A blue logo with text

Description automatically generated

**ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ**

**ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ**

**ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ**

**ΠΟΙΟΤΗΤΑ ΚΑΙ ΑΞΙΟΠΙΣΤΙΑ ΛΟΓΙΣΜΙΚΟΥ**

**Ηθικά πλαίσια τεχνητής νοημοσύνης**

**Δημήτριος Στεφάνου**

**20390221**

**Δημήτριος Γουρδομιχάλης**

**20390043**

**Μαρίνο Τσελάνι**

**20390241**

**ΑΘΗΝΑ 2024**

1. **Εισαγωγή**

«Η αξιόπιστη τεχνητή νοημοσύνη αποτελείται από τρία στοιχεία: (1) θα πρέπει να είναι νόμιμη, εξασφαλίζοντας συμμόρφωση με όλους τους ισχύοντες νόμους και κανονισμούς (2) θα πρέπει να είναι ηθική, δείχνοντας σεβασμό και να διασφαλίζει την τήρηση των ηθικών αρχών και αξιών και (3) θα πρέπει να είναι εύρωστη, τόσο από τεχνική όσο και από κοινωνική άποψη, δεδομένου ότι, ακόμη και με καλές προθέσεις, τα συστήματα ΤΝ μπορούν να προκαλέσουν ακούσια βλάβη. Η αξιόπιστη τεχνητή νοημοσύνη δεν αφορά μόνο την αξιοπιστία του ίδιου του συστήματος ΤΝ, αλλά περιλαμβάνει επίσης την αξιοπιστία όλων των διαδικασιών και φορέων που αποτελούν μέρος του κύκλου ζωής του συστήματος ΤΝ». Αυτός είναι ο ορισμός που έδωσε η Ευρωπαϊκή Επιτροπή για την αξιόπιστη τεχνητή νοημοσύνη στη Λίστα Αξιολόγησης για Αξιόπιστη Τεχνητή Νοημοσύνη (ALTAI) το 2020. Η λίστα αυτή βασίζεται σε τέσσερις δεοντολογικές αρχές των οποίων η τήρηση είναι απαραίτητη, εάν επιθυμούμε να διασφάλιση αξιόπιστων AI συστημάτων.

* Σεβασμός της ανθρώπινης αυτονομίας
* Πρόληψη βλάβης
* Δικαιοσύνη
* Επεξηγησιμότητα

Συνοπτικά, οι τέσσερις αυτές αρχές κάνουν λόγο για μία ανθρωποκεντρική προσέγγιση όσον αφορά την ανάπτυξη, την εγκατάσταση και τη χρήση των AI συστημάτων. Επισημαίνεται πως τα εν λόγω συστήματα, καθώς και όλα τα ενδιαφερόμενα μέρη, είναι αναγκαίο να σέβονται την ελευθερία, την αυτονομία, την ιδιωτικότητα και την αξιοπρέπεια του ανθρώπου, αλλά και την προστασία του περιβάλλοντος και του κοινωνικού συνόλου (Ευρωπαϊκή Επιτροπή, 2019; McCormack & Bendechache, 2024).

Σήμερα, η τεχνητή νοημοσύνη είναι ίσως πιο επίκαιρη από ποτέ, αφού πλέον έχει ενσωματωθεί ταχύτατα σε πολλές πτυχές της καθημερινής μας ζωής. Έτσι, γεννάται η επείγουσα ανάγκη για δημιουργία και χρήση πλαισίων αξιολόγησης αξιόπιστης τεχνητής νοημοσύνης (McCormack & Bendechache, 2024; Xia κ.ά., 2024). Ένα τέτοιο πλαίσιο, αν και εξαιρετικά απλοποιημένο, μπορεί να θεωρηθεί και η Λίστα Αξιολόγησης για Αξιόπιστη Τεχνητή Νοημοσύνη. Παρά την ύπαρξη τέτοιων προτύπων, η διαδικασία αξιολόγησης συστημάτων τεχνητής νοημοσύνης ως προς την αξιοπιστία τους παραμένει δύσκολη (McCormack & Bendechache, 2024).

Στην εργασία αυτή, θα παρουσιάσουμε ένα σύνολο πλαισίων αξιολόγησης, ευθυγραμμισμένων με την ανθρωποκεντρική προσέγγιση που περιγράφει το ALTAI, θα αναφέρουμε διάφορα εργαλεία αξιολόγησης αξιόπιστης ΤΝ, θα θέσουμε ορισμένα κριτήρια για τα πλαίσια με βάση τα οποία θα μετρηθεί η απόδοσή τους και θα πραγματοποιήσουμε μία συγκριτική ανάλυση μεταξύ αυτών. Τέλος, θα συζητήσουμε τα αποτελέσματα της έρευνας απαριθμώντας τα πλεονεκτήματα και τις αδυναμίες των πλαισίων και θα καταλήξουμε σε ορισμένα συμπεράσματα.

**2. Ανασκόπηση Βιβλιογραφίας**

Στο κεφάλαιο αυτό θα παρουσιάσουμε τις βασικές έννοιες των καθιερωμένων πλαισίων και προτύπων, καθώς και προτάσεις που προκύπτουν από τη βιβλιογραφία. Οι προτάσεις σχετίζονται τόσο με πλαίσια αξιολόγησης για αξιόπιστη τεχνητή νοημοσύνη όσο και με πρότυπα ταξινόμησης ήδη υπαρχόντων πλαισίων βάσει των κύριων χαρακτηριστικών τους. Θα αναφερθούν διαφόρων ειδών προσεγγίσεις και μέθοδοι για την αξιολόγηση κάθε σταδίου ανάπτυξης του κύκλου ζωής (ανάπτυξη, εγκατάσταση και χρήση) των AI μοντέλων, καθώς και εργαλεία που αποσκοπούν στην εκτίμησή τους ως προς την αξιοπιστία τους (Ευρωπαϊκή Επιτροπή, 2019).

**2.1 Καθιερωμένα εργαλεία, πλαίσια και πρότυπα αξιολόγησης για αξιόπιστα AI συστήματα**

Οι τέσσερις δεοντολογικές αρχές που περιγράφονται στην εισαγωγή, μπορούν να μεταφραστούν σε επτά απαιτήσεις των οποίων η ικανοποίηση συνεπάγεται την τήρησή τους. Οι απαιτήσεις αυτές αποτελούν το κύριο μέρος της Λίστας Αξιολόγησης για Αξιόπιστη Τεχνητή Νοημοσύνη (ALTAI) και ισχύουν για όλα τα ενδιαφερόμενα μέρη που εμπλέκονται στον κύκλο ζωής των συστημάτων. Ενδιαφερόμενα μέρη αποτελούν οι προγραμματιστές, οι εγκαταστάτες, καθώς και οι τελικοί χρήστες των συστημάτων. Ακολουθεί μια συνοπτική ανάλυση των επτά απαιτήσεων:

1. **Ανθρώπινη παρέμβαση και εποπτεία:** Η απαίτηση αυτή σχετίζεται με τη λειτουργία της τεχνητής νοημοσύνης ως ένα μέσο ανάπτυξης μίας δημοκρατικής και ισότιμης κοινωνίας, όπου οι άνθρωποι διατηρούν την αυτονομία τους και την ικανότητά τους να πάρουν ελεύθερα αποφάσεις. Οι χρήστες πρέπει επίσης να διαθέτουν τις κατάλληλες γνώσεις και μέσα που θα συμβάλλουν στη διευκόλυνση της κατανόησης και χρήσης των AI συστημάτων. Η δυνατότητα τόσο της αμφισβήτησης και παρακολούθησης της λειτουργίας των συστημάτων όσο και η δυνατότητα παρέμβασης σε αυτήν, όταν αυτό κρίνεται απαραίτητο, είναι εξίσου σημαντικές (Ευρωπαϊκή Επιτροπή, 2019; Ευρωπαϊκή Επιτροπή, 2020).
2. **Τεχνική στιβαρότητα και ασφάλεια:** Προϋποθέτει την ανάπτυξη των συστημάτων μέσω προσεγγίσεων που διασφαλίζουν την πρόληψη κινδύνων και την ελαχιστοποίηση της πιθανότητας μη αναμενόμενης βλάβης. Πρέπει να εξασφαλίζεται η προστασία των συστημάτων από κακόβουλους χρήστες, καθώς και η σωματική και διανοητική ακεραιότητα του ανθρώπου. Ο όρος “βλάβη” χρησιμοποιείται για να περιγράψει, πέρα από υλικές και σωματικές βλάβες, αυτές που μπορούν να προκληθούν σε κοινωνικά, πολιτισμικά και πολιτιστικά περιβάλλοντα, καθώς και στο ίδιο το φυσικό περιβάλλον. Επίσης, υπαγορεύει πως τα συστήματα πρέπει να είναι ακριβή στις προβλέψεις, συστάσεις και αποφάσεις που λαμβάνει (Ευρωπαϊκή Επιτροπή, 2019; Ευρωπαϊκή Επιτροπή, 2020).
3. **Ιδιωτική ζωή και διακυβέρνηση δεδομένων:** Οι χρήστες χρειάζεται να είναι βέβαιοι πως η ιδιωτική τους ζωή και τα προσωπικά τους δεδομένα δεν διακυβεύονται και δεν επρόκειτο να χρησιμοποιηθούν εναντίον τους. Ένα σύστημα τεχνητής νοημοσύνης πρέπει να εγγυάται για την προστασία τους. Επιπλέον, η ακεραιότητα των δεδομένων εκπαίδευσης των AI συστημάτων είναι απαραίτητη, εφόσον αλλοιωμένα, ανακριβή ή προκατειλημμένα δεδομένα μπορούν να επηρεάσουν σε μεγάλο βαθμό την απόδοση και τη συμπεριφορά τους (Ευρωπαϊκή Επιτροπή, 2019; Ευρωπαϊκή Επιτροπή, 2020).
4. **Διαφάνεια:** Σε πολλές περιπτώσεις, οι αποφάσεις που λαμβάνει ένα σύστημα τεχνητής νοημοσύνης είναι ικανές να προκαλέσουν αναταραχή στη ζωή των χρηστών. Στις περιπτώσεις αυτές, είναι αναγκαίο να παρέχονται επεξηγήσεις σχετικά με τη διαδικασία λήψης αποφάσεων του συστήματος, με σκοπό την τεκμηρίωση της ορθότητάς της. Κρίνεται απαραίτητη η δυνατότητα εντοπισμού των αιτιών που οδήγησαν το σύστημα σε εσφαλμένες αποφάσεις. Τέλος, οι χρήστες δικαιούνται να γνωρίζουν πως αλληλεπιδρούν με ένα AI σύστημα και όχι με άλλο ανθρώπινο ον. Έτσι, τα συστήματα πρέπει να παρουσιάζονται ως τέτοια (Ευρωπαϊκή Επιτροπή, 2019; Ευρωπαϊκή Επιτροπή, 2020).
5. **Πολυμορφία, απαγόρευση των διακρίσεων και δικαιοσύνη:** Τα συστήματα πρέπει να είναι συμπεριληπτικά, δηλαδή να διασφαλίζουν την ισότιμη πρόσβαση σε αυτό από όλους, αλλά και να προωθούν την ίση μεταχείριση όλων. Είναι σημαντικό να αποφεύγονται τάσεις διακρίσεων και προκαταλήψεων εις βάρος κοινωνικών ομάδων, που συνεπάγεται τη λεπτομερή εξέταση και αξιολόγηση των συνόλων δεδομένων εκπαίδευσης των συστημάτων (Ευρωπαϊκή Επιτροπή, 2019; Ευρωπαϊκή Επιτροπή, 2020).
6. **Κοινωνική και περιβαλλοντική ευημερία:** Για ακόμη μια φορά επισημαίνεται η ανάγκη λειτουργίας των συστημάτων με τρόπους που δεν είναι επιβλαβείς για το φυσικό περιβάλλον. Η κατανάλωση ενέργειας και πόρων από τα συστήματα πρέπει να βρίσκεται υπό παρακολούθηση κατά τη διάρκεια της εκπαίδευσής τους. Εκτός αυτού, τα συστήματα ενδέχεται να είναι υπεύθυνα για τον υποβιβασμό των κοινωνικών σχέσεων και δεξιοτήτων των ανθρώπων, καθώς η υπερβολική έκθεση σε αυτά μπορεί να επηρεάσει την αντίληψή μας σχετικά με αυτές τις έννοιες (Ευρωπαϊκή Επιτροπή, 2019; Ευρωπαϊκή Επιτροπή, 2020).
7. **Λογοδοσία:** Αφορά την ανάπτυξη μηχανισμών που θα καθιστούν δυνατή την υπευθυνότητα και τη λογοδοσία όσον αφορά τα AI συστήματα και τα αποτελέσματα που παράγουν. Οι πιθανές αρνητικές επιπτώσεις της χρήσης τους πρέπει να προσδιορίζονται και να γνωστοποιούνται σε όλα τα ενδιαφερόμενα μέρη, όπως και οι τρόποι αντιμετώπισής τους. Οι αλγόριθμοι, τα δεδομένα και οι διαδικασίες σχεδιασμού τους πρέπει να μπορούν να υπόκεινται σε αξιολόγηση. Τέλος, συνετή θα ήταν η ανάπτυξη μηχανισμών διασφάλισης έννομης προστασίας σε περίπτωση δημιουργίας δυσμενών καταστάσεων εξαιτίας ενός συστήματος τεχνητής νοημοσύνης. Τα παραπάνω συμβάλλουν στη διασφάλιση της εμπιστοσύνης απέναντι στα AI συστήματα (Ευρωπαϊκή Επιτροπή, 2019; Ευρωπαϊκή Επιτροπή, 2020).

