέρουν μια δομημένη προσέγγιση για την αντιμετώπιση αυτών των ζητημάτων με την αξιοποίηση της τεχνικής εμπειρογνωμοσύνης του ιδιωτικού τομέα.

Οι ΑΕΑ εξυπηρετούν δύο πρωταρχικούς σκοπούς. Πρώτον, ενθαρρύνουν τις οντότητες του ιδιωτικού τομέα να διεξάγουν μεθοδικές αξιολογήσεις των συστημάτων ΤΝ τους σε πρώιμο στάδιο της διαδικασίας ανάπτυξης, επιτρέποντάς τους να εντοπίζουν και να μετριάζουν πιθανές βλάβες πριν αυτές γίνουν δαπανηρές ή μη αναστρέψιμες. Δεύτερον, οι ΑΕΑ επιβάλλουν την εκτενή τεκμηρίωση των διαδικασιών λήψης αποφάσεων πίσω από τα συστήματα ΤΝ. Αυτή η τεκμηρίωση όχι μόνο ενισχύει τη λογοδοσία αλλά παρέχει επίσης πολύτιμες πληροφορίες για τους φορείς χάραξης πολιτικής και το κοινό, επιτρέποντας πιο τεκμηριωμένες και αποτελεσματικές μελλοντικές ρυθμίσεις. Με την απαίτηση διαφάνειας και προβληματισμού, οι ΑΕΑ δημιουργούν έναν βρόχο ανατροφοδότησης που μπορεί να ενημερώσει τόσο τις άμεσες όσο και τις μακροπρόθεσμες πολιτικές αποφάσεις.

Υπάρχουν τρεις βασικοί τύποι ΑΕΑ: αυτές που διαμορφώνονται σύμφωνα με τον Εθνικό Νόμο περί Περιβαλλοντικής Πολιτικής (NEPA), αυτές που βασίζονται στις εκτιμήσεις αντικτύπου για την προστασία των δεδομένων (DPIA) του Γενικού Κανονισμού για την Προστασία Δεδομένων (GDPR) και η προσέγγιση βάσει ερωτηματολογίου που χρησιμοποιείται στον Καναδά. Παρόλο που υπάρχουν εναλλακτικές λύσεις όπως οι έλεγχοι και η αυτορρύθμιση, **οι ΑΕΑ είναι ιδιαίτερα πλεονεκτικές λόγω της έμφασής τους στην έγκαιρη παρέμβαση, τις ανοιχτές ερωτήσεις και την προσβασιμότητα της τεκμηρίωσης που προκύπτει.** Τα χαρακτηριστικά αυτά καθιστούν τις ΑΕΑ

πιο προσαρμόσιμες και πιο ολοκληρωμένες από άλλους μηχανισμούς εποπτείας, επιτρέποντας βαθύτερη κατανόηση των πιθανών επιπτώσεων των συστημάτων ΤΝ.

Για την αποτελεσματική εφαρμογή των ΑΕΑ, είναι ζωτικής σημασίας να κατανοήσουμε πώς οι δημόσιοι θεσμοί διαμορφώνουν τη νομοθεσία μέσω συγκεκριμένων πλαισίων, ανάμεσά τους η συνεργατική διακυβέρνηση και ο νομικός διαχειριστισμός. Ταυτόχρονα, η αποδοχή του ιδιωτικού τομέα ως εταίρου είναι απαραίτητη, καθώς οι τεχνικές γνώσεις του είναι απαραίτητες για την ουσιαστική συμμόρφωση. Με την ευθυγράμμιση των κανονιστικών απαιτήσεων με τις διαδικασίες και τα κίνητρα του ιδιωτικού τομέα, οι ΑΕΑ μπορούν να γίνουν ένα πιο βιώσιμο και αποτελεσματικό εργαλείο για τη ρύθμιση των συστημάτων ΤΝ. Αν και οι ΑΕΑ μπορεί να μην αποτελούν τέλεια λύση, αποτελούν ένα απαραίτητο ενδιάμεσο βήμα προς μια πιο ισχυρή και ολοκληρωμένη ρύθμιση στο μέλλον.

Κεφάλαιο 4: Ανάπτυξη του Ερωτηματολογίου

4.1 Κατηγορίες ανάλυσης της Αλγοριθμικής Εκτίμησης Αντικτύπου

4.1.1 Κατηγοριοποίηση capAI

Η βασική δομή του εργαλείου αλγοριθμικής αξιολόγησης που αναπτύχθηκε στα πλαίσια αυτής της εργασίας βασίζεται στη βασική δομή του εγγράφου IRP όπως το ορίζει το capAI. Για λόγους πληρότητας λοιπόν θα παραθέσουμε παρακάτω τη δομή του προτεινόμενου IRP :

1. **Στάδιο Σχεδιασμού :** Είναι στο στάδιο ανάπτυξης μιας ΤΝ στο οποίο ζητείται η βασική ιδέα, το ορίζεται το πρόβλημα το οποίο καλείται η ΤΝ να λύση και εξετάζεται η εφικτότητα τον στόχων που τίθενται. Οι ερωτήσεις που αφορούν αυτό το στάδιο χωρίζονται σε δύο κατηγορίες:
    1.1. **Οργάνωση :** Εδώ υπάγονται ερωτήσεις που αφορούν την οργανωτική διακυβέρνηση της ανάπτυξης της συγκεκριμένης ΤΝ. Πιο συγκεκριμένα, δίνεται έμφαση στο κατά πόσο ορισμένες ηθικές αρχές και αξίες έχουν επικοινωνηθεί και συμφωνηθεί με τα ενδιαφερόμενη μέλη και τους εργαζόμενους στο έργο ΤΝ αλλά και στο κατά πόσο υπάρχει η απαραίτητη υπευθυνότητα και λογοδοσία σε περίπτωση μη συμμόρφωσής τους.
    1.2. **Περίπτωση χρήσης :** Εδώ υπάγονται ερωτήσεις που αφορούν τις απαιτήσεις περίπτωσης χρήσης και πιο συγκεκριμένα πράγματα όπως οι στόχοι του έργου, πως αυτοί αλληλεπιδρούν με τις ηθικές αξίες αλλά και πιθανά κριτήρια απόδοσης της εφαρμογής που θα χρησιμοποιηθούν σε μελλοντικά στάδια.
2. **Στάδιο Ανάπτυξης :** Σε αυτό το στάδιο οι προγραμματιστές χρησιμοποιούν μια σειρά από δεδομένα για να εκπαιδεύσουν έναν αλγόριθμο ΤΝ και έτσι να δημιουργήσουν το τελικό προϊόν ΤΝ. Οι ερωτήσεις του σταδίου αυτού χωρίζονται σε δύο κατηγορίες/υποστάδια :
    2.1. **Δεδομένα :** Εδώ οι ερωτήσεις επικεντρώνεται στην εξασφάλιση ότι τα δεδομένα που θα χρησιμοποιηθούν στην εκπαίδευση καταγράφονται επαρκώς, τηρούν τα νομικά πλαίσια και είναι “ποιοτικά” (δεν περιέχουν μεροληψίες, σφάλματα, είναι αντιπροσωπευτικά του προβλήματος, δεν έχουν ελλείψεις, μπορεί να εντοπιστεί εύκολα η πηγή τους κλπ)
    2.2. **Μοντέλο :** Αυτό το στάδιο περιστρέφεται γύρω από τη διασφάλιση ότι το μοντέλο εκπαιδεύεται ώστε να δίνει αξιόπιστα συμπεράσματα και περιλαμβάνει ερωτήσεις σχετικά με την πηγή, τη δικαιοσύνη, την επεξηγηματικότητα, τη στιβαρότητα, τους πιθανούς κινδύνους και τη συμπεριφορά εκπαίδευσης του μοντέλου.
3. **Στάδιο Αξιολόγησης :** Σε αυτό το στάδιο η αξιολόγηση πρέπει να διασφαλίσουν ότι οι επιδόσεις της ΤΝ έχουν υποβληθεί σε ενδελεχείς δοκιμές και ότι με βάση αυτές κρίθηκε ορθώς ασφαλές να εισέλθει στην αγορά. Αυτό το στάδιο χωρίζεται σε δύο φάσεις :
    3.1. **Testing :** Σε αυτή τη φάση οι ερωτήσεις αφορούν τις στρατηγικές testing στις οποίες υποβάλλεται το μοντέλο και την αντίστοιχη απόδοσή του.

3.2. **Εγκατάσταση :** Στη φάση αυτή αντικείμενο εξέτασης είναι οι τεχνικές εγκατάστασης του μοντέλου ΤΝ δοκιμαστικά στο περιβάλλον παραγωγής και τα αποτελέσματα αυτής της διαδικασίας.

4. **Στάδιο Λειτουργίας :** Αυτό το στάδιο αφορά ερωτήσεις που διασφαλίζουν ότι έχουν ληφθεί τα κατάλληλα μέτρα για να αποφευχθεί η σταδιακή αποσύνθεση του μοντέλου με την πάροδο του χρόνου. Το στάδιο αυτό αποτελείται από δύο φάσεις:
   4.1. **Διατήρηση :** Αυτή η φάση αναφέρεται σε όλες τις δραστηριότητες που διατηρούν τη λειτουργικότητα του συστήματος.
   4.2. **Συντήρηση :** Εδώ ο στόχος είναι η διασφάλιση ότι γίνονται συχνές ενημερώσεις προς τη διατήρηση των ποιοτικών αποτελεσμάτων αλλά και τη βελτίωση τους.
5. **Στάδιο Απόσυρσης :** Αυτό το στάδιο αφορά τη σωστή διαδικασία απόσυρσης ενός συστήματος ΤΝ η οποία θα πρέπει να αποτελείται από σωστή αξιολόγηση των ρίσκων απόσυρσης και εύρεση των αντίστοιχων λύσεών τους.

4.1.2 Τελική κατηγοριοποίηση

Για το ερωτηματολόγιο που κατασκευάστηκε σαν μέρος αυτής της εργασίας ακολουθήθηκε η εξής μεθοδολογία : Αρχικά, συλλέχθηκαν διάφορα άλλα ερωτηματολόγια αλγοριθμικής εκτίμησης αντικτύπου, τα οποία είναι τα εξής :

- Εργαλείο Αλγοριθμική εκτίμησης αντικτύπου του Καναδά [1]
- Εργαλείο Αλγοριθμική Εκτίμησης αντικτύπου του Ada Lovelace Institute [28]
- Η τυποποιημένη και εκτεταμένη εκδοχή του ερωτηματολογίου για αλγοριθμική εκτίμηση αντικτύπου του European Law Institute [29]
- Ελεγκτής συμμόρφωσης με το AI Act του Future of Life Institute [34]
- Διάγραμμα μορφής & λήψης αποφάσεων από την αναφορά για την αξιολόγηση αλγοριθμικών επιπτώσεων σε συστήματα τεχνητής νοημοσύνης και συστήματα αυτόματης λήψης αποφάσεων του πολωνικού Moje Państwo Foundation [30]
- Ερωτηματολόγιο εκτίμησης ρίσκου της Interpol και της Unicri [32]
- Οδηγός αυτοαξιολόγησης για συστήματα τεχνητής νοημοσύνης του Commission nationale de l'informatique et des libertés [31]
- Εργαλείο αξιολόγησης AI της Al4Belgium [35]
- Αξιολόγηση ετοιμότητας ΤΝ της CISCO [36]
- Υπεύθυνο πρότυπο εκτίμησης επιπτώσεων ΤΝ της MICROSOFT [37]

Από αυτά τα ερωτηματολόγια αποσπάστηκαν οι ερωτήσεις και συγκεντρώθηκαν σε ένα αρχείο δημιουργώντας μια τράπεζα περίπου των 700 ερωτήσεων. Ύστερα αυτές οι ερωτήσεις κατηγοριοποιήθηκαν με βάση τα 5 στάδια του κύκλου ανάπτυξης ΤΝ που αναφέρει το capAI (ανάλογα με το σε ποια φάση της διαδικασίας αναφέρονται). Μετά, ανά στάδιο, ομαδοποιήθηκαν οι βασικές θεματικές στις οποίες επικεντρώνονται οι ερωτήσεις, με βάση αυτές τις θεματικές έγινε και η απαραίτητη περαιτέρω έρευνα και έτσι καταλήξαμε στη δομή που αναφέρεται παρακάτω που αποτελείται από περίπου 300 ερωτήσεις (ο αριθμός του εξαρτάται από τις επιλογές που θα γίνουν καθώς το ερωτηματολόγιο έχει δυναμικό χαρακτήρα) :

1. **Γενικές πληροφορίες**
   Αυτό το στάδιο επικεντρώνεται σε κάποιες γενικές πληροφορίες που αφορούν την εταιρία που ανέπτυξε την ΤΝ, το άτομο/άτομα που συμπληρώνουν το ερωτηματολόγιο και το ίδιο το σύστημα ΤΝ. Πιο συγκεκριμένα μερικά βασικά σημεία αυτού του σταδίου είναι τα εξής:

a. Λεπτομέρειες οργανισμού: Προσδιορίζει τον τύπο του οργανισμού (π.χ. πάροχος, φορέας ανάπτυξης, διανομέας, εισαγωγέας κλπ.) Ζητά το όνομα του οργανισμού, την περιγραφή, τον ιστότοπο, τη δήλωση αποστολής και τα στοιχεία επικοινωνίας.
b. Απαντητές του ερωτηματολογίου: Καθορίζει αν το ερωτηματολόγιο έχει συμπληρωθεί από ένα ή περισσότερα άτομα, τη σχέση τους με το σύστημα και τον τρόπο με τον οποίο απέκτησαν τις πληροφορίες.
c. Λεπτομέρειες και περιγραφή του συστήματος: Ζητά το όνομα του συστήματος, τη φάση του έργου, τα κριτήρια του πεδίου εφαρμογής, την κατηγοριοποίηση του συστήματος με βάση τους χαρακτηρισμούς ρίσκου του AI Act, την περιγραφή του σκοπού του συστήματος, των προβλεπόμενων χρήσεων και των συνδέσμων συμπληρωματικών πληροφοριών.

2. **Στάδιο Σχεδιασμού : Οργάνωση**
Αυτό το στάδιο χρησιμοποιεί τις ερωτήσεις του ερωτηματολογίου της CISCO πάνω στη στρατηγική, την υποδομή, την ασφάλεια, τα δεδομένα, τη διακυβέρνηση, το προσωπικό και την κουλτούρα για εξετάσει την οργανωτική διακυβέρνηση της εταιρείας πάνω στην ηθικές αξίες και στην πρακτική εφαρμογή τους ανεξαρτήτως του συγκεκριμένου έργου ΤΝ. Πιο συγκεκριμένα έχουμε τα εξής:
a. Στρατηγική: Αξιολογεί τη στρατηγική ΤΝ του οργανισμού, την ηγεσία, τις διαδικασίες μέτρησης του αντίκτυπου, τη χρηματοοικονομική στρατηγική και την κατανομή του προϋπολογισμού.
b. Υποδομή/ασφάλεια: Αξιολογεί την επεκτασιμότητα της υποδομής ΤΝ του οργανισμού, τους πόρους GPU, την απόδοση του δικτύου, την ευαισθητοποίηση σε θέματα κυβερνοασφάλειας και τα μέτρα προστασίας δεδομένων.
c. Δεδομένα: Αξιολογεί τη συγκέντρωση δεδομένων, την προεπεξεργασία, την προσβασιμότητα, την ενσωμάτωση με εργαλεία ανάλυσης και την επάρκεια του προσωπικού στη χρήση αυτών των εργαλείων.
d. Διακυβέρνηση: Ελέγχει αν ο οργανισμός έχει καθορίσει τις αξίες που καθοδηγούν την ανάπτυξη της ΤΝ, ένα συμβούλιο ελέγχου δεοντολογίας της ΤΝ και μηχανισμούς για τον εντοπισμό και τη διόρθωση των προκαταλήψεων.
e. Προσωπικό: Αξιολογεί τους εργατικούς πόρους του οργανισμού, την επάρκεια του προσωπικού στις τεχνολογίες ΤΝ, τα προγράμματα κατάρτισης και τις εκτιμήσεις προσβασιμότητας για εργαζόμενους με αναπηρία.
f. Κουλτούρα: Αξιολογεί την επείγουσα ανάγκη του οργανισμού να αγκαλιάσει την ΤΝ, τη δεκτικότητα της ηγεσίας και των εργαζομένων και την ύπαρξη σχεδίου διαχείρισης αλλαγών.

3. **Στάδιο Σχεδιασμού : Περίπτωση Χρήσης**
Σε αυτό το στάδιο οι ερωτήσεις χωρίζονται σε 3 βασικές υποθεματικές:
a. Στόχοι της εφαρμογής : Εδώ εξετάζονται οι λόγοι για τους οποίους δημιουργήθηκε η εφαρμογή, ποιο πρόβλημα σκοπεύει να δώσει, ποια άτομα αφορά/επηρεάζει και τέλος ποια θα είναι τα κριτήρια απόδοσής της.
b. Αυτοματισμός : Αυτή η κατηγορία εξετάζει το δίλημμά του αν θα πρέπει να χρησιμοποιηθεί ΤΝ για αυτή την εφαρμογή και αν η επιλογή να αναπτυχθεί το εν λόγω σύστημα έχει γίνει εν γνώσει των μην αυτοματοποιημένων εναλλακτικών.
c. Ηθική αξιολόγηση : Χρησιμοποιώντας κυρίως την κατηγοριοποίηση των κινδύνων της ΤΝ που προάγεται από την εργασία “A Collaborative, Human-Centred Taxonomy of AI, Algorithmic, and Automation Harms” αξιολογείται η πιθανή επικινδυνότητα του συστήματος προς την κοινωνία και το περιβάλλον. [38]

4. **Στάδιο Ανάπτυξης : Δεδομένα**

   Σε αυτό το στάδιο οι ερωτήσεις χωρίζονται σε 3 βασικές υποθεματικές:

   a. Πηγή και τεκμηρίωση: Ερωτήσεις σχετικά με τον τύπο των δεδομένων που χρησιμοποιούνται, τη διαδικασία επιλογής δεδομένων, τις πηγές δεδομένων και αν τα δεδομένα επαναχρησιμοποιούνται ή συλλέγονται ειδικά για την εφαρμογή.
   b. Απόρρητο και νομιμότητα: Αξιολογεί τη χρήση προσωπικών δεδομένων, την ανωνυμοποίηση/ ψευδωνυμοποίηση δεδομένων, τη συμμόρφωση με τον κανονισμό ΓΚΠΔ και τις εκτιμήσεις αντικτύπου σχετικά με την προστασία των δεδομένων. [39]
   c. Ποιότητα και μεροληψία: Αξιολογεί τις μεθόδους συλλογής δεδομένων, την ανίχνευση μεροληψίας, την πληρότητα των δεδομένων, την ακρίβεια, την επικαιρότητα, τη συνέπεια, τη συνάφεια και την αντιπροσωπευτικότητα.

