AEQUITAS

D2.1 Έκθεση για τις προκαταλήψεις

Co-funded by
the European Union

Co-funded by the European Union

AEQUITAS

Περιεχόμενα

Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης. Οι απόψεις και οι γνώμες που διατυπώνονται εκφράζουν αποκλειστικά τις απόψεις των συντακτών και δεν αντιπροσωπεύουν κατ'ανάγκη τις απόψεις της Ευρωπαϊκής Ένωσης ή του Ευρωπαϊκού Εκτελεστικού Οργανισμού Εκπαίδευσης και Πολιτισμού (EACEA). Η Ευρωπαϊκή Ένωση και ο EACEA δεν μπορούν να θεωρηθούν υπεύθυνοι για τις εκφραζόμενες απόψεις. Κωδικός έργου: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης. Οι απόψεις και οι γνώμες που διατυπώνονται εκφράζουν αποκλειστικά τις απόψεις των συντακτών και δεν αντιπροσωπεύουν κατ'ανάγκη τις απόψεις της Ευρωπαϊκής Ένωσης ή του Ευρωπαϊκού Εκτελεστικού Οργανισμού Εκπαίδευσης και Πολιτισμού (EACEA). Η Ευρωπαϊκή Ένωση και ο EACEA δεν μπορούν να θεωρηθούν υπεύθυνοι για τις εκφραζόμενες απόψεις. Κωδικός έργου: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Με τη συγχρηματοδότηση της Ευρωπαϊκής Ένωσης

AEQUITAS

Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης. Οι απόψεις και οι γνώμες που διατυπώνονται εκφράζουν αποκλειστικά τις απόψεις των συντακτών και δεν αντιπροσωπεύουν κατ'ανάγκη τις απόψεις της Ευρωπαϊκής Ένωσης ή του Ευρωπαϊκού Εκτελεστικού Οργανισμού Εκπαίδευσης και Πολιτισμού (EACEA). Η Ευρωπαϊκή Ένωση και ο EACEA δεν μπορούν να θεωρηθούν υπεύθυνοι για τις εκφραζόμενες απόψεις. Κωδικός έργου: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Εισαγωγή

Μέρος του έργου AEQUITAS είναι η δημιουργία μιας βάσης δεδομένων σχετικά με τις έμφυλες και φυλετικές προκαταλήψεις στις ιατρικές εφαρμογές της Τεχνητής Νοημοσύνης (AI), με ιδιαίτερη έμφαση σε τρεις ασθένειες: καρδιαγγειακές παθήσεις, διαβήτη και κατάθλιψη.

Για να ολοκληρωθεί αυτή η δραστηριότητα, οι εταίροι της κοινοπραξίας πρέπει πρώτα να συλλέξουν διάφορες πηγές σχετικά με τις προκαταλήψεις που αναφέρονται παραπάνω. Το Πανεπιστημιακό Νοσοκομείο της Κολωνίας (Universitätsklinikum Köln, UKK), ως επικεφαλής της δραστηριότητας και εμπειρογνώμονας στον τομέα, οργάνωσε τη δραστηριότητα συλλογής πληροφοριών και παρείχε το πρότυπο χαρτογράφησης, το οποίο χρησιμοποίησαν οι εταίροι για να χαρτογραφήσουν τις πηγές, διασφαλίζοντας ότι οι σχετικές πληροφορίες θα μπορούσαν να μεταφερθούν εύκολα στη βάση δεδομένων.

Η παρούσα έκθεση παρουσιάζει τα θεωρητικά και επιστημονικά θεμέλια που καθοδήγησαν την επιλογή της δραστηριότητας συλλογής και του προτύπου χαρτογράφησης, συνοδευόμενα από μελέτες περιπτώσεων που παρουσιάζουν τα διάφορα είδη μεροληψίας, τις επιπτώσεις των πολιτικών που έχουν οι μεροληψίες που προκαλεί η βιοϊατρική τεχνητή νοημοσύνη στα δικαιώματα που προστατεύονται από τον Χάρτη των Θεμελιωδών Δικαιωμάτων της Ευρωπαϊκής Ένωσης, μια περιγραφή της δραστηριότητας συλλογής δεδομένων, το πρότυπο χαρτογράφησης, έναν κατάλογο πηγών που συλλέχθηκαν από τους εταίρους της AEQUITAS και άλλο υποστηρικτικό υλικό. Το υπόλοιπο της έκθεσης έχει την ακόλουθη δομή:

Αρχικά, παρουσιάζουμε το θεωρητικό υπόβαθρο της εργασίας μας σε μια εισαγωγή σχετικά με τις εφαρμογές της τεχνητής νοημοσύνης στην ιατρική και την έννοια της μεροληψίας στα υπολογιστικά συστήματα και την ιατρική. Ξεκινάμε εστιάζοντας στην ιατρική, παρουσιάζοντας πρώτα πώς εκδηλώνεται η μεροληψία λόγω φυλής και

Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης. Οι απόψεις και οι γνώμες που διατυπώνονται εκφράζουν αποκλειστικά τις απόψεις των συντακτών και δεν αντιπροσωπεύουν κατ'ανάγκη τις απόψεις της Ευρωπαϊκής Ένωσης ή του Ευρωπαϊκού Εκτελεστικού Οργανισμού Εκπαίδευσης και Πολιτισμού (EACEA). Η Ευρωπαϊκή Ένωση και ο EACEA δεν μπορούν να θεωρηθούν υπεύθυνοι για τις εκφραζόμενες απόψεις. Κωδικός έργου: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

φύλου στην ιατρική περίθαλψη και, δεύτερον, πώς αντιμετωπίζονται τα ηθικά ζητήματα που προκύπτουν στην ιατρική πρακτική και τη βιοϊατρική έρευνα από τη βιοηθική, προσφέροντας μια σύντομη εισαγωγή στις τέσσερις αρχές της βιοηθικής (αυτονομία, μη βλάβη, ευεργεσία και δικαιοσύνη).

Στη συνέχεια, στρέφουμε την προσοχή μας στον τομέα της τεχνητής νοημοσύνης, παρουσιάζοντας τους τύπους μεροληψίας που μπορούν να παρατηρηθούν στα συστήματα τεχνητής νοημοσύνης, όπως εκδηλώνονται στη διαδικασία της μηχανικής μάθησης και της τεχνητής νοημοσύνης (ML/AI). Κάθε τύπος μεροληψίας συνοδεύεται από παραδείγματα και μια μελέτη περίπτωσης σχετικά με τις μεροληψίες λόγω φύλου και φυλής και τον αντίκτυπό τους σε κοινωνικό επίπεδο όσον αφορά τις τρεις ασθένειες στις οποίες επικεντρώνεται το έργο AEQUITAS (καρδιαγγειακές παθήσεις, διαβήτης και κατάθλιψη), που προέρχονται από πηγές που συλλέχθηκαν από τους εταίρους του AEQUITAS μετά την ολοκλήρωση του T2.2. Όταν αυτό δεν ήταν δυνατό επειδή το συλλεγμένο υλικό δεν κατέδειχνε σαφώς τον συγκεκριμένο τύπο μεροληψίας της τεχνητής νοημοσύνης που εξετάζονταν, παρουσιάστηκε μια εναλλακτική περίπτωση από έναν άλλο ιατρικό τομέα που ήταν εύκολα γενικεύσιμη στις ασθένειες-στόχους του AEQUITAS. Οι περιγραφές των περιπτωσιολογικών μελετών, μαζί με τις συλλεγμένες πηγές, βασίζονται σε πρόσθετους επιστημονικούς πόρους, όπως απαιτείται, για την υποστήριξή τους.

Τέλος, στην ενότητα Επιπτώσεις στην πολιτική, καταδεικνύεται ο τρόπος με τον οποίο οι διάφοροι τύποι μεροληψιών της τεχνητής νοημοσύνης επηρεάζουν τα θεμελιώδη δικαιώματα που προστατεύονται από τον Χάρτη της ΕΕ, ιδίως τις αρχές της ανθρώπινης αξιοπρέπειας, της ισότητας ενώπιον του νόμου, της μη διάκρισης, καθώς και το δικαίωμα στην ακεραιότητα του ατόμου, το δικαίωμα στην υγειονομική περίθαλψη, την προστασία των δεδομένων και το δικαίωμα σε αποτελεσματική έννομη προστασία, και καταλήγει στις διασφαλίσεις που μπορούν να τεθούν σε

Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης. Οι απόψεις και οι γνώμες που διατυπώνονται εκφράζουν αποκλειστικά τις απόψεις των συντακτών και δεν αντιπροσωπεύουν κατ'ανάγκη τις απόψεις της Ευρωπαϊκής Ένωσης ή του Ευρωπαϊκού Εκτελεστικού Οργανισμού Εκπαίδευσης και Πολιτισμού (EACEA). Η Ευρωπαϊκή Ένωση και ο EACEA δεν μπορούν να θεωρηθούν υπεύθυνοι για τις εκφραζόμενες απόψεις. Κωδικός έργου: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

εφαρμογή στις αξιολογήσεις συμμόρφωσης, στην παρακολούθηση μετά τη διάθεση στην αγορά και στις δημόσιες συμβάσεις.

Η έκθεση ολοκληρώνεται με τις Αναφορές και τα ακόλουθα Παραρτήματα:

Παράρτημα 1: Μέθοδος συλλογής και χαρτογράφησης πηγών, το οποίο περιέχει το πρότυπο χαρτογράφησης και περιγράφει τη διαδικασία συλλογής, χαρτογράφησης και αξιολόγησης πληροφοριών που πραγματοποιήθηκε κατά τη διάρκεια των εργασιών T2.1 και T2.2.

Παράρτημα 2: Περιέχει υποστηρικτικό υλικό για τις εργασίες T2.1 και T2.2, δηλαδή διαφάνειες από συναντήσεις εταίρων που περιγράφουν τη διαδικασία, που παρουσιάστηκαν από την UKK.

Παράρτημα 3: Κατάλογος πηγών που συγκέντρωσαν οι εταίροι του AEQUITAS.

Ιατρικές εφαρμογές τεχνητής νοημοσύνης

Η άνοδος των εφαρμογών τεχνητής νοημοσύνης τα τελευταία χρόνια έχει επηρεάσει σημαντικά την ιατρική, συμπεριλαμβανομένης της ψηφιακής συλλογής δεδομένων, της μηχανικής μάθησης και της υπολογιστικής υποδομής (Yu et al., 2018). Ιδιαίτερα η εισαγωγή αλγορίθμων βαθιάς μάθησης σε τομείς όπως η υπολογιστική όραση και η επεξεργασία φυσικής γλώσσας έχει φέρει επανάσταση στις υπολογιστικές εφαρμογές στην ακτινολογία, την παθολογία, την καρδιολογία, τη διαβητολογία, την ψυχιατρική, την ογκολογία κ.λπ. (Esteva et al., 2019; Koteluk et al., 2021; Rajpurkar et al., 2022; Gou et al., 2024). Ο Παγκόσμιος Οργανισμός Υγείας (ΠΟΥ) απαριθμεί τους ακόλουθους τομείς εφαρμογής των συστημάτων τεχνητής νοημοσύνης στην υγειονομική περίθαλψη: διάγνωση και διάγνωση βάσει πρόβλεψης, κλινική περίθαλψη, έρευνα και ανάπτυξη φαρμάκων, διαχείριση και σχεδιασμός συστημάτων

Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης. Οι απόψεις και οι γνώμες που διατυπώνονται εκφράζουν αποκλειστικά τις απόψεις των συντακτών και δεν αντιπροσωπεύουν κατ'ανάγκη τις απόψεις της Ευρωπαϊκής Ένωσης ή του Ευρωπαϊκού Εκτελεστικού Οργανισμού Εκπαίδευσης και Πολιτισμού (EACEA). Η Ευρωπαϊκή Ένωση και ο EACEA δεν μπορούν να θεωρηθούν υπεύθυνοι για τις εκφραζόμενες απόψεις. Κωδικός έργου: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

υγείας, δημόσια υγεία και επιτήρηση της δημόσιας υγείας, προαγωγή της υγείας, πρόληψη ασθενειών, επιτήρηση βάσει πρόβλεψης, ετοιμότητα για καταστάσεις έκτακτης ανάγκης και αντίδραση σε επιδημίες (Παγκόσμιος Οργανισμός Υγείας, 2021).

Ωστόσο, η έλευση των εφαρμογών τεχνητής νοημοσύνης στην ιατρική συνοδεύεται από μια σειρά προκλήσεων, όπως προκλήσεις υλοποίησης, συμπεριλαμβανομένων περιορισμών όσον αφορά την αξιοπιστία των μοντέλων και τα δεδομένα, ζητήματα λογοδοσίας, τα οποία περιλαμβάνουν προκλήσεις σε επίπεδο κανονιστικών ρυθμίσεων και σωστή απόδοση ευθυνών, καθώς και διασφάλιση της δικαιοσύνης μέσω της ηθικής χρήσης των δεδομένων, της δίκαιης κατανομής των οφελών και της ανίχνευσης και μετριασμού των προκαταλήψεων (Rajpurkar et al., 2022).