Στην καθιέρωση μιας ενιαίας και τυποποιημένης διαδικασίας αξιολόγησης συμβάλλει και το πρότυπο ISO/IEC 42001:2023. Πρόκειται περί ενός Management System Standard, το οποίο σημαίνει πως περιλαμβάνει συγκεκριμένες απαιτήσεις για την εγκαθίδρυση πολιτικών και διαδικασιών που αποσκοπούν στην επίτευξη συγκεκριμένων στόχων που αφορούν τη διακυβέρνηση της τεχνητής νοημοσύνης. Το ISO/IEC 42001 παρέχει οδηγίες για την υλοποίηση, τη διατήρηση και τη βελτίωση των AI συστημάτων βάσει διαχειριστικών διαδικασιών που δίνουν έμφαση στην ηθική, την ασφάλεια και τη διαφάνεια. Η αξιοπιστία, η διαχείριση κινδύνων, η διακυβέρνηση δεδομένων και η προσπάθεια συνεχούς βελτίωσης των AI συστημάτων αποτελούν τις κύριες αρχές του προτύπου, πάνω στις οποίες είναι βασισμένο. Το ISO/IEC 42001 συνοδεύεται από συμπληρωματικά πρότυπα, που επικεντρώνονται στη διαχείριση συγκεκριμένων πτυχών της διακυβέρνησης συστημάτων τεχνητής νοημοσύνης. Ένα τέτοιο συμπληρωματικό πρότυπο είναι το ISO/IEC 23894:2022, που παρέχει μια δομημένη προσέγγιση για την αναγνώριση και μείωση των κινδύνων που αφορούν τα AI συστήματα. To ISO/IEC 42001 αποτελείται από δέκα ρήτρες, εκ των οποίων οι επτά τελευταίες διαμορφώνουν τον πυρήνα του προτύπου. Αυτές είναι:

* 4η ρήτρα: Αφορά τον προσδιορισμό του πλαισίου της εκάστοτε οργάνωσης, που προϋποθέτει πως οι τελευταίες κατανοούν τους παράγοντες που επηρεάζουν το AIMS (Artificial Intelligence Management System Standard) προς ανάπτυξη. Έναν παράγοντα υψίστης σημασίας αποτελούν οι ανάγκες και οι προσδοκίες των ενδιαφερόμενων μερών.
* 5η ρήτρα (Ηγεσία): Απαιτεί την απόλυτη δέσμευση της διοίκησης της οργάνωσης, την εγκαθίδρυση μιας συγκεκριμένης πολιτικής όσον αφορά την τεχνητή νοημοσύνη, καθώς και την υιοθέτηση μιας κουλτούρας με κέντρο την υπεύθυνη χρήση του AI.
* 6η ρήτρα (Σχεδιασμός): Αφορά το σχεδιασμό διαδικασιών που αποσκοπούν στην αντιμετώπιση των κινδύνων, καθώς και στην εκμετάλλευση των ευκαιριών που ενδέχεται να προκύψουν. Επίσης, η οργάνωση θα πρέπει να προσδιορίσει τους στόχους της και να είναι έτοιμη να διαχειριστεί δυνητικές αλλαγές που αφορούν το AIMS.
* 7η ρήτρα (Υποστήριξη): Σχετίζεται με την εξασφάλιση των απαραίτητων πόρων και ικανοτήτων, όπως η ικανότητα της επικοινωνίας, καθώς και των κατάλληλων εγχειριδίων που χρειάζονται για την ανάπτυξη του AIMS.
* 8η ρήτρα (Λειτουργίες): Περιλαμβάνει προϋποθέσεις σχετικές με την υλοποίηση, το σχεδιασμό και τον έλεγχο του AI συστήματος, όπως ο προσδιορισμός των επιπτώσεών του.
* 9η ρήτρα (Αξιολόγηση απόδοσης): Δίνει έμφαση σε απαιτήσεις που αφορούν την παρακολούθηση, την ανάλυση, την αξιολόγηση και τη μέτρηση της ποιότητας του AIMS.
* 10η ρήτρα (Βελτίωση): Σχετίζεται με την ανάγκη για συνεχή βελτίωση του AIMS. Αυτό επιτυγχάνεται μέσω διάφορων διαδικασιών, όπως η διόρθωση αποκλίσεων και η διατήρηση πληροφοριών με σκοπό την παρακολούθηση της προόδου.

Επίσης, το πρότυπο παρέχει τέσσερα παραρτήματα που προσφέρουν περαιτέρω καθοδήγηση:

* Παράρτημα A: Περιγράφει 39 ελέγχους που σχετίζονται με υπεύθυνες πρακτικές τεχνητής νοημοσύνης. Οι έλεγχοι αυτοί αφορούν πεδία όπως η διαχείριση των δεδομένων, η διαφάνεια και η ηθική. Ένα σημαντικό χαρακτηριστικό του παραρτήματος είναι πως παροτρύνει τις οργανώσεις να συμπεριλάβουν ορισμένα ενδιαφερόμενα μέρη στη διαδικασία ανάπτυξης και εφαρμογής των AI συστημάτων, όπως υπαλλήλους και πελάτες.
* Παράρτημα B: Παρέχει οδηγίες σχετικά με τη διαχείριση δεδομένων, την εκτίμηση δυνητικών κινδύνων και τον προσδιορισμό των πιθανών επιπτώσεων του συστήματος.
* Παράρτημα C: Αναλύει πηγές κινδύνων και πιθανούς επιχειρησιακούς στόχους της τεχνητής νοημοσύνης.
* Παράρτημα D: Εξετάζει σενάρια χρήσης της τεχνητής νοημοσύνης ανάλογα με το πλαίσιο κάθε βιομηχανίας.

Ακολουθώντας τις παραπάνω οδηγίες, μια οργάνωση έχει τη δυνατότητα να αναπτύξει ένα AIMS που ικανοποιεί τις ανάγκες της και εξασφαλίζει την υπεύθυνη ανάπτυξη, εφαρμογή και διαχείριση των AI συστημάτων. Μέσω αυστηρών και λεπτομερών διαδικασιών αξιολόγησης του AIMS της εκάστοτε οργάνωσης, η τελευταία έχει τη δυνατότητα να αποκτήσει πιστοποίηση ISO/IEC 42001. Η κατοχή πιστοποίησης ISO/IEC 42001 εγγυάται πως μια οργάνωση είναι αφοσιωμένη στην επιθυμία της για αξιόπιστη τεχνητή νοημοσύνη και της παρέχει ανταγωνιστικό πλεονέκτημα.

**2.2 Προτάσεις πλαισίων και προτύπων για αξιόπιστα AI συστήματα**

Η πρώτη πρόταση αφορά το πλαίσιο των Xia κ.ά. (2024). Οι συγγραφείς υποστηρίζουν πως η αξιολόγηση πρέπει να εστιάζει τόσο στα μέρη που απαρτίζουν το AI μοντέλο που ενσωματώνεται στο σύστημα όσο και στα στοιχεία του συστήματος που δεν αφορούν το AI μοντέλο (π.χ. διεπαφές χρήστη και APIs). Προτείνεται, δηλαδή, μια καθολική εκτίμηση του συστήματος και όχι απλά του τμήματος που αφορά την τεχνητή νοημοσύνη. Το πλαίσιο αποτελείται από τρεις βασικές φάσεις:

* Εναρμόνιση της ορολογίας: Καθώς εμπλέκονται πολλές κοινότητες για την ανάπτυξη των εν λόγω συστημάτων (π.χ. τεχνητή νοημοσύνη, software engineering κ.ά.) είναι απαραίτητη η εγκαθίδρυση μια εναρμονισμένης ορολογίας που θα χρησιμοποιείται από όλους. Έτσι, δεν θα εμποδίζεται η συνεργασία μεταξύ των κοινοτήτων και θα προωθείται μια ενοποιημένη προσέγγιση. Για παράδειγμα, το πλαίσιο προτείνει πως ο όρος “Narrow AI” αναφέρεται σε έναν τύπο τεχνητής νοημοσύνης που είναι σχεδιασμένος για την εκτέλεση μίας συγκεκριμένης εργασίας εντός στενά ορισμένων πλαισίων. Αντίθετα, ο όρος “General AI” αναφέρεται σε τύπους τεχνητής νοημοσύνης που μπορούν να ανταπεξέλθουν σε μία μεγάλη ποικιλία εργασιών που αφορούν διαφορετικές περιπτώσεις χρήσεων.
* Ταξινομημένη αξιολόγηση: Αφορά την αξιολόγηση σε επίπεδο στοιχείων, αλλά και σε επίπεδο συστήματος. Η αξιολόγηση σε επίπεδο στοιχείων αφορά τόσο την αξιολόγηση των AI τμημάτων όσο και την αξιολόγηση των τμημάτων που δεν σχετίζονται με το AI μοντέλο. Η αξιολόγηση σε επίπεδο συστήματος αναφέρεται στην καθολική αξιολόγηση, που περιλαμβάνει την πολυπλοκότητα του συστήματος, τις δυνατότητες και τους περιορισμούς που προκύπτουν από τις περιβαλλοντικές παραμέτρους και τη συνολική αλληλεπίδραση των στοιχείων του. Κάθε στοιχείο, ανάλογα με την κατηγορία στην οποία ανήκει, αξιολογείται βάσει συγκεκριμένων πεδίων και μετρικών (π.χ. accuracy, quality), η σημασιολογία των οποίων έχει επεξηγηθεί στην πρώτη φάση. Επιπλέον, οι συγγραφείς παρέχουν αναλυτικές κατευθυντήριες γραμμές για την εφαρμογή της αξιολόγησης σύμφωνα με κάθε πεδίο ή μετρική.
* Χαρτογράφηση του κύκλου ζωής του AI συστήματος: Σε αυτή τη φάση, εντοπίζονται κρίσιμες αξιολογήσεις για κάθε φάση του κύκλου ζωής του συστήματος και για διαφορετικά ενδιαφερόμενα μέρη. Κάθε οντότητα, ανάλογα με το στάδιο του κύκλου ζωής του συστήματος, χρειάζεται να αξιολογεί το σύστημα με διαφορετικό τρόπο.

Στην έρευνά τους, οι Oveisi κ.ά. (2024) αναλύουν πάνω από 200 πρότυπα και μελέτες. Διαχωρίζουν τα AI συστήματα σε στρώματα και παρέχουν οδηγίες για την αξιολόγηση ή ανάπτυξη κάθε στρώματος, που αποσκοπούν στη δημιουργία ενός στιβαρού και αξιόπιστου συνολικού συστήματος. Επιπλέον, τονίζουν την ανάγκη για συνεχή αξιολόγηση καθ’ όλη τη διάρκεια ζωής του. Τα στρώματα είναι:

* Υλικό: Χρησιμοποιεί το μοντέλο RAMI, που αποσκοπεί στην ανάπτυξη προϊόντων τεχνητής νοημοσύνης. Για κάθε φάση ανάπτυξης, παραπέμπει σε διάφορα τυποποιημένα πρότυπα που διασφαλίζουν την υψηλή ποιότητα και στιβαρότητα του προϊόντος.
* Λογισμικό: Εξετάζονται ποσοτικές και ποιοτικές μετρικές αξιολόγησης. Για τις ποσοτικές αξιολογήσεις, οι συγγραφείς προτείνουν δοκιμαστικές εκτελέσεις του κώδικα του συστήματος και μέτρηση της απόδοσης βάσει μαθηματικών τύπων. Η ποιοτική αξιολόγηση χωρίζεται σε πέντε κατηγορίες: 1) Διαφάνεια και επεξηγησιμότητα, 2) Κυβερνοασφάλεια και ιδιωτικότητα, 3) Αμεροληψία, 4) Στιβαρότητα και 5) Ασφάλεια και αξιοπιστία. Για κάθε μία από της κατηγορίες, οι συγγραφείς παραπέμπουν τον αναγνώστη σε ένα σύνολο προτύπων και οδηγιών για την αξιολόγηση του λογισμικού με βάση κάθε κατηγορία. Τέλος, δίνεται έμφαση στην ανάγκη δοκιμής και αξιολόγησης “παραδοσιακού” λογισμικού που δεν σχετίζεται με το πεδίο της τεχνητής νοημοσύνης.
* Σύνολα δεδομένων: Η έρευνα εστιάζει στην ανάγκη για ακεραιότητα, ποιότητα και ασφάλεια των δεδομένων. Μέσω προτύπων διακυβέρνησης δεδομένων εξασφαλίζεται η ιδιωτικότητα του χρήστη, καθώς και η διαφάνεια. Παρέχονται καθιερωμένα πλαίσια και πρότυπα διακυβέρνησης δεδομένων για την ανάπτυξη κατάλληλων διαδικασιών διαχείρισης δεδομένων σε μία οργάνωση.
* Μοντέλα μηχανικής μάθησης: Σε αυτό το στρώμα, οι συγγραφείς εστιάζουν κυρίως σε μετρικές απόδοσης του μοντέλου. Διαχωρίζουν τα μοντέλα μηχανικής μάθησης σε δύο κατηγορίες: 1) Επιβλεπόμενης μάθησης και 2) Μη επιβλεπόμενης μάθησης. Ανάλογα με την κάθε κατηγορία, προτείνουν μετρικές για την μέτρηση της απόδοσης, όπως είναι, για παράδειγμα, ο δείκτης Jaccard, η ακρίβεια και η ανάκλιση.

Οι McCormack και Bendechache (2024) προτείνουν μία μέθοδο ταξινόμησης υφιστάμενων πλαισίων αξιολόγησης αξιόπιστης τεχνητής νοημοσύνης. Τα πλαίσια που εξετάσθηκαν, ανάλογα με τον τύπο της αξιολόγησης που πρότειναν, ταξινομήθηκαν σε τέσσερις κύριες κατηγορίες:

* Εννοιολογική αξιολόγηση: Μέθοδοι υψηλού επιπέδου που δεν είναι δοκιμασμένες και δεν παρέχουν λεπτομέρειες όσον αφορά την υλοποίησή τους.
* Χειρωνακτική αξιολόγηση: Είναι βασισμένη σε ερωτηματολόγια αυτοαξιολόγησης, όπως το ALTAI. Βασικό της μειονέκτημα αποτελεί το γεγονός πως είναι υπερβολικά χρονοβόρα.
* Αυτόματη αξιολόγηση: Αφορά αυτόματες μεθόδους αξιολόγησης. Μεγάλο τους πλεονέκτημα αποτελεί η συνέπεια και η σταθερότητα της διαδικασίας αξιολόγησης. Παρόλα αυτά, βασίζονται σε προκαθορισμένες μετρικές που μπορεί να μην είναι όλες κατάλληλες για κάθε πτυχή του αξιόπιστου AI.
* Ημιαυτόματη αξιολόγηση: Περιλαμβάνει αυτοματοποιημένα, καθώς και χειρωνακτικά βήματα. Απαιτείται η ανθρώπινη παρέμβαση σε ορισμένα στάδια, συνδυάζοντας τα πλεονεκτήματα του αυτοματισμού και τη χειρωνακτική αποτελεσματικότητα.

Στη συνέχεια, πραγματοποιήθηκε μία υπό-ταξινόμηση των πλαισίων σε τέσσερις υποκατηγορίες, ανάλογα με το θέμα ως προς το οποίο αξιολογούνται. Οι υποκατηγορίες είναι:

* Αξιολόγηση ως προς τη δικαιοσύνη και τη συμμόρφωση
* Αξιολόγηση ως προς τη διαφάνεια
* Αξιολόγηση ως προς τους κινδύνους και τη λογοδοσία
* Αξιολόγηση ως προς την αξιοπιστία και την ασφάλεια

Η πλειοψηφία των πλαισίων κατατάσσεται στην κατηγορία της εννοιολογικής αξιολόγησης, το οποίο υποδεικνύει το περιορισμένο εύρος της έρευνας σε αυτόν τον τομέα. Για κάθε κύρια κατηγορία εξετάζεται κάθε υποκατηγορία αξιολόγησης και αναλύονται οι ανάλογες έννοιες, καθώς και τα κατάλληλα εργαλεία αξιολόγησης. Για παράδειγμα, πρότυπα και εργαλεία χειρωνακτικής αξιολόγησης ως προς τους κινδύνους και τη λογοδοσία, αποτελούν το ISO\IEC 42001 και το NIST (National Institute of Standards and Technology) AI Management Risk Framework. Οι εννοιολογικές μέθοδοι αναλύουν έννοιες όπως πολιτικές εντοπισμού παραβιάσεων συμμόρφωσης και χρήση της τεχνητής νοημοσύνης για τον προσδιορισμό ηθικών αρχών. Όσον αφορά εργαλεία για αυτοματοποιημένες διαδικασίες, αναφέρεται το AI Fairness 360 της IBM και το Microsoft Fairlearn, που χρησιμοποιούνται για την αξιολόγηση της δικαιοσύνης, της αμεροληψίας και της διαφάνειας σε AI μοντέλα. Επιπλέον, μία μέθοδος που ενσωματώνει ημιαυτόματες διαδικασίες είναι το FairHIL (Fair Human-in-the-Loop). Το συγκεκριμένο εργαλείο προσφέρει μια γραφική διεπαφή χρήστη, η οποία περιλαμβάνει ένα σύνολο οπτικοποιήσεων, όπως η απεικόνιση των χαρακτηριστικών των αποτελεσμάτων εξόδου του μοντέλου. Οι χρήστες έχουν τη δυνατότητα να ρυθμίσουν τη βαρύτητα των χαρακτηριστικών, προκειμένου να εκπαιδεύσουν εκ νέου το μοντέλο σύμφωνα με τις απαιτήσεις τους. Οι συγγραφείς επισημαίνουν την προτίμησή τους στις ημιαυτόματες μεθόδους, καθώς η ανθρώπινη παρέμβαση στη διαδικασία αξιολόγησης θεωρείται ότι προσφέρει σημαντικά πλεονεκτήματα, ενισχύοντας την αξιοπιστία και την εγκυρότητα των αποτελεσμάτων.

Το άρθρο των Nasir κ.ά. (2024) εστιάζει στη δημιουργία ενός ηθικού πλαισίου για την χρήση και τη λειτουργία της τεχνητής νοημοσύνης αλλά με επίκεντρο τον τομέα της υγειονομικής περίθαλψης. Για την ανάλυση του θέματος αναφέρονται διάφορες έννοιες, όπως είναι η διαφάνεια και ερμηνεία, η ασφάλεια δεδομένων, το δίκαιο και η ανθρωποκεντρική χρήση του ΑΙ.

Αναλυτικότερα, η έννοια της διαφάνειας αναφέρεται στην παροχή συναφών εξηγήσεων σχετικά με τα μοντέλα της τεχνητής νοημοσύνης για όλες τις αποφάσεις που λαμβάνουν, το οποίο είναι ιδιαίτερα σημαντικό σε αυτό το πεδίο καθώς αυτές οι αποφάσεις μπορεί να είναι ζήτημα ζωής ή θανάτου για τους ασθενείς. Όσον αφορά την ασφάλεια των δεδομένων οι συγγραφείς εστιάζουν στην εξασφάλιση της ιδιωτικότητας και πιο σημαντικά στην προστασία από κακόβουλους χρήστες, καθώς πρόκειται για ευαίσθητα δεδομένα. Άλλη μια σημαντική έννοια που αναφέρεται είναι το δίκαιο, δηλαδή, η αποφυγή οποιασδήποτε διάκρισης και προκατάληψης σε επίπεδο δεδομένων αλλά και των ιδίων των αλγορίθμων. Τέλος, η έννοια του ανθρωποκεντρισμού, έχει να κάνει με τη λειτουργεία της τεχνητής νοημοσύνης έτσι ώστε αυτή να εξυπηρετεί τις ανθρώπινες ανάγκες.