5. **Στάδιο ανάπτυξης: Μοντέλο**

   Σε αυτό το στάδιο οι ερωτήσεις χωρίζονται σε 6 βασικές υποθεματικές:

   a. **Αλγόριθμος:** Ζητά περιγραφή της τεχνολογίας που χρησιμοποιείται, της εφαρμογής του αλγορίθμου και του βαθμού αυτοματοποίησης.
   b. **Δικαιοσύνη:** Αξιολογεί τους τύπους δικαιοσύνης που εφαρμόζονται, τα εργαλεία που χρησιμοποιούνται για τη δικαιοσύνη και τα ζητήματα προσβασιμότητας για χρήστες με αναπηρίες. [40][41][42][43]
   c. **Εξηγησιμότητα και διαφάνεια:** Αξιολογεί κατά πόσον ο αλγόριθμος είναι ανοικτού κώδικα, παρέχει εξηγήσεις και προσαρμόζει τις επεξηγήσεις ανάλογα με το επίπεδο γνώσεων του χρήστη. [44]
   d. **Ανιχνευσιμότητα:** Αξιολογεί την καταγραφή των αλλαγών, την ιχνηλασιμότητα των αποφάσεων και τη συνολική ιχνηλασιμότητα του συστήματος.
   e. **Στιβαρότητα και Testing:** Αξιολογεί τα είδη των test στιβαρότητας που διεξάγονται.[45][46][47]
   f. **Ηθικά ζητήματα:** Αξιολογεί τα σενάρια καλύτερης και χειρότερης περίπτωσης, τις κοινωνικοπεριβαλλοντικές απαιτήσεις και τους κινδύνους για την αυτονομία, τη σωματική βλάβη, την ψυχολογική βλάβη, τη βλάβη της φήμης, την οικονομική βλάβη, τα ανθρώπινα δικαιώματα, τον κοινωνικό/πολιτισμικό αντίκτυπο, τον πολιτικό/οικονομικό αντίκτυπο και τον περιβαλλοντικό αντίκτυπο. Για τις ερωτήσεις κινδύνου, οι κατηγορίες των κινδύνων ορίζονται με βάση την εργασία "A Collaborative, Human-Centred Taxonomy of AI, Algorithmic, and Automation Harms" ενώ οι άξονες ανάλυσής τους (σημαντικότητα, αναγκαιότητα, δυσκολία, ανιχνευσιμότητα και πιθανότητα) βασίζονται στους άξονες ανάλυσης των κινδύνων που προάγει το Εργαλείο Αλγοριθμική Εκτίμησης αντικτύπου του Ada Lovelace Institute και το Ερωτηματολόγιο εκτίμησης ρίσκου της Interpol. [38][28][32]

6. **Στάδιο Αξιολόγησης**

   Σε αυτό το στάδιο οι ερωτήσεις χωρίζονται σε 4 βασικές υποθεματικές:

   a. **Testing:** Αξιολογεί τη στρατηγική του testing, την τεκμηρίωση των επιδόσεων του μοντέλου, τα tests για ακραίες τιμές και τον εντοπισμό μοτίβων αποτυχίας.
   b. **Εγκατάσταση:** Αξιολογεί τις στρατηγικές εγκατάστασης και εξυπηρέτησης, τον εντοπισμό και τον μετριασμό των κινδύνων.
   c. **Έλεγχοι συμμόρφωσης:** Αξιολογεί τη συμμόρφωση με τα πρότυπα IEEE και ISO, τις εκτιμήσεις επιπτώσεων στην προστασία των δεδομένων και την τήρηση κωδίκων δεοντολογίας ή βέλτιστων πρακτικών.[48]

d. **Πιστοποίηση και έλεγχοι:** Αξιολογεί τη δυνατότητα ελέγχου (auditability) του συστήματος, τις διαδικασίες ελέγχου από τρίτους και την έρευνα του νομικού πλαισίου.

7. **Στάδιο Λειτουργίας**
   Σε αυτό το στάδιο οι ερωτήσεις χωρίζονται σε 2 βασικές υποθεματικές:
   a. **Συνεχής παρακολούθηση και ανατροφοδότηση:** Αξιολογεί τη στρατηγική ενημέρωσης, τις περιοδικές αναθεωρήσεις των ηθικών αξιών, την επικοινωνία των κινδύνων στους τελικούς χρήστες και τις διαδικασίες αναφοράς ευπαθειών.
   b. **Επιθέσεις και ασφάλεια:** Αξιολογεί την πολυπλοκότητα του περιβάλλοντος ανάπτυξης, τα μέτρα κυβερνοασφάλειας, τους μηχανισμούς εφεδρείας, την ανίχνευση εχθρικών επιθέσεων και την προστασία από δηλητηρίαση δεδομένων. [49]
8. **Στάδιο Απόσυρσης**
   Αυτό το στάδιο αξιολογεί τη διαδικασία παροπλισμού του συστήματος ΤΝ, συμπεριλαμβανομένης της εκτίμησης κινδύνου, της συμμετοχής των ενδιαφερομένων μερών, του χειρισμού δεδομένων και των περιβαλλοντικών εκτιμήσεων. Πιο συγκεκριμένα μερικά βασικά σημεία αυτού του σταδίου είναι τα εξής [50]:
   a. Αξιολόγηση κινδύνου: Ερωτάται αν διενεργήθηκε εκτίμηση κινδύνου πριν από τον παροπλισμό.
   b. Συμμετοχή των ενδιαφερομένων μερών: Αξιολογεί την έκταση της συμμετοχής των ενδιαφερομένων στη διαδικασία παροπλισμού.
   c. Χειρισμός δεδομένων: Αξιολογεί τον τρόπο χειρισμού των προσωπικών δεδομένων και των ευαίσθητων πληροφοριών κατά τον παροπλισμό.
   d. Περιβαλλοντικές εκτιμήσεις: Αξιολογεί την προσέγγιση του οργανισμού για την ελαχιστοποίηση των περιβαλλοντικών επιπτώσεων κατά τον παροπλισμό.
   e. Τεκμηρίωση και διαφάνεια: Αξιολογεί την τεκμηρίωση και τη διαφάνεια της διαδικασίας παροπλισμού.

4.2 Βάρη

Η αξιοπιστία ενός συστήματος έχει οριστεί με πολλούς διαφορετικούς τρόπους από πολλούς διαφορετικούς ερευνητές. Για το CapAI ένα αξιόπιστο σύστημα πρέπει να είναι νόμιμο, ηθικό και τεχνικά στιβαρό (capAI) ενώ μια άλλη προσέγγιση από το National Institute of Standards and Technology θέλει μια αξιόπιστη ΤΝ να είναι έγκυρη και αξιόπιστη, ασφαλής, ασφαλής και ανθεκτική, υπεύθυνη και διαφανής, επεξηγήσιμη και ερμηνεύσιμη, με αυξημένη προστασία της ιδιωτικής ζωής και δίκαιη και απαλλαγμένη από επιβλαβείς προκαταλήψεις [1][51].

Για τις ανάγκες αριθμητικής αξιολόγησης των απαντήσεων του εργαλείου Αλγοριθμική Εκτίμησης Αντικτύπου που κατασκευάζεται στα πλαίσια αυτής της εργασίας θα χρησιμοποιηθούν οι εξής μετρικές αξιοπιστίας όπως ορίζονται στην εργασία “A Unified Framework of Five Principles for AI in Society” από τους by Luciano Floridi and Josh Cowls [52]:

- **Ευεργεσία(Beneficence):** Προώθηση της ευημερίας, διατήρηση της αξιοπρέπειας και συντήρηση του πλανήτη. Η τεχνητή νοημοσύνη θα πρέπει να αναπτύσσεται προς όφελος της ανθρωπότητας και του περιβάλλοντος, διασφαλίζοντας ότι συμβάλλει θετικά στην κοινωνία και τον πλανήτη.
- **Μην κακοήθεια(Non-maleficence):** Αποφυγή βλάβης, ιδίως όσον αφορά την προστασία της ιδιωτικής ζωής, την ασφάλεια και την κακή χρήση της ΤΝ. Τα συστήματα τεχνητής νοημοσύνης θα πρέπει να σχεδιάζονται έτσι ώστε να αποφεύγονται αρνητικές συνέπειες,

όπως παραβιάσεις της ιδιωτικής ζωής, απειλές για την ασφάλεια και κακή χρήση που θα μπορούσαν να βλάψουν τα άτομα ή την κοινωνία.

- **Αυτονομία(Autonomy):** Διασφάλιση ότι οι άνθρωποι διατηρούν τον έλεγχο των διαδικασιών λήψης αποφάσεων. Η τεχνητή νοημοσύνη θα πρέπει να ενισχύει την ανθρώπινη αυτονομία και όχι να την υπονομεύει. Οι άνθρωποι θα πρέπει να έχουν τη δυνατότητα να αποφασίζουν πότε και πώς να αναθέτουν αποφάσεις σε συστήματα ΤΝ, και οι αποφάσεις αυτές θα πρέπει να είναι αναστρέψιμες.
- **Δικαιοσύνη(Justice):** Προώθηση της δικαιοσύνης, της ισότητας και της αλληλεγγύης. Η τεχνητή νοημοσύνη θα πρέπει να σχεδιαστεί για να εξαλείψει τις διακρίσεις, να προωθήσει την ποικιλομορφία και να διασφαλίσει ότι τα οφέλη της τεχνητής νοημοσύνης κατανέμονται δίκαια σε όλη την κοινωνία. Θα πρέπει επίσης να αποφεύγεται η δημιουργία νέων μορφών αδικίας ή προκατάληψης.
- **Εξηγησιμότητα(Explicability):** Διασφάλιση ότι τα συστήματα ΤΝ είναι διαφανή, κατανοητά και υπεύθυνα. Η τεχνητή νοημοσύνη θα πρέπει να σχεδιάζεται έτσι ώστε οι διαδικασίες λήψης αποφάσεων να είναι κατανοητές στους χρήστες και θα πρέπει να υπάρχει σαφής λογοδοσία για τα αποτελέσματα των συστημάτων τεχνητής νοημοσύνης.

Αυτές οι 5 μετρικές αποτελούν τη βάση της αριθμητικής αξιολόγησης του ερωτηματολογίου. Πιο συγκεκριμένα κάθε απάντηση σε ερώτηση κλειστού τύπου βαθμολογείται με 5 βάρη, ένα για κάθε μετρική, εύρους από -5 μέχρι 5 ανάλογα με το πόσο "σύμφωνη με τις καλές πρακτικές ανάπτυξης ΤΝ" ή όχι είναι αυτή. Τα βάρη αυτά αθροίζονται για να είναι δυνατόν στο τέλος να παρουσιαστεί το τελικό αποτέλεσμα ανά μετρική, το οποίο όσο μικρότερο είναι τόσο πιο σύμφωνο με τις καλές πρακτικές είναι το αξιολογούμενο σύστημα. Σε ερωτήσεις μοναδικής επιλογής, το μικρότερο βάρος αντιστοιχεί στην καλύτερη περίπτωση (δηλαδή, την πιο σύμφωνη με τις καλές πρακτικές), ενώ το μεγαλύτερο βάρος αντιστοιχεί στη χειρότερη περίπτωση. Σε ερωτήσεις πολλαπλών επιλογών, τα βάρη προστίθενται. Εάν κάθε επιπλέον απάντηση αυξάνει τον κίνδυνο, οι μετρικές είναι θετικές και το συνολικό βάρος αυξάνεται. Αντίθετα, εάν κάθε επιπλέον απάντηση μειώνει τον κίνδυνο ή προσφέρει μέσο προφύλαξης ή μετριασμού, οι μετρικές είναι αρνητικές και έτσι το συνολικό βάρος μειώνεται.Εάν η απάντηση είναι ουδέτερη, το βάρος παραμένει 0. Μοναδική εξαίρεση αυτού αποτελούν οι πίνακες αξιολόγησης ηθικών κινδύνων στο "Στάδιο Ανάπτυξης : Μοντέλο**"** από τους οποίους για κάθε συγκεκριμένη κατηγορία κινδύνου (σειρά στον πίνακα) υπολογίζεται το γινόμενο των Importance, Urgency, Difficulty, Detectability, Probability το οποίο αθροίζεται (πολλαπλασιασμένο με έναν συντελεστή κανονικοποίησης 0.5) στη μετρική μη-κακοήθειας. Στο τέλος υπολογίζεται και ένα ποσοστό ανά μετρική πέρα από το απόλυτο νούμερο για να είναι ξεκάθαρο πόσο συμβατό είναι το αξιολογούμενο σύστημα με το "ιδανικό" σύστημα της κάθε μετρικής.

Κεφάλαιο 5: Τεχνική Υλοποίηση

5.1 Τεχνολογίες που χρησιμοποιήθηκαν

5.1.1 SurveyJS

Το SurveyJS είναι μια σειρά βιβλιοθηκών JavaScript ανοικτού κώδικα, αδειοδοτημένη από το MIT, που έχουν σχεδιαστεί για να διευκολύνουν τη δημιουργία και τη διαχείριση δυναμικών ερευνών, φορμών και κουίζ [53]. Προσφέρει ένα εργαλείο δημιουργίας ερευνών (σε μορφή json), το οποίο διαθέτει ένα φιλικό προς το χρήστη περιβάλλον drag-and-drop, επιτρέποντας στους χρήστες να σχεδιάζουν έρευνες, να συλλέγουν απαντήσεις και να στέλνουν όλα τα δεδομένα της φόρμας σε μία βάση δεδομένων. Οι υποστηριζόμενοι τύποι ερωτήσεων περιλαμβάνουν:

- Εισαγωγή κειμένου(Text input): Πεδία μονής γραμμής για σύντομες απαντήσεις.
- Σχόλιο(Comment): Πεδία πολλαπλών γραμμών για λεπτομερείς απαντήσεις.
- Dropdown/Tagbox: Μενού που παρουσιάζουν επιλέξιμες επιλογές.
- Checkbox: Λίστες που επιτρέπουν πολλαπλές επιλογές.
- Radiogroup: Επιλογές που επιτρέπουν μία μόνο επιλογή.
- Boolean: Εναλλασσόμενοι διακόπτες για δυαδικές απαντήσεις (π.χ. αληθές/ψευδές).
- Βαθμολόγηση(Rating): Κλίμακες για αξιολογικές αξιολογήσεις.
- Πίνακας(Matrix/Matrixdropdown): Πλέγματα που θέτουν πολλαπλές ερωτήσεις με ομοιόμορφες επιλογές απαντήσεων.
- Πολλαπλό κείμενο (Multiple text): Ομάδες πεδίων εισόδου κάτω από ένα ενιαίο ερώτημα.

Επιπλέον, το SurveyJS προσφέρει υποστήριξη για την απρόσκοπτη ενσωμάτωση του ερωτηματολογίου json σε πλαίσια javascript όπως το Vue, το Angular ή το React.

5.1.2 React

Το React είναι μια βιβλιοθήκη JavaScript ανοικτού κώδικα που αναπτύχθηκε από το Facebook και χρησιμοποιείται κυρίως για την κατασκευή διεπαφών χρήστη, ειδικά σε εφαρμογές μίας σελίδας όπου η αποτελεσματική παρουσίαση δεδομένων είναι απαραίτητη. Επιτρέπει στους προγραμματιστές να κατασκευάζουν επαναχρησιμοποιήσιμα στοιχεία UI, προωθώντας συντηρήσιμες και κλιμακούμενες δομές κώδικα [53][54].

Το React ενθαρρύνει τη δημιουργία ενθυλακωμένων συστατικών που διαχειρίζονται τη δική τους κατάσταση, διευκολύνοντας έτσι την επαναχρησιμοποίηση. Επίσης βελτιώνει την απόδοση ενημερώνοντας μόνο τα τμήματα του Μοντέλου Αντικειμένου Εγγράφου (DOM) που έχουν αλλάξει, αντί για την επανεμφάνιση ολόκληρης της διεπαφής, ενώ απλοποιεί τη διαχείριση δεδομένων και την αποσφαλμάτωση, διασφαλίζοντας ότι τα δεδομένα κινούνται προς μία μόνο κατεύθυνση. Τέλος συγχωνεύει τη σύνταξη JavaScript και HTML, καθιστώντας τον κώδικα πιο διαισθητικό και πιο εύκολο στη συγγραφή.

5.1.3 Nginx server

Ο Nginx (προφέρεται «engine-x») είναι ένας διακομιστής ιστού υψηλής απόδοσης και αντίστροφος διακομιστής μεσολάβησης που φημίζεται για τη σταθερότητά του, το πλούσιο σύνολο

χαρακτηριστικών του και την αποδοτική χρήση των πόρων του. [56][57] Χρησιμοποιείται εκτενώς για την εξυπηρέτηση στατικού περιεχομένου, τη διαχείριση αντίστροφης διαμεσολάβησης, την εξισορρόπηση φορτίων και την εκτέλεση διαφόρων άλλων λειτουργιών διακομιστή ιστού.