Το έργο AEQUITAS επικεντρώνεται σε περιπτώσεις μεροληψίας λόγω φύλου και φυλής σε καρδιαγγειακές παθήσεις, διαβήτη και κατάθλιψη. Οι ιατρικές εφαρμογές τεχνητής νοημοσύνης υποστηρίζουν την καρδιαγγειακή περίθαλψη μέσω της υποστήριξης κλινικών αποφάσεων, της τηλεϊατρικής, της εκτίμησης κινδύνου, της εξατομικευμένης θεραπείας, της προγνωστικής ανάλυσης και της απομακρυσμένης παρακολούθησης (Bernstein et al., 2025; Naskar et al., 2025), βελτιώνουν τη διαχείριση του διαβήτη (συμπεριλαμβανομένης της παρακολούθησης των ασθενών και της αυτοδιαχείρισης), τη διάγνωση, τη θεραπεία και την πρόληψη (Contreras & Vehi, 2018; Khalifa & Albadawy, 2024; Naskar et al., 2025; Sheng et al., 2024). Όσον αφορά την κατάθλιψη, συμμετέχουν στον έλεγχο, τη διάγνωση και τη θεραπεία (Alhuwaydi, 2024) με ιδιαίτερη έμφαση στην ανίχνευση και τον έλεγχο με τη χρήση μεγάλων γλωσσικών μοντέλων (LLM) (Cao et al., 2025; Kumari et al., 2025; Mao et al., 2023; Wang et al., 2025). Σε όλους τους παραπάνω τομείς, υπάρχουν προκλήσεις μεροληψίας, για παράδειγμα, όσον αφορά τις καρδιαγγειακές παθήσεις, τον διαβήτη και την κατάθλιψη, βλ. (van Assen et al. 2024), (Cronjé et al. 2023), (Dang et al. 2024), αντίστοιχα.

Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης. Οι απόψεις και οι γνώμες που διατυπώνονται εκφράζουν αποκλειστικά τις απόψεις των συντακτών και δεν αντιπροσωπεύουν κατ'ανάγκη τις απόψεις της Ευρωπαϊκής Ένωσης ή του Ευρωπαϊκού Εκτελεστικού Οργανισμού Εκπαίδευσης και Πολιτισμού (EACEA). Η Ευρωπαϊκή Ένωση και ο EACEA δεν μπορούν να θεωρηθούν υπεύθυνοι για τις εκφραζόμενες απόψεις. Κωδικός έργου: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Προκλήσεις όπως η μεροληψία, είτε στην ιατρική είτε στην τεχνητή νοημοσύνη, αντιμετωπίζονται μέσω ενός συνδυασμού βιοηθικής και ηθικής τεχνητής νοημοσύνης. Στην επόμενη ενότητα, παρέχουμε μια σύντομη επισκόπηση αυτών των δύο τύπων εφαρμοσμένων τομέων ηθικής που χρησίμευσαν ως θεωρητική και επιστημονική βάση για την ανάπτυξη του προτύπου χαρτογράφησης μεροληψίας.

Ηθική και μεροληψία στην ιατρική και την τεχνητή νοημοσύνη

Βιοηθική και μεροληψία στην ιατρική

Η μεροληψία στην ιατρική είναι καλά τεκμηριωμένη: βλ. για παράδειγμα, (Hammond et al. 2021) σχετικά με τη γνωστική μεροληψία, η οποία συνίσταται σε συστηματικά σφάλματα στη σκέψη λόγω των περιορισμών της ανθρώπινης επεξεργασίας ή ακατάλληλων νοητικών μοντέλων, και (FitzGerald και Hurst 2017) για την έμμεση μεροληψία που περιλαμβάνει συσχετίσεις εκτός της συνειδητής αντίληψης που οδηγούν σε αρνητική αξιολόγηση ενός ατόμου με βάση άσχετα χαρακτηριστικά, όπως η φυλή ή το φύλο.

Η φυλετική μεροληψία στην ιατρική έχει μελετηθεί καλά στην περίπτωση των ΗΠΑ, για παράδειγμα, όπου έχει τεκμηριωθεί ότι οι Αφροαμερικανοί, καθώς και τα μέλη άλλων μειονοτικών ομάδων, υποβάλλονται σε λιγότερες επεμβάσεις και λαμβάνουν ιατρική περίθαλψη χαμηλότερης ποιότητας, καθώς λαμβάνουν λιγότερο επιθετική θεραπεία, υποβάλλονται σε χειρουργικές επεμβάσεις σε μικρότερο ποσοστό και παραπέμπονται σε ειδικούς σε μικρότερο βαθμό σε σύγκριση με τα λευκά άτομα (Bowser, 2001; Williams & Wyatt, 2015).

Η έμφυλη μεροληψία μπορεί να αποδοθεί στην έλλειψη ευαισθητοποίησης σε θέματα φύλου και στις στερεοτυπικές προκαταλήψεις για τους άνδρες και τις γυναίκες (Hamberg, 2008), σε συνδυασμό με μια γενικευμένη έλλειψη γνώσης σχετικά με τη λειτουργία του γυναικείου σώματος και τις βιολογικές του διαφορές

Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης. Οι απόψεις και οι γνώμες που διατυπώνονται εκφράζουν αποκλειστικά τις απόψεις των συντακτών και δεν αντιπροσωπεύουν κατ'ανάγκη τις απόψεις της Ευρωπαϊκής Ένωσης ή του Ευρωπαϊκού Εκτελεστικού Οργανισμού Εκπαίδευσης και Πολιτισμού (EACEA). Η Ευρωπαϊκή Ένωση και ο EACEA δεν μπορούν να θεωρηθούν υπεύθυνοι για τις εκφραζόμενες απόψεις. Κωδικός έργου: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

από το ανδρικό σώμα. Για παράδειγμα, οι γυναίκες ηλικίας 50 ετών και άνω που βρίσκονται σε κρίσιμη κατάσταση είχαν λιγότερες πιθανότητες από τους άνδρες να εισαχθούν σε μονάδα εντατικής θεραπείας (ΜΕΘ) (Bierman, 2007), και ακόμη και τα αρσενικά μοντέλα ποντικών είναι συνολικά πιο αντιπροσωπευτικά από τα θηλυκά μοντέλα στη βασική, προκλινική και χειρουργική βιοϊατρική έρευνα (Yoon et al., 2014).

Είναι επίσης σημαντικό να σημειωθεί ότι τα άτομα LGBT+ υφίστανται διακρίσεις όσον αφορά την πρόσβαση στην υγειονομική περίθαλψη και υπόκεινται σε στερεότυπα που δεν επηρεάζουν τον ετεροφυλόφιλο πληθυσμό. Αυτοί οι κοινωνικοί και πολιτισμικοί παράγοντες διαιωνίζουν τις διακρίσεις και έχουν αντίκτυπο στην υγεία. Για παράδειγμα, μια μελέτη στις ΗΠΑ, βασισμένη σε δεδομένα από την Εθνική Έρευνα Υγείας (NHIS) 2013-14, διαπίστωσε ότι οι ενήλικες LGB ανέφεραν υψηλότερα επίπεδα κακής υγείας, λειτουργικών περιορισμών, σοβαρής ψυχολογικής δυσφορίας και δυσκολιών στην πρόσβαση στην υγειονομική περίθαλψη σε σύγκριση με τους ετεροφυλόφιλους ομολόγους τους. Αυτές οι ανισότητες οφείλονται στο άγχος των μειονοτικών ομάδων και στην πολύπλευρη κοινωνική περιθωριοποίηση (Liu et al., 2023).

Από την άλλη πλευρά, η ιατρική ως επιστήμη υπόκειται σε υψηλά ηθικά πρότυπα από την αρχαιότητα έως σήμερα (Baker & McCullough, 2008). Για αιώνες, υπάρχει η κοινωνική προσδοκία ότι ένας γιατρός θα ακολουθεί τους ηθικούς κανόνες επαγγελματικής ευθύνης που καθορίζονται από τα πρότυπα του επαγγέλματός του, όπως εκφράζονται μέσω επαγγελματικών κανόνων που κυμαίνονται από τον Όρκο του Ιπποκράτη από το 400 π.Χ. (Miles, 2005) έως τις Διακηρύξεις της Γενεύης και του Ελσίνκι (Tröhler, 2008). Όπως επισημαίνουν οι (Vevaina et al., 1993), οι γιατροί θεωρούνται υπεύθυνοι να συμμορφώνονται με τον ηθικό κώδικα του επαγγέλματός τους λόγω της επένδυσης που κάνει η κοινωνία στην εκπαίδευσή τους (χρηματική και η χρήση των μελών της ως εκπαιδευτικό υλικό καθ' όλη τη διάρκεια της εκπαίδευσης

Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης. Οι απόψεις και οι γνώμες που διατυπώνονται εκφράζουν αποκλειστικά τις απόψεις των συντακτών και δεν αντιπροσωπεύουν κατ'ανάγκη τις απόψεις της Ευρωπαϊκής Ένωσης ή του Ευρωπαϊκού Εκτελεστικού Οργανισμού Εκπαίδευσης και Πολιτισμού (EACEA). Η Ευρωπαϊκή Ένωση και ο EACEA δεν μπορούν να θεωρηθούν υπεύθυνοι για τις εκφραζόμενες απόψεις. Κωδικός έργου: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

και της σταδιοδρομίας του γιατρού) και του ουσιαστικού μονοπωλίου που απολαμβάνει το επάγγελμά τους μέσω της αδειοδότησης.

Η βιοϊατρική ηθική (ή βιοηθική) είναι ένας τομέας της πρακτικής (ή εφαρμοσμένης) ηθικής που ασχολείται με τα ηθικά ζητήματα που προκύπτουν από την άσκηση της ιατρικής και τη βιοϊατρική έρευνα (Vevaina et al., 1993). Κεντρική θέση στη βιοϊατρική ηθική κατέχουν οι τέσσερις αρχές που ορίζονται από τους Beauchamp και Childress (Beauchamp & Childress, 2019):

1. Αυτονομία: σεβασμός της ικανότητας λήψης αποφάσεων των αυτόνομων ατόμων. Δύο γενικές προϋποθέσεις είναι απαραίτητες για την αυτονομία: η ελευθερία, που εκδηλώνεται ως ανεξαρτησία από ελεγκτικές επιρροές, και η ικανότητα δράσης, δηλαδή η ικανότητα για σκόπιμη δράση.

2. Μη βλάβη: αποφυγή της πρόκλησης βλάβης.

3. Ευεργεσία: λήψη θετικών μέτρων για να βοηθήσουμε τους άλλους, συγκεκριμένα, αποτρέποντας το κακό ή τη βλάβη, απομακρύνοντας το κακό ή τη βλάβη και προωθώντας το καλό.

4. Δικαιοσύνη: δίκαιη κατανομή των οφελών, των κινδύνων και του κόστους. Η δικαιοσύνη ερμηνεύεται ως δίκαιη, ισότιμη και κατάλληλη μεταχείριση των ατόμων και των ομάδων, δεδομένων των πολλών ανισοτήτων στην υγειονομική περίθαλψη και την έρευνα με βάση τη φυλή, την εθνικότητα, το φύλο και την κοινωνική θέση.

Ηθική και μεροληψία στην τεχνητή νοημοσύνη

Η εισαγωγή της τεχνητής νοημοσύνης και η ραγδαία ανάπτυξη των εφαρμογών της έχουν εγείρει μια σειρά από ηθικά ζητήματα (Christoforaki & Beyan, 2022), με την μεροληψία και τη διάκριση να κατέχουν εξέχουσα θέση μεταξύ αυτών.

Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης. Οι απόψεις και οι γνώμες που διατυπώνονται εκφράζουν αποκλειστικά τις απόψεις των συντακτών και δεν αντιπροσωπεύουν κατ'ανάγκη τις απόψεις της Ευρωπαϊκής Ένωσης ή του Ευρωπαϊκού Εκτελεστικού Οργανισμού Εκπαίδευσης και Πολιτισμού (EACEA). Η Ευρωπαϊκή Ένωση και ο EACEA δεν μπορούν να θεωρηθούν υπεύθυνοι για τις εκφραζόμενες απόψεις. Κωδικός έργου: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Έτσι, η ηθική της τεχνητής νοημοσύνης αναπτύχθηκε ως ένας τομέας πρακτικής (ή εφαρμοσμένης) ηθικής που περιλαμβάνει «ένα σύνολο αξιών, αρχών και τεχνικών που χρησιμοποιούν ευρέως αποδεκτά πρότυπα του σωστού και του λάθους για να καθοδηγήσουν την ηθική συμπεριφορά στην ανάπτυξη και χρήση των τεχνολογιών τεχνητής νοημοσύνης» (Leslie, 2019, σ. 3).

Η ηθική της τεχνητής νοημοσύνης αντλεί στοιχεία τόσο από τη βιοηθική (τις τέσσερις αρχές που παρουσιάστηκαν παραπάνω) όσο και από τη συζήτηση για τα ανθρώπινα δικαιώματα, η οποία περιλαμβάνει, μεταξύ άλλων, το δικαίωμα στην ισότητα, την ελευθερία και την αξιοπρέπεια σύμφωνα με το νόμο, την προστασία των πολιτικών, κοινωνικών και πολιτικών δικαιωμάτων, την καθολική αναγνώριση της προσωπικότητας και το δικαίωμα στην ελεύθερη και ανεμπόδιστη συμμετοχή στη ζωή της κοινότητας (Leslie, 2019).

Οι τέσσερις αρχές της βιοηθικής που τροποποιήθηκαν με την Επεξηγησιμότητα αποδίδονται για την ΤΝ στο (Floridi et al., 2018) ως εξής:

1. Αυτονομία, ως η δύναμη των ανθρώπων να αποφασίζουν αν θα αποφασίσουν, και που ενέχει τον κίνδυνο να αναθέσουν πάρα πολλά στις μηχανές.

2. Η μη βλάβη, ως πρόληψη των βλαβών που προκύπτουν είτε από την πρόθεση των ανθρώπων είτε από την απρόβλεπτη συμπεριφορά των μηχανών.

3. Ευεργεσία, ως προώθηση της ευημερίας, διατήρηση της αξιοπρέπειας και προστασία του πλανήτη.

4. Δικαιοσύνη, ως πρόληψη και εξάλειψη ήδη υφιστάμενων αδικαιολόγητων διακρίσεων, καθώς και νέων βλαβών, και διασφάλιση της ίσης κατανομής των οφελών της τεχνητής νοημοσύνης.

Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης. Οι απόψεις και οι γνώμες που διατυπώνονται εκφράζουν αποκλειστικά τις απόψεις των συντακτών και δεν αντιπροσωπεύουν κατ'ανάγκη τις απόψεις της Ευρωπαϊκής Ένωσης ή του Ευρωπαϊκού Εκτελεστικού Οργανισμού Εκπαίδευσης και Πολιτισμού (EACEA). Η Ευρωπαϊκή Ένωση και ο EACEA δεν μπορούν να θεωρηθούν υπεύθυνοι για τις εκφραζόμενες απόψεις. Κωδικός έργου: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

5. Επεξηγησιμότητα, που ορίζεται ως η κατανόηση και η λογοδοσία των διαδικασιών λήψης αποφάσεων της τεχνητής νοημοσύνης.

Όσον αφορά τα ανθρώπινα δικαιώματα, σύμφωνα με μια έκθεση του 2018 που χρηματοδοτήθηκε από το Συμβούλιο της Ευρώπης (Επιτροπή εμπειρογνωμόνων για τους διαμεσολαβητές του διαδικτύου (MSI-NET), 2018), τα ανθρώπινα δικαιώματα που επηρεάζονται ιδιαίτερα από τους αλγόριθμους και τις τεχνικές αυτοματοποιημένης επεξεργασίας δεδομένων περιλαμβάνουν:

- Δωρεάν δοκιμή και δίκαιη δίκη
- Προστασία της ιδιωτικής ζωής και των δεδομένων
- Ελευθερία έκφρασης
- Αποτελεσματική αποκατάσταση
- Ελευθερία συνάθροισης και συνεταιρισμού
- Απαγόρευση διακρίσεων
- Κοινωνικά δικαιώματα και πρόσβαση σε δημόσιες υπηρεσίες
- Δικαίωμα σε ελεύθερες εκλογές

Οι μεροληπτικοί αλγόριθμοι αναφέρονται ρητά ως πιθανοί παράγοντες διάκρισης κατά κοινωνικών ομάδων με βάση την ηλικία, τον σεξουαλικό προσανατολισμό, τη φυλή, το φύλο ή την κοινωνικοοικονομική θέση (Επιτροπή εμπειρογνωμόνων για τους διαμεσολαβητές του διαδικτύου (MSI-NET), 2018, σ. 27). Επιπλέον, η Σύμβαση-Πλαίσιο του Συμβουλίου της Ευρώπης για την Τεχνητή Νοημοσύνη και τα Ανθρώπινα Δικαιώματα, τη Δημοκρατία και το Κράτος Δικαίου αναφέρει συγκεκριμένα ότι τα κράτη μέλη «θα θεσπίζουν ή θα διατηρούν μέτρα με σκοπό να διασφαλίζουν ότι οι δραστηριότητες στο πλαίσιο του κύκλου ζωής των συστημάτων τεχνητής νοημοσύνης

Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης. Οι απόψεις και οι γνώμες που διατυπώνονται εκφράζουν αποκλειστικά τις απόψεις των συντακτών και δεν αντιπροσωπεύουν κατ'ανάγκη τις απόψεις της Ευρωπαϊκής Ένωσης ή του Ευρωπαϊκού Εκτελεστικού Οργανισμού Εκπαίδευσης και Πολιτισμού (EACEA). Η Ευρωπαϊκή Ένωση και ο EACEA δεν μπορούν να θεωρηθούν υπεύθυνοι για τις εκφραζόμενες απόψεις. Κωδικός έργου: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

σέβονται την ισότητα, συμπεριλαμβανομένης της ισότητας των φύλων, και την απαγόρευση των διακρίσεων, όπως προβλέπεται από το εφαρμοστέο διεθνές και εθνικό δίκαιο» (Σύμβαση-Πλαίσιο του Συμβουλίου της Ευρώπης για την Τεχνητή Νοημοσύνη και τα Ανθρώπινα Δικαιώματα, τη Δημοκρατία και το Κράτος Δικαίου, 2024, σ. 4).

Έτσι, η ηθική της τεχνητής νοημοσύνης έχει επίσης συγκλίνει σε ένα σύνολο αρχών που βασίζονται στις τέσσερις κλασικές αρχές της ιατρικής ηθικής, καθώς και σε άλλες προσεγγίσεις, που συνοψίζονται στο (Christoforaki & Beyan, 2022). Ωστόσο, όπως σημειώνεται στο (Mittelstadt, 2019), σε σύγκριση με την ιατρική, η ανάπτυξη της τεχνητής νοημοσύνης στερείται: (1) κοινών στόχων και καταπιστευματικών υποχρεώσεων, (2) επαγγελματικής ιστορίας και κανόνων, (3) αποδεδειγμένων μεθόδων για τη μετατροπή των αρχών σε πράξη και (4) ισχυρών μηχανισμών νομικής και επαγγελματικής λογοδοσίας, γεγονός που υπονομεύει την επιτυχία της προσέγγισης βάσει αρχών. Φυσικά, υπάρχει επίσης ένα πολύπλοκο κανονιστικό πλαίσιο που διέπει την ανάπτυξη και τη χρήση της τεχνητής νοημοσύνης στην ΕΕ, συμπεριλαμβανομένων των νόμων κατά των διακρίσεων, ένα θέμα που, ωστόσο, δεν εμπίπτει στο πεδίο εφαρμογής της παρούσας έκθεσης.

Οι οργανώσεις της κοινωνίας των πολιτών (ΟΚΠ) ως ενδιαφερόμενοι φορείς στο οικοσύστημα υγειονομικής περίθαλψης (Vayena et al., 2018) μπορούν να διαδραματίσουν σημαντικό ρόλο στον εντοπισμό και την αντιμετώπιση της μεροληψίας της ΤΝ και της διακυβέρνησης της ΤΝ γενικότερα, μέσω της υπεράσπισης της ηθικής ανάπτυξης της ΤΝ, της λογοδοσίας των ενδιαφερόμενων μερών, της εκπαίδευσης του κοινού, της εκπροσώπησης των περιθωριοποιημένων κοινοτήτων, της διαμόρφωσης πολιτικών και κανονιστικών πλαισίων και της ενίσχυσης της συνεργασίας μεταξύ κυβερνήσεων, τεχνολογικών εταιρειών και του κοινού (Korir, 2024).

Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης. Οι απόψεις και οι γνώμες που διατυπώνονται εκφράζουν αποκλειστικά τις απόψεις των συντακτών και δεν αντιπροσωπεύουν κατ'ανάγκη τις απόψεις της Ευρωπαϊκής Ένωσης ή του Ευρωπαϊκού Εκτελεστικού Οργανισμού Εκπαίδευσης και Πολιτισμού (EACEA). Η Ευρωπαϊκή Ένωση και ο EACEA δεν μπορούν να θεωρηθούν υπεύθυνοι για τις εκφραζόμενες απόψεις. Κωδικός έργου: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Σε αυτό το θεωρητικό πλαίσιο, αναπτύσσεται μια ποικιλία τεχνικών λύσεων για την αντιμετώπιση της μεροληψίας. Στην επόμενη ενότητα, παρουσιάζουμε μια ταξινόμηση των προκαταλήψεων που προκαλούνται από την ΤΝ, η οποία χρησίμευσε ως βάση για το πρότυπο χαρτογράφησης μας, εστιάζοντας στον αντίκτυπό τους στο φύλο και στις φυλετικές διακρίσεις. Οι ανθρώπινες ψυχικές προκαταλήψεις (Hofmann, 2023), για παράδειγμα, οι γνωστικές προκαταλήψεις, όπως η μεροληψίας επιβεβαίωσης ή διαθεσιμότητας, αν και έχουν μεγάλο αντίκτυπο στην Ιατρική, θεωρούνται εκτός πεδίου εφαρμογής του τρέχοντος έργου.

Μεροληψία στα συστήματα Τεχνητής Νοημοσύνης

Η μεροληψία στα συστήματα υπολογιστών ορίζεται στους (Friedman & Nissenbaum, 1996, σελ. 332) ως όρος «[που αναφέρεται] σε συστήματα υπολογιστών που συστηματικά και άδικα κάνουν διακρίσεις εις βάρος ορισμένων ατόμων ή ομάδων ατόμων υπέρ άλλων. Ένα σύστημα κάνει άδικες διακρίσεις εάν αρνείται μια ευκαιρία ή ένα αγαθό ή εάν αποδίδει ένα ανεπιθύμητο αποτέλεσμα σε ένα άτομο ή μια ομάδα ατόμων για λόγους που είναι παράλογοι ή ακατάλληλο.»

Σύμφωνα με τους (Friedman & Nissenbaum, 1996), η μεροληψία στα συστήματα υπολογιστών μπορεί να διακριθεί σε τρεις κατηγορίες: προϋπάρχουσα μεροληψία, τεχνική μεροληψία και αναδυόμενη μεροληψία. Στην επόμενη υποενότητα, εξετάζουμε κάθε είδος μεροληψίας και το παρουσιάζουμε με μελέτες περιπτώσεων, όπως εκδηλώνονται στην επιστημονική βιβλιογραφία.

Προϋπάρχουσα μεροληψία

Η προϋπάρχουσα μεροληψία πηγάζει από προκαταλήψεις σε κοινωνικούς θεσμούς, πρακτικές και συμπεριφορές που υπάρχουν ήδη και είναι ανεξάρτητες και συνήθως υπάρχουν πριν από τη δημιουργία του συστήματος. Αυτό το είδος μεροληψίας

Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης. Οι απόψεις και οι γνώμες που διατυπώνονται εκφράζουν αποκλειστικά τις απόψεις των συντακτών και δεν αντιπροσωπεύουν κατ'ανάγκη τις απόψεις της Ευρωπαϊκής Ένωσης ή του Ευρωπαϊκού Εκτελεστικού Οργανισμού Εκπαίδευσης και Πολιτισμού (EACEA). Η Ευρωπαϊκή Ένωση και ο EACEA δεν μπορούν να θεωρηθούν υπεύθυνοι για τις εκφραζόμενες απόψεις. Κωδικός έργου: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

ενσωματώνεται στο σύστημα είτε συνειδητά είτε ασυνείδητα, μερικές φορές ακόμη και όταν οι δημιουργοί του συστήματος προσπαθούν να το αποφύγουν.

Μελέτη περίπτωσης: Διάγνωση καρδιαγγειακών παθήσεων στις γυναίκες

Οι καρδιαγγειακές παθήσεις (CVD) θεωρούνται συνήθως «ασθένειες των ανδρών», γεγονός που έχει συμβάλει στην υποδιάγνωση και την ανεπαρκή θεραπεία των γυναικών. Όπως φαίνεται στο (Al Hamid et al., 2024), μια συστηματική ανασκόπηση του θέματος, οι CVD αναφέρθηκαν λιγότερο συχνά σε γυναίκες που είτε παρουσίαζαν ηπιότερα συμπτώματα από τους άνδρες είτε τα συμπτώματά τους είχαν διαγνωστεί εσφαλμένα ως γαστρεντερικά ή συμπτώματα που σχετίζονταν με άγχος. Ως εκ τούτου, στις γυναίκες προσφέρθηκαν λιγότερες διαγνωστικές εξετάσεις και φάρμακα και παραπέμφθηκαν σε καρδιολόγους ή/και νοσηλεύτηκαν λιγότερο συχνά. Επιπλέον, σε περίπτωση νοσηλείας, οι γυναίκες είχαν λιγότερες πιθανότητες να υποβληθούν σε στεφανιαία επέμβαση. Ως εκ τούτου, οι παράγοντες κινδύνου των γυναικών υποτιμήθηκαν από τους γιατρούς, ιδίως από τους άνδρες γιατρούς. Δεδομένου ότι οι γυναίκες εξακολουθούν να υποεκπροσωπούνται στον τομέα της καρδιολογίας (Fatunde et al., 2025), μπορεί να συναχθεί το συμπέρασμα ότι οι γυναίκες έχουν λιγότερες πιθανότητες να λάβουν κατάλληλη υγειονομική περίθαλψη λόγω των ήδη υφιστάμενων προκαταλήψεων.

Τα συστήματα τεχνητής νοημοσύνης εκπαιδεύονται χρησιμοποιώντας δεδομένα που συλλέγονται από υπάρχουσες πρακτικές, οπότε ένα σύστημα διάγνωσης καρδιαγγειακών παθήσεων με τεχνητή νοημοσύνη θα ενσωματώνει αυτή την μεροληψία, δημιουργώντας διακρίσεις εις βάρος των γυναικών, ανεξάρτητα από τις επιλογές κατά την τεχνική υλοποίηση.

Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης. Οι απόψεις και οι γνώμες που διατυπώνονται εκφράζουν αποκλειστικά τις απόψεις των συντακτών και δεν αντιπροσωπεύουν κατ'ανάγκη τις απόψεις της Ευρωπαϊκής Ένωσης ή του Ευρωπαϊκού Εκτελεστικού Οργανισμού Εκπαίδευσης και Πολιτισμού (EACEA). Η Ευρωπαϊκή Ένωση και ο EACEA δεν μπορούν να θεωρηθούν υπεύθυνοι για τις εκφραζόμενες απόψεις. Κωδικός έργου: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Με τη συγχρηματοδότηση της Ευρωπαϊκής Ένωσης

AEQUITAS

Τεχνική μεροληψία

Η τεχνική μεροληψία προκύπτει από τεχνικούς περιορισμούς ή παραμέτρους, ιδίως όταν οι δημιουργοί συστημάτων προσπαθούν να καταστήσουν τις ανθρώπινες κατασκευές συμβατές με τους υπολογιστές, όπως η ποσοτικοποίηση του ποιοτικού, η διακριτοποίηση του συνεχούς ή η τυποποίηση του μη τυποποιημένο. Επιπλέον, η αποσύνδεση των αλγορίθμων από τα περιβάλλοντα στα οποία λειτουργούν μπορεί να οδηγήσει σε αδυναμία δίκαιης μεταχείρισης όλων των ομάδων υπό όλες τις σημαντικές συνθήκες.