Για την περεταίρω αξιολόγηση του πλαισίου αυτού γίνεται μια γενική αναφορά σε μελέτες χωρίς να ονομάζει κάποια συγκεκριμένη. Οι μελέτες αυτές αναδεικνύουν διάφορα ηθικά ζητήματα όπως οι προκαταλήψεις στα δεδομένα. Άλλα θέματα που αναδεικνύονται είναι αυτά της διαφάνειας αλλά και της ακρίβειας των αποτελεσμάτων μέσω μελετών που γίνονται σε εργαλεία όπως το GPT-4 και τα συστήματα Watson της IBM τα οποία παράγουν αποτελέσματα που χρειάζονται επεξήγηση. Πέρα από τις μελέτες, χρησιμοποιούνται διάφορες μέθοδοι και εργαλεία για την αξιολόγηση. Πιο συγκεκριμένα, χρησιμοποιούνται εργαλεία όπως το LIME και το SHAP που αφορούν το explainable AI (XAI) και αναλύουν τα χαρακτηριστικά που μπορούν να επηρεάσουν μια πρόβλεψη σε κάποια εξέταση ενός ασθενή. Άλλες μέθοδοι είναι διάφορες διαδικασίες που εντοπίζουν και αναλύουν τις τυχόν προκαταλήψεις σε δεδομένα ,όπως είναι το Aggregation και το Measurement Bias, και διάφοροι μηχανισμοί ελέγχου παραβιάσεων ασφάλειας όπως είναι τα blockchain-based audits.

Το άρθρο της Μυλώση κ.ά. (2021) αναφέρεται στην ηθική της τεχνητής νοημοσύνης (AI) και εστιάζει ιδιαίτερα στον αλγοριθμικό ντετερμινισμό και την αυτονομίας, όπως αυτά περιγράφονται στο πλαίσιο του GDPR. Οι βασικές έννοιες που παρουσιάζονται στο GDPR είναι, η διαφάνεια, η δικαιοσύνη, η ωφέλεια, η αποφυγή βλάβης και η αυτονομία. Πιο αναλυτικά, η διαφάνεια επικεντρώνεται στην παροχή εξηγήσιμων αποφάσεων σχετικά με την τεχνητή νοημοσύνη, έτσι ώστε οι χρήστες να είναι επαρκώς ενημερωμένοι σχετικά με τη λειτουργία του μοντέλου τεχνητής νοημοσύνης. Όσον αφορά την δικαιοσύνη, οι συγγραφείς εξηγούν πως συνδέεται με τη δίκαιη μεταχείριση και αποφυγή των προκαταλήψεων έναντι των ανθρώπων. Η ωφέλεια αναφέρει πως το AI πρέπει να είναι ανθρωποκεντρικό, δηλαδή να εξυπηρετεί τις ανθρώπινες ανάγκες, να ελέγχεται από του ανθρώπους και τελικά να προάγει το κοινό καλό. Έπειτα, αναφέρεται η αποφυγή βλάβης, που αφορά την ελαχιστοποίηση της πιθανότητας να υπάρξουν αρνητικές επιπτώσεις λόγω της χρήσης αυτών των μοντέλων της τεχνητής νοημοσύνης. Τέλος, η αυτονομία, που είναι παρόμοια με την ωφέλεια, παρέχει την επιβεβαίωση πως οι χρήστες θα κατέχουν πλήρη έλεγχο πάνω στα δεδομένα τους αλλά και στον τρόπο που αυτά χρησιμοποιούνται.

Οι συγγραφείς αναφέρουν τις παρακάτω μελέτες: Η πρώτη είναι η AI4People (2018), που με βάση αυτή τέθηκαν οι πέντε βασικές αρχές ηθικής χρήση της AI που αναφέρθηκαν παραπάνω. Η επόμενη είναι αυτή από την Cambridge Analytica (2018), που αναφέρεται σε ένα παράδειγμα μη τήρησης της ηθικής χρήσης δεδομένων πολλών χρηστών, τονίζοντας την ανάγκη για διαφάνεια και λογοδοσία. Άλλη μία είναι η μελέτη της Ευρωπαϊκής Επιτροπής (2020) που σχετίζεται με το GDPR και πώς αυτό εφαρμόζεται σε τεχνολογίες AI, με βασικό γνώμονα την αύξηση της προστασίας των δεδομένων. Τελευταία είναι η Amazon Recruiting Tool (2018). Αυτή χρησιμοποιεί ένα παράδειγμα που αφορά έναν αλγόριθμο ο οποίος διαιώνιζε προκαταλήψεις κατά των γυναικών, που υπογραμμίζει την αναγκαιότητα βελτίωσης της δικαιοσύνης των αλγορίθμων.

Για την περεταίρω αξιολόγηση του πλαισίου χρησιμοποιούνται διάφορες μέθοδοι και εργαλεία. Και εδώ γίνεται λόγος για τα explainable AI (XAI) εργαλεία LIME και SHAP, που χρησιμοποιούνται για την επεξήγηση της λογικής των αλγοριθμικών αποφάσεων. Στην συνέχεια γίνεται λόγος για την μέθοδο Privacy by Design, δηλαδή η δημιουργία των AI συστημάτων βάσει της προστασίας των δεδομένων των χρηστών. Άλλη μέθοδος είναι η συμμόρφωση με τις αρχές του GDPR. Η τελευταία μέθοδος που αναφέρεται είναι η ανάλυση κινδύνου, ώστε να προστατεύονται τα δικαιώματα και τα δεδομένα των χρηστών.

Το άρθρο των O’ Sullivan κ.ά. (2018) επικεντρώνεται στα ηθικά πλαίσια του AI και της αυτόνομης χειρουργικής. Οι βασικές έννοιες που παρουσιάζονται είναι η υπευθυνότητα, δηλαδή η καταγραφή και η ανάλυση των αποφάσεων και ενεργειών με τρόπο τέτοιο ώστε να μπορεί να γίνει η κατάλληλη απόδοση ευθυνών σε περίπτωση λάθους (νομική ευθύνη). Πέρα από την υπευθυνότητα και την νομική ευθύνη αναφέρεται και η διαφάνεια, δηλαδή η κατανόηση της λειτουργίας των ΑΙ συστημάτων αλλά και την επεξήγηση του τρόπου που λαμβάνουν αποφάσεις.

Επίσης, στο άρθρο γίνονται αναφορές στα πρότυπα ISO/IEC 15408 και IEC 60601 για την ασφάλεια ιατρικού εξοπλισμού και μελέτες όπως των Bonaci et al. (2017) για τους κινδύνους επιθέσεων σε τέτοιου τύπου συστήματα. Για την αξιολόγηση του πλαισίου χρησιμοποιούνται διάφορες μέθοδοι και εργαλεία, όπως η Explainable AI (XAI) για την κατανόηση της λειτουργίας και των αποφάσεων των αλγορίθμων, η χρήση συστημάτων μαύρου κουτιού, που επιτρέπει την μελέτη των δεδομένων του συστήματος με σκοπό την εξέταση τυχόν αποτυχιών των συστημάτων. Τέλος, αναφέρονται τεχνικές της κυβερνοασφάλειας ως εργαλεία πρόληψης και αντιμετώπισης επιθέσεων κατά των συστημάτων.

Οι Peters κ.ά. (2020) θίγουν την έννοια της υπευθυνότητας και εστιάζουν στον ηθικό σχεδιασμό και ανάπτυξη των AI συστημάτων, ώστε να αποφευχθούν από την αρχή ανεπιθύμητες επιπτώσεις χρήσης τους. Επίσης, οι συγγραφείς αναφέρουν τη διαφάνεια, εφόσον θεωρούν σημαντική την κατανόηση των αποφάσεων των μηχανισμών του ΑΙ αλλά και των επιπτώσεών τους. Αναφέρονται, επιπλέον, στην ανθρωποκεντρική προσέγγιση ανάπτυξης των συστημάτων, που απαιτεί τη συμμετοχή των ανθρώπων από όλες τις σκοπιές, όπως χρήστες, σχεδιαστές και εμπειρογνώμονες.

Το άρθρο αναφέρεται σε διάφορες μελέτες. Μία από αυτές, είναι η IEEE P7000 Standards (2019) που κάνει λόγο για την ηθική σχεδίαση αυτόνομων και έξυπνων συστημάτων, καθώς και η AI4People Framework (2018), που πρόκειται για μία μελέτη που ορίζει βασικές αρχές για τα ηθικά AI συστήματα. Η επόμενη μελέτη είναι των Jobin et al. (2019) που συστηματικά αναλύουν 84 ηθικές κατευθυντήριες γραμμές που προσδιορίζουν αρχές όπως η διαφάνεια, η δικαιοσύνη, η υπευθυνότητα και η ιδιωτικότητα. Τελευταία, αλλά εξίσου σημαντική, είναι η μελέτη Datasheets for Datasets από τους Gebru et al.(2018), η οποία αναφέρεται στην ενίσχυση της διαφάνειας και την ευθύνης με τη χρήση τεκμηριωμένων συνόλων δεδομένων.

Μια μεθοδολογία ανάπτυξης που χρησιμοποιείται είναι η κατά στάδια υλοποίηση, που εισάγεται από το πλαίσιο Responsible Design Process. Κατά το πλαίσιο αυτό, αξιολογείται ανά στάδια ανάπτυξης η ευθυγράμμιση του συστήματος με ηθικές αρχές όπως η διαφάνεια και η δικαιοσύνη. Επίσης, χρησιμοποιούνται και εργαλεία αξιολόγησης, όπως το IBM 360 Fairness Toolkit, που επιτρέπουν τη μέτρηση της δικαιοσύνης αναλύοντας και επανεκτιμώντας τους αλγορίθμους και τα σύνολα δεδομένων. Τέλος, χρησιμοποιούνται Ethical Checklists που περιέχουν τα καίρια ηθικά ζητήματα και χρησιμοποιούνται ως οδηγοί αξιολόγησης.