Το Nginx είναι γνωστό για την αποτελεσματική εξυπηρέτηση στατικών στοιχείων όπως HTML, CSS, JavaScript και εικόνες. Λειτουργεί ως αντίστροφος μεσάζων προωθώντας τα αιτήματα των πελατών σε άλλους διακομιστές, κατανέμοντας έτσι το φορτίο και ενισχύοντας την επεκτασιμότητα. Επιπλέον, το Nginx λειτουργεί ως εξισορροπιστής φορτίου, κατανέμοντας την εισερχόμενη κίνηση σε πολλούς διακομιστές για να αποτρέψει την εμφάνιση συμφόρησης σε οποιονδήποτε διακομιστή. Διαχειρίζεται επίσης τον τερματισμό SSL/TLS, αναλαμβάνοντας την κρυπτογράφηση και αποκρυπτογράφηση της κυκλοφορίας για την ενίσχυση της ασφάλειας. Κατά την ανάπτυξη εφαρμογών, συμπεριλαμβανομένων εκείνων που έχουν κατασκευαστεί με πλαίσια όπως το React, το Nginx εξυπηρετεί τα μεταγλωττισμένα στατικά αρχεία, εξασφαλίζοντας γρήγορη και αξιόπιστη παράδοση περιεχομένου στους χρήστες.

5.2 Μέθοδος ανάπτυξης

Η ανάπτυξη αυτού του έργου ακολούθησε μια δομημένη προσέγγιση για τη δημιουργία μιας διαδικτυακής πλατφόρμας εκτίμησης επιπτώσεων. Η διαδικασία περιέλαβε τέσσερα κύρια στάδια: προετοιμασία ερωτηματολογίου, στάθμιση μετρικών, υλοποίηση front-end και ανάπτυξη διακομιστή για να καταστεί το εργαλείο δημόσια προσβάσιμο.

Το πρώτο στάδιο περιελάμβανε την κατασκευή του ερωτηματολογίου αξιολόγησης. Αντλώντας έμπνευση από το καναδικό εργαλείο εκτίμησης επιπτώσεων, χρησιμοποιήθηκε το SurveyJS Form Creator. Μέσα σε αυτό το περιβάλλον, εισήχθηκε το σύνολο των περίπου 300 ερωτήσεων, εξασφαλίζοντας τη σωστή δόμηση και κατηγοριοποίηση τους. Στη συνέχεια, το SurveyJS δημιούργησε αυτόματα ένα αρχείο JSON, το οποίο αποτέλεσε τη δομημένη βάση για το ερωτηματολόγιο και διευκόλυνε τη μετέπειτα ενσωμάτωσή του στο σύστημα.

Αφού προετοιμάστηκε το ερωτηματολόγιο, αναπτύχθηκε ένα αρχείο κώδικα Python για τον εμπλουτισμό του με τις πέντε μετρικές ηθικής αξιολόγησης (ευεργεσία, μη κακοήθεια, εξηγησιμότητα, δικαιοσύνη και αυτονομία). Αυτό το αρχείο επεξεργάστηκε το αρχείο JSON, προσθέτοντας μηδενικές τιμές βαρύτητας σε κάθε τύπο ερώτησης κλειστού τύπου (checkbox, radiogroup και boolean). Ύστερα, αφότου είχαν δημιουργηθεί τα κατάλληλα πεδία για τις μετρικές, το αρχείο υπέστει μη αυτόματη επεξεργασία έτσι ώστε να προστεθούν οι κατάλληλες τιμές μετρικών ανά απάντηση. Για τις αναδιπλούμενες ερωτήσεις πίνακα (matrix dropdown), το αρχείο Python απέδωσε βάρη που αντιστοιχούσαν στις επιλογές απάντησης (π.χ. η επιλογή του να είναι το Importance 3 αντιστοιχεί σε βάρος 3), διασφαλίζοντας ότι κάθε απάντηση μπορούσε να αξιολογηθεί ποσοτικά. Αυτό το βήμα μετέτρεψε την αρχική έρευνα σε ένα εργαλείο σταθμισμένης αξιολόγησης, έτοιμο για ενσωμάτωση σε ένα διαδραστικό σύστημα.

Μετά την προετοιμασία του σταθμισμένου ερωτηματολογίου, ξεκίνησε η ανάπτυξη μιας διαδικτυακής διεπαφής για την αλληλεπίδραση του χρήστη. Χρησιμοποιώντας ως βάση τα demo του SurveyJS, υλοποιήθηκε μια εφαρμογή βασισμένη σε JavaScript που φορτώνει το ερωτηματολόγιο JSON, αποδίδοντάς μέσω γραφική διεπαφής σε ένα πρόγραμμα περιήγησης. Αυτή η διεπαφή επιτρέπει στους προγραμματιστές ΤΝ και στους αξιολογητές συστημάτων να αλληλεπιδρούν με το ερωτηματολόγιο με δομημένο, φιλικό προς τον χρήστη τρόπο. Με την ολοκλήρωση της έρευνας, το σύστημα επεξεργάζεται τις απαντήσεις και υπολογίζει τις πέντε βαθμολογίες αξιολόγησης των επιπτώσεων, παρουσιάζοντάς τες τόσο σε απόλυτες αριθμητικές τιμές όσο και σε ποσοστιαίες αναπαραστάσεις. Στη συνέχεια δημιουργείται μια αναφορά PDF, η οποία περιέχει τόσο τις

υπολογισμένες βαθμολογίες όσο και το πλήρες σύνολο των απαντήσεων, παρέχοντας ένα δομημένο και διαφανές αποτέλεσμα αξιολόγησης.

Στο τελικό στάδιο, το σύστημα εγκαταστάθηκε σε κατάλληλο διακομιστή για να καταστεί δημόσια προσβάσιμο. Ο κώδικας που αναπτύχθηκε μεταφορτώθηκε σε έναν διακομιστή του Εθνικού Μετσόβιου Πολυτεχνείου, όπου διαμορφώθηκε ένας διακομιστής ιστού Nginx για τη φιλοξενία της εφαρμογής. Τα απαραίτητα αρχεία μεταφέρθηκαν στον διακομιστή και ο Nginx ρυθμίστηκε για να παρέχει το ερωτηματολόγιο στους χρήστες. Ως αποτέλεσμα, η πλατφόρμα αξιολόγησης είναι πλέον πλήρως λειτουργική και δημόσια προσβάσιμη στη διεύθυνση https://147.102.74.45/ ή https://aia.dslab.ece.ntua.gr/ , επιτρέποντας στους προγραμματιστές τεχνητής νοημοσύνης και στους ενδιαφερόμενους φορείς να αξιολογούν τα συστήματα τεχνητής νοημοσύνης με δομημένο και αποτελεσματικό τρόπο. (ο κώδικας είναι διαθέσιμος στο https://github.com/ntua-el19016/vscode)

5.3 Ανάλυση του κώδικα

Ο βασικός κώδικας του εργαλείου βρίσκεται στο αρχείο index.js το οποίο πρακτικά είναι ένα React Component το οποίο αναπαράγει μια έρευνα χρησιμοποιώντας τη βιβλιοθήκη SurveyJS, υπολογίζει τελικό σκορ ηθικών μετρικών με βάση τις απαντήσεις της έρευνας και δημιουργεί μια αναφορά PDF με τα αποτελέσματα. Ακολουθεί μια ανάλυση των βασικών τμημάτων:

1. **SurveyComponent**
    a. Αρχικοποίηση της έρευνας :

```
const survey = new Survey.Model(json_questionnaire);
const [surveyData, setSurveyData] = React.useState(null);
const [isSurveyCompleted, setIsSurveyCompleted] = React.useState(false);
```

Η survey αρχικοποιείται με το json_questionnaire που περιέχει τη δομή της έρευνας. Η κατάσταση surveyData περιέχει τα υπολογισμένα αποτελέσματα της έρευνας. Η κατάσταση isSurveyCompleted παρακολουθεί αν η έρευνα έχει ολοκληρωθεί.

    b. Χειριστής ολοκλήρωσης της έρευνας :

```
survey.onComplete.add((sender, options) => {
        const { beneficence, non_maleficence, autonomy, justice, explicability, totalScore,
questionDetails } = calculateScore(sender.data, json_questionnaire);
        setSurveyData({ beneficence, non_maleficence, autonomy, justice, explicability,
totalScore, questionDetails });
        setIsSurveyCompleted(true);
});
```

Όταν η έρευνα ολοκληρωθεί, καλείται η calculateScore για τον υπολογισμό των βαθμολογιών με βάση τις απαντήσεις. Τα αποτελέσματα αποθηκεύονται στην κατάσταση surveyData και η κατάσταση isSurveyCompleted τίθεται σε true.

    c. Χειριστής αποθήκευσης PDF:

```
const handleSavePdf = () => {
        if (surveyData) {
        savePdf(surveyData);
        }
};
```

Αυτή η συνάρτηση καλεί την savePdf για να δημιουργήσει και να αποθηκεύσει το αρχείο PDF, εάν είναι διαθέσιμα τα surveyData.

d. Χειριστής κλεισίματος του Modal :

```
const handleCloseModal = () => {
        setIsSurveyCompleted(false);
};
```

Αυτή η συνάρτηση κλείνει το modal θέτοντας το isSurveyCompleted σε false.

e. Απεικόνιση του component:

```
return (
  <>
        <SurveyReact.Survey model={survey} />

        {isSurveyCompleted && (
        <div id="savePdfModal">
        <div className="modal-content">
        <h3>Thank you for completing the survey!</h3>
        <p>Would you like to save your results as a PDF?</p>
        <button onClick={handleSavePdf} id="savePdfButton">Save as PDF</button>
        <button onClick={handleCloseModal} className="closeButton">Close</button>
        </div>
        </div>
        )}
  </>
);
```

Η απόδοση της έρευνας γίνεται με τη χρήση του SurveyReact.Survey. Εάν η έρευνα ολοκληρωθεί, εμφανίζεται ένα modal με επιλογές αποθήκευσης των αποτελεσμάτων ως PDF ή κλεισίματος του modal.

2. Συνάρτηση calculateScore

Αυτή η συνάρτηση υπολογίζει βαθμολογίες για διάφορες ηθικές αρχές με βάση τις απαντήσεις της έρευνας:

a. Αρχικοποίηση:

```
let beneficence = 0;
let non_maleficence = 0;
let autonomy = 0;
let justice = 0;
let explicability = 0;
const questionDetails = [];
```

b. Βοηθητική συνάρτηση:

```
const findQuestion = (elements, questionName) => {
        if (!elements) return null;
        for (const element of elements) {
        if (element.name === questionName) {
        return element;
        }
        if (element.type === "panel" && element.elements) {
```

```
            const found = findQuestion(element.elements, questionName);
            if (found) return found;
            }
            }
            return null;
};
```

- Αυτή η συνάρτηση αναζητά αναδρομικά μια ερώτηση μέσα σε ένθετα πάνελ.

c. Επανάληψη σε όλες τις απαντήσεις:

```
Object.keys(surveyData).forEach(question => {
            const response = surveyData[question];
            const questionElement = surveyJson.pages.flatMap(page => findQuestion(page.elements,
question)).find(el => el !== null);
            if (!questionElement) {
            console.warn(`Question ${question} not found in survey JSON`);
            return;
            }
            ...
});
```

Για κάθε απάντηση, βρίσκεται το αντίστοιχο στοιχείο ερώτησης. Οι βαθμολογίες για τις διάφορες μετρικές υπολογίζονται με βάση τον τύπο της απάντησης.

d. Επιστροφή των υπολογισμένων βαθμολογιών:

```
return { beneficence, non_maleficence, autonomy, justice, explicability, totalScore: beneficence +
non_maleficence + autonomy + justice + explicability, questionDetails };
```

Η συνάρτηση επιστρέφει τις υπολογισμένες βαθμολογίες και τις λεπτομέρειες κάθε ερώτησης.

3. Συνάρτηση savePdf

Αυτή η συνάρτηση παράγει μια αναφορά PDF με τα αποτελέσματα της έρευνας:

a. Αρχικοποίηση:

```
const { jsPDF } = window.jspdf;
const surveyPdf = new jsPDF();
const titleFontSize = 16;
const subtitleFontSize = 14;
const textFontSize = 12;
const margin = 20;
const lineHeight = 10;
const maxYPosition = 250;
const pageWidth = 180;
```

b. Συνάρτηση προσθήκης περιτυλιγμένου κειμένου:

```
const addWrappedText = (text, fontSize = textFontSize, color = 'black') => {
            surveyPdf.setFontSize(fontSize);
            surveyPdf.setTextColor(color);
            const lines = surveyPdf.splitTextToSize(text, pageWidth);
            lines.forEach(line => {
            if (yPosition + lineHeight > maxYPosition) {
            surveyPdf.addPage();
```

```
            yPosition = 20;
            }
            surveyPdf.text(line, margin, yPosition);
            yPosition += lineHeight;
            });
};
```

- Αυτή η συνάρτηση χειρίζεται το τύλιγμα του κειμένου και τα διαλείμματα σελίδων.

c. Λειτουργία υπολογισμού ποσοστού:

```
const calculatePercentage = (score, min, max) => {
            return ((score - min) / (max - min)) * 100;
}
```

- Αυτή η συνάρτηση υπολογίζει το ποσοστό μιας βαθμολογίας εντός ενός συγκεκριμένου εύρους.

d. Δημιουργία περιεχομένου PDF:

```
surveyData.questionDetails.forEach((question) => {
            addWrappedText(`${question.name}: ${question.title}`, subtitleFontSize, 'blue');
            addWrappedText(`Type: ${question.type}`, textFontSize, 'gray');
            addWrappedText(`Beneficence: ${question.beneficence}`, textFontSize, 'green');
            addWrappedText(`Non-maleficence: ${question.non_maleficence}`, textFontSize, 'red');
            addWrappedText(`Autonomy: ${question.autonomy}`, textFontSize, 'purple');
            addWrappedText(`Justice: ${question.justice}`, textFontSize, 'orange');
            addWrappedText(`Explicability: ${question.explicability}`, textFontSize, 'brown');
            if (Array.isArray(question.answer)) {
            question.answer.forEach((answer) => {
            addWrappedText(`Answer: ${answer}`, textFontSize, 'black');
            });
            } else {
            addWrappedText(`Answer: ${question.answer}`, textFontSize, 'black');
            }
            yPosition += lineHeight;
});
```

Οι λεπτομέρειες κάθε ερώτησης προστίθενται στο PDF.

e. Προσθήκη συνολικής βαθμολογίας:

```
const totalScore = surveyData.beneficence + surveyData.non_maleficence +
surveyData.autonomy + surveyData.justice + surveyData.explicability;
addWrappedText(`Total Score: ${totalScore}`, titleFontSize, 'black');
```

Οι συνολικές βαθμολογίες για κάθε ηθική αρχή και η συνδυασμένη συνολική βαθμολογία προστίθενται στο PDF.

f. Αποθήκευση PDF:

```
surveyPdf.save("survey_results.pdf");
```

Το PDF αποθηκεύεται με το όνομα αρχείου survey_results.pdf.

Κεφάλαιο 6: Συγκριτική Μελέτη και Βελτιστοποίηση

6.1 Βασικός Στόχος

Η κύρια έμπνευσή για το σχεδιασμό αυτού του ερωτηματολογίου εκτίμησης επιπτώσεων ήταν η διαδικασία capAI και το καναδικό εργαλείο εκτίμησης επιπτώσεων. [1][2] Και τα δύο πλαίσια προσφέρουν διακριτά πλεονεκτήματα και αδυναμίες, για τα οποία έγινε προσπάθεια να εξισορροπηθούν και να ενσωματωθούν σε μια πιο ολοκληρωμένη και πρακτική προσέγγιση.

Η διαδικασία **capAI** είναι ένα ιδιαίτερα ηθικά στιβαρό, οργανωμένο και μεθοδικό πλαίσιο που αντιμετωπίζει συστηματικά τους πιθανούς κινδύνους από την ανάπτυξη μοντέλων τεχνητής νοημοσύνης. Υπερέχει για τη σχολαστικότητά της και την ικανότητά της να καλύπτει ένα ευρύ φάσμα ηθικών προβληματισμών, διασφαλίζοντας ότι καμία κρίσιμη πτυχή της ανάπτυξης της τεχνητής νοημοσύνης δεν παραβλέπεται. Ωστόσο, ένας από τους περιορισμούς της είναι η εξάρτησή της από ανοιχτές ερωτήσεις που συχνά είναι πολύ γενικές. Ενώ οι ερωτήσεις αυτές ενθαρρύνουν τον βαθύ προβληματισμό σχετικά με τις ηθικές αρχές, στερούνται της εξειδίκευσης που απαιτείται για να μεταφραστούν αυτές οι ηθικές αξίες σε πρακτικά, εφαρμόσιμα βήματα στον τομέα της ανάπτυξης της ΤΝ. Με άλλα λόγια, το capAI παρέχει ένα ισχυρό ηθικό θεμέλιο, αλλά υπολείπεται στο να προσφέρει συγκεκριμένες οδηγίες για το πώς να λειτουργήσουν αυτές οι αρχές σε έργα τεχνητής νοημοσύνης στον πραγματικό κόσμο.

Από την άλλη πλευρά, το **καναδικό εργαλείο αξιολόγησης** επιπτώσεων ακολουθεί μια πιο ρεαλιστική προσέγγιση. Περιλαμβάνει συγκεκριμένες, πρακτικές ερωτήσεις -όπως το αν ο αλγόριθμος στον οποίο βασίζεται η ελεγχόμενη ΤΝ είναι εμπορικό μυστικό ή οχι - και προσφέρει απαντήσεις πολλαπλής επιλογής που επιτρέπουν τον υπολογισμό μιας αριθμητικής βαθμολογίας κινδύνου και μετριασμού. Η προσέγγιση αυτή μετατρέπει τις ασαφείς, ιδεολογικές έννοιες σε μετρήσιμα μεγέθη, αξιοποιώντας τις εφαρμοσμένες γνώσεις της επιστήμης των υπολογιστών για να παρέχει μια σαφή, ποσοτικοποιήσιμη αξιολόγηση των κινδύνων. Ωστόσο, ενώ το καναδικό μοντέλο υπερέχει σε πρακτικότητα και εξειδίκευση, δεν διαθέτει το βάθος και την ολοκληρωμένη κατανόηση των πολύπλευρων κινδύνων της ΤΝ που προσφέρει το capAI. Δεν αποτυπώνει πλήρως την ηθική πολυπλοκότητα ή τις ευρύτερες κοινωνικές επιπτώσεις των συστημάτων ΤΝ, εστιάζοντας περισσότερο στις τεχνικές και διαδικαστικές πτυχές.