Μελέτη περίπτωσης: Προγνωστική ακρίβεια των μοντέλων πρόβλεψης κινδύνου εγκεφαλικού επεισοδίου σε μαύρους και λευκούς πληθυσμούς

(Hong et al., 2023) πραγματοποίησαν μια αναδρομική μελέτη της προβλεπτικής ακρίβειας του κινδύνου εγκεφαλικού επεισοδίου, συγκρίνοντας τα υπάρχοντα μοντέλα πρόβλεψης κινδύνου εγκεφαλικού επεισοδίου και τις νέες τεχνικές μηχανικής μάθησης που περιλαμβάνουν, μεταξύ άλλων κριτηρίων, τη φυλή των ασθενών. Όλοι οι αλγόριθμοι παρουσίασαν χειρότερη διάκριση στα μαύρα άτομα σε σύγκριση με τα λευκά άτομα. Αυτή η κατάσταση, σύμφωνα με τους συγγραφείς, μπορεί να αποδοθεί σε παράγοντες κινδύνου που δεν καταγράφονται στα δεδομένα, όπως ο τύπος ασφάλισης, τα γλωσσικά εμπόδια και άλλοι παράγοντες που προκύπτουν από τη διαφοροποιημένη πρόσβαση σε υπηρεσίες υγειονομικής περίθαλψης, δηλαδή τα δεδομένα είναι αποσπασμένα από το κοινωνικοοικονομικό περιβάλλον στο οποίο παράχθηκαν. Ταυτόχρονα, όλοι οι προαναφερθέντες παράγοντες κινδύνου είναι κατασκευάσματα που είναι δύσκολο να αναπαρασταθούν σε μορφή κατάλληλη για υπολογιστές. Σε όλα τα παραπάνω, θα μπορούσαμε επίσης να προσθέσουμε ότι οι αλγόριθμοι τεχνητής νοημοσύνης τελευταίας τεχνολογίας είναι από τη φύση τους αδιαφανείς όσον αφορά τα χαρακτηριστικά που επιλέγουν για να επιτύχουν υψηλή ακρίβεια (Knight, 2017), με αποτέλεσμα ακόμη και οι

Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης. Οι απόψεις και οι γνώμες που διατυπώνονται εκφράζουν αποκλειστικά τις απόψεις των συντακτών και δεν αντιπροσωπεύουν κατ'ανάγκη τις απόψεις της Ευρωπαϊκής Ένωσης ή του Ευρωπαϊκού Εκτελεστικού Οργανισμού Εκπαίδευσης και Πολιτισμού (EACEA). Η Ευρωπαϊκή Ένωση και ο EACEA δεν μπορούν να θεωρηθούν υπεύθυνοι για τις εκφραζόμενες απόψεις. Κωδικός έργου: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

δημιουργοί τους να μην είναι σε θέση να εξηγήσουν πώς λειτουργούν και, συνεπώς, να ελέγξουν εάν κάποιος από τους προαναφερθέντες κοινωνικοοικονομικούς παράγοντες λαμβάνεται πραγματικά υπόψη στην εσωτερική λειτουργία του συστήματος τεχνητής νοημοσύνης.

Αναδυόμενη μεροληψία

Η αναδυόμενη μεροληψία εκδηλώνεται σε ένα πλαίσιο χρήσης με πραγματικούς χρήστες, συνήθως μετά την ολοκλήρωση ενός σχεδιασμού, ως αποτέλεσμα της αλλαγής των κοινωνικών γνώσεων που δεν μπορούν ή δεν ενσωματώνονται στο σχεδιασμό του συστήματος, ή ενός πληθυσμού με διαφορετικές γνώσεις ή πολιτισμικές αξίες από αυτές που υποτίθεται ότι υπάρχουν στο σχεδιασμό.

Μελέτη περίπτωσης: Μεταβολές στο σύνολο δεδομένων

Η μετατόπιση του συνόλου δεδομένων είναι μια αναντιστοιχία μεταξύ των κατανομών των συνόλων δεδομένων εκπαίδευσης και δοκιμής κατά την ανάπτυξη του αλγορίθμου και μπορεί να οδηγήσει σε διαφορετικές επιδόσεις σε επίπεδο υποομάδας (Chen et al., 2023).

Στην ανίχνευση του καρκίνου του δέρματος, για παράδειγμα, πολλά σύνολα δεδομένων απεικόνισης που χρησιμοποιούνται για την εκπαίδευση αλγορίθμων τεχνητής νοημοσύνης για την ανίχνευση του καρκίνου του δέρματος προέρχονται από χώρες με πληθυσμούς με ανοιχτόχρωμη επιδερμίδα (Guo et al., 2021), με αποτέλεσμα να υποεκπροσωπούνται ορισμένες δημογραφικές ομάδες. Οι αλγόριθμοι τεχνητής νοημοσύνης που έχουν εκπαιδευτεί με αυτά τα σύνολα δεδομένων έχουν χαμηλή απόδοση όταν εφαρμόζονται σε χώρες με πιο ποικιλόμορφο πληθυσμό, με αποτέλεσμα να εισάγουν διακρίσεις σε βάρος των ατόμων με σκούρο δέρμα. Η συλλογή, η σχολιασμός και η επικύρωση των συνόλων δεδομένων είναι δύσκολη και δαπανηρή, με αποτέλεσμα τα συστήματα τεχνητής νοημοσύνης που αναπτύσσονται

Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης. Οι απόψεις και οι γνώμες που διατυπώνονται εκφράζουν αποκλειστικά τις απόψεις των συντακτών και δεν αντιπροσωπεύουν κατ'ανάγκη τις απόψεις της Ευρωπαϊκής Ένωσης ή του Ευρωπαϊκού Εκτελεστικού Οργανισμού Εκπαίδευσης και Πολιτισμού (EACEA). Η Ευρωπαϊκή Ένωση και ο EACEA δεν μπορούν να θεωρηθούν υπεύθυνοι για τις εκφραζόμενες απόψεις. Κωδικός έργου: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

σε χώρες με χαμηλό και μεσαίο εισόδημα να βασίζονται σε δημόσια διαθέσιμα σύνολα δεδομένων που ενδέχεται να μην αντικατοπτρίζουν την κατανομή του πληθυσμού τους, με αποτέλεσμα να υπάρχει αναντιστοιχία μεταξύ του πληθυσμού προέλευσης και του πληθυσμού-στόχου. Το ίδιο μπορεί να συμβεί και σε χώρες με υψηλό εισόδημα, για παράδειγμα, λόγω μεταβολών στον πληθυσμό από την αύξηση της μετανάστευσης ή λόγω διακυμάνσεων στην αυτοαναφερόμενη φυλή. Όπως σημειώνεται στο (Chen et al., 2023), «δεδομένου ότι είναι πλέον αποδεκτό ότι η φυλή είναι ένα κοινωνικό κατασκεύασμα και ότι υπάρχει μεγαλύτερη γενετική ποικιλομορφία εντός μιας συγκεκριμένης φυλής από ό,τι μεταξύ των φυλών» […] «η ιατρική κοινότητα έχει αρχίσει να συνειδητοποιεί ότι οι ταξινομίες του παρελθόντος δεν αντιπροσωπεύουν επαρκώς τις ομάδες ανθρώπων που υποτίθεται ότι αντιπροσωπεύουν» και «μπορούν να συσκοτίσουν τον πολιτισμό, την ιστορία, την κοινωνικοοικονομική κατάσταση και άλλους παράγοντες που επηρεάζουν την ισότητα..”

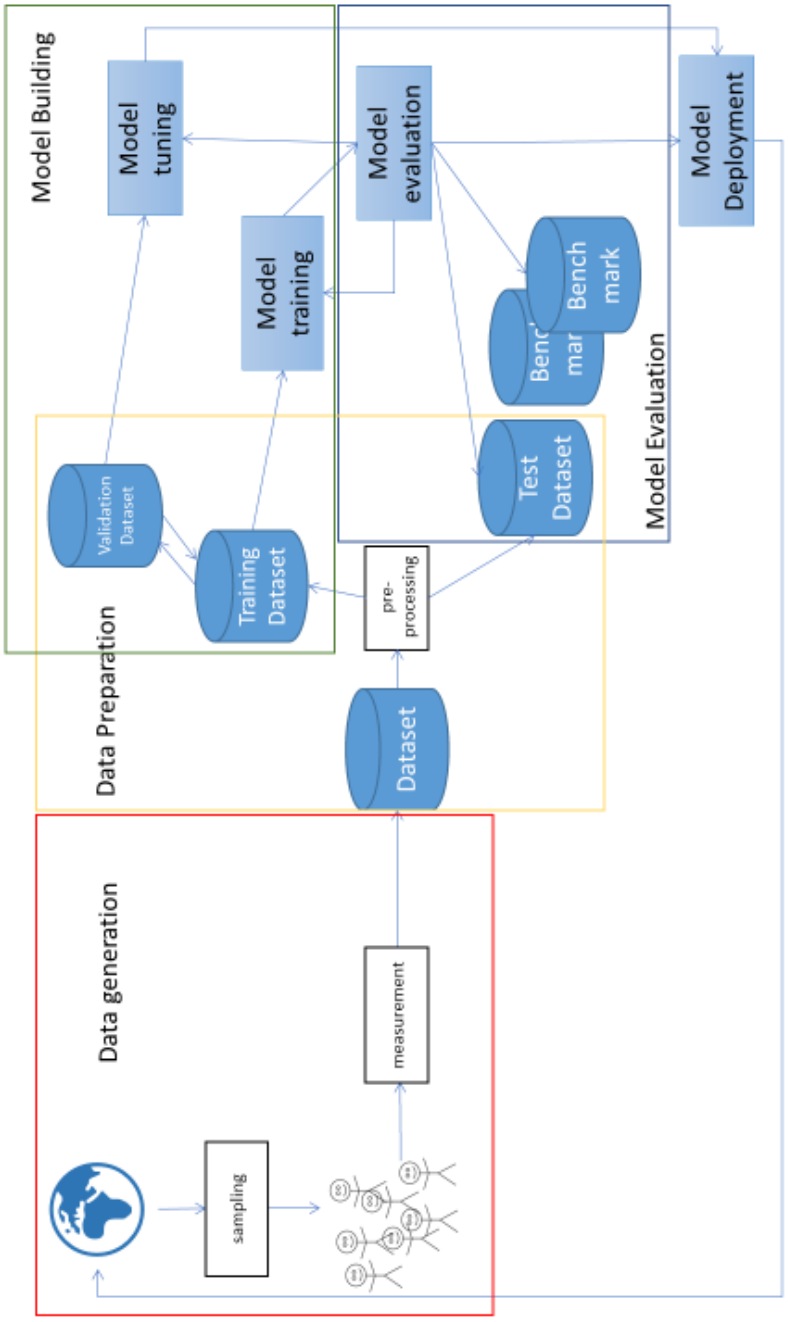
Είδη μεροληψίας που αφορούν συγκεκριμένα τη διαδικασία ML/AL

Ενώ τα παραπάνω ισχύουν για όλα τα συστήματα υπολογιστών, οι εφαρμογές τεχνητής νοημοσύνης έχουν πιο συγκεκριμένες απαιτήσεις, οπότε χρειαζόμασταν μια πιο λεπτομερή ταξινόμηση. Κατά συνέπεια, αποφασίσαμε να ακολουθήσουμε την ταξινόμηση μεροληψίας που παρουσιάζεται στο (Suresh & Guttag, 2020), καθώς προσδιορίζει τους τύπους μεροληψίας σε κάθε βήμα της διαδικασίας ML/AI, όπως απεικονίζεται στο Σχήμα 1.

Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης. Οι απόψεις και οι γνώμες που διατυπώνονται εκφράζουν αποκλειστικά τις απόψεις των συντακτών και δεν αντιπροσωπεύουν κατ'ανάγκη τις απόψεις της Ευρωπαϊκής Ένωσης ή του Ευρωπαϊκού Εκτελεστικού Οργανισμού Εκπαίδευσης και Πολιτισμού (EACEA). Η Ευρωπαϊκή Ένωση και ο EACEA δεν μπορούν να θεωρηθούν υπεύθυνοι για τις εκφραζόμενες απόψεις. Κωδικός έργου: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Με τη συγχρηματοδότηση της Ευρωπαϊκής Ένωσης

AEQUITAS



Σχήμα 1 ML/AI Pipeline. Εικόνα προσαρμοσμένη από (Suresh & Guttag, 2020)

Ένας τυπικός αγωγός ML/AI μπορεί να περιγραφεί ως εξής:

Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης. Οι απόψεις και οι γνώμες που διατυπώνονται εκφράζουν αποκλειστικά τις απόψεις των συντακτών και δεν αντιπροσωπεύουν κατ'ανάγκη τις απόψεις της Ευρωπαϊκής Ένωσης ή του Ευρωπαϊκού Εκτελεστικού Οργανισμού Εκπαίδευσης και Πολιτισμού (EACEA). Η Ευρωπαϊκή Ένωση και ο EACEA δεν μπορούν να θεωρηθούν υπεύθυνοι για τις εκφραζόμενες απόψεις. Κωδικός έργου: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

- **Δημιουργία δεδομένων.** Η δημιουργία ενός συστήματος ML/AI ξεκινά με τη δημιουργία δεδομένων. Αυτό περιλαμβάνει πρώτα τη συλλογή και την προετοιμασία δεδομένων για τη σύνταξη ενός συνόλου δεδομένων για το σύστημα AI. Τα υπάρχοντα δεδομένα στον κόσμο πρέπει να συλλέγονται με τον προσδιορισμό ενός δείγματος πληθυσμού-στόχου. Το επόμενο βήμα είναι ο ορισμός και η μέτρηση των χαρακτηριστικών που σχετίζονται με την εφαρμογή που θα υλοποιηθεί και/ή η σχολιασμός των δεδομένων με τις κατάλληλες ετικέτες. Πρόκειται για μια δαπανηρή και χρονοβόρα διαδικασία, οπότε οι επαγγελματίες του AI χρησιμοποιούν συνήθως υπάρχοντα σύνολα δεδομένων (δημόσια ή αγορασμένα).