Ο Prem (2023) εστιάζει σε έννοιες όπως η διαφάνεια, η λογοδοσία, η ιδιωτικότητα, η δικαιοσύνη και η επεξηγησιμότητα στα πλαίσια του αξιόπιστου ΑΙ. Αναφέρει σημαντικές μελέτες σχετικές με την ηθική στην τεχνητή νοημοσύνη. Η πιο σημαντική μελέτη είναι αυτή της Ευρωπαϊκής Επιτροπής, που θεσμοθετεί την ηθική των ΑΙ συστημάτων, δημιουργώντας μια λίστα με βασικές ηθικές απαιτήσεις που πρέπει να πληρούν τα AI συστήματα. Αυτές είναι η ανθρώπινη παρέμβαση και εποπτεία, η τεχνική ανθεκτικότητα και ασφάλεια, η ιδιωτικότητα και διακυβέρνηση δεδομένων, η διαφάνεια, η ποικιλομορφία, η κοινωνική και περιβαλλοντική ευημερία, και η λογοδοσία. Άλλη μια μελέτη που αναφέρεται είναι αυτή των Morley et al. (2020) με τίτλο “From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices” που αναλύει σε βάθος τα ηθικά πλαίσια και αξιολογεί ένα μεγάλο εύρος πρακτικών εργαλείων χρήσης της τεχνητής νοημοσύνης. Απώτερος σκοπός της έρευνας είναι η πρόταση μιας εκτενούς διαδικασίας σχεδιασμού AI συστημάτων, αναγνωρίζοντας τις φάσεις όπου απαιτούνται ηθικές παρεμβάσεις. Τέλος, αναφέρεται η μελέτη των Floridi et al. (2018), στην οποία πραγματοποιείται μια σειρά αναλύσεων 47 αρχών ηθικής για την ΤΝ μέσα από διαφορά έγγραφα και οδηγίες από 6 βασικές πήγες. Η έρευνα αναζητά τις κυριότερες ηθικές αρχές της βιβλιογραφίας με σκοπό να παραχθεί μια πιο ενιαία οπτική.

Στο paper του Pastor-Dark κ.ά. (2022), που έχει ως κύριο θέμα τα Ηθικά Πλαίσια Τεχνητής Νοημοσύνης για τη δημιουργία βιώσιμων πόλεων, γίνεται λεπτομερής ανάλυση των SDG (Στόχοι Βιώσιμης Ανάπτυξης ΟΗΕ). Οι Στόχοι Βιώσιμης Ανάπτυξης και οι ηθικές αρχές που συνοδεύουν αποτελούν τα θεμέλια ενός ολοκληρωμένου πλαισίου. Οι SDG οφείλουν να υιοθετηθούν από όλες τις χώρες-μέλη του ΟΗΕ, ανεξάρτητα της πολιτικοοικονομικής κατάστασης στην οποία βρίσκονται. Σύμφωνα με τον ΟΗΕ, η εφαρμογή των 17 SDG και των υπο-στόχων τους θα έχει ως αποτέλεσμα την κοινωνική ευημερία, δηλαδή τη βελτίωση των συνθηκών ζωής των πολιτών όλων των χωρών του κόσμου. Αναλυτικότερα, στο συγκεκριμένο paper αποτυπώνονται διάφοροι κοινωνικοί και οικονομικοί στόχοι, με σκοπό τη μετατροπή των πόλεων σε βιώσιμες, σύγχρονες και ανθεκτικές. Από τους συγγραφείς παρουσιάζεται πλήθος ηθικών αρχών. Οι βασικότερες είναι οι εξής: ισότητα, ισονομία, προάσπιση των ατομικών δικαιωμάτων, προστασία των δεδομένων, ελευθερία, αλληλεγγύη, συνεργασία μηχανών και ανθρώπων, ψηφιακή ένταξη, πολυπολιτισμικότητα, αυτονομία της κοινωνίας και κλιματική εξέλιξη. Καταλαβαίνουμε ότι τα ψηφιακά δικαιώματα πρέπει να μεταφραστούν σε ηθικές αρχές. Αξίζει να τονιστεί ότι μέσω του συνδυασμού των παραπάνω ηθικών αρχών με το ψηφιακό μετασχηματισμό, με το πέρασμα των ετών, θα εντοπιστεί η ατομική βελτίωση των πολιτών και συνολικά η κοινωνική βιωσιμότητα. Για παράδειγμα, το SDG-11 περιγράφει τις πόλεις ως ασφαλείς, χωρίς αποκλεισμούς και βιώσιμες. Ταυτόχρονα, συνδέει αρχές που πρέπει να ακολουθούν οι πολίτες με τη ψηφιακή ανάπτυξη. Αναλυτικότερα, κατά τη συλλογή δεδομένων, με τη χρήση εργαλείων Τεχνητής Νοημοσύνης, πρέπει να λαμβάνονται υπ’ όψη όλοι οι έλεγχοι που έχουν να κάνουν με την προστασία των προσωπικών δεδομένων, όπως αποτυπώνονται στους σχετικούς κανονισμούς. Επίσης, η χρήση δεδομένων για τη λήψη αποφάσεων από μοντέλα AI πρέπει να προϋποθέτει τη διαφάνεια σε κάθε στάδιο της διαδικασίας (χρήση αλγορίθμων αξιολόγησης) και την κατανόηση της κοινωνίας από το μοντέλο. Άρα, η Τεχνητή Νοημοσύνη και πλαίσια ΤΝ με «ηθικό χαρακτήρα» έχουν τη δυνατότητα να βοηθήσουν στην επίτευξη των Στόχων Βιώσιμης Ανάπτυξης από τεχνική άποψη. Σύμφωνα με τους συγγραφείς, η χρήση μοντέλων Τεχνητής Νοημοσύνης και υποκατηγοριών της, όπως η Μηχανική Μάθηση, είναι αρκετά πιθανό πως θα οδηγήσουν σε ταχύτερη και προς το καλύτερο αλλαγή της μορφής της κοινωνίας και των πόλεων, και στην εύκολη προσαρμογή των ανθρώπων στη ψηφιακή κοινωνία.

Η έρευνα των Floridi κ.ά. (2018) αποτελεί μία σύνθεση συνόλων αρχών, δηλαδή υπαρχόντων καθιερωμένων πλαισίων. Αναλυτικότερα, σύμφωνα με τους συγγραφείς, στόχος του paper είναι να αποτυπώσει τις βάσεις για μια ποιοτική AI κοινωνία. Σημειώνουν ότι έχει γίνει αντιληπτό ότι η Τεχνητή Νοημοσύνη έχει αντίκτυπο στην κοινωνία. Σημαντικό ζήτημα είναι να γίνει πρόβλεψη των κινδύνων της ΤΝ, δηλαδή της κακής χρήσης της και συνεπώς να αξιοποιηθεί μόνο το θετικό αντίκτυπο της. Από την άλλη πλευρά, ο φόβος για τις συνέπειες της ΤΝ μπορεί να οδηγήσει στην υποχρησιμοποίηση της Τεχνολογίας και των εργαλείων που παρέχονται. Για αυτό και οι συγγραφείς πιστεύουν ότι η σύνθεση αρχών διαφορετικών πλαισίων θα διασφαλίσει την παραγωγή αποτελεσμάτων με θετικό πρόσημο για την κοινωνία και την αποφυγή της κακής χρήσης της Τεχνητής Νοημοσύνης. Ειδικότερα, εξετάστηκαν διάφορα πλαίσια όπως οι πέντε βασικές αρχές για έναν κώδικα τεχνητής νοημοσύνης της Επιτροπής ΤΝ της Βουλής των Λόρδων του Ηνωμένου Βασιλείου, η Διακήρυξη του Μόντρεαλ για την Υπεύθυνη ΤΝ, οι αρχές του Asilomar, και άλλα. Τα παραπάνω πλαίσια αποδίδουν συνολικά 47 αρχές. Οι πιο βασικές είναι: δικαιοσύνη, αυτονομία, ευεργεσία, μη-κακοήθεια και επεξήγηση. Η δικαιοσύνη περιγράφεται αναλυτικά στην αρχή της συγκεκριμένης ενότητας. Η αυτονομία έχει να κάνει με το δικαίωμα του ατόμου να αποφασίσει για τον εαυτό του. Με την ένταξη της ΤΝ στις ζωές μας έχουμε δώσει κομμάτι της λήψης αποφάσεων σε εκείνη. Οπότε, είναι κρίσιμο να βρεθεί σχετική ισορροπία. Όσον αφορά την ευεργεσία, εννοείται η διατήρηση της αξιοπρέπειας και η προστασία της ευημερίας. Η ανάπτυξη μοντέλων AI πρέπει να δίνει προτεραιότητα στην ανθρώπινη ευημερία. Παράλληλα, η μη-κακοήθεια καλύπτει ένα μεγάλο τμήμα αρχών, όπως η ασφάλεια και η προστασία του προσωπικού απορρήτου. Αξίζει να τονιστεί η σημασία της προστασίας των προσωπικών δεδομένων και της ιδιωτικής ζωής, αντιμετωπίζοντας τους κινδύνους που ελλοχεύει η τεχνολογική καινοτομία. Τέλος, η επεξήγηση συμπληρώνει τις προηγούμενες 4 αρχές, σύμφωνα με τους συγγραφείς, και σηματοδοτεί την κατανόηση των αποτελεσμάτων της χρήσης AI για την κοινωνία και γενικότερα για τον τρόπο που ενεργεί, αλλά και για το ποιος (άνθρωπος, οργανισμός ή υπηρεσία) είναι υπεύθυνος για εκείνη. Οπότε, είναι απαραίτητη η ενσωμάτωση των παραπάνω συνόλων ηθικών αρχών στο σχεδιασμό μοντέλων Τεχνητής Νοημοσύνης, ώστε να δημιουργηθεί μια καλή κοινωνία AI. Με αυτό τον τρόπο, θα αυξηθεί και η εμπιστοσύνη των ανθρώπων για την Τεχνητή Νοημοσύνη και θα κατανοήσουν τα οφέλη της χρήσης της. Για την αξιολόγηση του λογισμικού και την εφαρμογή των παραπάνω είναι απαραίτητη η ύπαρξη εποπτείας από οργανισμούς όπως ο ΟΗΕ και η Ευρωπαϊκή Ένωση, με τη βοήθεια επιστημόνων. Επίσης, πρέπει να δοκιμαστούν σε πρακτικό επίπεδο μηχανισμοί ελέγχου των συνεπειών συγκεκριμένων AI μοντέλων. Τέλος, αναμφίβολα οφείλουν να ακολουθούνται οι κατάλληλες νομικές διαδικασίες σε περίπτωση όπου ανιχνευθεί επιζήμια χρήση Τεχνητής Νοημοσύνης.