Όπως έχουμε ήδη αναφέρει με βάση την έρευνα "**An institutional view of Algorithmic Impact Assessments**" κατανοούμε ότι οι αξιολογήσεις των επιπτώσεων της ΤΝ μπορούν να χωριστούν σε τρία είδη: 1) αξιολογήσεις με βάση τη NEPA, οι οποίες αποτελούνται από λεπτομερείς, ανοικτού τύπου ερωτήσεις 2) μοντέλα DPIA (Data Protection Impact Assessment), τα οποία περιλαμβάνουν συστηματικές περιγραφές της επεξεργασίας, αιτιολογήσεις και σχέδια για τον μετριασμό των κινδύνων και 3) μοντέλα με βάση ερωτηματολόγιο και συστήματα βαθμολόγησης, όπως η καναδική AEA [33]. Στόχος αυτής της εργασίας είναι ο συνδυασμός των πλεονεκτημάτων αυτών των προσεγγίσεων σε ένα ενιαίο, συνεκτικό πλαίσιο.

Το ερωτηματολόγιο που αναπτύχθηκε στα πλαίσια αυτής της εργασίας είναι μια προσπάθεια να συγχωνευτεί η εξαντλητική, ηθικά εύρωστη προσέγγιση των μοντέλων NEPA/DPIA -ιδίως της μεθόδου capAI- με την πρακτικότητα και την εξειδίκευση της καναδικής AEA. Στόχος ήταν να δημιουργηθεί ένα εργαλείο το οποίο όχι μόνο θα προσδιορίζει και θα μετριάζει τους κινδύνους ολοκληρωμένα, αλλά και θα παρέχει έναν δομημένο, μετρήσιμο τρόπο για την αξιολόγηση των εν λόγω κινδύνων. Για να επιτευχθεί αυτό, σχεδιάστηκε ένα ερωτηματολόγιο που περιλαμβάνει τόσο

ερωτήσεις ανοικτού τύπου για να ενθαρρύνει τον βαθύ ηθικό προβληματισμό όσο και ερωτήσεις κλειστού τύπου για να παρέχει σαφείς, μετρήσιμες μετρήσεις. Επιπλέον, το σύστημα βαθμολόγησης για την αξιολόγηση των επιπέδων κινδύνου και των προσπαθειών μετριασμού που αναπτύχθηκε είναι παρόμοιο με αυτό του καναδικού μοντέλου, αλλά με ευρύτερο πεδίο εφαρμογής που καταγράφει τις ηθικές και κοινωνικές διαστάσεις της ανάπτυξης της ΤΝ.

Στην ουσία, το ερωτηματολόγιο αυτό είναι μια εκτεταμένη, πιο ολοκληρωμένη έκδοση του καναδικού AEA, εμπλουτισμένη με την ηθική αυστηρότητα και τη διεξοδικότητα της διαδικασίας capAI. Συνδυάζοντας αυτά τα στοιχεία, στοχεύθηκε η δημιουργία ενός εργαλείου που είναι τόσο πρακτικό όσο και ηθικά ισχυρό, ικανό να καθοδηγήσει τους προγραμματιστές ΤΝ μέσα στο σύνθετο τοπίο των κινδύνων και των ευθυνών, παρέχοντας παράλληλα μετρήσιμα αποτελέσματα για την ενημέρωση της λήψης αποφάσεων.

6.2 Συγκριτική Μελέτη Μεθοδολογίας

Εκτός από τη διαδικασία capAI και το καναδικό εργαλείο εκτίμησης επιπτώσεων, το ερωτηματολόγιο που αναπτύχθηκε στο πλαίσιο του παρόντος εγγράφου ενημερώθηκε από ένα ευρύ φάσμα υφιστάμενων πλαισίων εκτίμησης επιπτώσεων. Κάθε ένα από αυτά τα πλαίσια συνεισέφερε μοναδικά πλεονεκτήματα, τα οποία ενσωματώθηκαν προσεκτικά στο σχεδιασμό του παρόντος ερωτηματολογίου, ενώ οι περιορισμοί τους ανέδειξαν την ανάγκη για μια πιο ολοκληρωμένη και καινοτόμο προσέγγιση. Ακολουθεί ανάλυση αυτών των πλαισίων, των συνεισφορών τους και του τρόπου με τον οποίο επηρέασαν την ανάπτυξη του παρόντος ερωτηματολογίου.

Όπως έχουμε ήδη αναφέρει η αξιολόγηση του **Ινστιτούτου Ada Lovelace** είναι ένα εξειδικευμένο εργαλείο που επικεντρώνεται σε δεδομένα υγείας, γεγονός που περιορίζει εγγενώς τη δυνατότητα εφαρμογής του σε ευρύτερα πλαίσια ΤΝ. [28] Ωστόσο, εισάγει μια πολύτιμη άσκηση αναστοχασμού για τους συμμετέχοντες, ενθαρρύνοντάς τους να δώσουν πιο συγκεκριμένες και ολοκληρωμένες απαντήσεις. Παρόλο που αυτό το αναστοχαστικό στοιχείο δεν ενσωματώθηκε άμεσα στο ερωτηματολόγιο που αναπτύχθηκε εδώ, η προσέγγιση του πλαισίου Ada Lovelace για την αξιολόγηση των βλαβών ήταν ιδιαίτερα διορατική. Αξιολογεί κάθε πιθανή βλάβη από την άποψη της σημασίας, του έκτακτου χαρακτήρα, της δυσκολίας και της ανιχνευσιμότητας, προσφέροντας μια διαφοροποιημένη προοπτική για τον κίνδυνο(Importance, Urgency, Difficulty, Detectability). Ωστόσο, δεν διαθέτει αριθμητικό σύστημα βαθμολόγησης για την ποσοτικοποίηση αυτών των αξιολογήσεων, γεγονός που περιορίζει την πρακτική του χρησιμότητα. Αυτό το κενό αντιμετωπίστηκε στην ενότητα Development Stage : Model : Ethical Considerations στο ερωτηματολόγιο που αναπτύχθηκε εδώ, όπου εισήχθη ένα αριθμητικό σύστημα βαθμολόγησης για την παροχή μετρήσιμων αποτελεσμάτων.

Το Τυποποιημένο και το Εκτεταμένο Ερωτηματολόγιο του **European Law Institute** αποτελούνται σε μεγάλο βαθμό από ερωτήσεις ανοικτού τύπου, οι οποίες, όπως συζητήθηκε προηγουμένως, μπορεί να είναι υπερβολικά γενικές και να στερούνται εξειδίκευσης [29]. Ενώ αυτά τα ερωτηματολόγια χρησίμευσαν ως πολύτιμη πηγή έμπνευσης για το περιεχόμενο των επιμέρους ερωτήσεων, η εξάρτησή τους από ερωτήσεις ανοικτού τύπου χωρίς δομημένο μηχανισμό βαθμολόγησης τα κατέστησε λιγότερο κατάλληλα για πρακτική, εφαρμοσμένη χρήση στην ανάπτυξη ΤΝ.

Το EU AI Act Checker, που αναπτύχθηκε από το **Future of Life Institute**, εισάγει την έννοια του δυναμικού ερωτηματολογίου, το οποίο προσαρμόζεται με βάση τις απαντήσεις των χρηστών [34] . Αν και η καινοτομία αυτή είναι πολλά υποσχόμενη, το εργαλείο είναι περιορισμένο σε πεδίο εφαρμογής, εστιάζοντας κυρίως στη βασική συμμόρφωση με τον νόμο της ΕΕ για την ΤΝ

και παραλείποντας ουσιαστικές ερωτήσεις αξιολόγησης κινδύνων ή μετριασμού. Η έννοια του δυναμικού ερωτηματολογίου, ωστόσο, ενσωματώθηκε στο εργαλείο που αναπτύχθηκε εδώ, ιδίως σε περιπτώσεις όπου μια ερώτηση κλειστού τύπου απαιτεί πρόσθετες διευκρινίσεις. Για παράδειγμα, αν ένας προγραμματιστής απαντήσει «ναι» σε μια ερώτηση σχετικά με την ύπαρξη μιας συγκεκριμένης διαδικασίας, μια επακόλουθη ανοιχτή ερώτηση τον καλεί να περιγράψει τη διαδικασία αυτή. Αυτή η δυναμική προσέγγιση ενισχύει την ευελιξία και το βάθος του ερωτηματολογίου χωρίς να θυσιάζει τη σαφήνεια.

Η αξιολόγηση επιπτώσεων για συστήματα τεχνητής νοημοσύνης και συστήματα αυτόματης λήψης αποφάσεων του **Ιδρύματος Moje Państwo** είναι ένα άλλο παράδειγμα πλαισίου που βασίζεται σε μεγάλο βαθμό σε ερωτήσεις ανοικτού τύπου, με λίγες μόνο ερωτήσεις πολλαπλής επιλογής ή ερωτήσεις boolean, οπότε λειτούργησε και αυτο σε μεγάλο βαθμό σαν πηγή έμπνευσης για μεμονωμένες ερωτήσεις χωρίς να συμμετέχει ιδιαίτερα στην διαμόρφωση της δομής του ερωτηματολογίου του παρόντος εγγράφου [30].

Μια σημαντική καινοτομία εντοπίστηκε στο ερωτηματολόγιο της **INTERPOL και της UNICRI**, το οποίο ορίζει τον κίνδυνο ως Κίνδυνος = Πιθανότητα × Επιπτώσεις, αποδίδοντας βαθμολογίες από 1 έως 5 για κάθε παράγοντα [32]. Η προσέγγιση αυτή παρέχει μια σαφή, ποσοτικοποιήσιμη μέθοδο για την αξιολόγηση των κινδύνων, η οποία προσαρμόστηκε και επεκτάθηκε στην ενότητα Development Stage : Model : Ethical Considerations στο ερωτηματολόγιο που αναπτύχθηκε εδώ. Με βάση αυτό το πλαίσιο σε συνδυασμό με την προσέγγιση του Ada Lovelace Institute που αναφέρθηκε παραπάνω, ο κίνδυνος βελτιώθηκε περαιτέρω ως εξής: Κίνδυνος = Σημασία × Έκτακτο × Δυσκολία × Ανιχνευσιμότητα × Πιθανότητα, ορίζοντας ουσιαστικά τον αντίκτυπο ως συνδυασμό πολλαπλών παραγόντων. Αυτή η πιο λεπτομερής προσέγγιση επιτρέπει μια πιο ακριβή και πρακτική αξιολόγηση των κινδύνων, αντιμετωπίζοντας έναν βασικό περιορισμό του μοντέλου της INTERPOL και της UNICRI.

Το εργαλείο αξιολόγησης επιπτώσεων της **AI4Belgium** εισήγαγε μια καινοτόμο προσέγγιση για κλιμακούμενες απαντήσεις, μετατρέποντας τις ερωτήσεις που παραδοσιακά θα ήταν ερωτήσεις τύπου ναι ή όχι σε πιο διαφοροποιημένες, πολυεπίπεδες απαντήσεις [35]. Αυτές οι κλιμακούμενες απαντήσεις συχνά ακολουθούνταν από ανοικτές ερωτήσεις αιτιολόγησης, συνδυάζοντας τα πλεονεκτήματα των φιλικών προς τις μετρήσεις απαντήσεων με τη σαφήνεια των λεπτομερών εξηγήσεων. Η τεχνική αυτή ενσωματώθηκε σε αρκετές ενότητες του ερωτηματολογίου που αναπτύχθηκε εδώ, ιδίως στις ενότητες Development Stage : Model : Ethical Considerations και Development Stage : Model : Traceability ενισχύοντας τόσο την πρακτικότητα όσο και το βάθος του.

Ομοίως, το ερωτηματολόγιο **CISCO** επέκτεινε την προσέγγιση των κλιμακούμενων απαντήσεων, προσφέροντας λεπτομερείς επιλογές απάντησης για τις ερωτήσεις «ναι ή όχι» [36]. Για παράδειγμα, αντί για μια απλή απάντηση ναι ή όχι στην ερώτηση αν υπάρχει κάποια οργανωμένη στρατηγική αντιμετώπισης κάποιου κινδύνου, παρείχε επιλογές όπως: «Ναι - Υπάρχει μια βραχυπρόθεσμη και μακροπρόθεσμη στρατηγική», “Ναι - Υπάρχει μόνο μια βραχυπρόθεσμη οικονομική στρατηγική”, “Όχι - Αλλά αναπτύσσουμε επί του παρόντος μια οικονομική στρατηγική” και “Όχι - Δεν σχεδιάζουμε να αναπτύξουμε μια οικονομική στρατηγική”. Η μέθοδος αυτή ήταν ιδιαίτερα χρήσιμη για ερωτήσεις που στοχεύουν τον οργανισμό πίσω από την ανάπτυξη του συστήματος ΤΝ και ενσωματώθηκε στο Design Stage : Organisation του ερωτηματολογίου που αναπτύχθηκε εδώ. Η προσέγγιση των κλιμακούμενων απαντήσεων εφαρμόστηκε επίσης σε άλλες ενότητες, όπως το Retirement Stage , το Operation Stage και το Evaluation Stage, αντλώντας περιεχόμενο από διάφορες πηγές για να διασφαλιστεί η ολοκληρωμένη κάλυψη.

Τέλος, η αξιολόγηση της **Microsoft** αποτελείται κυρίως από ερωτήσεις ανοικτού τύπου, με λίγες μόνο ερωτήσεις πολλαπλής επιλογής, ενώ το ερωτηματολόγιο της **CNIL** κυρίως έχει ομοιότητες με το καναδικό εργαλείο εκτίμησης επιπτώσεων, ιδίως όσον αφορά την πρακτική μορφή

του, με πολλαπλές επιλογές [37][31]. Και τα δύο εργαλεία αυτά λόγω της έλλειψης καινοτομίας στη μορφή και τη δομή τους χρησιμοποιήθηκαν απλά σταν πηγή για το περιεχόμενο μεμονωμένων ερωτήσεων.

Συνοψίζοντας, ενώ καθένα από αυτά τα πλαίσια συνέβαλε με πολύτιμες γνώσεις και μεθοδολογίες, αποκάλυψε επίσης σημαντικά κενά στον τομέα των αξιολογήσεων επιπτώσεων της ΤΝ.Το ερωτηματολόγιο που αναπτύχθηκε στο πλαίσιο του παρόντος εγγράφου επιδιώκει να αντιμετωπίσει αυτά τα κενά συνδυάζοντας τα πλεονεκτήματα των υφιστάμενων πλαισίων -όπως τον έξυπνο ορισμό κινδύνου του Ινστιτούτου Ada Lovelace, τη δυναμική προσαρμοστικότητα του EU AI Act Checker και την κλιμακούμενη προσέγγιση των απαντήσεων των Al4Belgium και CISCO- ενώ παράλληλα εισάγει καινοτόμα στοιχεία όπως η εκλεπτυσμένη φόρμουλα αξιολόγησης κινδύνου και ένα ισορροπημένο μείγμα ερωτήσεων ανοικτού και κλειστού τύπου.Με την ενσωμάτωση αυτών των χαρακτηριστικών, το παρόν ερωτηματολόγιο αποτελεί ένα σημαντικό βήμα προς τα εμπρός στον τομέα, προσφέροντας ένα πιο ολοκληρωμένο, πρακτικό και ηθικά ισχυρό εργαλείο για την αξιολόγηση των επιπτώσεων των συστημάτων τεχνητής νοημοσύνης.

6.3 Συγκριτική Μελέτη Περιεχομένου

Είναι λογικό αφού το περιεχόμενο του ερωτηματολογίου που αναπτύχθηκε σε αυτή την εργασία έχει προκύψει από το σύνολο των ερωτήσεων των εργαλείων που αναλύσαμε και παραπάνω να μην έχει νόημα η συγκριτική μελέτη περιεχομένου αυτών, καθώς η νοηματική πληρότητα ήταν ο βασικός λόγος για τον οποίο μελετήθηκαν τα 10 αυτά εργαλεία εξ αρχής. Έτσι θα στραφούμε στην έρευνα «A comparative review of 10 Fundamental Rights Impact Assessments (FRIA) for AI-systems» όπου 10 διαφορετικά (μοναδική επικάλυψη με τη δική μας μελέτη είναι το εργαλείο του Καναδά) ΑΕΑ αναλύονται με τον εξής τρόπο [58] : Τα δέκα FRIA για συστήματα ΤΝ, αξιολογούνται τα με βάση δώδεκα κριτήρια που καλύπτουν νομικές, οργανωτικές, τεχνικές και κοινωνικές διαστάσεις νοηματικής πληρότητας με βάση πάντα τις απαιτήσεις του νόμου περί ΤΝ. Η αξιολόγηση αυτή γίνεται με την εφαρμογή μιας δυαδικής τιμής (ναι ή όχι) για να προσδιοριστεί κατά πόσον κάθε FRIA πληροί το κάθε κριτήριο αξιολόγησης. Το έγγραφο διαπιστώνει ότι καμία από τις δέκα FRIA δεν πληροί πλήρως και τα δώδεκα κριτήρια αξιολόγησης, αναδεικνύοντας σημαντικά κενά στα υφιστάμενα πλαίσια. Ενώ ορισμένες FRIA βαθμολογούνται σχετικά υψηλά -όπως αυτές του Ινστιτούτου Alan Turing- καμία δεν επιτυγχάνει τέλεια βαθμολογία, υποδεικνύοντας περιθώρια βελτίωσης, ιδίως στην τεχνική και κοινωνική διάσταση. Για λόγους συγκριτικής μελέτης λοιπόν θα εξετάσουμε πόσα κριτήρια μπορεί το ερωτηματολόγιο αυτής της εργασίας να περάσει.

Το 1ο νομικό κριτήριο αφορά τα θεμελιώδη δικαιώματα του ανθρώπου που μπορεί να επηρεαστούν από τη χρήση ενός συστήματος ΤΝ. Αυτό το κριτήριο καλύπτεται από το ερωτηματολόγιο καθώς η περιγραφή διαδικασιών, χρονικών περιθωρίων χρήσης, ανάλυση των φυσικών προσώπων που επηρεάζονται από τις αποφάσεις και με ποιους συγκεκριμένου τρόπους μπορεί να προκληθεί βλάβη σε αυτά τα άτομα καλύπτονται από την ενότητα Design stage: Use Case. Επίσης, ερωτήματα ανθρώπινης παρεμβατικότητας/ελέγχου της λειτουργίας και μηχανισμών υποβολής καταγγελιών καλύπτονται στην ενότητα Operation Stage. Τέλος, συγκεκριμένες αναφορές σε πιθανές μεροληψίες του μοντέλου γίνονται και στην ενότητα Development Stage : Data : Data Quality and Bias αλλά και στην ενότητα Development Stage : Model : Fairness.