- **Προετοιμασία δεδομένων.** Σε αυτό το στάδιο, το σύνολο δεδομένων χωρίζεται σε τρία σύνολα, και συγκεκριμένα: το σύνολο δεδομένων εκπαίδευσης - το πραγματικό σύνολο δεδομένων που χρησιμοποιείται για την εκπαίδευση του μοντέλου, το σύνολο δεδομένων επικύρωσης, ένα δείγμα δεδομένων που χρησιμοποιείται για την αξιολόγηση της καταλληλότητας ενός μοντέλου στο σύνολο δεδομένων εκπαίδευσης, ενώ ταυτόχρονα ρυθμίζονται οι υπερπαραμέτροι του μοντέλου (παράμετροι του μοντέλου που δεν μπορούν να αντληθούν από τα δεδομένα, π.χ. ο αριθμός των επιπέδων και των νευρώνων σε ένα μοντέλο νευρωνικού δικτύου). Σε αυτή τη φάση, τα δεδομένα ενδέχεται να χρειαστεί να υποστούν προεπεξεργασία (π.χ. καθαρισμός, κανονικοποίηση) και το σύνολο δεδομένων δοκιμής, το μέρος των δεδομένων που χρησιμοποιείται για την αξιολόγηση του τελικού μοντέλου, παρέχοντας ένα χρυσό πρότυπο μόλις το μοντέλο εκπαιδευτεί πλήρως.

- **Κατασκευή μοντέλων.** Σε αυτή τη φάση, το μοντέλο εκπαιδεύεται με βάση τα δεδομένα εκπαίδευσης και τελειοποιείται με την προσαρμογή των υπερπαραμέτρων στο σύνολο δεδομένων επικύρωσης.

Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης. Οι απόψεις και οι γνώμες που διατυπώνονται εκφράζουν αποκλειστικά τις απόψεις των συντακτών και δεν αντιπροσωπεύουν κατ'ανάγκη τις απόψεις της Ευρωπαϊκής Ένωσης ή του Ευρωπαϊκού Εκτελεστικού Οργανισμού Εκπαίδευσης και Πολιτισμού (EACEA). Η Ευρωπαϊκή Ένωση και ο EACEA δεν μπορούν να θεωρηθούν υπεύθυνοι για τις εκφραζόμενες απόψεις. Κωδικός έργου: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

- **Αξιολόγηση μοντέλου.** Το εκπαιδευμένο μοντέλο αξιολογείται χρησιμοποιώντας το σύνολο δεδομένων δοκιμής και, σε ορισμένες περιπτώσεις, σύνολα δεδομένων αναφοράς, τα οποία είναι ανεξάρτητα συγκεντρωμένα σύνολα δεδομένων που χρησιμοποιούνται για να αποδείξουν την ανθεκτικότητα του μοντέλου και/ή να επιτρέψουν τη σύγκριση με άλλες μεθόδους.

- **Ανάπτυξη μοντέλου.** Εφαρμογή του μοντέλου σε πραγματικές συνθήκες. Αυτό μπορεί να οδηγήσει σε αλλαγές ανάλογα με τα αποτελέσματα και μπορεί επίσης να δημιουργήσει έναν κύκλο ανατροφοδότησης στην αρχή της διαδικασίας.

Λαμβάνοντας υπόψη τις φάσεις της διαδικασίας ML/AI που περιγράφονται παραπάνω, υιοθετούμε την ταξινόμηση των προκαταλήψεων των (Suresh & Guttag, 2020). Συγκεκριμένα, προσδιορίζουν τις ακόλουθες κατηγορίες προκαταλήψεων: ιστορικές, προκαταλήψεις αναπαράστασης, μέτρησης, συγκέντρωσης, μάθησης, αξιολόγησης και ανάπτυξης. Στις ακόλουθες υποενότητες, ορίζουμε τις παραπάνω προκαταλήψεις και παρουσιάζουμε μελέτες περιπτώσεων από τις πηγές που συλλέχθηκαν για το έργο.

Ιστορική Μεροληψία

Η ιστορική μεροληψία αντιστοιχεί σε προϋπάρχουσα μεροληψία όπως ορίζεται από τους Friedman και Nissenbaum (1996), η οποία ενσωματώνει ήδη υπάρχουσες προκαταλήψεις και στερεότυπα στα δεδομένα. Ένα παράδειγμα παρουσιάζεται στο έργο του Calderone (1990), το οποίο εξετάζει εάν η συχνότητα χορήγησης αναλγητικών και κατασταλτικών φαρμάκων σε μετεγχειρητικούς ασθενείς που υποβλήθηκαν σε επέμβαση αορτοστεφανιαίας παράκαμψης *(coronary artery bypass graft /CABG)* διαφέρει ανάλογα με το φύλο και την ηλικία του/της ασθενούς. Τα αποτελέσματα έδειξαν ότι στους άνδρες ασθενείς και στους ασθενείς ηλικίας 61 ετών ή νεότερους χορηγούνταν αναλγητικά πολύ συχνότερα σε σύγκριση με τις γυναίκες ασθενείς και ασθενείς ηλικίας 62 ετών και άνω, στις οποίες, αντιθέτως, χορηγούνταν κατασταλτικά

Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης. Οι απόψεις και οι γνώμες που διατυπώνονται εκφράζουν αποκλειστικά τις απόψεις των συντακτών και δεν αντιπροσωπεύουν κατ'ανάγκη τις απόψεις της Ευρωπαϊκής Ένωσης ή του Ευρωπαϊκού Εκτελεστικού Οργανισμού Εκπαίδευσης και Πολιτισμού (EACEA). Η Ευρωπαϊκή Ένωση και ο EACEA δεν μπορούν να θεωρηθούν υπεύθυνοι για τις εκφραζόμενες απόψεις. Κωδικός έργου: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

φάρμακα πολύ συχνότερα. Η μελέτη περίπτωσης σχετικά με την προγνωστική ακρίβεια μοντέλων πρόβλεψης κινδύνου εγκεφαλικού επεισοδίου σε Μαύρους και Λευκούς πληθυσμούς παρουσιάζεται παραπάνω στην υποενότητα της για την Προϋπάρχουσα Μεροληψία· ωστόσο, εδώ θα παρουσιάσουμε μια άλλη μελέτη περίπτωσης που αναδεικνύει την ιστορική μεροληψία όσον αφορά τη χρήση της Τεχνητής Νοημοσύνης στην ψυχική υγεία.

Μελέτη περίπτωσης: Τεχνητή Νοημοσύνη στην ψυχική υγεία και οι μεροληψίες των γλωσσικών μοντέλων

Οι Straw και Callison-Burch (2020) παρουσιάζουν μια συστηματική ανασκόπηση της βιβλιογραφίας σχετικά με τις χρήσεις της Επεξεργασίας Φυσικής Γλώσσας *(natural language processing/NLP)* στην ψυχική υγεία, με στόχο τον εντοπισμό του τρόπου με τον οποίο αυτές οι μεροληψίες ενδέχεται να διευρύνουν τις ανισότητες στην υγεία. Τα μοντέλα ΤΝ που χρησιμοποιούν NLP για την ανάλυση και σκιαγράφηση του ψυχικού προφίλ συλλέγουν μεγάλα σύνολα δεδομένων εκφραστικής γλώσσας, τα οποία συνήθως προέρχονται από μέσα κοινωνικής δικτύωσης, διαδικτυακά φόρουμ, ιστολόγια και δωμάτια συνομιλίας (*chat rooms*). Ωστόσο, τα δεδομένα αυτά είναι ήδη επηρεασμένα από το προσωπικό υπόβαθρο και το κοινωνικό πλαίσιο του ατόμου.

Συγκεκριμένα, όσον αφορά το φύλο και τη γλώσσα, υπάρχει εκτενής (αγγλόφωνη) βιβλιογραφία, η οποία συνοψίζεται από τους Pennebaker et al. (2003), και αναδεικνύει διαφορές στη χρήση λέξεων μεταξύ γυναικών και ανδρών. Για παράδειγμα, οι γυναίκες χρησιμοποιούν λιγότερο κατηγορηματικό λόγο, γεγονός που εκδηλώνεται με μεγαλύτερη ευγένεια, λιγότερη χρήση υβριστικών εκφράσεων, περισσότερους ενισχυτές (π.χ. «πραγματικά», «τόσο») και περισσότερες επιφυλάξεις (δηλαδή προσδιορισμούς ή λέξεις αβεβαιότητας όπως «κάπως», «ίσως»). Οι άνδρες, από την άλλη πλευρά, περιγράφονται ως πιο κατευθυντικοί και ακριβείς, καθώς και λιγότερο συναισθηματικοί στη χρήση της γλώσσας, η οποία χαρακτηρίζεται από αναφορές σε

Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης. Οι απόψεις και οι γνώμες που διατυπώνονται εκφράζουν αποκλειστικά τις απόψεις των συντακτών και δεν αντιπροσωπεύουν κατ'ανάγκη τις απόψεις της Ευρωπαϊκής Ένωσης ή του Ευρωπαϊκού Εκτελεστικού Οργανισμού Εκπαίδευσης και Πολιτισμού (EACEA). Η Ευρωπαϊκή Ένωση και ο EACEA δεν μπορούν να θεωρηθούν υπεύθυνοι για τις εκφραζόμενες απόψεις. Κωδικός έργου: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

ποσότητα, αξιολογικά επίθετα (π.χ. «καλός», «ανόητος»), ελλειπτικές προτάσεις («Υπέροχη εικόνα») και αναφορές στο «εγώ». Όπως επισημαίνουν οι συγγραφείς, οι διαφορές αυτές συνάδουν με ένα κοινωνιολογικό πλαίσιο ερμηνείας των έμφυλων διαφορών, αλλά μπορούν επίσης να αποδοθούν και σε εναλλακτικές εξηγήσεις, όπως για παράδειγμα η μεγαλύτερη εμπλοκή των γυναικών σε κοινωνικές σχέσεις.

Όσον αφορά την ψυχική υγεία, άνδρες και γυναίκες γράφουν σημειώματα αυτοκτονίας εκφράζοντας τη συναισθηματική τους οδύνη με διαφορετικό τρόπο· οι γυναίκες τείνουν να εσωτερικεύουν τα αρνητικά συναισθήματα, ενώ οι άνδρες εκφράζουν αυξανόμενο θυμό (Straw & Callison-Burch, 2020). Ένα σύστημα ΤΝ για ζητήματα ψυχικής υγείας που είναι σχεδιασμένο να ελέγχει (*screening*) με βάση τα γλωσσικά πρότυπα του ενός φύλου ενδέχεται να είναι ακατάλληλο για το άλλο (και αυτό λαμβάνοντας υπόψη το φύλο σε δυαδικό πλαίσιο, το οποίο αποκλείει ένα μεγάλο μέρος του πληθυσμού).

Μεροληψία αναπαράστασης

Η μεροληψία αναπαράστασης προκύπτει όταν το δείγμα ανάπτυξης υποεκπροσωπεί κάποιο τμήμα του πληθυσμού κατά τη φάση συλλογής δεδομένων. Αυτό μπορεί να συμβεί με τους εξής τρόπους: κατά τον ορισμό του πληθυσμού-στόχου, εάν αυτός δεν αντανακλά τον πληθυσμό χρήσης· κατά τον ορισμό του πληθυσμού-στόχου, εάν περιλαμβάνει υποεκπροσωπούμενες ομάδες· ή κατά τη δειγματοληψία από τον πληθυσμό-στόχο, εάν η μέθοδος δειγματοληψίας είναι περιορισμένη ή άνιση. Η μεροληψία αναπαράστασης οδηγεί σε λάθος γενίκευση για ένα υποσύνολο του πληθυσμού χρηστών. Ένα χαρακτηριστικό παράδειγμα αφορά την ανίχνευση καρκίνου του δέρματος, καθώς πολλά σύνολα δεδομένων ιατρικής απεικόνισης υποεκπροσωπούν ορισμένες δημογραφικές ομάδες, με αποτέλεσμα τα μοντέλα μηχανικής μάθησης (*machine learning models*) να εκπαιδεύονται κυρίως σε εικόνες ατόμων με ανοιχτόχρωμο δέρμα (Guo et al., 2021). Λαμβάνοντας υπόψη τις

Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης. Οι απόψεις και οι γνώμες που διατυπώνονται εκφράζουν αποκλειστικά τις απόψεις των συντακτών και δεν αντιπροσωπεύουν κατ'ανάγκη τις απόψεις της Ευρωπαϊκής Ένωσης ή του Ευρωπαϊκού Εκτελεστικού Οργανισμού Εκπαίδευσης και Πολιτισμού (EACEA). Η Ευρωπαϊκή Ένωση και ο EACEA δεν μπορούν να θεωρηθούν υπεύθυνοι για τις εκφραζόμενες απόψεις. Κωδικός έργου: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

στοχευόμενες ασθένειες με τις οποίες ασχολείται το έργο AEQUITAS, παρουσιάζουμε μια μελέτη περίπτωσης σχετικά με τη μεροληψία αναπαράστασης με βάση τη φυλή στην περίπτωση του σακχαρώδη διαβήτη τύπου 2.