Η επόμενη πρόταση που θα εξεταστεί προέρχεται από την ερευνητική εργασία των Mylrea και Robinson (2023), και αφορά ένα πλαίσιο που έχει ως σκοπό τη βελτίωση της εμπιστοσύνης στο σχεδιασμό και τη διαχείριση συστημάτων Τεχνητής Νοημοσύνης. Βασικό εργαλείο του πλαισίου είναι ο φακός εντροπίας (εφευρέθηκε από τον Conant το 1976). Αναμφίβολα, χωρίς την κατάλληλη προσαρμογή του φακού εντροπίας στα χαρακτηριστικά της ΤΝ σε συνάρτηση και με τις προκλήσεις της θα ήταν μία κίνηση υψηλού ρίσκου. Αξίζει σε αυτό το σημείο να αναφερθεί ότι η υψηλή εντροπία ενός συστήματος Τεχνητής Νοημοσύνης είναι αρκετά πιθανό να έχει ως συνέπεια τη μείωση της εμπιστοσύνης και της διαφάνειας. Σύμφωνα με τους συγγραφείς, είναι δυνατή η πρόβλεψη της συμπεριφοράς ενός συστήματος AI μέσω της μεγιστοποίησης της εντροπίας παραγωγής. Άρα, η έξοδος με τη μεγαλύτερη εντροπία φανερώνει την κατάσταση που είναι πιθανότερο να συμβεί. Είναι σημαντικό να κατανοήσουμε τον τρόπο με τον οποίο εμπλουτίζεται με ηθικές αρχές το AI Trust Framework and Maturity Model(AI-TMM), καθώς αρχές όπως η εμπιστοσύνη, η ασφάλεια, το απόρρητο των δεδομένων, οι έλεγχοι πρόσβασης με τη βοήθεια της κρυπτογράφησης και η συναίνεση του χρήστη, είναι απαραίτητες για την ομαλή λειτουργία του συστήματος. Η εμπιστοσύνη για παράδειγμα μειώνεται όταν οι έξοδοι του AI συστήματος είναι αναξιόπιστες. Ένας από τους στόχους της πρότασης είναι η κατανόηση της εισόδου του χρήστη και η προσαρμογή του συστήματος με βάση τις ανάγκες του χρήστη (εξατομίκευση). Επίσης, πρέπει οι διαδικασίες λήψης αποφάσεων να χαρακτηρίζονται από διαφάνεια και επεξηγησιμότητα. Το AI Trust Framework and Maturity Model αξιολογεί το σύστημα ως προς τα ηθικά του χαρακτηριστικά μέσω επαναλαμβανόμενων μετρήσεων. Αναλυτικότερα, η αξιολόγηση της ασφάλειας του συστήματος ελέγχεται σε κάθε στάδιο τόσο του σχεδιασμού όσο και της εφαρμογής. Επιπλέον, η μέτρηση της διαφάνειας πραγματοποιείται μέσω της ερμηνευσιμότητας, δηλαδή της επεξήγησης βάσει κανόνων. Όσον αφορά τα δεδομένα, οι συγγραφείς σημειώνουν ότι στην πρόταση εμπεριέχονται τεχνικές αξιολόγησης της ηθικής προέλευσης των δεδομένων και η δυνατότητα εντοπισμού προκαταλήψεων στα δεδομένα εκπαίδευσης. Οπότε, καταλαβαίνουμε ότι η συλλογή, η αποθήκευση και η διαχείριση δεδομένων ενσωματώνουν όλες τις ηθικές αρχές που αναφέραμε παραπάνω. Αναντίρρητα, τα συστήματα Τεχνητής Νοημοσύνης είναι συνυφασμένα με αρκετές προκλήσεις. Ωστόσο, με τη χρήση του φακού εντροπίας, είναι δυνατή η βελτίωση του τρόπου με τον οποίο αξιολογούμε και διαχειριζόμαστε την εμπιστοσύνη στην Τεχνητή Νοημοσύνη. Συνεπώς, θα μειωθούν οι κίνδυνοι που έχουμε αναφέρει σε αρκετές ενότητες της συγκεκριμένης εργασίας, διασφαλίζοντας ότι τα συστήματα AI λειτουργούν αξιόπιστα.

Στη συνέχεια θα αναφέρουμε την πρόταση των Korobenko κ.ά. (2024). Όπως σημειώνουν στο σχετικό paper, η ανάγκη για τη δημιουργία της συγκεκριμένης πρότασης προέκυψε από την όλο και μεγαλύτερη ανησυχία των ερευνητών για τις επιπτώσεις της Τεχνητής Νοημοσύνης στην ασφάλεια και την ιδιωτική ζωή. Αναλυτικότερα, τονίζουν ότι υπάρχει ανησυχία για τη δράση της AI στα αρχικά στάδια της διαχείρισης των δεδομένων, αναφέροντας τη συλλογή, την αποθήκευση, αλλά και την επεξεργασία προσωπικών δεδομένων. Δεν πρέπει να παραλείψουμε ότι η κακή διαχείριση AI τεχνολογίας θα έχει αρνητικό και επικίνδυνο αντίκτυπο συνολικά στην κοινωνία. Θεωρούν ότι το εννοιολογικό πλαίσιο που προτείνουν θα ενδιαφέρει τους οργανισμούς, τα πανεπιστήμια και τους κυβερνητικούς φορείς. Συνοπτικά, κύρια ιδέα του πλαισίου είναι η ανάπτυξη AI τεχνολογίας που ευθυγραμμίζεται με τις βασικές ηθικές αρχές. Τα σημαντικότερα ζητήματα που αναφέρουν είναι το απόρρητο και η ασφάλεια, σημειώνοντας ότι συνδέονται μεταξύ τους. Οι συγγραφείς επέλεξαν ένα πρωτότυπο τρόπο παρουσίασης του πλαισίου. Ειδικότερα, έχουν διακρίνει 4 διαστάσεις-θεμέλια του πλαισίου: δεδομένα, τεχνολογία, άνθρωποι και διαδικασίες. Αρχικά, όσον αφορά τα δεδομένα, τα χαρακτηρίζουν ως υπόστρωμα της Τεχνητής Νοημοσύνης και τονίζουν τη σημασία τήρησης ορισμένων ηθικών παραγόντων από τους οργανισμούς, όπως η ανωνυμία και η συγκατάθεση, για τη μείωση σχετικών κινδύνων και την αποφυγή καταπάτησης ανθρώπινων δικαιωμάτων. Η τεχνολογία έχει να κάνει με τους μηχανισμούς ασφαλείας που υπάρχουν σε ένα σύστημα Τεχνητής Νοημοσύνης, για να προστατεύει τις βασικές ηθικές αρχές. Οι άνθρωποι αφορούν τον ανθρώπινο παράγοντα (ατομικό και κοινωνικό) που συναντάμε σε συστήματα AI. Η ανθρώπινη διάσταση, σύμφωνα με τους συγγραφείς, κατέχει τις πιο επικίνδυνες επιπτώσεις. Τέλος, η διάσταση της διαδικασίας έχει το ρόλο ελεγκτή της ευθυγράμμισης των συστημάτων AI με τις ηθικές αρχές. Παράδειγμα της συγκεκριμένης διάστασης αποτελεί η ηθική επίβλεψη σε έναν οργανισμό και η στρατηγική αντιμετώπισης περιστατικών. Μέσω της διάκρισης των διαστάσεων, είναι ευκολότερη η ανίχνευση κινδύνων και η απαγόρευση συστημάτων Τεχνητής Νοημοσύνης τα οποία είναι επικίνδυνα για τη δημόσια ασφάλεια και άλλα θεμελιώδη δικαιώματα. Συμπεραίνουμε ότι το προτεινόμενο πλαίσιο έχει τη δυνατότητα καθοδήγησης των οργανισμών σχετικά με τη συστηματική αξιολόγηση και την προστασία του απορρήτου και της ασφάλειας των συστημάτων Τεχνητής Νοημοσύνης που διαθέτουν. Επιπλέον, άποψη των συγγραφέων είναι ότι το συγκεκριμένο πλαίσιο μπορεί να βοηθήσει τους οργανισμούς στη λήψη αποφάσεων ευθυγραμμισμένων με τις ηθικές αρχές.

Το άρθρο των Olorunfemi κ.ά. (2024) καθοδηγεί την ανάπτυξη της AI τεχνολογίας και την ενσωμάτωσή της σε συστήματα Πληροφορικής. Σύμφωνα με τους συγγραφείς, πηγή των ιδεών τους δεν αποτέλεσε αποκλειστικά η επιστήμη της Πληροφορικής, αλλά και τομείς όπως η Ηθική, η Ψυχολογία, το Δίκαιο και η Φιλοσοφία. Βασικές αρχές του πλαισίου είναι η διαφάνεια, η δικαιοσύνη, η λογοδοσία, το απόρρητο, η ασφάλεια, η προστασία της ιδιωτικής ζωής και περιπτώσεις μη εναρμόνισης των συστημάτων με αυτές είναι αρκετά πιθανό να έχουν σοβαρές επιπτώσεις τόσο για το άτομα, όσο και για την κοινωνία. Η απόδοση ευθυνών σε περίπτωση παραβιάσεων ασφαλείας, η κατανόηση του τρόπου με τον οποίο ενεργούν τα συστήματα Τεχνητής Νοημοσύνης και λαμβάνουν αποφάσεις, και η προστασία των προσωπικών δεδομένων είναι κάποια από τα θέματα που οδήγησαν του συγγραφείς στη συγκεκριμένη μελέτη. Επιπλέον, οι συγγραφείς (Olorunfemi, Amoo, Atadoga, Fayayola, Abrahams, Shoetan) τονίζουν ότι το νομικό πλαίσιο που αφορά την Τεχνητή Νοημοσύνη είναι αρκετά ασαφές, λόγω της συνεχούς εξέλιξης του συγκεκριμένου πεδίου. Η υιοθέτηση του εννοιολογικού πλαισίου που παρουσιάζεται έχει ως σκοπό την παροχή στήριξης σε προγραμματιστές, υπεύθυνους χάραξης πολιτικής και οργανισμούς, ώστε να αναπτύσσονται συστήματα Τεχνητής Νοημοσύνης που βασίζονται εκτός από την τελευταία τεχνολογία και σε ηθικές αρχές. Αρχικά, οι προγραμματιστές πρέπει να είναι σίγουροι σχετικά με την υποστήριξη των ηθικών αρχών των αλγορίθμων Τεχνητής Νοημοσύνης που αναπτύσσουν. Οπότε, κατά τη φάση του σχεδιασμού, πρέπει να διασφαλίζεται ότι οι ηθικές αξίες ενσωματώνονται στο σύστημα. Σημαντικός είναι αναντίρρητα και ο ρόλος των υπεύθυνων χάραξης πολιτικής στη συγκεκριμένη φάση, παρέχοντας βοήθεια στους εργαζόμενους των οποίων το έργο είναι καθαρά πρακτικό, μέσω της θέσπισης ρυθμιστικών πλαισίων. Στη συνέχεια, κατά τη φάση της υλοποίησης, είναι απαραίτητο να χαρακτηρίζονται οι χρησιμοποιούμενοι μέθοδοι και εργαλεία από διαφάνεια. Κατά τη φάση της ανάπτυξης, οφείλουν όποια ηθικά ζητήματα εμφανιστούν, να αντιμετωπιστούν με επιτυχία, χωρίς να προκαλέσουν προβλήματα στο σύστημα και στους χρήστες του. Τέλος, κατά τη φάση της συντήρησης, και σκεπτόμενοι το συνεχώς μεταβαλλόμενο νομικό πλαίσιο του τομέα της Τεχνητής Νοημοσύνης, είναι υποχρεωτική η ενημέρωση της ηθικής πολιτικής του οργανισμού, με βάση τις αλλαγές που υφίστανται οι σχετικοί νόμοι/κανονισμοί. Συμπερασματικά, οι οργανισμοί είναι απαραίτητο να δίνουν προτεραιότητα σε ηθικά ζητήματα και να προωθούν την ηθική ανάπτυξη συστημάτων Τεχνητής Νοημοσύνης.