Το 2ο νομικό κριτήριο αφορά την προστασία δεδομένων και από την άποψη της ιδιωτικότητας και τη συμβατότητα με τον κανονισμό ΓΚΠΔ αλλά και από την άποψη αποφυγή μεροληψιών, Τα ερωτήματα αυτά καλύπτονται από το ερωτηματολόγιο σε όλη την ενότητα

Development Stage : Data αλλά και στην Development Stage : Model : Fairness με συγκεκριμένες αναφορές σε DPIA να γίνονται και στην ενότητα Evaluation Stage : Compliance Checks.

Το 3ο νομικό κριτήριο αφορά ερωτήσεις Διοικητικού Δικαίου. Οι ανησυχίες διαφάνειας καλύπτονται από το ερωτηματολόγιο με τις ερωτήσεις στις ενότητες Development Stage : Explainability and Transparency και Development Stage : Model :Traceability, οι διαβεβαιώσεις μη-μεροληψίας γινονται στην ενότητα Development Stage : Model : Fairness και η μη-κακοήθεια ελέγχεται σε διάφορα σημεία των ενοτήτων Development Stage : Model, Evaluation Stage και Operation Stage.

Το 4ο νομικό κριτήριο αφορά την εξισορρόπηση πιθανών αρνητικών επιπτώσεων που μπορεί να έχει μια ΤΝ με το γενικό καλό που μπορεί να προσφέρει. Αυτό καλύπτεται από το ερωτηματολόγιο αυτής της εργασίας καθώς περιλαμβάνει και ερωτήσεις αξιολόγησης των κινδύνων που μπορεί να δημιουργήσει η χρήση του (Ενότητα Development Stage : Model : Ethical Considerations) αλλά και ερωτήσεις για το πόσο αναγκαίο/χρήσιμο είναι το εργαλείο στην ενότητα Design Stage : Use case. Επίσης, αυτή η εξισορρόπηση μοντελοποιείται με τις μετρικές beneficence και non-maleficence και έτσι μετά τη συμπλήρωση του ερωτηματολογίου μπορεί να υπάρξει άμεσα μια πρώτη εικόνα του πόσο καλά τα πάει η εξεταζόμενη ΤΝ στο δίπολο που εκφράζεται από αυτό το κριτήριο. Επίσης, σημαντικό είναι να αναφέρουμε ότι και ερωτήσεις ανθρώπινης επίβλεψης και παρεμβατικότητας στις αποφάσεις του μοντέλου καλύπτονται στις ενότητες Design Stage : Use Case και Operation Stage καθώς και αυτό είναι ένα μέτρο διασφάλισης του κριτηρίου αυτού.

Το 5ο και 6ο είναι οργανωτικά κριτήρια που αφορούν τον σκοπό του εργαλείου ΤΝ και την οργανωτική διακυβέρνηση κατά τη διάρκεια ανάπτυξής του τα οποία καλύπτονται από τις ενότητες του ερωτηματολογίου Design Stage : Use Case και Design Stage : Organisation αντίστοιχα.

Το 7ο και το 8ο είναι τεχνικά κριτήρια που αφορούν την τεκμηρίωση (documentation) και την ποιότητα των δεδομένων τα οποία καλύπτονται από τις ενότητες Development Stage : Model : Traceability και Development Stage : Data αντίστοιχα.

Το 9ο και το 10ο είναι επίσης τεχνικά κριτήρια που αφορούν την ακρίβεια μιας εφαρμογής ΤΝ και τις μεθόδους ελέγχου αντίστοιχα. Αυτά τα κριτήρια καλύπτονται από το ερωτηματολόγιο με τις ερωτήσεις των ενοτήτων Development Stage : Model : Robustness and testing the tool ,Evaluation Stage : Testing, Operation Stage : Attacks and Security και σε μερικά σημεία των Development Stage : Explainability and Transparency, Development Stage : Model : Fairness και Development Stage : Continuous Monitoring and feedback.

Το 11ο και το 12ο είναι κοινωνικά κριτήρια που αφορούν την ανάμειξη ποικίλων και πολλών stakeholder στη διαδικασία σχεδιασμού της αξιολόγησης αντικτύπου αλλά και στη διαδικασία επίλυσης των προβλημάτων που προκύπτουν από την αξιολόγηση. Το ερωτηματολόγιο που αναπτύχθηκε στην πορεία αυτής της εργασίας, παρά των δυναμικών χαρακτηριστικών του και των ερωτήσεων ανοιχτού τύπου του, έχει γενικά περιορισμένη δυνατότητα να εντάξει περαιτέρω ερωτήσεις/άξονες ανά περίπτωση συμπεριλαμβάνοντας τους stakeholders κάθε συστήματος ΤΝ στηδιαμόρφωση και απάντησή τους. Από την άλλη όμως η προσέγγιση που έχει ακολουθηθεί για τον σχεδιασμό των ερωτήσεων δεν ακολουθεί ούτε μια αμιγώς τεχνοκρατική λογική αλλά ούτε μια αμιγώς ανθρωπιστική λογική χωρίς απτές αναφορές σε τεχνικά εργαλεία, πράγμα που μπορεί να μιμείται τα αποτελέσματα ανάμειξης ποικίλων ενδιαφερόμενων στη διαδικασία σχεδιασμού του. Επίσης, το ερωτηματολόγιο μπορεί να συμπληρωθεί συνεργατικά, με τη συμβολή ατόμων που εντάσσονται στους stakeholders ανά περίπτωση, ενώ ενθαρρύνεται μέσα από τις ερωτήσεις του σε διάφορα σημεία η συμπερίληψη στην ανάπτυξη της ΤΝ ποικιλίας επηρεαζόμενων ατόμων. Τέλος, μια τμηματική-ανά στάδιο συμπλήρωση του ερωτηματολογίου συνδυασμένη συμβολή και αξιολόγηση των αποτελεσμάτων από ενδιαφερόμενα για την αξιολογούμενη ΤΝ άτομα μπορεί να θεωρηθεί ότι εκπληρώνει πλήρως και τα δύο κοινωνικά κριτήρια που αξιολογούνται σε αυτή την παράγραφο.

Συμπερασματικά, παρατηρούμε αρχικά στο κριτήριο 10 που κάθε ένα από τα 10 αξιολογούμενα ΑΕΑ απέτυχε, το ερωτηματολόγιο αυτής της εργασίας κατάφερε να επιτύχει. Επίσης, ακόμα και αν προσμετρηθούν ως "όχι" τα δύο τελευταία κριτήρια το ερωτηματολόγιο καταφέρνει να κερδίσει μια βαθμολογία 10/12 που μόνο άλλα 2 ΑΕΑ κατάφεραν να πετύχουν ή να ξεπεράσουν. Είναι, λοιπόν, προφανής η αξία του σαν εργαλείο ελέγχου συμβατότητας ΤΝ με τις αντίστοιχες νομοθεσίας αλλά και η πρακτική συμβολή του στη γενικότερη προσπάθεια ανάπτυξης σωστών, στιβαρών και διεξοδικών ΑΕΑ.

Κριτήρια	Alan Turing Institute	Algorithm Watch	Aligner	BSR	Danish Institute for Human Rights	Demos Helsinki	ForHumanity	Government of Canada	UNESCO	Utrecht University-Dutch Ministry of the Interior	ΕΜΠ
Fundamental rights	✅	✅	✅	✅	✅	✅	✅	❌	✅	✅	✅
Data protection	✅	✅	✅	❌	✅	✅	❌	✅	✅	✅	✅
National administrative law	✅	✅	❌	❌	❌	❌	❌	❌	❌	✅	✅
Proportionality test	✅	❌	❌	❌	❌	✅	❌	❌	✅	✅	✅
Intended purpose of algorithm	✅	✅	✅	❌	✅	✅	✅	✅	✅	✅	✅
Organizational measures	✅	✅	✅	❌	✅	✅	✅	✅	✅	✅	✅
Documentation instructions	✅	✅	✅	❌	✅	✅	❌	✅	✅	✅	✅
Data quality	✅	✅	✅	❌	✅	✅	❌	✅	✅	✅	✅
Accuracy specifications	✅	✅	✅	❌	❌	✅	❌	❌	✅	✅	✅
Methodological control measure	❌	❌	❌	❌	❌	❌	❌	❌	❌	❌	✅
Stakeholder panel to identify normative questions	✅	✅	❌	✅	✅	✅	✅	✅	✅	✅	✅*
Stakeholder panel to resolve normative questions	✅	❌	❌	❌	✅	❌	✅	❌	❌	❌	✅*
Total	11/12	9/12	7/12	3/12	8/12	9/12	5/12	6/12	9/12	10/12	12*/12

Κεφάλαιο 7: Συμπεράσματα και Επεκτάσεις

7.1 Συμπεράσματα

Η Τεχνητή Νοημοσύνη έχει γίνει αναπόσπαστο μέρος της σύγχρονης κοινωνίας, επαναστατώντας τις βιομηχανίες, ενισχύοντας την παραγωγικότητα και προσφέροντας καινοτόμες λύσεις σε σύνθετα προβλήματα. Ωστόσο, η ταχεία διάδοση των τεχνολογιών τεχνητής νοημοσύνης έχει επίσης εγείρει σημαντικές ανησυχίες σχετικά με τους πιθανούς κινδύνους τους. Αυτοί οι κίνδυνοι είναι πολυδιάστατοι, περιλαμβάνοντας απειλές για τα ανθρώπινα δικαιώματα, την κοινωνική ακεραιότητα και τη βιωσιμότητα του περιβάλλοντος. Συγκεκριμένα, τα συστήματα τεχνητής νοημοσύνης έχουν εμπλακεί σε ζητήματα όπως η χειραγώγηση, οι παραβιάσεις ιδιωτικότητας, η απώλεια αυτονομίας, οι αλγοριθμικές προκαταλήψεις και ακόμη και οι απειλές προς τις δημοκρατικές διαδικασίες. Επιπλέον, ο περιβαλλοντικός αντίκτυπος της τεχνητής νοημοσύνης, ιδιαίτερα όσον αφορά την κατανάλωση ενέργειας και τις εκπομπές άνθρακα, δεν μπορεί να παραβλεφθεί.

Σε απάντηση σε αυτές τις προκλήσεις, έχουν αρχίσει να αναδύονται ρυθμιστικά πλαίσια, με στόχο την μείωση των κινδύνων που συνδέονται με την τεχνητή νοημοσύνη με κορυφαίο παράδειγμα τέτοιας ρύθμισης είναι ο Κανονισμός για την Τεχνητή Νοημοσύνη της Ευρωπαϊκής Ένωσης (AI Act), ο οποίος επιδιώκει να καθιερώσει ένα ολοκληρωμένο νομικό πλαίσιο για την ανάπτυξη και την εφαρμογή συστημάτων τεχνητής νοημοσύνης. Ως αποτέλεσμα, οι Αξιολογήσεις Επιπτώσεων Αλγορίθμων (ΑΕΑ) έχουν αποκτήσει σημασία ως ένα πρακτικό εργαλείο για τη διασφάλιση της συμμόρφωσης με αυτές τις ρυθμιστικές απαιτήσεις. Οι ΑΕΑ παρέχουν μια δομημένη προσέγγιση για την αξιολόγηση των πιθανών επιπτώσεων των συστημάτων τεχνητής νοημοσύνης, βοηθώντας τις οργανώσεις να εντοπίσουν και να αντιμετωπίσουν τους κινδύνους πριν από την ανάπτυξη.

Αυτή η εργασία έχει εξερευνήσει διάφορα πλαίσια ΑΕΑ που αναπτύχθηκαν από διάφορα ιδρύματα και οργανισμούς, καθένα με τη δική του μοναδική δομή και προσέγγιση. Συνδυάζοντας τα πλεονεκτήματα αυτών των ποικίλων μεθοδολογιών, η εργασία έχει προτείνει ένα ολοκληρωμένο πλαίσιο ΑΕΑ σχεδιασμένο να ευθυγραμμίζεται με τις αρχές του AI Act. Ο στόχος αυτού του πλαισίου είναι να παρέχει ένα πρακτικό, προσαρμόσιμο και ανθεκτικό εργαλείο που μπορεί να υιοθετηθεί ευρέως σε διάφορους κλάδους, συμβάλλοντας τελικά στην καθιέρωση ενός βιομηχανικού προτύπου ΑΕΑ.

Η προσέγγιση που ακολουθείται σε αυτή τη εργασία τονίζει τη σημασία μιας ολιστικής διαδικασίας αξιολόγησης, ενσωματώνοντας παράγοντες όπως η διαφάνεια, η λογοδοσία, η δικαιοσύνη και ο μετριασμός ατομικής κοινωνικής και περιβαλλοντικής επικινδυνότητας. Ενσωματώνοντας αυτά τα στοιχεία σε ένα ενιαίο, συνεκτικό πλαίσιο, η ΑΕΑ που αναπτύχθηκε εδώ στοχεύει στη γεφύρωση του χάσματος μεταξύ των κανονιστικών απαιτήσεων και της πρακτικής εφαρμογής, προάγοντας την υπεύθυνη ανάπτυξη και εφαρμογή των τεχνολογιών τεχνητής νοημοσύνης.

7.2 Μελλοντικές Βελτιώσεις

Ενώ το πλαίσιο ΑΕΑ που προτείνεται σε αυτή την εργασία αντιπροσωπεύει ένα σημαντικό βήμα προς τα εμπρός, υπάρχουν αρκετοί τομείς όπου θα μπορούσαν να γίνουν περαιτέρω βελτιώσεις για να αυξηθεί η αποτελεσματικότητα και η εφαρμοσιμότητά του. Αυτές οι βελτιώσεις

μπορούν να κατηγοριοποιηθούν ευρέως σε τεχνολογικές εξελίξεις, δυναμική προσαρμοστικότητα και συνεχιζόμενη έρευνα.

7.2.1 Τεχνολογικές Προόδους

Μία πιθανή βελτίωση περιλαμβάνει την ενσωμάτωση διαδικτυακών υπηρεσιών στο ερωτηματολόγιο ΑΕΑ. Αυτές οι υπηρεσίες θα μπορούσαν αυτόματα να αναλύσουν τον κώδικα του συστήματος τεχνητής νοημοσύνης που εξετάζεται, παρέχοντας μια αντικειμενική αξιολόγηση παραγόντων όπως η ποιότητα τεκμηρίωσης, η εξηγησιμότητα και η συμμόρφωση με τις βέλτιστες πρακτικές. Πολλά υπάρχοντα εργαλεία, όπως το What-If Tool της Google και το AI Fairness 360 της IBM, προσφέρουν ήδη λειτουργίες που θα μπορούσαν να αξιοποιηθούν για αυτόν τον σκοπό [59][60]. Ενσωματώνοντας τέτοια εργαλεία, η ΑΕΑ θα μπορούσε να παρέχει μια πιο ακριβή και ολοκληρωμένη αξιολόγηση της τεχνικής ανθεκτικότητας του συστήματος AI και της συμμόρφωσής του με τα ηθικά πρότυπα.

7.2.2 Δυναμική Προσαρμοστικότητα

Η δυναμική φύση του ερωτηματολογίου ΑΕΑ θα μπορούσε να διερευνηθεί περαιτέρω για να ενισχυθεί η ευελιξία και η σημασία του σε διάφορα συμφραζόμενα. Τρείς βασικοί τομείς για βελτίωση περιλαμβάνουν:

- **Δυναμική Ρύθμιση Βαρών:** Η βαρύτητα διαφορετικών κριτηρίων εντός του ερωτηματολογίου θα μπορούσε να προσαρμοστεί με βάση τον συγκεκριμένο τύπο συστήματος τεχνητής νοημοσύνης που αξιολογείται. Για παράδειγμα, στην περίπτωση της αεροπορικής τεχνολογίας, όπου η ασφάλεια και η αυτονομία είναι πρωταρχικής σημασίας, αυτοί οι παράγοντες θα μπορούσαν να αποδοθούν μεγαλύτερη βαρύτητα σε σύγκριση με άλλα κριτήρια όπως η εξηγησιμότητα. Αυτό θα διασφάλιζε ότι η αξιολόγηση είναι προσαρμοσμένη στους μοναδικούς κινδύνους και τις απαιτήσεις κάθε τομέα εφαρμογής.
- **Απελευθέρωση Ερωτήσεων με Προϋποθέσεις:** Το ερωτηματολόγιο θα μπορούσε να σχεδιαστεί ώστε να ξεκλειδώνει συγκεκριμένα τμήματα βάσει των απαντήσεων που παρέχονται. Για παράδειγμα, αν διαπιστωθεί ότι ένα σύστημα τεχνητής νοημοσύνης χρησιμοποιεί δεδομένα υγείας, ένα ειδικό σύνολο ερωτήσεων που σχετίζονται με την ιδιωτικότητα και την ασφάλεια των δεδομένων υγείας θα μπορούσε να ενεργοποιηθεί αυτόματα. Ομοίως, τα συστήματα με δυναμικές πηγές δεδομένων θα μπορούσαν να ενεργοποιήσουν ερωτήσεις που σχετίζονται με την απόκλιση δεδομένων και την επανεκπαίδευση του μοντέλου. Αυτή η προσέγγιση θα έκανε το ερωτηματολόγιο πιο αποτελεσματικό και στοχευμένο, μειώνοντας το βάρος στους ερωτώμενους ενώ διασφαλίζει ότι όλα τα σχετικά ζητήματα θα αντιμετωπιστούν.
- **Δυναμική Προσθήκη Ερωτήσεων και Βαρών από Ενδιαφερόμενους:** Το ερωτηματολόγιο θα μπορούσε να περιλαμβάνει ένα εξειδικευμένο τμήμα που ξεκλειδώνεται μόνο όταν υπάρχει εγγύηση ότι στην διαδικασία συμπλήρωσης έχουν ενταχθεί αντιπρόσωποι από όλες τις ενδιαφερόμενες ομάδες (stakeholders). Σε αυτό το τμήμα, οι συμμετέχοντες θα μπορούσαν να προσθέτουν νέες ερωτήσεις και βάρη ανάλογα με τις ανάγκες και τις προτεραιότητες της συγκεκριμένης εφαρμογής ΤΝ. Αυτή η δυναμική προσέγγιση θα εξασφάλιζε ότι το ερωτηματολόγιο παραμένει

ευέλικτο και προσαρμοστικό, ενώ ταυτόχρονα θα ενίσχυε τη συμμετοχή και την ευθύνη όλων των ενδιαφερομένων μερών στη διαδικασία αξιολόγησης.