Μελέτη περίπτωσης: Αξιολόγηση φυλετικής μεροληψίας σε αλγορίθμους πρόβλεψης κινδύνου για σακχαρώδη διαβήτη τύπου 2

Σύμφωνα με τους Cronjé et al. (2023), όσον αφορά τον πληθυσμό των ΗΠΑ, παρά τον συγκριτικά χαμηλότερο κίνδυνο, οι μη ισπανόφωνοι Λευκοί παραμένουν υπερεκπροσωπούμενοι στη βιβλιογραφία για την πρόβλεψη κινδύνου διαβήτη. Σε μια διαφορετική ανασκόπηση σχετικά με τη φυλετική/εθνοτική ισότητα στην τεχνητή νοημοσύνη για τη διαχείριση του διαβήτη, στα άρθρα που ανέφεραν φυλή, η μέση κατανομή ήταν 69,5% Λευκοί, 17,1% Μαύροι και 3,7% Ασιάτες, ενώ μόνο 2 άρθρα ανέφεραν τη συμπερίληψη ιθαγενών Αμερικανών συμμετεχόντων/συμμετεχουσών (Pham et al., 2021).

Είναι ευρέως τεκμηριωμένο ότι οι ανισότητες στα αποτελέσματα του διαβήτη οφείλονται σε μεγάλο βαθμό σε σύνθετους και αλληλένδετους κοινωνικούς προσδιοριστές της υγείας, συμπεριλαμβανομένης της πρόσβασης σε υγιεινή διατροφή, της ποιοτικής υγειονομικής περίθαλψης, της ασφαλιστικής κάλυψης, των εκπαιδευτικών εμποδίων και των διαφορετικών ποσοστών χρήσης της τεχνολογίας. Τα αποτελέσματα αυτά περιλαμβάνουν υψηλότερα ποσοστά επιπλοκών και χειρότερο γλυκαιμικό έλεγχο μεταξύ μειονοτικών και χαμηλού εισοδήματος πληθυσμών (Alipour & Alipour, 2025).

Ως αποτέλεσμα, ένα σύστημα τεχνητής νοημοσύνης που εκπαιδεύεται σε υπάρχοντα σύνολα δεδομένων ενδέχεται να παρουσιάζει περιορισμένη γενίκευση, οδηγώντας σε μεροληπτικά προγνωστικά μοντέλα που μπορεί να ευνοούν άτομα συγκεκριμένων φυλετικών ομάδων, για παράδειγμα, σε προληπτικές παρεμβάσεις.

Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης. Οι απόψεις και οι γνώμες που διατυπώνονται εκφράζουν αποκλειστικά τις απόψεις των συντακτών και δεν αντιπροσωπεύουν κατ'ανάγκη τις απόψεις της Ευρωπαϊκής Ένωσης ή του Ευρωπαϊκού Εκτελεστικού Οργανισμού Εκπαίδευσης και Πολιτισμού (EACEA). Η Ευρωπαϊκή Ένωση και ο EACEA δεν μπορούν να θεωρηθούν υπεύθυνοι για τις εκφραζόμενες απόψεις. Κωδικός έργου: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Μεροληψία μέτρησης

Η μεροληψία μέτρησης προκύπτει κατά την επιλογή, συλλογή ή τον υπολογισμό των χαρακτηριστικών και των ετικετών που χρησιμοποιούνται σε ένα πρόβλημα πρόβλεψης, ιδίως όταν χρησιμοποιείται ένας πληρεξούσιος δείκτης (*proxy*), δηλαδή η προσέγγιση μιας έννοιας που δεν κωδικοποιείται ή δεν είναι άμεσα παρατηρήσιμη. Ένα παράδειγμα παρουσιάζεται σε μελέτη των Obermeyer et al., (2019), όπου το κόστος υγειονομικής περίθαλψης χρησιμοποιήθηκε ως πληρεξούσιος δείκτης για την πρόβλεψη και ιεράρχηση των ασθενών που θα ωφελούνταν περισσότερο από πρόσθετη φροντίδα, ο οποίος οδήγησε σε φυλετικές διακρίσεις. Ως εκ τούτου, το κόστος υγείας αποτελεί ακατάλληλο δείκτη των πραγματικών αναγκών υγείας, καθώς οι Μαύροι ασθενείς, αντιμετωπίζοντας δυσανάλογα υψηλά επίπεδα φτώχειας, συχνά δαπανούν λιγότερα για υπηρεσίες υγείας σε σύγκριση με τους Λευκούς. Εξαιτίας αυτής της μεροληψίας, ο αλγόριθμος κατέληξε λανθασμένα στο συμπέρασμα ότι οι Μαύροι ασθενείς ήταν υγιέστεροι από εξίσου ασθενείς Λευκούς, με αποτέλεσμα να τους αποδίδεται χαμηλότερη προτεραιότητα κατά την πρόσβαση σε υπηρεσίες υγείας.

Άλλες πηγές μεροληψίας μέτρησης μπορεί να προκύψουν όταν η μέθοδος μέτρησης διαφέρει μεταξύ ομάδων, για παράδειγμα όταν δύο ομάδες παρακολουθούνται για την ίδια συμπεριφορά, αλλά η μία παρακολουθείται αυστηρότερα ή συχνότερα από την άλλη. Παρομοίως, η ακρίβεια της μέτρησης μπορεί να διαφέρει μεταξύ ομάδων, γεγονός που σε ιατρικές εφαρμογές μπορεί να οδηγήσει σε συστηματικά υψηλότερα ποσοστά λανθασμένης διάγνωσης ή σε υποδιάγνωση σε ορισμένες ομάδες. Για παράδειγμα, οι γιατροί είναι πιθανότερο να υποτιμούν τον πόνο των Μαύρων ασθενών σε σύγκριση με μη Μαύρους ασθενείς, λόγω εσφαλμένων πεποιθήσεων σχετικά με βιολογικές διαφορές μεταξύ Μαύρων και Λευκών, με αποτέλεσμα οι Μαύροι ασθενείς να είναι λιγότερο πιθανό να λάβουν αναλγητική αγωγή και, ακόμη και όταν λάβουν, να τους χορηγούνται μικρότερες δόσεις (Hoffman et al., 2016).

Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης. Οι απόψεις και οι γνώμες που διατυπώνονται εκφράζουν αποκλειστικά τις απόψεις των συντακτών και δεν αντιπροσωπεύουν κατ'ανάγκη τις απόψεις της Ευρωπαϊκής Ένωσης ή του Ευρωπαϊκού Εκτελεστικού Οργανισμού Εκπαίδευσης και Πολιτισμού (EACEA). Η Ευρωπαϊκή Ένωση και ο EACEA δεν μπορούν να θεωρηθούν υπεύθυνοι για τις εκφραζόμενες απόψεις. Κωδικός έργου: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Μελέτη περίπτωσης: Φυλετικές και εθνοτικές διαφορές στη συσχέτιση μεταξύ μέσης γλυκόζης και αιμοσφαιρίνης A1c

Η εξέταση A1C μετρά τη μέση ποσότητα γλυκόζης (σακχάρου) στο αίμα και χρησιμοποιείται για την ανίχνευση προδιαβήτη ή για τη διάγνωση σακχαρώδους διαβήτη τύπου 2. Ωστόσο, η A1C αποτελεί μόνο έμμεσο δείκτη και δεν συνδέεται αιτιωδώς με τα κλινικά αποτελέσματα υγείας, καθώς υπάρχουν πολλοί τρόποι με τους οποίους μπορεί να μεταβληθεί η σχέση μεταξύ των άμεσων μετρήσεων της γλυκαιμίας (δηλαδή της συγκέντρωσης γλυκόζης στο αίμα) και της A1C. Παρατηρείται μάλιστα σημαντική διακύμανση στη σχέση γλυκαιμίας–A1C τόσο μεταξύ διαφορετικών ατόμων όσο και στο ίδιο άτομο με την πάροδο του χρόνου. Επιπλέον, μελέτες έχουν αναφέρει σημαντικά υψηλότερα επίπεδα αιμοσφαιρίνης A1C σε Αφροαμερικανούς ασθενείς σε σύγκριση με Λευκούς ασθενείς με την ίδια μέση γλυκόζη (Karter et al., 2023).

Εάν ένα σύστημα τεχνητής νοημοσύνης που έχει σχεδιαστεί για τη διάγνωση του διαβήτη εκπαιδευτεί να χρησιμοποιεί τα αποτελέσματα της εξέτασης A1C ως πληρεξούσιο δείκτη της γλυκαιμίας, χωρίς να λαμβάνει υπόψη άλλους παράγοντες, όπως η φυλή του/της ασθενούς, αυτό μπορεί να οδηγήσει σε πρόωρες διαγνώσεις διαβήτη και ακατάλληλη θεραπευτική αντιμετώπιση, με αποτέλεσμα μεροληπτική ποιότητα φροντίδας και ανισότητες στην υγεία. Ωστόσο, όπως επισημαίνεται στους Alipour & Alipour (2025), σε μια συστηματική ανασκόπηση των μεροληψιών που ενδέχεται να επηρεάζουν την ισότητα των μοντέλων TN/MM στον διαβήτη (συμπεριλαμβανομένης της μεροληψίας μέτρησης), παρότι οι μελέτες που εξετάστηκαν αναφέρουν ρητά ότι η μεροληψία μέτρησης μπορεί να μεταφερθεί και να ενισχυθεί μέσω των μοντέλων τεχνητής νοημοσύνης εάν δεν διορθωθεί, καθώς καμία από αυτές δεν έλαβε υπόψη τέτοιες μεροληψίες κατά την ανάπτυξη του μοντέλου, ούτε τις αντιμετώπισε ρητά ή ανέφερε διόρθωση διαφορών στην ακρίβεια των μετρήσεων.

Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης. Οι απόψεις και οι γνώμες που διατυπώνονται εκφράζουν αποκλειστικά τις απόψεις των συντακτών και δεν αντιπροσωπεύουν κατ'ανάγκη τις απόψεις της Ευρωπαϊκής Ένωσης ή του Ευρωπαϊκού Εκτελεστικού Οργανισμού Εκπαίδευσης και Πολιτισμού (EACEA). Η Ευρωπαϊκή Ένωση και ο EACEA δεν μπορούν να θεωρηθούν υπεύθυνοι για τις εκφραζόμενες απόψεις. Κωδικός έργου: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Μεροληψία συγκέντρωσης

Η μεροληψία συγκέντρωσης προκύπτει όταν εφαρμόζεται ένα ενιαίο μοντέλο (*"one-size-fits-all")* σε ένα σύνολο δεδομένων που περιλαμβάνει ετερογενείς ομάδες ανθρώπων ή αντικειμένων. Ένα χαρακτηριστικό παράδειγμα είναι η αντιστοίχιση δεδομένων εισόδου (π.χ. το εισόδημα ενός ατόμου) σε κατηγορίες/ετικέτες που το περιγράφουν (π.χ. χαμηλό, μεσαίο, υψηλό), με την υπόθεση ότι η ερμηνεία αυτών των κατηγοριών είναι συνεπής σε όλα τα υποσύνολα των δεδομένων. Στην πράξη, όμως, το κοινωνικό, πολιτισμικό ή γεωγραφικό υπόβαθρο ενός ατόμου μπορεί να μεταβάλλει το πραγματικό νόημα αυτών των αριθμών. Για παράδειγμα, ένα «υψηλό» εισόδημα σε μια μικρή αγροτική πόλη ή σε μια χώρα χαμηλού ή μεσαίου εισοδήματος μπορεί να έχει εντελώς διαφορετική αγοραστική δύναμη και κοινωνική σημασία σε σύγκριση με ένα «υψηλό» εισόδημα σε μια μεγάλη μητρόπολη ή σε μια χώρα υψηλού εισοδήματος. Όταν ένα μοντέλο αγνοεί αυτές τις διαφοροποιήσεις και εφαρμόζει ενιαία όρια ή κανόνες, ενδέχεται να παράγει συστηματικά λανθασμένες ή άδικες προβλέψεις για ορισμένες ομάδες.

Μελέτη περίπτωσης: Ψηφιακά εργαλεία υγείας για την παθητική παρακολούθηση της κατάθλιψης

Η χρήση ψηφιακών εργαλείων για τη μέτρηση φυσιολογικών και συμπεριφορικών μεταβλητών με σκοπό την παθητική παρακολούθηση της κατάθλιψης εξετάζεται από τους De Angel et al., (2022), σε μια συστηματική ανασκόπηση του θέματος. Στο πλαίσιο αυτό εξετάστηκαν άρθρα που διερευνούν συσχετίσεις μεταξύ της κατάθλιψης και αντικειμενικών συμπεριφορικών δεδομένων που συλλέχθηκαν από αισθητήρες έξυπνων τηλεφώνων και φορητών συσκευών. Τα δεδομένα αυτά μετατράπηκαν σε χαρακτηριστικά (*features*) που χρησιμοποιήθηκαν από μοντέλα τεχνητής νοημοσύνης για την πραγματοποίηση προβλέψεων, και αντιστοιχούσαν σε δείκτες όπως ο ύπνος, η

Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης. Οι απόψεις και οι γνώμες που διατυπώνονται εκφράζουν αποκλειστικά τις απόψεις των συντακτών και δεν αντιπροσωπεύουν κατ'ανάγκη τις απόψεις της Ευρωπαϊκής Ένωσης ή του Ευρωπαϊκού Εκτελεστικού Οργανισμού Εκπαίδευσης και Πολιτισμού (EACEA). Η Ευρωπαϊκή Ένωση και ο EACEA δεν μπορούν να θεωρηθούν υπεύθυνοι για τις εκφραζόμενες απόψεις. Κωδικός έργου: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

σωματική δραστηριότητα, ο κιρκάδιος ρυθμός, η κοινωνικότητα, η τοποθεσία και η χρήση του τηλεφώνου.