**3. Μεθοδολογία Έρευνας**

Η μεθοδολογία έρευνας που ακολουθήθηκε αποτελείται από πέντε κύρια βήματα:

[1]: Αναγνώριση του κύριου ερωτήματος της έρευνας

[2]: Εντοπισμός σχετικών ερευνών

[3]: Επιλογή ερευνών

[4]: Ταξινόμηση απαραίτητων δεδομένων

[5]: Επεξεργασία, σύνοψη και περιγραφή των αποτελεσμάτων

Στο πρώτο βήμα, ήταν απαραίτητη η συνεργασία και η επικοινωνία των ερευνητών με σκοπό την εύρεση του κύριου ερωτήματος, λαμβάνοντας υπόψη ποια ήταν τα σημαντικότερα χαρακτηριστικά των ερωτήσεων που επιδιώκει να απαντήσει η έρευνα. Τελικά, το ερώτημα ήταν: *Ποια είναι τα πλαίσια αξιολόγησης για αξιόπιστη τεχνητή νοημοσύνη;* Η γενικευμένη μορφή της ερώτησης ήταν απαραίτητη, καθώς γνωρίζαμε πως θα μειωνόταν η πιθανότητα παράλειψης σημαντικών ερευνών από τα επιστρεφόμενα αποτελέσματα. Ο όρος ‘αξιόπιστη’ υποδεικνύει πως τα πλαίσια αξιολόγησης πρέπει να είναι ευθυγραμμισμένα με τις απαιτήσεις του ALTAI. Στα μεταγενέστερα βήματα, διαβάζοντας τη σύνοψη ορισμένων ερευνών, πραγματοποιήθηκαν διάφορες τροποποιήσεις του κύριου ερωτήματος με σκοπό την επιστροφή αποτελεσμάτων που απαντούν το ερώτημα με υψηλότερη ακρίβεια.

Για το δεύτερο βήμα, κρίθηκε απαραίτητο να καθοριστούν η γλώσσα των δημοσιεύσεων προς ανάλυση και το χρονικό διάστημα δημοσίευσής τους. Τελικά, επιλέχθηκε η αγγλική γλώσσα, καθώς ήταν αναμενόμενο πως θα εξασφάλιζε μεγαλύτερη διαθεσιμότητα δημοσιευμένου υλικού. Όσον αφορά το χρονικό διάστημα δημοσίευσης των ερευνών, αποφασίστηκε πως οι έρευνες προς εξέταση θα ήταν εκείνες που δημοσιεύθηκαν μεταξύ του 2017 και 2024, δεδομένου ότι τα πλαίσια αξιολόγησης αξιόπιστης τεχνητής νοημοσύνης αποτελούν ένα σχετικά νέο ερευνητικό πεδίο. Οι ηλεκτρονικές βάσεις δεδομένων στις οποίες έγιναν οι σχετικές αναζητήσεις ήταν οι IEEE Xplore, arXiv, Google Scholar, MDPI, ResearchGate και Springer. Όπως αναφέρθηκε παραπάνω, το κύριο ερευνητικό ερώτημα αναδιατυπώθηκε επανειλημμένα, καθώς κάθε βάση δεδομένων απαιτούσε προσαρμογές στο ερώτημα, ώστε να αποδώσει κατά το δυνατόν ακριβέστερα και πιο συναφή αποτελέσματα. Έτσι, το ερώτημα συμπτύχθηκε σε μικρότερου μήκους εκδοχές, όπως “Ethical AI frameworks” και “AI and ethics”. Μία ακόμη μέθοδος που ακολουθήθηκε για τον εντοπισμό κατάλληλων δημοσιεύσεων ήταν η αναζήτησή τους στις βιβλιογραφικές αναφορές σχετικών ερευνών. Η απόφαση σχετικά με το αν μια δημοσίευση θα κρινόταν υποψήφια για συμπερίληψη στη συγκεκριμένη έρευνα καθορίστηκε βάσει της παρουσίας λέξεων-κλειδιών στους τίτλους, τις περιλήψεις και τα συμπεράσματά της.

Στο τρίτο βήμα, ήταν απαραίτητη η διάκριση μεταξύ των δημοσιεύσεων που θα χρησιμοποιούνταν ως βιβλιογραφικές αναφορές και εκείνων που θα απορρίπτονταν. Αυτό επιτεύχθηκε διαβάζοντας εξ’ ολοκλήρου τις υποψήφιες έρευνες και ελέγχοντας το περιεχόμενό τους. Για να συμπεριληφθεί μια έρευνα στις βιβλιογραφικές αναφορές, έπρεπε να πληροί ένα από τα εξής κριτήρια: να προτείνει ένα πλαίσιο ή πρότυπο αξιολόγησης αξιόπιστης τεχνητής νοημοσύνης, να παρουσιάζει ένα σύνολο σχετικών προδιαγραφών για τέτοιου είδους μοντέλα ή να παρέχει μια ταξινόμηση υφιστάμενων πλαισίων βάσει συγκεκριμένων χαρακτηριστικών τους.

Στο τέταρτο βήμα, οι έρευνες ταξινομήθηκαν σε θεωρητικές και πρακτικές ανάλογα με το περιεχόμενό τους. Αν μια έρευνα προτείνει μια εννοιολογική προσέγγιση αξιολόγησης, τότε κατατάσσεται στις θεωρητικές έρευνες, ενώ εάν προτείνει μία πιο πρακτική προσέγγιση ή εργαλεία αξιολόγησης, τότε κατατάσσεται στις πρακτικές έρευνες. Επιπλέον, σημειώνεται εάν μια έρευνα ικανοποιεί ή όχι κάθε μία από τις επτά απαιτήσεις του ALTAI. Εάν μία απαίτηση ικανοποιείται, τότε σημειώνεται ο βαθμός ικανοποίησης στην κλίμακα 1-10.

Στο τελευταίο βήμα, αναφέρθηκαν οι τάσεις, τα κενά, τα πλεονεκτήματα και οι περιορισμοί των πλαισίων της βιβλιογραφίας, το οποίο βοήθησε στην εξαγωγή συμπερασμάτων.

**4. Αποτελέσματα Έρευνας**

Σε αυτό το κεφάλαιο, παρουσιάζονται τα θεωρητικά και τα πρακτικά πλαίσια αξιολόγησης μέσω δύο πινάκων, ανάλογα με το αν προτείνουν εννοιολογικές ή πρακτικές προσεγγίσεις αξιολόγησης. Για κάθε ένα από αυτά, σημειώνεται εάν ικανοποιούν ή όχι κάθε μία από τις απαιτήσεις του ALTAI. Εάν μία απαίτηση ικανοποιείται, τότε σημειώνεται ο βαθμός ικανοποίησής της. Έπειτα, πραγματοποιείται μία ανάλυση των αποτελεσμάτων στους πίνακες 3 και 4. Ακολουθεί ο πίνακας των θεωρητικών πλαισίων:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Πλαίσιο/Απαίτηση | Ανθρώπινη παρέμβαση και εποπτεία | Τεχνική στιβαρότητα και ασφάλεια | Ιδιωτική ζωή και διακυβέρνηση δεδομένων | Διαφάνεια | Πολυμορφία και δικαιοσύνη | Κοινωνική και περιβαλλοντική ευημερία | Λογοδοσία |
| Floridi κ.ά (2018) | ΝΑΙ (8) | ΝΑΙ (8) | ΝΑΙ (9) | ΝΑΙ (9) | ΝΑΙ (8) | ΝΑΙ (8) | ΝΑΙ (8) |
| Korobenko κ.ά. (2024) | ΟΧΙ | ΝΑΙ (8) | ΝΑΙ (9) | ΟΧΙ | ΝΑΙ (7) | ΟΧΙ | ΝΑΙ (7) |
| Pastor-Dark κ.ά. (2022) | ΝΑΙ (8) | ΝΑΙ (8) | ΝΑΙ (8) | ΝΑΙ (8) | ΝΑΙ (9) | ΝΑΙ (10) | ΝΑΙ (8) |
| Nasir κ.ά. (2024) | ΝΑΙ (8) | ΟΧΙ | ΝΑΙ (7) | ΝΑΙ (8) | ΝΑΙ (7) | ΟΧΙ | ΟΧΙ |
| Milossi κ.ά. (2021) | ΝΑΙ (9) | ΟΧΙ | ΝΑΙ (8) | ΝΑΙ (9) | ΝΑΙ (8) | ΟΧΙ | ΝΑΙ (7) |
| Prem (2023) | ΝΑΙ (9) | ΝΑΙ (8) | ΝΑΙ (9) | ΝΑΙ (9) | ΝΑΙ (8) | ΝΑΙ (7) | ΝΑΙ (8) |
| Xia κ.ά (2024) | ΝΑΙ (7) | ΝΑΙ (8) | ΟΧΙ | ΝΑΙ (7) | ΟΧΙ | ΟΧΙ | ΝΑΙ (6) |
| McCormack & Bendechache (2024) | ΝΑΙ (8) | ΝΑΙ (7) | ΟΧΙ | ΝΑΙ (9) | ΝΑΙ (8) | ΟΧΙ | ΝΑΙ (7) |

Πίνακας 1

Ακολουθεί ο πίνακας των πρακτικών πλαισίων:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Πλαίσιο/Απαίτηση | Ανθρώπινη παρέμβαση και εποπτεία | Τεχνική στιβαρότητα και ασφάλεια | Ιδιωτική ζωή και διακυβέρνηση δεδομένων | Διαφάνεια | Πολυμορφία και δικαιοσύνη | Κοινωνική και περιβαλλοντική ευημερία | Λογοδοσία |
| Mylrea & Robinson (2023) | ΝΑΙ (8) | ΝΑΙ (8) | ΝΑΙ (9) | ΝΑΙ (8) | ΝΑΙ (8) | ΝΑΙ (8) | ΝΑΙ (8) |
| Olorunfemi κ.ά. (2024) | ΝΑΙ (7) | ΝΑΙ (8) | ΝΑΙ (8) | ΝΑΙ (7) | ΝΑΙ (7) | ΝΑΙ (7) | ΝΑΙ (7) |
| O'Sullivan κ.ά. (2018) | ΝΑΙ (7) | ΝΑΙ (8) | ΝΑΙ (7) | ΝΑΙ (8) | ΟΧΙ | ΟΧΙ | ΝΑΙ (8) |
| Peters κ.ά. (2020) | ΝΑΙ (8) | ΝΑΙ (8) | ΝΑΙ (8) | ΝΑΙ (9) | ΝΑΙ (8) | ΝΑΙ (8) | ΝΑΙ (8) |
| Oveisi κ.ά. (2024) | ΝΑΙ (8) | ΝΑΙ (9) | ΝΑΙ (9) | ΝΑΙ (8) | ΝΑΙ (7) | ΝΑΙ (7) | ΝΑΙ (9) |

Πίνακας 2

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Μέσος Όρος** | | | | | | |
| Ανθρώπινη παρέμβαση και εποπτεία | Τεχνική στιβαρότητα και ασφάλεια | Ιδιωτική ζωή και διακυβέρνηση δεδομένων | Διαφάνεια | Πολυμορφία και δικαιοσύνη | Κοινωνική και περιβαλλοντική ευημερία | Λογοδοσία |
| 7.3 | 6.7 | 7.0 | 7.6 | 6.5 | 4.2 | 7.0 |

Στους πίνακες 3 και 4 υπολογίζεται ο μέσος όρος των βαθμολογιών της κάθε απαίτησης, καθώς και το πλήθος των πλαισίων της βιβλιογραφίας που τις ικανοποιούν αντίστοιχα.