7.2.3 Συνεχιζόμενη Έρευνα και Βελτίωση

Το πλαίσιο ΑΕΑ που αναπτύχθηκε σε αυτή την εργασία έχει σχεδιαστεί ώστε να είναι προσαρμόσιμο, επιτρέποντας την ενσωμάτωση νέων ερωτήσεων και μετρήσεων καθώς εξελίσσεται το πεδίο της τεχνητής νοημοσύνης. Ωστόσο, η συνεχής έρευνα είναι απαραίτητη για να διασφαλιστεί ότι το πλαίσιο παραμένει ενημερωμένο και αποτελεσματικό. Το μελλοντικό έργο θα μπορούσε να επικεντρωθεί στην ανάπτυξη πιο συγκεκριμένων και εφαρμοσμένων ερωτήσεων, καθώς και στη βελτιστοποίηση των μετρικών που χρησιμοποιούνται για την αξιολόγηση των συστημάτων τεχνητής νοημοσύνης. Για παράδειγμα, θα μπορούσαν να διεξαχθούν επιπλέον έρευνες για τον εντοπισμό βέλτιστων πρακτικών για την αξιολόγηση του περιβαλλοντικού αντίκτυπου των συστημάτων τεχνητής νοημοσύνης, ή για την ανάπτυξη πιο λεπτομερών μετρικών για την αξιολόγηση της δικαιοσύνης και της προκατάληψης στα μοντέλα μηχανικής μάθησης.

Επιπλέον, η συνεργασία με τους ενδιαφερόμενους φορείς της βιομηχανίας, τους πολιτικούς υπεύθυνους και τους ακαδημαϊκούς ερευνητές θα είναι κρίσιμη για τη συνεχιζόμενη βελτίωση του πλαισίου ΑΕΑ. Εμπλέκοντας μια ποικιλία από διαφορετικές προοπτικές, το πλαίσιο μπορεί να βελτιώνεται συνεχώς για να αντιμετωπίζει τις αναδυόμενες προκλήσεις και να ενσωματώνει τις τελευταίες εξελίξεις στην ηθική και ρύθμιση της τεχνητής νοημοσύνης.

7.3 Τελικές Παρατηρήσεις

Η ανάπτυξη ενός ολοκληρωμένου και πρακτικού πλαισίου ΑΕΑ αντιπροσωπεύει ένα κρίσιμο βήμα προς την εξασφάλιση της υπεύθυνης ανάπτυξης και εφαρμογής των τεχνολογιών AI. Αντιμετωπίζοντας τους πολυδιάστατους κινδύνους που συνδέονται με την τεχνητή νοημοσύνη και ευθυγραμμίζοντας με τις αναδυόμενες κανονιστικές απαιτήσεις, το πλαίσιο που προτείνεται σε αυτή τη εργασία έχει τη δυνατότητα να γίνει ένα πολύτιμο εργαλείο για οργανισμούς σε διάφορους κλάδους. Ωστόσο, η δουλειά δεν τελειώνει εδώ. Καθώς η τεχνητή νοημοσύνη συνεχίζει να εξελίσσεται, έτσι πρέπει να εξελίσσονται και τα εργαλεία και οι μεθοδολογίες που χρησιμοποιούνται για την αξιολόγηση του αντίκτυπού της. Μέσω συνεχούς έρευνας, τεχνολογικής καινοτομίας και συνεργασίας, το πλαίσιο ΑΕΑ μπορεί να βελτιωθεί και να εξελιχθεί περαιτέρω, συμβάλλοντας τελικά σε ένα μέλλον όπου οι τεχνολογίες AI αναπτύσσονται και εφαρμόζονται με τρόπο ηθικό, διαφανή και ωφέλιμο για την κοινωνία στο σύνολό της.

ΠΑΡΑΡΤΗΜΑ Α

Α.1 Κώδικας Python για προσθήκη μετρικών

```
import json

def modify_question(question):
      if question["type"] == "radiogroup" or question["type"] ==
"checkbox":
      new_choices = []
      for choice in question["choices"]:
            if isinstance(choice, dict):
                  value = choice["value"]
            else:
                  value = choice
            new_choice = {
                  "text": choice["text"],
                  "value": value,
                  "beneficence": 0 ,
                  "non_maleficence": 0 ,
                  "autonomy": 0 ,
                  "justice": 0 ,
                  "explicability": 0
            }
            new_choices.append(new_choice)
      question["choices"] = new_choices
      elif question["type"] == "boolean":
      question.update({
            "beneficenceTrue": 0,
            "non_maleficenceTrue": 0,
            "autonomyTrue": 0,
            "justiceTrue": 0,
            "explicabilityTrue": 0,
            "beneficenceFalse": 0,
            "non_maleficenceFalse": 0,
            "autonomyFalse": 0,
            "justiceFalse": 0,
            "explicabilityFalse": 0
      })
      elif question["type"] == "matrixdropdown":
      question["choices"] = [
            {"value": 1, "weight": 1},
            {"value": 2, "weight": 2},
            {"value": 3, "weight": 3},
```

```
            {"value": 4, "weight": 4},
            {"value": 5, "weight": 5}
      ]
      return question

def modify_survey(json_data):
      for page in json_data["pages"]:
      for element in page["elements"]:
            if "elements" in element:  # For panels
                  for sub_element in element["elements"]:
                  modify_question(sub_element)
            else:
                  modify_question(element)
      return json_data

# Load the JSON file
with open('json_demo.json', 'r') as file:
      survey_data = json.load(file)

# Modify the survey data
modified_survey = modify_survey(survey_data)

# Save the modified JSON to a new file
with open('modified_survey.json', 'w') as file:
      json.dump(modified_survey, file, indent=4)

print("Survey has been modified and saved to 'modified_survey.json'.")
```

A.2 Κώδικας Python για υπολογισμό μεγίστου/ελαχίστου μετρικών

```
import json

# Load the questionnaire JSON
with open('modified_survey.json', 'r') as file:
      questionnaire = json.load(file)

# Initialize metrics
metrics = ['beneficence', 'non_maleficence', 'explicability',
'justice', 'autonomy']
min_metrics = {metric: 0 for metric in metrics}
max_metrics = {metric: 0 for metric in metrics}

def calculate_metrics(page):
      for element in page['elements']:
      if element['type'] == 'panel':
            # Recursively process panels
```

```
            calculate_metrics(element)
      elif element['type'] == 'radiogroup':
            # For radiogroup, only one answer can be selected
            for metric in metrics:
                  # Find the max and min values for the current metric
in this question
                  max_value = max(choice.get(metric, 0) for choice in
element['choices'])
                  min_value = min(choice.get(metric, 0) for choice in
element['choices'])
                  # Add to the total max and min metrics
                  max_metrics[metric] += max_value
                  min_metrics[metric] += min_value
      elif element['type'] == 'checkbox':
            # For checkbox, multiple answers can be selected
            for metric in metrics:
                  # Sum all positive values for max and all negative
values for min
                  max_value = sum(choice.get(metric, 0) for choice in
element['choices'] if choice.get(metric, 0) > 0)
                  min_value = sum(choice.get(metric, 0) for choice in
element['choices'] if choice.get(metric, 0) < 0)
                  # Add to the total max and min metrics
                  max_metrics[metric] += max_value
                  min_metrics[metric] += min_value
      elif element['type'] == 'boolean':
            # For boolean, only one answer (True or False) can be
selected
            for metric in metrics:
                  # Compare True and False values for the current
metric
                  true_value = element.get(f'{metric}True', 0)
                  false_value = element.get(f'{metric}False', 0)
                  # Add the max and min values to the totals
                  max_metrics[metric] += max(true_value, false_value)
                  min_metrics[metric] += min(true_value, false_value)

      elif element['type'] == 'matrixdropdown':
            num_rows = len(element['rows'])
            num_columns = len(element['columns'])
            num_choices = len(element['choices'])
            additional_value = (num_rows * num_columns * num_choices) /
2
            max_metrics['non_maleficence'] += additional_value
            min_metrics['non_maleficence'] += 0
```

```
# Iterate through each page and calculate metrics
for page in questionnaire['pages']:
      calculate_metrics(page)

# Print results
print("Maximum Metrics:", max_metrics)
print("Minimum Metrics:", min_metrics)
```

A.3 Κώδικας εφαρμογής

```
//code for example questionnare with 2 weights and all question types
/*const calculateScore = function(surveyData, surveyJson) {
      let score1 = 0;
      let score2 = 0;
      const questionDetails = [];  // To store question details for the
PDF report

      // Iterate through the survey data responses
      Object.keys(surveyData).forEach(question => {
      const response = surveyData[question];
      const questionElement = surveyJson.pages.flatMap(page =>
page.elements).find(element => element.name === question);

      if (!questionElement) {
            console.warn(`Question ${question} not found in survey
JSON`);
            return;
      }

      let questionScore1 = 0;
      let questionScore2 = 0;
      let questionAnswer = '';

      // Handle different question types using a switch-case
      switch (questionElement.type) {
            case "radiogroup":
            case "checkbox":
                  if (Array.isArray(response)) {
                  questionAnswer = response.join(', ');
                  response.forEach(answer => {
                        const choice =
questionElement.choices.find(choice => choice.value === answer);
                        if (choice) {
                              questionScore1 += choice.weight1;
                              questionScore2 += choice.weight2;
                        }
```

```
                });
                } else {
                questionAnswer = response;
                const choice = questionElement.choices.find(choice =>
choice.value === response);
                if (choice) {
                      questionScore1 += choice.weight1;
                      questionScore2 += choice.weight2;
                }
                }
                questionDetails.push({
                name: question,
                title: questionElement.title,
                type: questionElement.type,
                answer: questionAnswer,
                score1: questionScore1,
                score2: questionScore2
                });

                score1 += questionScore1;
                score2 += questionScore2;

                break;

          case "rating":
                const ratingAnswer = response;
                questionScore1 = ratingAnswer;
                questionScore2 = ratingAnswer;  // If needed, adjust
scoring
                questionAnswer = ratingAnswer;

                questionDetails.push({
                name: question,
                title: questionElement.title,
                type: questionElement.type,
                answer: questionAnswer,
                score1: questionScore1,
                score2: questionScore2
                });

                score1 += questionScore1;
                score2 += questionScore2;

                break;

          case "dropdown":
```

```
                  const dropdownChoice =
questionElement.choices.find(choice => choice.value === response);
                  if (dropdownChoice) {
                  questionScore1 = dropdownChoice.weight1;
                  questionScore2 = dropdownChoice.weight2;
                  questionAnswer = response;
                  }
                  questionDetails.push({
                  name: question,
                  title: questionElement.title,
                  type: questionElement.type,
                  answer: questionAnswer,
                  score1: questionScore1,
                  score2: questionScore2
                  });

                  score1 += questionScore1;
                  score2 += questionScore2;

                  break;

            case "tagbox":
                  if (Array.isArray(response)) {
                  questionAnswer = response.join(', ');
                  response.forEach(answer => {
                        const tagboxChoice =
questionElement.choices.find(choice => choice.value === answer);
                        if (tagboxChoice) {
                              questionScore1 += tagboxChoice.weight1;
                              questionScore2 += tagboxChoice.weight2;
                        }
                  });
                  } else {
                  const tagboxChoice =
questionElement.choices.find(choice => choice.value === response);
                  if (tagboxChoice) {
                        questionScore1 = tagboxChoice.weight1;
                        questionScore2 = tagboxChoice.weight2;
                        questionAnswer = response;
                  }
                  }
                  questionDetails.push({
                  name: question,
                  title: questionElement.title,
                  type: questionElement.type,
                  answer: questionAnswer,
```

```
                score1: questionScore1,
                score2: questionScore2
                });

                score1 += questionScore1;
                score2 += questionScore2;
                break;

            case "boolean":
                if (response === true) {
                questionScore1 += questionElement.weight1true;
                questionScore2 += questionElement.weight2true; // You
can adjust these values as per your needs
                questionAnswer = "Yes";
                } else if (response === false) {
                questionScore1 += questionElement.weight1false
                questionScore2 += questionElement.weight2false
                questionAnswer = "No";
                }
                questionDetails.push({
                name: question,
                title: questionElement.title,
                type: questionElement.type,
                answer: questionAnswer,
                score1: questionScore1,
                score2: questionScore2
                });

                score1 += questionScore1;
                score2 += questionScore2;
                break;

            case "imagepicker":
                const imageChoice =
questionElement.choices.find(choice => choice.value === response);
                if (imageChoice) {
                questionScore1 = imageChoice.weight1;
                questionScore2 = imageChoice.weight2;
                questionAnswer = response;
                }
                questionDetails.push({
                name: question,
                title: questionElement.title,
                type: questionElement.type,
                answer: questionAnswer,
                score1: questionScore1,
```

```
                  score2: questionScore2
                  });

                  score1 += questionScore1;
                  score2 += questionScore2;
                  break;

            case "ranking":
                  if (Array.isArray(response)) {
                  questionAnswer = response.join(', ');
                  response.forEach((item, index) => {
                        const rankingChoice =
questionElement.choices.find(choice => choice.value === item);
                        if (rankingChoice) {
                              questionScore1 += rankingChoice.weight1 *
(index + 1);
                              questionScore2 += rankingChoice.weight2 *
(index + 1);
                        }
                  });
                  }
                  questionDetails.push({
                  name: question,
                  title: questionElement.title,
                  type: questionElement.type,
                  answer: questionAnswer,
                  score1: questionScore1,
                  score2: questionScore2
                  });

                  score1 += questionScore1;
                  score2 += questionScore2;
                  break;

            case "text":
            case "comment":
                  questionAnswer = response;
                  questionDetails.push({
                  name: question,
                  title: questionElement.title,
                  type: questionElement.type,
                  answer: questionAnswer,
                  score1: questionScore1,
                  score2: questionScore2
                  });
```

```
                score1 += questionScore1;
                score2 += questionScore2;
                break;

            case "multipletext":
                let multipleTextScore1 = 0;
                let multipleTextScore2 = 0;
                let multipleTextAnswer = '';

                questionElement.items.forEach(item => {
                const itemAnswer = response[item.name];
                multipleTextAnswer += `${item.title}: ${itemAnswer ||
'No answer'}\n`;

                if (itemAnswer) {
                      multipleTextScore1 += 1;
                      multipleTextScore2 += 1; // Adjust as per your
rules
                }
                });
                questionScore1 = multipleTextScore1;
                questionScore2 = multipleTextScore2;
                questionDetails.push({
                name: question,
                title: questionElement.title,
                type: questionElement.type,
                answer: multipleTextAnswer,
                score1: questionScore1,
                score2: questionScore2
                });

                score1 += questionScore1;
                score2 += questionScore2;
                break;

            case "matrix":
                let matrixScore1 = 0;
                let matrixScore2 = 0;
                let matrixAnswer = '';
                Object.keys(response).forEach(row => {
                const column = response[row];
                matrixAnswer += `${row}: ${column || 'No answer'}\n`;
                const columnIndex =
questionElement.columns.indexOf(column);
                if (columnIndex !== -1) {
                      matrixScore1 += columnIndex + 1;  // Customize
```

```
scoring
                        matrixScore2 += columnIndex + 1;
                  }
                  });
                  questionScore1 = matrixScore1;
                  questionScore2 = matrixScore2;
                  questionDetails.push({
                  name: question,
                  title: questionElement.title,
                  type: questionElement.type,
                  answer: matrixAnswer,
                  score1: questionScore1,
                  score2: questionScore2
                  });

                  score1 += questionScore1;
                  score2 += questionScore2;
                  break;

            case "matrixdropdown":
                  let matrixDropdownScore1 = 0;
                  let matrixDropdownScore2 = 0;
                  let matrixDropdownAnswer = ''
                  Object.keys(response).forEach(row => {
                  const columnResponses = response[row];
                  matrixDropdownAnswer += `${row}:\n`;
                  Object.keys(columnResponses).forEach(column => {
                        const columnValue = columnResponses[column];
                        matrixDropdownAnswer += `  ${column}:
${columnValue || 'No answer'}\n`;
                        const choice =
questionElement.choices.find(choice => choice.value === columnValue);
                        if (choice) {
                              matrixDropdownScore1 += choice.weight1;
                              matrixDropdownScore2 += choice.weight2;
                        }
                  });
                  });
                  questionScore1 = matrixDropdownScore1;
                  questionScore2 = matrixDropdownScore2;
                  questionDetails.push({
                  name: question,
                  title: questionElement.title,
                  type: questionElement.type,
                  answer: matrixDropdownAnswer,
                  score1: questionScore1,
```

```
                  score2: questionScore2
                  });

                  score1 += questionScore1;
                  score2 += questionScore2;
                  break;

            default:
                  console.warn(`Unknown question type:
${questionElement.type}`);
                  break;
      }
      });

      return { score1, score2, totalScore: score1 + score2,
questionDetails };
}; */
/* const savePdf = function(surveyData) {
      const { jsPDF } = window.jspdf;
      const surveyPdf = new jsPDF();
      surveyPdf.text("Survey Results", 20, 20);

      let yPosition = 30;
      const margin = 20;
      const lineHeight = 10;
      const maxYPosition = 250;

      surveyData.questionDetails.forEach((question) => {
      if (yPosition + lineHeight > maxYPosition) {
            surveyPdf.addPage();
            yPosition = 20;
      }

      surveyPdf.text(`${question.name}: ${question.title}`, margin,
yPosition);
      yPosition += lineHeight;
      surveyPdf.text(`Type: ${question.type}`, margin, yPosition);
      yPosition += lineHeight;
      surveyPdf.text(`Score1: ${question.score1}`, margin, yPosition);
      yPosition += lineHeight;
      surveyPdf.text(`Score2: ${question.score2}`, margin, yPosition);
      yPosition += lineHeight;

      if (Array.isArray(question.answer)) {
            question.answer.forEach((answer) => {
                  if (yPosition + lineHeight > maxYPosition) {
```

```
                surveyPdf.addPage();
                yPosition = 20;
                }
                surveyPdf.text(`Answer: ${answer}`, margin,
yPosition);
                yPosition += lineHeight;
          });
     } else {
          if (yPosition + lineHeight > maxYPosition) {
                surveyPdf.addPage();
                yPosition = 20;
          }
          surveyPdf.text(`Answer: ${question.answer}`, margin,
yPosition);
          yPosition += lineHeight;
     }

     yPosition += 2 * lineHeight;
     });

     yPosition += 4 * lineHeight;
     if (surveyData.score1 !== undefined) {
     if (yPosition + lineHeight > maxYPosition) {
          surveyPdf.addPage();
          yPosition = 20;
     }
     surveyPdf.text(`Total Score 1: ${surveyData.score1}`, margin,
yPosition);
     } else {
     surveyPdf.text("Total Score 1: 0", margin, yPosition);
     }


     yPosition += 2 * lineHeight;
     if (surveyData.score2 !== undefined) {
     if (yPosition + lineHeight > maxYPosition) {
          surveyPdf.addPage();
          yPosition = 20;
     }
     surveyPdf.text(`Total Score 2: ${surveyData.score2}`, margin,
yPosition);
     } else {
     surveyPdf.text("Total Score 2: 0", margin, yPosition);
     }

     yPosition += 2 * lineHeight;
```

```
      if (surveyData.totalScore !== undefined) {
      if (yPosition + lineHeight > maxYPosition) {
            surveyPdf.addPage();
            yPosition = 20;
      }
      surveyPdf.text(`Total Score: ${surveyData.totalScore}`, margin,
yPosition);
      } else {
      surveyPdf.text("Total Score: 0", margin, yPosition);
      }

      surveyPdf.save("survey_results.pdf");
}; */

//maxmin for demo
// const maxBeneficence = 22
// const minBeneficence = -26
// const maxNonMaleficence = 134.5
// const minNonMaleficence = -31
// const maxAutonomy = 10
// const minAutonomy = -10
// const maxJustice = 7
// const minJustice = -13
// const maxExplicability = 12
// const minExplicability = -12


//maxmin for actual questionnaire
const maxBeneficence = 101
const minBeneficence = -106
const maxNonMaleficence = 459
const minNonMaleficence = -199
const maxAutonomy = 94
const minAutonomy = -61
const maxJustice = 110
const minJustice = -81
const maxExplicability = 44
const minExplicability = -26


const calculateScore = function(surveyData, surveyJson) {
      let beneficence = 0;
      let non_maleficence = 0;
      let autonomy = 0;
      let justice = 0;
      let explicability = 0;
```

```
      const questionDetails = [];  // To store question details for the
PDF report




      // Helper function to search for a question recursively inside
panels
      const findQuestion = (elements, questionName) => {
      if (!elements) return null;
      for (const element of elements) {
            if (element.name === questionName) {
                  return element;
            }
            if (element.type === "panel" && element.elements) {
                  const found = findQuestion(element.elements,
questionName);
                  if (found) return found;
            }
      }
      return null;
      };


      // Iterate through the survey data responses
      Object.keys(surveyData).forEach(question => {
      const response = surveyData[question];
      // const questionElement = surveyJson.pages.flatMap(page =>
page.elements).find(element => element.name === question);
      const questionElement = surveyJson.pages.flatMap(page =>
findQuestion(page.elements, question)).find(el => el !== null);


      if (!questionElement) {
            console.warn(`Question ${question} not found in survey
JSON`);
            return;
      }

      let questionBeneficence = 0;
      let questionNonMaleficence = 0;
      let questionAutonomy = 0;
      let questionJustice = 0;
      let questionExplicability = 0;
      let questionAnswer = '';
```

```
      // Handle different question types using a switch-case
      switch (questionElement.type) {
            case "radiogroup":
            case "checkbox":
                  if (questionElement.choices &&
Array.isArray(response)) {
                  questionAnswer = response.map(answer => {
                        const choice =
questionElement.choices.find(choice => choice.value === answer);
                        return choice ? choice.text : answer;  // Store
text instead of value
                  }).join(', ');

                  response.forEach(answer => {
                        const choice =
questionElement.choices.find(choice => choice.value === answer);
                        if (choice) {
                              questionBeneficence += choice.beneficence;
                              questionNonMaleficence +=
choice.non_maleficence;
                              questionAutonomy += choice.autonomy;
                              questionJustice += choice.justice;
                              questionExplicability +=
choice.explicability;
                        }
                  });
                  } else {
                  const choice = questionElement.choices.find(choice =>
choice.value === response);
                  questionAnswer = choice ? choice.text : response;  //
Store text instead of value

                  if (choice) {
                        questionBeneficence += choice.beneficence;
                        questionNonMaleficence +=
choice.non_maleficence;
                        questionAutonomy += choice.autonomy;
                        questionJustice += choice.justice;
                        questionExplicability += choice.explicability;
                  }
                  }
                  questionDetails.push({
                  name: question,
                  title: questionElement.title,
                  type: questionElement.type,
                  answer: questionAnswer,
```

```
                beneficence: questionBeneficence,
                non_maleficence: questionNonMaleficence,
                autonomy: questionAutonomy,
                justice: questionJustice,
                explicability: questionExplicability
                });

                beneficence += questionBeneficence;
                non_maleficence += questionNonMaleficence;
                autonomy += questionAutonomy;
                justice += questionJustice;
                explicability += questionExplicability;

                break;

        case "rating":
                const ratingAnswer = response;
                questionBeneficence = ratingAnswer;
                questionNonMaleficence = ratingAnswer;
                questionAutonomy = ratingAnswer;
                questionJustice = ratingAnswer;
                questionExplicability = ratingAnswer;
                questionAnswer = ratingAnswer;

                questionDetails.push({
                name: question,
                title: questionElement.title,
                type: questionElement.type,
                answer: questionAnswer,
                beneficence: questionBeneficence,
                non_maleficence: questionNonMaleficence,
                autonomy: questionAutonomy,
                justice: questionJustice,
                explicability: questionExplicability
                });

                beneficence += questionBeneficence;
                non_maleficence += questionNonMaleficence;
                autonomy += questionAutonomy;
                justice += questionJustice;
                explicability += questionExplicability;

                break;

        case "dropdown":
                const dropdownChoice =
```

```
questionElement.choices.find(choice => choice.value === response);
                if (dropdownChoice) {
                questionBeneficence = dropdownChoice.beneficence;
                questionNonMaleficence =
dropdownChoice.non_maleficence;
                questionAutonomy = dropdownChoice.autonomy;
                questionJustice = dropdownChoice.justice;
                questionExplicability = dropdownChoice.explicability;
                questionAnswer = response;
                }
                questionDetails.push({
                name: question,
                title: questionElement.title,
                type: questionElement.type,
                answer: questionAnswer,
                beneficence: questionBeneficence,
                non_maleficence: questionNonMaleficence,
                autonomy: questionAutonomy,
                justice: questionJustice,
                explicability: questionExplicability
                });

                beneficence += questionBeneficence;
                non_maleficence += questionNonMaleficence;
                autonomy += questionAutonomy;
                justice += questionJustice;
                explicability += questionExplicability;

                break;

            case "tagbox":
                if (Array.isArray(response)) {
                questionAnswer = response.join(', ');
                response.forEach(answer => {
                      const tagboxChoice =
questionElement.choices.find(choice => choice.value === answer);
                      if (tagboxChoice) {
                            questionBeneficence +=
tagboxChoice.beneficence;
                            questionNonMaleficence +=
tagboxChoice.non_maleficence;
                            questionAutonomy += tagboxChoice.autonomy;
                            questionJustice += tagboxChoice.justice;
                            questionExplicability +=
tagboxChoice.explicability;
                      }
```

```
                });
                } else {
                const tagboxChoice =
questionElement.choices.find(choice => choice.value === response);
                if (tagboxChoice) {
                      questionBeneficence = tagboxChoice.beneficence;
                      questionNonMaleficence =
tagboxChoice.non_maleficence;
                      questionAutonomy = tagboxChoice.autonomy;
                      questionJustice = tagboxChoice.justice;
                      questionExplicability =
tagboxChoice.explicability;
                      questionAnswer = response;
                }
                }
                questionDetails.push({
                name: question,
                title: questionElement.title,
                type: questionElement.type,
                answer: questionAnswer,
                beneficence: questionBeneficence,
                non_maleficence: questionNonMaleficence,
                autonomy: questionAutonomy,
                justice: questionJustice,
                explicability: questionExplicability
                });

                beneficence += questionBeneficence;
                non_maleficence += questionNonMaleficence;
                autonomy += questionAutonomy;
                justice += questionJustice;
                explicability += questionExplicability;
                break;

          case "boolean":
                if (response === true) {
                questionBeneficence +=
questionElement.beneficenceTrue;
                questionNonMaleficence +=
questionElement.non_maleficenceTrue;
                questionAutonomy += questionElement.autonomyTrue;
                questionJustice += questionElement.justiceTrue;
                questionExplicability +=
questionElement.explicabilityTrue;
                questionAnswer = "True/Yes";
                } else if (response === false) {
```

```
                questionBeneficence +=
questionElement.beneficenceFalse
                questionNonMaleficence +=
questionElement.non_maleficenceFalse
                questionAutonomy += questionElement.autonomyFalse
                questionJustice += questionElement.justiceFalse
                questionExplicability +=
questionElement.explicabilityFalse
                questionAnswer = "False/No";
                }
                questionDetails.push({
                name: question,
                title: questionElement.title,
                type: questionElement.type,
                answer: questionAnswer,
                beneficence: questionBeneficence,
                non_maleficence: questionNonMaleficence,
                autonomy: questionAutonomy,
                justice: questionJustice,
                explicability: questionExplicability
                });

                beneficence += questionBeneficence;
                non_maleficence += questionNonMaleficence;
                autonomy += questionAutonomy;
                justice += questionJustice;
                explicability += questionExplicability;
                break;

            case "imagepicker":
                const imageChoice =
questionElement.choices.find(choice => choice.value === response);
                if (imageChoice) {
                questionBeneficence = imageChoice.beneficence;
                questionNonMaleficence = imageChoice.non_maleficence;
                questionAutonomy = imageChoice.autonomy;
                questionJustice = imageChoice.justice;
                questionExplicability = imageChoice.explicability;
                questionAnswer = response;
                }
                questionDetails.push({
                name: question,
                title: questionElement.title,
                type: questionElement.type,
                answer: questionAnswer,
                beneficence: questionBeneficence,
```

```
                  non_maleficence: questionNonMaleficence,
                  autonomy: questionAutonomy,
                  justice: questionJustice,
                  explicability: questionExplicability
                  });

                  beneficence += questionBeneficence;
                  non_maleficence += questionNonMaleficence;
                  autonomy += questionAutonomy;
                  justice += questionJustice;
                  explicability += questionExplicability;
                  break;

            case "ranking":
                  if (Array.isArray(response)) {
                  questionAnswer = response.join(', ');
                  response.forEach((item, index) => {
                        const rankingChoice =
questionElement.choices.find(choice => choice.value === item);
                        if (rankingChoice) {
                              questionBeneficence +=
rankingChoice.beneficence * (index + 1);
                              questionNonMaleficence +=
rankingChoice.non_maleficence * (index + 1);
                              questionAutonomy += rankingChoice.autonomy
* (index + 1);
                              questionJustice += rankingChoice.justice *
(index + 1);
                              questionExplicability +=
rankingChoice.explicability * (index + 1);
                        }
                  });
                  }
                  questionDetails.push({
                  name: question,
                  title: questionElement.title,
                  type: questionElement.type,
                  answer: questionAnswer,
                  beneficence: questionBeneficence,
                  non_maleficence: questionNonMaleficence,
                  autonomy: questionAutonomy,
                  justice: questionJustice,
                  explicability: questionExplicability
                  });

                  beneficence += questionBeneficence;
```

```
                non_maleficence += questionNonMaleficence;
                autonomy += questionAutonomy;
                justice += questionJustice;
                explicability += questionExplicability;
                break;

            case "text":
            case "comment":
                questionAnswer = response;
                questionDetails.push({
                name: question,
                title: questionElement.title,
                type: questionElement.type,
                answer: questionAnswer,
                beneficence: questionBeneficence,
                non_maleficence: questionNonMaleficence,
                autonomy: questionAutonomy,
                justice: questionJustice,
                explicability: questionExplicability
                });

                beneficence += questionBeneficence;
                non_maleficence += questionNonMaleficence;
                autonomy += questionAutonomy;
                justice += questionJustice;
                explicability += questionExplicability;
                break;

            case "multipletext":
                let multipleTextBeneficence = 0;
                let multipleTextNonMaleficence = 0;
                let multipleTextAutonomy = 0;
                let multipleTextJustice = 0;
                let multipleTextExplicability = 0;
                let multipleTextAnswer = '';

                questionElement.items.forEach(item => {
                const itemAnswer = response[item.name];
                multipleTextAnswer += `${item.title}: ${itemAnswer ||
'No answer'}\n`;

                if (itemAnswer) {
                    multipleTextBeneficence += 0;
                    multipleTextNonMaleficence += 0;
                    multipleTextAutonomy += 0;
                    multipleTextJustice += 0;
```

```
                        multipleTextExplicability += 0;
                    }
                    });
                    questionBeneficence = multipleTextBeneficence;
                    questionNonMaleficence = multipleTextNonMaleficence;
                    questionAutonomy = multipleTextAutonomy;
                    questionJustice = multipleTextJustice;
                    questionExplicability = multipleTextExplicability;
                    questionDetails.push({
                    name: question,
                    title: questionElement.title,
                    type: questionElement.type,
                    answer: multipleTextAnswer,
                    beneficence: questionBeneficence,
                    non_maleficence: questionNonMaleficence,
                    autonomy: questionAutonomy,
                    justice: questionJustice,
                    explicability: questionExplicability
                    });

                    beneficence += questionBeneficence;
                    non_maleficence += questionNonMaleficence;
                    autonomy += questionAutonomy;
                    justice += questionJustice;
                    explicability += questionExplicability;
                    break;

                case "matrix":
                    let matrixBeneficence = 0;
                    let matrixNonMaleficence = 0;
                    let matrixAutonomy = 0;
                    let matrixJustice = 0;
                    let matrixExplicability = 0;
                    let matrixAnswer = '';
                    Object.keys(response).forEach(rowValue => {
                    const columnValue = response[rowValue];

                    // Find row text
                    const rowObj = questionElement.rows.find(r => r.value
=== rowValue);
                    const rowText = rowObj ? rowObj.text : rowValue;

                    // Find column text
                    const columnObj = questionElement.columns.find(col =>
col.name === columnValue);
                    const columnText = columnObj ? columnObj.title :
```

```
columnValue;

                  matrixAnswer += `${rowText}: ${columnText || 'No
answer'}\n`;

                  // Calculate scores
                  const columnIndex =
questionElement.columns.findIndex(col => col.name === columnValue);
                  if (columnIndex !== -1) {
                        matrixBeneficence += columnIndex + 1;
                        matrixNonMaleficence += columnIndex + 1;
                        matrixAutonomy += columnIndex + 1;
                        matrixJustice += columnIndex + 1;
                        matrixExplicability += columnIndex + 1;
                  }
                  });
                  questionBeneficence = matrixBeneficence;
                  questionNonMaleficence = matrixNonMaleficence;
                  questionAutonomy = matrixAutonomy;
                  questionJustice = matrixJustice;
                  questionExplicability = matrixExplicability;
                  questionDetails.push({
                  name: question,
                  title: questionElement.title,
                  type: questionElement.type,
                  answer: matrixAnswer,
                  beneficence: questionBeneficence,
                  non_maleficence: questionNonMaleficence,
                  autonomy: questionAutonomy,
                  justice: questionJustice,
                  explicability: questionExplicability
                  });

                  beneficence += questionBeneficence;
                  non_maleficence += questionNonMaleficence;
                  autonomy += questionAutonomy;
                  justice += questionJustice;
                  explicability += questionExplicability;
                  break;

            case "matrixdropdown":
                  let matrixDropdownBeneficence = 0;
                  let matrixDropdownNonMaleficence = 0;
                  let matrixDropdownAutonomy = 0;
                  let matrixDropdownJustice = 0;
                  let matrixDropdownExplicability = 0;
```

```
                    let matrixDropdownAnswer = ''
                    let matrixWeight = 0;
                    Object.keys(response).forEach(row => {
                    const columnResponses = response[row];
                    // Find row text
                    const rowObj = questionElement.rows.find(r => r.value
=== row);
                    const rowText = rowObj ? rowObj.text : row;

                    matrixDropdownAnswer += `${rowText}:\n`;

                    let rowWeight = 1;
                    Object.keys(columnResponses).forEach(columnName => {

                          const columnValue =
columnResponses[columnName];

                          // Find column text
                          const columnObj =
questionElement.columns.find(col => col.name === columnName);
                          const columnText = columnObj ? columnObj.title
: columnName;

                          matrixDropdownAnswer += `  ${columnText}:
${columnValue || 'No answer'}\n`;

                          const choice =
questionElement.choices.find(choice => choice.value === columnValue);


                          if (choice) {
                                rowWeight *= choice.weight;
                          }
                    });
                    matrixWeight += rowWeight;
                    });
                    questionBeneficence = 0;
                    questionNonMaleficence = matrixWeight * 0.5;
                    questionAutonomy = 0;
                    questionJustice = 0;
                    questionExplicability = 0;
                    questionDetails.push({
                    name: question,
                    title: questionElement.title,
                    type: questionElement.type,
                    answer: matrixDropdownAnswer,
```

```
                    beneficence: questionBeneficence,
                    non_maleficence: questionNonMaleficence,
                    autonomy: questionAutonomy,
                    justice: questionJustice,
                    explicability: questionExplicability
                    });

                    beneficence += questionBeneficence;
                    non_maleficence += questionNonMaleficence;
                    autonomy += questionAutonomy;
                    justice += questionJustice;
                    explicability += questionExplicability;
                    break;

            default:
                    console.warn(`Unknown question type:
${questionElement.type}`);
                    break;
      }

      // Log question data to the console
      console.log(`Question: ${questionElement.title}`);
      console.log(`Answer Type: ${questionElement.type}`);
      console.log(`Answer: ${questionAnswer}`);
      console.log(`Beneficence: ${questionBeneficence}`);
      console.log(`Non-maleficence: ${questionNonMaleficence}`);
      console.log(`Autonomy: ${questionAutonomy}`);
      console.log(`Justice: ${questionJustice}`);
      console.log(`Explicability: ${questionExplicability}`);
      console.log("------------------------------");
      });



      return { beneficence, non_maleficence, autonomy, justice,
explicability, totalScore: beneficence + non_maleficence + autonomy +
justice + explicability, questionDetails };
};


const savePdf = function(surveyData) {
      const { jsPDF } = window.jspdf;
      const surveyPdf = new jsPDF();
      const titleFontSize = 16;
      const subtitleFontSize = 14;
      const textFontSize = 12;
```

```
      const margin = 20;
      const lineHeight = 10;
      const maxYPosition = 250;
      const pageWidth = 180;

      surveyPdf.setFontSize(titleFontSize);
      surveyPdf.text("Survey Results", margin, 20);

      let yPosition = 30;

      // Function to handle text wrapping and page breaks
      const addWrappedText = (text, fontSize = textFontSize, color =
'black') => {
      surveyPdf.setFontSize(fontSize);
      surveyPdf.setTextColor(color);
      const lines = surveyPdf.splitTextToSize(text, pageWidth);
      lines.forEach(line => {
            if (yPosition + lineHeight > maxYPosition) {
                  surveyPdf.addPage();
                  yPosition = 20;
            }
            surveyPdf.text(line, margin, yPosition);
            yPosition += lineHeight;
      });
      };

      // Function to calculate percentage of total metric
      const calculatePercentage = (score, min, max) => {
      return ((score - min) / (max - min)) * 100;
      }

      surveyData.questionDetails.forEach((question) => {
      addWrappedText(`${question.name}: ${question.title}`,
subtitleFontSize, 'blue');
      addWrappedText(`Type: ${question.type}`, textFontSize, 'gray');
      addWrappedText(`Beneficence: ${question.beneficence}`,
textFontSize, 'green');
      addWrappedText(`Non-maleficence: ${question.non_maleficence}`,
textFontSize, 'red');
      addWrappedText(`Autonomy: ${question.autonomy}`, textFontSize,
'purple');
      addWrappedText(`Justice: ${question.justice}`, textFontSize,
'orange');
      addWrappedText(`Explicability: ${question.explicability}`,
textFontSize, 'brown');
      if (Array.isArray(question.answer)) {
```

```
            question.answer.forEach((answer) => {
                  addWrappedText(`Answer: ${answer}`, textFontSize,
'black');
            });
      } else {
            addWrappedText(`Answer: ${question.answer}`, textFontSize,
'black');
      }

      yPosition += lineHeight;
      });

      yPosition += 2 * lineHeight;

      // Total scores for the five new categories
      if (surveyData.beneficence !== undefined) {
      if (yPosition + lineHeight > maxYPosition) {
            surveyPdf.addPage();
            yPosition = 20;
      }
      addWrappedText(`Total Beneficence: ${surveyData.beneficence} or
${calculatePercentage(surveyData.beneficence, minBeneficence,
maxBeneficence)}%`, subtitleFontSize, 'green');
      } else {
      addWrappedText("Total Beneficence: 0", subtitleFontSize,
'green');
      }

      if (surveyData.non_maleficence !== undefined) {
      addWrappedText(`Total Non-maleficence:
${surveyData.non_maleficence} or
${calculatePercentage(surveyData.non_maleficence, minNonMaleficence,
maxNonMaleficence)}%`, subtitleFontSize, 'red');
      } else {
      addWrappedText("Total Non-maleficence: 0", subtitleFontSize,
'red');
      }

      if (surveyData.autonomy !== undefined) {
      addWrappedText(`Total Autonomy: ${surveyData.autonomy} or
${calculatePercentage(surveyData.autonomy, minAutonomy,
maxAutonomy)}%`, subtitleFontSize, 'purple');
      } else {
      addWrappedText("Total Autonomy: 0", subtitleFontSize, 'purple');
      }
```

```
      if (surveyData.justice !== undefined) {
      addWrappedText(`Total Justice: ${surveyData.justice} or
${calculatePercentage(surveyData.justice, minJustice, maxJustice)}%`,
subtitleFontSize, 'orange');
      } else {
      addWrappedText("Total Justice: 0", subtitleFontSize, 'orange');
      }

      if (surveyData.explicability !== undefined) {
      addWrappedText(`Total Explicability: ${surveyData.explicability}
or ${calculatePercentage(surveyData.explicability, minExplicability,
maxExplicability)}%`, subtitleFontSize, 'brown');
      } else {
      addWrappedText("Total Explicability: 0", subtitleFontSize,
'brown');
      }

      // The combined total score across all five categories
      yPosition += lineHeight;
      const totalScore = surveyData.beneficence +
surveyData.non_maleficence + surveyData.autonomy + surveyData.justice +
surveyData.explicability;
      addWrappedText(`Total Score: ${totalScore}`, titleFontSize,
'black');

      surveyPdf.save("survey_results.pdf");
};


function SurveyComponent() {
      const survey = new Survey.Model(json_questionnaire);
      const [surveyData, setSurveyData] = React.useState(null); // Set
to null initially
      const [isSurveyCompleted, setIsSurveyCompleted] =
React.useState(false); // Track survey completion

      survey.onComplete.add((sender, options) => {
      const { beneficence, non_maleficence, autonomy, justice,
explicability, totalScore, questionDetails } =
calculateScore(sender.data, json_questionnaire);
      setSurveyData({ beneficence, non_maleficence, autonomy, justice,
explicability, totalScore, questionDetails }); // Set both total score
and question details
      setIsSurveyCompleted(true); // Mark the survey as completed
      });
```

```
      const handleSavePdf = () => {
      if (surveyData) {
            savePdf(surveyData);  // Pass the whole surveyData to
savePdf
      }
      };

      const handleCloseModal = () => {
      setIsSurveyCompleted(false); // Close the modal
      };

      return (
      <>
      <SurveyReact.Survey model={survey} />

      {isSurveyCompleted && (
            <div id="savePdfModal">
            <div className="modal-content">
            <h3>Thank you for completing the survey!</h3>
            <p>Would you like to save your results as a PDF?</p>
            <button onClick={handleSavePdf} id="savePdfButton">Save as
PDF</button>
            <button onClick={handleCloseModal}
className="closeButton">Close</button>
            </div>
            </div>
      )}
      </>
      );
}

// Rendering the component
const root =
ReactDOM.createRoot(document.getElementById("surveyElement"));
root.render(<SurveyComponent />);
```