Πίνακας 3

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Πόσα πλαίσια ικανοποιούν την κάθε απαίτηση;** | | | | | | |
| Ανθρώπινη παρέμβαση και εποπτεία | Τεχνική στιβαρότητα και ασφάλεια | Ιδιωτική ζωή και διακυβέρνηση δεδομένων | Διαφάνεια | Πολυμορφία και δικαιοσύνη | Κοινωνική και περιβαλλοντική ευημερία | Λογοδοσία |
| 12/13 | 11/13 | 11/13 | 12/13 | 11/13 | 7/13 | 12/13 |

Πίνακας 4

**5. Συζήτηση Αποτελεσμάτων**

Αρχικά, οι πίνακες 1 και 2 δείχνουν πως τα θεωρητικά πλαίσια είναι περισσότερα από τα πρακτικά. Επιπλέον, με βάση τα αποτελέσματα που συγκεντρώθηκαν από τους πίνακες 3 και 4, γίνεται φανερό πως το σύνολο των πλαισίων της βιβλιογραφίας έχει την τάση να εστιάζει στην ικανοποίηση των απαιτήσεων της διαφάνειας, της ανθρώπινης παρέμβασης και εποπτείας και της λογοδοσίας. Οι απαιτήσεις αυτές ικανοποιούνται από 12 πλαίσια και έχουν τους υψηλότερους μέσους όρους βαθμολογιών: 7.6, 7.3 και 7.0 αντίστοιχα.

Αντίθετα, φαίνεται πως υπάρχει έλλειψη κυρίως όσον αφορά την απαίτηση της κοινωνικής και περιβαλλοντικής ευημερίας, και κατόπιν, σε μικρότερο βαθμό, τις απαιτήσεις της τεχνικής στιβαρότητας και ασφάλειας και της πολυμορφίας και δικαιοσύνης. Η κοινωνική ευημερία και ασφάλεια διαθέτει μέσο όρο 4.2 και ικανοποιείται από μόλις 7 πλαίσια της βιβλιογραφίας. Οι απαιτήσεις της τεχνικής στιβαρότητας και ασφάλειας και της πολυμορφίας και δικαιοσύνης διαθέτουν μέσο όρο 6.7 και 6.5 αντίστοιχα και ικανοποιούνται από 11 πλαίσια η κάθε μία.

Τα πλεονεκτήματα των πλαισίων που περιέχονται στη βιβλιογραφία είναι τα εξής:

* Παροχή σαφών κριτηρίων και οδηγιών: Όλα τα πλαίσια προσφέρουν ένα σύνολο σαφών κριτηρίων για την αξιολόγηση των AI συστημάτων, ενώ τα πιο πρακτικά πλαίσια παρέχουν οδηγίες για την αξιόπιστη ανάπτυξή τους.
* Είναι θεμελιωμένα στις ανθρωποκεντρικές αρχές του ALTAI: Η τήρηση των επτά απαιτήσεων του ALTAI συνεπάγεται την ανάπτυξη αξιόπιστων συστημάτων ΤΝ, που έχουν στο κέντρο τους τον άνθρωπο και λειτουργούν με στόχο την ικανοποίηση των αναγκών του ενώ παράλληλα διασφαλίζουν την προστασία του, καθώς και την προστασία το περιβάλλοντος και του κοινωνικού συνόλου.
* Παροχή μεθοδολογιών, εργαλείων και μετρικών που αναγνωρίζονται από την επιστημονική κοινότητα: Τα πλαίσια χρησιμοποιούν εργαλεία και μετρικές αξιολόγησης που έχουν αναπτυχθεί, χρησιμοποιηθεί και αξιολογηθεί από εμπειρογνώμονες.

Οι περιορισμοί των πλαισίων που περιέχονται στη βιβλιογραφία είναι οι εξής:

* Η πολυπλοκότητα εφαρμογής τους είναι υψηλή, αφού για μερικά πλαίσια απαιτείται η κατοχή εξειδικευμένων γνώσεων και υπολογιστικών πόρων. Επιπλέον, η υψηλή πολυπλοκότητα καθιστά τα πλαίσια δυσνόητα για μερικούς χρήστες.
* Ορισμένα πλαίσια παρέχουν αποκλειστικά θεωρητικές προσεγγίσεις, χωρίς να αναφέρουν πρακτικά εργαλεία αξιολόγησης. Αυτό δυσκολεύει τη διαδικασία αξιολόγησης στην πράξη.
* Υπάρχει έλλειψη κοινωνικοπολιτικών και κυρίως περιβαλλοντικών κριτηρίων αξιολόγησης και τεχνικών ανάπτυξης.

**6. Συμπεράσματα**

Τα ευρήματα της έρευνας καθιστούν ευκολότερο τον προσδιορισμό της κατεύθυνσης προς την οποία πρέπει να κινηθούν σχετικές μελλοντικές έρευνες. Συνοψίζοντας, κάθε πλαίσιο εστιάζει στη σημασία της ανάπτυξης συστημάτων ΤΝ με κέντρο τον άνθρωπο. Παρόλα αυτά, τα περισσότερα πλαίσια παρέχουν θεωρητικές προσεγγίσεις αξιολόγησης και ανάπτυξης, το οποίο υποδεικνύει πως η έρευνα στον τομέα είναι ακόμη αρκετά περιορισμένη. Επίσης, παρότι το ALTAI διευκρινίζει πως η τήρηση και των επτά απαιτήσεων είναι απαραίτητη, πολλά πλαίσια τείνουν να αγνοούν την απαίτηση της κοινωνικής και περιβαλλοντικής ευημερίας και να εστιάζουν σε μικρότερο βαθμό στις απαιτήσεις της τεχνικής στιβαρότητας και ασφάλειας και της πολυμορφίας και δικαιοσύνης. Από την άλλη, δίνεται ιδιαίτερη έμφαση στη διαφάνεια, στην ανθρώπινη παρέμβαση και εποπτεία και στη λογοδοσία. Είναι, λοιπόν, ασφαλές το συμπέρασμα πως στο μέλλον είναι αναγκαίο να πραγματοποιηθούν έρευνες που εστιάζουν κυρίως σε πρακτικές προσεγγίσεις αξιολόγησης και στην ικανοποίηση όλων των απαιτήσεων του ALTAI, εάν επιθυμούμε την ανάπτυξη πραγματικά αξιόπιστων συστημάτων ΤΝ που δεν αποτελούν κίνδυνο για τον άνθρωπο, την κοινωνία και το περιβάλλον.

**Βιβλιογραφία**

[1] European Comission. (2019). *HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE SET UP BY THE EUROPEAN COMMISSION ETHICS GUIDELINES FOR TRUSTWORTHY AI*.

[2] European Commission. (2020, July 17). *Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment | Shaping Europe’s digital future*. Digital-Strategy.ec.europa.eu. <https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment>

[3] Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*(4), 689–707. <https://doi.org/10.1007/s11023-018-9482-5>

[4] Korobenko, D., Nikiforova, A., & Sharma, R. (2024). Towards a Privacy and Security-Aware Framework for Ethical AI: Guiding the Development and Assessment of AI Systems. *Proceedings of the 25th Annual International Conference on Digital Government Research*, 740–753. <https://doi.org/10.1145/3657054.3657141>

[5] Mccormack, L., & Bendechache, M. (2024). *Ethical AI Governance: Methods for Evaluating Trustworthy AI*.

[6] Milossi, M., Alexandropoulou-Egyptiadou, E., & Psannis, K. E. (2021). AI Ethics: Algorithmic Determinism or Self-Determination? The GPDR Approach. *IEEE Access*, *9*, 58455–58466. <https://doi.org/10.1109/access.2021.3072782>

[7] Mylrea, M., & Robinson, N. (2023). Artificial Intelligence (AI) Trust Framework and Maturity Model: Applying an Entropy Lens to Improve Security, Privacy, and Ethical AI. *Entropy*, *25*(10), 1429. <https://doi.org/10.3390/e25101429>

[8] Nasir, S., Khan, R. A., & Bai, S. (2024). Ethical Framework for Harnessing the Power of AI in Healthcare and Beyond. *IEEE Access*, *12*, 31014–31035. <https://doi.org/10.1109/access.2024.3369912>

[9] O’Sullivan, S., Nevejans, N., Allen, C., Blyth, A., Leonard, S., Pagallo, U., Holzinger, K., Holzinger, A., Sajid, M. I., & Ashrafian, H. (2019). Legal, regulatory, and ethical frameworks for development of standards in artificial intelligence (AI) and autonomous robotic surgery. *The International Journal of Medical Robotics and Computer Assisted Surgery*, *15*(1). <https://doi.org/10.1002/rcs.1968>

[10] Oluwabukunmi Latifat Olorunfemi, Olukunle Oladipupo Amoo, Akoh Atadoga, Oluwatoyin Ajoke Fayayola, Temitayo Oluwaseun Abrahams, & Philip Olaseni Shoetan. (2024). TOWARDS A CONCEPTUAL FRAMEWORK FOR ETHICAL AI DEVELOPMENT IN IT SYSTEMS. *Computer Science & IT Research Journal*, *5*(3), 616–627. <https://doi.org/10.51594/csitrj.v5i3.910>

[11] Pastor-Escuredo, D., Treleaven, P., & Vinuesa, R. (2022). An Ethical Framework for Artificial Intelligence and Sustainable Cities. *AI*, *3*(4), 961–974. <https://doi.org/10.3390/ai3040057>

[12] Peters, D., Vold, K., Robinson, D., & Calvo, R. A. (2020). Responsible AI—Two Frameworks for Ethical Design Practice. *IEEE Transactions on Technology and Society*, *1*(1), 34–47. https://doi.org/10.1109/tts.2020.2974991

Prem, E. (2023). From ethical AI frameworks to tools: a review of approaches. *AI and Ethics*, *3*(3), 699–716. <https://doi.org/10.1007/s43681-023-00258-9>

[13] Xia, B., Lu, Q., Zhu, L., & Xing, Z. (2024). An AI System Evaluation Framework for Advancing AI Safety: Terminology, Taxonomy, Lifecycle Mapping. *Proceedings of the 1st ACM International Conference on AI-Powered Software*, 74–78. https://doi.org/10.1145/3664646.3664766