Βιβλιογραφία

[1] L. Floridi, M. Holweg, M. Taddeo, J. A. Silva, J. Mökander, and Y. Wen, "capAI - A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act," *SSRN Electronic Journal*, Jan. 2022, doi: 10.2139/ssrn.4064091.

[2] "Algorithmic Impact Assessment Tool," *Algorithmic Impact Assessment Tool*. https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html

[3] J. McCarthy, M. L. Minsky, N. Rochester, and C. E. Shannon, "A proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955," *AI Magazine*, vol. 27, no. 4, p. 12, Dec. 2006, doi: 10.1609/aimag.v27i4.1904.

[4] V. A. P. by T. Tietz, "John McCarthy and the raise of artificial intelligence | SCIHI blog," Sep. 04, 2020. http://scihi.org/john-mccarthy-artificial-intelligence

[5] E. Rich, *Artificial Intelligence 3E (SIE)*. Tata McGraw-Hill Education, 2019.

[6] NASA, "What is artificial intelligence?" https://www.nasa.gov/what-is-artificial-intelligence/

[7] S. J. Russell, P. Norvig, and E. Davis, *Artificial intelligence: A Modern Approach*. Prentice Hall, 2010.

[8] V. C. Müller, "Ethics of Artificial Intelligence and Robotics," *The Stanford Encyclopedia of Philosophy*, (Fall 2023 Edition), Apr. 2020, [Online]. Available: https://plato.stanford.edu/archives/fall2023/entries/ethics-ai/

[9] United Nations Information Centre, Greece, "Universal Declaration of Human Rights - Greek," *United Nations*, Dec. 1948. https://www.ohchr.org/en/human-rights/universal-declaration/translations/greek-ellinika

[10] L. Sweeney, "Discrimination in online ad delivery," *Communications of the ACM*, vol. 56, no. 5, pp. 44–54, Apr. 2013, doi: 10.1145/2447976.2447990.

[11] J. Dressel and H. Farid, "The accuracy, fairness, and limits of predicting recidivism," *Science Advances*, vol. 4, no. 1, Jan. 2018, doi: 10.1126/sciadv.aao5580.

[12] B. Friedman and H. Nissenbaum, "Bias in computer systems," *ACM Transactions on Office Information Systems*, vol. 14, no. 3, pp. 330–347, Jul. 1996, doi: 10.1145/230538.230561.

[13] J. Kahn, "The world just blew a 'historic opportunity' to stop killer robots—and that might be a good thing," *Fortune*, Dec. 2021. [Online]. Available:

https://fortune.com/2021/12/22/killer-robots-ban-fails-un-artificial-intelligence-laws/

[14] A. Sharkey, "Autonomous weapons systems, killer robots and human dignity," *Ethics and Information Technology*, vol. 21, no. 2, pp. 75–87, Dec. 2018, doi: 10.1007/s10676-018-9494-0.

[15] M. Goos, A. Manning, and A. Salomons, "Job polarization in Europe," *American Economic Review*, vol. 99, no. 2, pp. 58–63, Apr. 2009, doi: 0.1257/aer.99.2.58.

[16] G. Bolsover and P. Howard, "Computational propaganda and political big data: moving toward a more critical research agenda," *Big Data*, vol. 5, no. 4, pp. 273–276, doi: 10.1089/big.2017.29024.cpr.

[17] Clifford Chance, "An overview of the newly adopted EU Data Governance Act," Sep. 2022. https://www.cliffordchance.com/insights/resources/blogs/talking-tech/en/articles/2022/09/an-overview-of-the-newly-adopted-eu-data-governance-act.html

[18] "Regulation - 2022/868 - EN - EUR-LEX," May 2022. https://eur-lex.europa.eu/eli/reg/2022/868/oj/eng

[19] "Regulation - EU - 2023/2854 - EN - EUR-LEX," Dec. 2023. https://eur-lex.europa.eu/eli/reg/2023/2854/oj/eng

[20] European Commission, "Commission publishes frequently asked questions about the Data Act," Sep. 2024. https://digital-strategy.ec.europa.eu/en/library/commission-publishes-frequently-asked-questions-about-data-act

[21] "Regulation - EU - 2024/1689 - EN - EUR-LEX," Jun. 2024. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

[22] European Commission, "AI Act Shaping Europe's digital future." https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

[23] "Νόμος 4961/2022 (Κωδικοποιημένος) - ΦΕΚ Α 146/27.07.2022," *kodiko.gr*. https://www.kodiko.gr/nomothesia/document/810877/nomos-4961-2022

[24] S. Srinath, "AI legislation in the US: A 2025 overview," *Software Improvement Group*. https://www.softwareimprovementgroup.com/us-ai-legislation-overview/

[25] Li and Zhou, "Navigating the Complexities of AI Regulation in China | Perspectives | Reed Smith LLP," *Reed Smith*. https://www.reedsmith.com/en/perspectives/2024/08/navigating-the-complexities-of-ai-regulation-in-china

[26] D. Atanasovska and D. Trombevski, "China's Evolving AI Regulations and Compliance for Companies," *GDPR Local*.

https://gdprlocal.com/chinas-evolving-ai-regulations-and-compliance-for-companies/

[27] B. C. Stahl *et al.*, "A Systematic Review of Artificial Intelligence Impact Assessments," *Artificial Intelligence Review*, vol. 56, no. 11, pp. 12799–12831, doi: 10.1007/s10462-023-10420-8.

[28] Ada Lovelace Institute, "Algorithmic Impact Assessment: a case study in healthcare," Feb. 2022. https://www.adalovelaceinstitute.org/report/algorithmic-impact-assessment-case-study-healthcare/

[29] European Law Institute, "Model rules on impact assessment of Algorithmic Decision-Making Systems used by Public Administration." [Online]. Available: https://www.europeanlawinstitute.eu/fileadmin/user_upload/p_eli/Publications/ELI_Model_Rules_on_Impact_Assessment_of_ADMSs_Used_by_Public_Administration.pdf

[30] Fundacja Moje Państwo, "Algorithmic Impact Assessment Artificial Intelligence Systems and Automatic Decision-Making Systems – Proposal for the public sector." [Online]. Available: https://mojepanstwo.pl/pliki/algorithmic-impact-assessment-ai-adm.pdf

[31] Commission nationale de l'informatique et des libertés, "Self-assessment guide for Artificial intelligence (AI) systems," *CNIL*. https://www.cnil.fr/en/self-assessment-guide-artificial-intelligence-ai-systems

[32] INTERPOL and UNICRI, "Risk Assessment Questionnaire." [Online]. Available: https://www.interpol.int/es/content/download/20929/file/Risk%20Assesment%20Questionnaire.pdf

[33] A. D. Selbst, "An Institutional View of Algorithmic impact Assessments," *SSRN Electronic Journal*, Jun. 2021, [Online]. Available: https://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID3908204_code1328346.pdf?abstractid=3867634&mirid=1

[34] Future of Life Institute, "EU AI Act Compliance Checker | EU Artificial Intelligence Act." https://artificialintelligenceact.eu/assessment/eu-ai-act-compliance-checker/

[35] Al4Belgium, "Tool: AI Assessment Tool." https://data-en-maatschappij.ai/en/tools/tool-ai-assessment-tool

[36] Cisco, "Cisco AI Readiness Assessment." https://www.cisco.com/c/m/en_us/solutions/ai/readiness-index/assessment-tool.html

[37] Microsoft, "Microsoft Responsible AI Impact Assessment Template." [Online]. Available: https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-RAI-Impact

-Assessment-Template.pdf

[38] G. Abercrombie *et al.*, "A collaborative, Human-Centred taxonomy of AI, algorithmic, and automation harms," *arXiv (Cornell University)*, doi: 10.48550/arxiv.2407.01294.

[39] "Article 9 GDPR," *GDPRhub*. https://gdprhub.eu/Article_9_GDPR

[40] O. Zdrok, "Fairness Metrics in AI: Your Step-by-Step Guide to Equitable Systems," *Shelf*, Jun. 2024. https://shelf.io/blog/fairness-metrics-in-ai/

[41] "Group vs. Individual Fairness in AI," *Lumenova AI*, Aug. 2022. https://www.lumenova.ai/blog/group-fairness-vs-individual-fairness/

[42] C. Saplicki, "Fairness Explained: Definitions and Metrics - IBM Data Science in Practice - Medium," *Medium*. [Online]. Available: https://medium.com/ibm-data-ai/fairness-explained-definitions-and-metrics-9690f8e0a4ea

[43] J. R. Foulds, R. Islam, K. N. Keya, and S. Pan, "An intersectional definition of fairness," *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pp. 1918–1921, Apr. 2020, doi: 10.1109/icde48307.2020.00203.

[44] P. J. Phillips *et al.*, "Four principles of explainable artificial intelligence," Sep.2021. doi: 10.6028/nist.ir.8312.

[45] "Robustness Testing: The Essential Guide | Nightfall AI Security 101," *Nightfall AI*. https://www.nightfall.ai/ai-security-101/robustness-testing

[46] GeeksforGeeks, "Robustness Testing," *GeeksforGeeks*. https://www.geeksforgeeks.org/robustness-testing/

[47] "AI Robustness | Deepgram," *Deepgram*. https://deepgram.com/ai-glossary/ai-robustness

[48] G. Mentzas, M. Fikardos, K. Lepenioti, and D. Apostolou, "Exploring the landscape of trustworthy artificial intelligence: Status and challenges," *Intelligent Decision Technologies*, vol. 18, no. 2, pp. 837–854, Jun. 2024, doi: 10.3233/idt-240366.

[49] Υπουργείο ψηφιακής διακυβέρνησης, "Εργαλείο αυτοαξιολόγησης της κυβερνοασφάλειας Οργανισμών." https://mindigital.gr/wp-content/uploads/2022/03/cybersecurity-self-assessment.xlsm

[50] "Best Practices For Decommissioning AI Systems | Restackio." https://www.restack.io/p/ai-system-decommissioning-practices-answer-best-practices-cat-ai

[51] National Institute of Standards and Technology, "AI Risks and Trustworthiness." https://airc.nist.gov/airmf-resources/airmf/3-sec-characteristics/

[52] L. Floridi and J. Cowls, “A unified framework of five principles for AI in society,” *Harvard Data Science Review*, doi: 10.1162/99608f92.8cd550d1.

[53] “SurveyJS - JavaScript Libraries for Surveys and Forms.” https://surveyjs.io/

[54] “React.” https://react.dev/

[55] GeeksforGeeks, “React Introduction,” *GeeksforGeeks*. https://www.geeksforgeeks.org/reactjs-introduction

[56] “nginx.” https://nginx.org/

[57] I. Rose, “#1 What is NGINX? Understanding how it works and its main use cases,” *VPS Jungle Tutorials*. https://www.vpsjungle.in/tutorials/what-is-nginx-understanding-how-it-work/

[58] Algorithm Audit, “A Comparative Review of 10 Fundamental Rights Impact Assessments (FRIA) for AI-systems,” Sep. 2024. [Online]. Available: https://algorithmaudit.eu/knowledge-platform/knowledge-base/comparative_review_10_frias/

[59] “What-If Tool.” https://pair-code.github.io/what-if-tool/

[60] “AI Fairness 360.” https://ai-fairness-360.org/