Ωστόσο, οι συγγραφείς υπογραμμίζουν την ετερογένεια που προκύπτει από την ποικιλία των μεθόδων που χρησιμοποιούνται για τη δημιουργία αυτών των χαρακτηριστικών. Για παράδειγμα, το χαρακτηριστικό «ποιότητα ύπνου» μπορεί να οριστεί με βάση τον αριθμό αφυπνίσεων, το συνολικό χρόνο εγρήγορσης ή το ποσοστό χρόνου ύπνου σε σχέση με τον χρόνο αφύπνισης κατά τη διάρκεια μιας συνεδρίας ύπνου. Επιπλέον, πρέπει να λαμβάνονται υπόψη οι διαφορές στον τρόπο με τον οποίο οι αισθητήρες διαφορετικών συσκευών καταγράφουν και ορίζουν ένα επεισόδιο ως «ύπνο». Όταν αυτές οι διαφοροποιήσεις δεν εξετάζονται ξεχωριστά αλλά συγχωνεύονται συνολικά υπό τον όρο «ποιότητα ύπνου», και εφόσον ένα σύνολο δεδομένων μπορεί να περιλαμβάνει άτομα ή ομάδες με διαφορετικά κοινωνικά, πολιτισμικά ή συμπεριφορικά πρότυπα, είναι σαφές ότι το συγκεκριμένο χαρακτηριστικό ενδέχεται να έχει διαφορετική σημασία για κάθε ομάδα ή άτομο.

Η συγκέντρωση (*aggregation*) τέτοιων δεδομένων σε ένα ενιαίο χαρακτηριστικό μπορεί να οδηγήσει σε ένα σύστημα το οποίο είτε δεν ανταποκρίνεται επαρκώς σε καμία ομάδα, είτε ευνοεί την κυρίαρχη πληθυσμιακή ομάδα, ιδίως όταν συνυπάρχει και μεροληψία αναπαράστασης. Για παράδειγμα, υπάρχουν ενδείξεις διαφορών στον ύπνο μεταξύ ανδρών και γυναικών, ενώ οι γυναίκες συχνά υποεκπροσωπούνται στην έρευνα για τον ύπνο. Επιπλέον, άλλοι παράγοντες που συνήθως δεν λαμβάνονται υπόψη στις μελέτες για τα πρότυπα και τις διαταραχές ύπνου περιλαμβάνουν τη μη διάκριση μεταξύ φύλου ως κοινωνικής κατασκευής και βιολογικού φύλου, καθώς και τη μη συνεκτίμηση διαθεματικών ταυτοτήτων που προσδιορίζονται από τη διασταύρωση ηλικίας, φυλής και κοινωνικοοικονομικής τάξης (Lok et al., 2024).

Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης. Οι απόψεις και οι γνώμες που διατυπώνονται εκφράζουν αποκλειστικά τις απόψεις των συντακτών και δεν αντιπροσωπεύουν κατ'ανάγκη τις απόψεις της Ευρωπαϊκής Ένωσης ή του Ευρωπαϊκού Εκτελεστικού Οργανισμού Εκπαίδευσης και Πολιτισμού (EACEA). Η Ευρωπαϊκή Ένωση και ο EACEA δεν μπορούν να θεωρηθούν υπεύθυνοι για τις εκφραζόμενες απόψεις. Κωδικός έργου: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Μεροληψία μάθησης

Η μεροληψία μάθησης προκύπτει όταν οι επιλογές μοντελοποίησης ενισχύουν τις διαφορές στην απόδοση μεταξύ διαφορετικών παραδειγμάτων μέσα στα δεδομένα. Ένα παράδειγμα αφορά τη διαφορική ιδιωτικότητα (*differential privacy*), έναν μηχανισμό που χρησιμοποιείται σε συστήματα τεχνητής νοημοσύνης και διασφαλίζει ότι εξετάζοντας την έξοδο ενός συστήματος, δεν είναι δυνατόν να προσδιοριστεί αν τα δεδομένα ενός συγκεκριμένου ατόμου περιλαμβάνονταν στο αρχικό σύνολο δεδομένων. Η διαφορική ιδιωτικότητα χρησιμοποιείται σε σύνολα δεδομένων υγείας για την προστασία ευαίσθητων πληροφοριών των ασθενών, για παράδειγμα στην περίπτωση σπάνιων ασθενειών, όπου η περίπτωση κάθε ασθενούς είναι λίγο-πολύ μοναδική σε μια περιορισμένη περιοχή που καλύπτεται από ένα νοσοκομείο· έτσι, ακόμη και αν τα δεδομένα είναι ανωνυμοποιημένα, δεν είναι ιδιαίτερα δύσκολο να συναχθεί η ταυτότητα του ατόμου.

Ωστόσο, έχει αποδειχθεί ότι η διαφορική ιδιωτικότητα μειώνει την επίδραση των υποεκπροσωπούμενων δεδομένων στο μοντέλο· συνεπώς, εάν το σύστημα τεχνητής νοημοσύνης είναι εξαρχής μεροληπτικό, η εφαρμογή ενός μέτρου ενίσχυσης της ιδιωτικότητας επιτείνει ακόμη περισσότερο αυτή τη μεροληψία (Bagdasaryan & Shmatikov, 2019).

Μελέτη περίπτωσης: Διαφορική ιδιωτικότητα και ανισότητες στην υγεία

Τον Σεπτέμβριο του 2018, το US Census Bureau ανακοίνωσε ότι θα εφαρμόσει τη διαφορική ιδιωτικότητα σε προϊόντα δεδομένων που προέρχονται από τα στοιχεία της απογραφής του 2020. Ωστόσο, οι Santos-Lozada κ.ά., (2020), διερεύνησαν τον τρόπο με τον οποίο η εφαρμογή της διαφορικής ιδιωτικότητας μπορεί να μεταβάλει τη γνώση σχετικά με τις ανισότητες στην υγεία ως προς τη θνησιμότητα, ιδίως για φυλετικές ή εθνοτικές μειονότητες σε μικρές περιοχές και λιγότερο αστικοποιημένα περιβάλλοντα. Τα αποτελέσματά τους έδειξαν ότι η διαφορική ιδιωτικότητα επηρεάζει εντονότερα τις

Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης. Οι απόψεις και οι γνώμες που διατυπώνονται εκφράζουν αποκλειστικά τις απόψεις των συντακτών και δεν αντιπροσωπεύουν κατ'ανάγκη τις απόψεις της Ευρωπαϊκής Ένωσης ή του Ευρωπαϊκού Εκτελεστικού Οργανισμού Εκπαίδευσης και Πολιτισμού (EACEA). Η Ευρωπαϊκή Ένωση και ο EACEA δεν μπορούν να θεωρηθούν υπεύθυνοι για τις εκφραζόμενες απόψεις. Κωδικός έργου: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

εκτιμήσεις των ποσοστών θνησιμότητας για τους μη Ισπανόφωνους Μαύρους και τους Ισπανόφωνους, σε σύγκριση με τις αντίστοιχες εκτιμήσεις για τους μη Ισπανόφωνους Λευκούς.

Τα ευρήματα αυτά υποστηρίχθηκαν από τους Kurz κ.ά., (2022), οι οποίοι έδειξαν ότι η εφαρμογή της διαφορικής ιδιωτικότητας στα ίδια δεδομένα μπορεί να οδηγήσει σε εσφαλμένη αποτύπωση των ποσοστών συμμετοχής στο Medicaid – το κρατικό πρόγραμμα υγειονομικής περίθαλψης για άτομα με χαμηλά εισοδήματα στις ΗΠΑ - μεταξύ ήδη περιθωριοποιημένων φυλετικών και εθνοτικών ομάδων. Συγκεκριμένα, τα ποσοστά αυτά για ορισμένους συνδυασμούς κομητείας, φυλής και εθνοτικής ταυτότητας διέφεραν μεταξύ των αποτελεσμάτων με διαφορική ιδιωτικότητα και των αρχικών δεδομένων, σε ορισμένες περιπτώσεις υπερβαίνοντας το 10%. Επιπλέον, τα μη Ισπανόφωνα Λευκά άτομα ήταν η μόνη φυλετική και εθνοτική υποομάδα για την οποία ο αλγόριθμος διαφορικής ιδιωτικότητας κατέγραψε με ακρίβεια τα ποσοστά συμμετοχής στο Medicaid. Το εύρημα αυτό ενδέχεται να έχει σημαντικές επιπτώσεις για την πολιτική υγείας, καθώς τα δεδομένα της απογραφής χρησιμοποιούνται για τον σχεδιασμό κυβερνητικών προγραμμάτων, την κατανομή πόρων, καθώς και για την αξιολόγηση και παρακολούθηση πολιτικών.

Μεροληψία αξιολόγησης

Η μεροληψία αξιολόγησης προκύπτει όταν τα δεδομένα αναφοράς (*benchmarks*) που χρησιμοποιούνται για μια συγκεκριμένη εργασία δεν αντιπροσωπεύουν τον πληθυσμό στον οποίο θα εφαρμοστεί το σύστημα. Τα benchmarks είναι τυποποιημένα σύνολα δεδομένων που χρησιμοποιούνται για τη μέτρηση της ποιότητας ενός μοντέλου, επιτρέποντας την ποσοτική σύγκριση μεταξύ διαφορετικών μοντέλων. Κατά συνέπεια, υπάρχει ο κίνδυνος να ενθαρρύνεται η ανάπτυξη και η εφαρμογή μοντέλων που αποδίδουν καλά μόνο στο υποσύνολο των δεδομένων που εκπροσωπείται στο benchmark. Έτσι, μπορεί να προκύψει διάκριση σε βάρος ευάλωτων υποομάδων ή

Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης. Οι απόψεις και οι γνώμες που διατυπώνονται εκφράζουν αποκλειστικά τις απόψεις των συντακτών και δεν αντιπροσωπεύουν κατ'ανάγκη τις απόψεις της Ευρωπαϊκής Ένωσης ή του Ευρωπαϊκού Εκτελεστικού Οργανισμού Εκπαίδευσης και Πολιτισμού (EACEA). Η Ευρωπαϊκή Ένωση και ο EACEA δεν μπορούν να θεωρηθούν υπεύθυνοι για τις εκφραζόμενες απόψεις. Κωδικός έργου: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

ατόμων, εάν το benchmark υπόκειται σε ιστορική, αναπαραστατική μεροληψία ή σε μεροληψία μέτρησης.

Στον τομέα της υγείας, οι λόγοι υποεκπροσώπησης συγκεκριμένων πληθυσμών στα σύνολα δεδομένων μπορεί να οφείλονται είτε στην απουσία ατόμων ή ομάδων από τα δεδομένα (για παράδειγμα, έγκυες γυναίκες, λόγω ηθικών περιορισμών), είτε στο ότι τα άτομα κατηγοριοποιούνται εσφαλμένα ή ακατάλληλα σε ομάδες (π.χ. κατηγορίες όπως «μικτή εθνοτική καταγωγή» ή «άλλο»). Οι βαθύτερες αιτίες μπορεί να περιλαμβάνουν κοινωνικούς, τεχνικούς ή νομικούς/ηθικούς λόγους, όπως δομικά εμπόδια στην πρόσβαση σε υπηρεσίες υγείας, τεχνικά εμπόδια στη συλλογή ή ψηφιοποίηση σχετικών δεδομένων υγείας, ατομικούς και θεσμικούς περιορισμούς σχετικά με τη συναίνεση για διαμοιρασμό δεδομένων, καθώς και νομικούς ή ηθικούς περιορισμούς που εμποδίζουν την προσβασιμότητα των δεδομένων, μεταξύ άλλων (Arora et al., 2023). Το αποτέλεσμα είναι ότι τα συστήματα τεχνητής νοημοσύνης που έχουν βαθμολογηθεί με βάση τέτοια benchmarks ενδέχεται να παρουσιάζουν χαμηλότερη απόδοση όταν εφαρμόζονται σε άτομα από υποεκπροσωπούμενες ομάδες. Ωστόσο, είναι σημαντικό να σημειωθεί ότι η εγκυρότητα των benchmarks αποτελεί ένα ευρύτερο ζήτημα και δεν περιορίζεται αποκλειστικά στη μεροληψία (Brooks, 2025).

Μελέτη περίπτωσης: Σύνολα δεδομένων δερματικών εικόνων

Τα σύνολα δεδομένων απεικόνισης δέρματος υποεκπροσωπούν ορισμένες δημογραφικές ομάδες, καθώς οι περισσότερες εικόνες σε αυτά προέρχονται από πληθυσμούς της Βόρειας Αμερικής ή της Ευρώπης και απεικονίζουν κυρίως άτομα με ανοιχτόχρωμο δέρμα (Guo et al., 2021). Λόγω του υψηλού κόστους και της δυσκολίας κατασκευής αυτών των συνόλων δεδομένων, πέρα από την εκπαίδευση μοντέλων, μπορούν επίσης να χρησιμοποιηθούν και ως benchmarks (σύνολα δεδομένων αναφοράς).

Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης. Οι απόψεις και οι γνώμες που διατυπώνονται εκφράζουν αποκλειστικά τις απόψεις των συντακτών και δεν αντιπροσωπεύουν κατ'ανάγκη τις απόψεις της Ευρωπαϊκής Ένωσης ή του Ευρωπαϊκού Εκτελεστικού Οργανισμού Εκπαίδευσης και Πολιτισμού (EACEA). Η Ευρωπαϊκή Ένωση και ο EACEA δεν μπορούν να θεωρηθούν υπεύθυνοι για τις εκφραζόμενες απόψεις. Κωδικός έργου: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Η μελέτη περίπτωσης που αναδεικνύει την αναδυόμενη μεροληψία (βλ. παραπάνω) — δηλαδή τα σύνολα εικόνων καρκίνου του δέρματος που χρησιμοποιούνται για την εκπαίδευση μοντέλων πρόβλεψης— αποτελεί παράδειγμα ακατάλληλου benchmark όταν ο πληθυσμός χρηστών/χρηστριών προέρχεται από υποεκπροσωπούμενες ομάδες (Guo et al., 2021). Ένα παρόμοιο παράδειγμα, αν και δεν σχετίζεται με την τεχνητή νοημοσύνη, καταδεικνύει τη γενικότητα του προβλήματος και αφορά τα οξύμετρα Pulse (συσκευές που μετρούν τον κορεσμό οξυγόνου στο αίμα και χρησιμοποιούνται, για παράδειγμα, σε περιπτώσεις καρδιακής προσβολής ή καρδιακής ανεπάρκειας), τα οποία έχει αποδειχθεί ότι λειτουργούν με μεγαλύτερη ακρίβεια σε άτομα με ανοιχτόχρωμη χρωστική δέρματος (Sjoding et al., 2020).

Η μεροληψία αναπαράστασης, μέτρησης, συγκέντρωσης (*aggregation*), μάθησης και αξιολόγησης μπορεί να χαρτογραφηθεί στην τεχνική μεροληψία (βλ. παραπάνω), όπως αυτή ορίζεται από τους Friedman & Nissenbaum (1996).

Μεροληψία κατά την ανάπτυξη

Η μεροληψία κατά την ανάπτυξη προκύπτει όταν υπάρχει αναντιστοιχία μεταξύ του προβλήματος το οποίο έχει σχεδιαστεί να επιλύει ένα μοντέλο και του τρόπου με τον οποίο χρησιμοποιείται στην πράξη. Η αναντιστοιχία αυτή μπορεί να προκαλέσει βλάβη, ιδίως όταν συνδυάζεται με γνωστικές μεροληψίες, όπως η μεροληψία επιβεβαίωσης (*confirmation bias*) και η μεροληψία αυτοματοποίησης (*automation bias*). Η μεροληψία κατά την ανάπτυξη ταυτίζεται με την αναδυόμενη μεροληψία (*emergent bias* – βλ. παραπάνω), όπως αυτή ορίζεται από τους Friedman & Nissenbaum (1996).

Μελέτη περίπτωσης: Μετατόπιση πεδίου

Η περίπτωση της μετατόπισης δεδομένων (*data shift*) τεκμηριώνεται στην υποενότητα περί αναδυόμενης μεροληψίας σχετικά με την ανίχνευση καρκίνου του δέρματος.

Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης. Οι απόψεις και οι γνώμες που διατυπώνονται εκφράζουν αποκλειστικά τις απόψεις των συντακτών και δεν αντιπροσωπεύουν κατ'ανάγκη τις απόψεις της Ευρωπαϊκής Ένωσης ή του Ευρωπαϊκού Εκτελεστικού Οργανισμού Εκπαίδευσης και Πολιτισμού (EACEA). Η Ευρωπαϊκή Ένωση και ο EACEA δεν μπορούν να θεωρηθούν υπεύθυνοι για τις εκφραζόμενες απόψεις. Κωδικός έργου: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Επιπλέον, μπορούμε να ορίσουμε την περίπτωση της μετατόπισης πεδίου (*domain shift*), η οποία προκύπτει όταν ένα σύστημα έχει υλοποιηθεί, έχει λάβει κανονιστική έγκριση και έχει αναπτυχθεί στην κλινική πράξη, αλλά εφαρμόζεται σε διαφορετική ομάδα ασθενών από εκείνη για την οποία είχε αρχικά εκπαιδευτεί. Για παράδειγμα, ένα σύστημα μπορεί να αναπτυχθεί για νοσοκομείο σε χώρα υψηλού εισοδήματος και να εφαρμοστεί σε χώρα χαμηλού ή μεσαίου εισοδήματος χωρίς να λαμβάνονται υπόψη παράγοντες όπως τα κοινωνικο-δημογραφικά χαρακτηριστικά των ασθενών ή το κατά πόσο οι ασθενείς έχουν το ίδιο συνολικό επίπεδο κινδύνου σε σύγκριση με εκείνες/εκείνους που περιλαμβάνονταν στα δεδομένα εκπαίδευσης (Vokinger et al., 2021).

Επιπτώσεις για τη χάραξη πολιτικής

Τα στοιχεία που χαρτογραφήθηκαν στο Παραδοτέο D2.1 καταδεικνύουν ότι οι έμφυλες και φυλετικές μεροληψίες στη βιοϊατρική τεχνητή νοημοσύνη δεν αποτελούν τυχαία ή μεμονωμένα τεχνικά σφάλματα, αλλά συστημικούς κινδύνους που αναδύονται σε ολόκληρο τον κύκλο ζωής των συστημάτων ΤΝ που χρησιμοποιούνται στην υγειονομική περίθαλψη. Στα καρδιαγγειακά νοσήματα, την κατάθλιψη και τον διαβήτη, η μεροληψία προκύπτει από ιστορικά στρεβλωμένα κλινικά σύνολα δεδομένων, άνισες διαγνωστικές πρακτικές, μεταβλητές-υποκατάστατα που ενσωματώνουν διαρθρωτικές ανισότητες, καθώς και από πλαίσια εφαρμογής που κατανέμουν άνισα τόσο τα οφέλη όσο και τις βλάβες. Τα ευρήματα αυτά επιβεβαιώνουν ότι η βιοϊατρική ΤΝ εμπλέκεται άμεσα με πολλαπλά δικαιώματα και αρχές που προστατεύονται από τον Χάρτη Θεμελιωδών Δικαιωμάτων της Ευρωπαϊκής Ένωσης, ιδίως τις αρχές της ανθρώπινης αξιοπρέπειας, της ισότητας ενώπιον του νόμου και της απαγόρευσης των διακρίσεων, καθώς και το δικαίωμα στην ακεραιότητα του προσώπου, το δικαίωμα στην υγειονομική περίθαλψη, την προστασία δεδομένων και το δικαίωμα πραγματικής προσφυγής.

Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης. Οι απόψεις και οι γνώμες που διατυπώνονται εκφράζουν αποκλειστικά τις απόψεις των συντακτών και δεν αντιπροσωπεύουν κατ'ανάγκη τις απόψεις της Ευρωπαϊκής Ένωσης ή του Ευρωπαϊκού Εκτελεστικού Οργανισμού Εκπαίδευσης και Πολιτισμού (EACEA). Η Ευρωπαϊκή Ένωση και ο EACEA δεν μπορούν να θεωρηθούν υπεύθυνοι για τις εκφραζόμενες απόψεις. Κωδικός έργου: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Υπό το πρίσμα αυτό, τα ενωσιακά και εθνικά πλαίσια πολιτικής που διέπουν την ΤΝ στην υγειονομική περίθαλψη οφείλουν να αντιμετωπίζουν τον μετριασμό της μεροληψίας όχι ως προαιρετική ηθική προσθήκη, αλλά ως δεσμευτικό στοιχείο νόμιμης και σύμφωνης με τα θεμελιώδη δικαιώματα ανάπτυξης και χρήσης συστημάτων ΤΝ. Οι ευρωπαϊκές και εθνικές ρυθμιστικές πρωτοβουλίες για την ΤΝ στην υγεία θα πρέπει να ιδωθούν στο πλαίσιο του ευρύτερου συστήματος προστασίας θεμελιωδών δικαιωμάτων που διέπει την ΤΝ (βλ. Novossiolova, 2025· Novossiolova et al., 2025· Kasapi, 2025).

Ο Κανονισμός για την Τεχνητή Νοημοσύνη της ΕΕ παρέχει ένα αναγκαίο ρυθμιστικό θεμέλιο, καθώς κατατάσσει τα περισσότερα συστήματα βιοϊατρικής ΤΝ ως συστήματα υψηλού κινδύνου. Ωστόσο, η αποτελεσματικότητά του στην πράξη θα εξαρτηθεί από το πώς οι εγγυήσεις θεμελιωδών δικαιωμάτων θα εφαρμοστούν στις διαδικασίες αξιολόγησης της συμμόρφωσης, στην παρακολούθηση μετά τη διάθεση στην αγορά και στις δημόσιες προμήθειες.

Καταρχάς, οι εγγυήσεις για ουσιαστική ανθρώπινη εποπτεία πρέπει να ενισχυθούν και να εξειδικευθούν για τα συστήματα βιοϊατρικής ΤΝ σε ολόκληρο τον κύκλο ζωής τους. Τα κλινικά εργαλεία ΤΝ που χρησιμοποιούνται για διάγνωση, διαστρωμάτωση κινδύνου, στον προσυμπτωματικό έλεγχο ή στην υποστήριξη θεραπείας δεν θα πρέπει σε καμία περίπτωση να λειτουργούν ως *de facto* αυτόνομοι λήπτες αποφάσεων. Η ανθρώπινη εποπτεία πρέπει να περιλαμβάνει όχι μόνο τη δυνατότητα παρέμβασης ή παράκαμψης (override) από επαγγελματίες υγείας, αλλά και σαφή θεσμική ευθύνη για την κατανόηση των περιορισμών του συστήματος, των γνωστών κινδύνων μεροληψίας και των αποκλίσεων απόδοσης μεταξύ των διαφορετικών υποομάδων. Σε συμφωνία με την προστασία της ανθρώπινης αξιοπρέπειας και της ακεραιότητας που κατοχυρώνεται στον Χάρτη Θεμελιωδών Δικαιωμάτων της ΕΕ, οι επαγγελματίες υγείας θα πρέπει να εκπαιδεύονται και να υποστηρίζονται θεσμικά ώστε να εξετάζουν κριτικά τα αποτελέσματα της ΤΝ, αντί να τα αποδέχονται άκριτα. Αυτό προϋποθέτει την

Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης. Οι απόψεις και οι γνώμες που διατυπώνονται εκφράζουν αποκλειστικά τις απόψεις των συντακτών και δεν αντιπροσωπεύουν κατ'ανάγκη τις απόψεις της Ευρωπαϊκής Ένωσης ή του Ευρωπαϊκού Εκτελεστικού Οργανισμού Εκπαίδευσης και Πολιτισμού (EACEA). Η Ευρωπαϊκή Ένωση και ο EACEA δεν μπορούν να θεωρηθούν υπεύθυνοι για τις εκφραζόμενες απόψεις. Κωδικός έργου: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

ενσωμάτωση της παιδείας για την ΤΝ, της επίγνωσης της μεροληψίας και της εκπαίδευσης στα θεμελιώδη δικαιώματα στην ιατρική εκπαίδευση και στη συνεχιζόμενη επαγγελματική κατάρτιση.

Οι υποχρεώσεις διαφάνειας θα πρέπει να ερμηνεύονται διασταλτικά στο πλαίσιο της υγειονομικής περίθαλψης. Οι ασθενείς και οι χρήστριες/χρήστες υπηρεσιών υγείας πρέπει να ενημερώνονται κάθε φορά που χρησιμοποιούνται συστήματα ΤΝ στη λήψη κλινικών αποφάσεων που τις/τους αφορούν, συμπεριλαμβανομένων διαδικασιών προσυμπτωματικού ελέγχου, ιεράρχησης ή βαθμολόγησης κινδύνου. Όταν αποτελέσματα που παράγονται από ΤΝ αξιοποιούνται σε δημόσιες υπηρεσίες υγείας, θα πρέπει να είναι σαφώς αναγνωρίσιμα ως τέτοια και να συνοδεύονται από κατανοητές εξηγήσεις σχετικά με τον ρόλο τους, τους περιορισμούς τους και τους γνωστούς κινδύνους μεροληψίας. Τα άτομα θα πρέπει επίσης να ενημερώνονται όταν τα προσωπικά τους δεδομένα χρησιμοποιούνται για εκπαίδευση, δοκιμή ή συνεχή μάθηση συστημάτων ΤΝ, ιδίως όταν πρόκειται για ευαίσθητα δεδομένα υγείας.

Αυτά τα μέτρα διαφάνειας είναι ουσιώδη για τη διασφάλιση των δικαιωμάτων στην προστασία δεδομένων και στην πραγματική προσφυγή, όπως κατοχυρώνονται στον Χάρτη Θεμελιωδών Δικαιωμάτων της ΕΕ, και για να καθίσταται δυνατή η ουσιαστική αμφισβήτηση αποφάσεων που ενδέχεται να επηρεάζουν δυσμενώς τα άτομα.

Δεύτερον, η εκτίμηση επιπτώσεων στα θεμελιώδη δικαιώματα πρέπει να καταστεί μια τακτική και δεσμευτική απαίτηση για τα βιοϊατρικά συστήματα ΤΝ, επεκτεινόμενη πέρα από τους ελέγχους πριν τη διάθεση στην αγορά σε συνεχή αξιολόγηση κατά την ανάπτυξη και χρήση τους. Τα εμπειρικά στοιχεία του Παραδοτέου D2.1 δείχνουν ότι πολλές βλάβες λόγω μεροληψίας καθίστανται ορατές μόνο όταν τα συστήματα ΤΝ αλληλεπιδρούν με πραγματικούς πληθυσμούς και κλινικές ροές εργασίας, ιδίως μέσω διατομεακών επιδράσεων που αφορούν το φύλο, τη φυλή, την ηλικία και την κοινωνικοοικονομική κατάσταση. Οι εκτιμήσεις επιπτώσεων που αφορούν τα

Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης. Οι απόψεις και οι γνώμες που διατυπώνονται εκφράζουν αποκλειστικά τις απόψεις των συντακτών και δεν αντιπροσωπεύουν κατ'ανάγκη τις απόψεις της Ευρωπαϊκής Ένωσης ή του Ευρωπαϊκού Εκτελεστικού Οργανισμού Εκπαίδευσης και Πολιτισμού (EACEA). Η Ευρωπαϊκή Ένωση και ο EACEA δεν μπορούν να θεωρηθούν υπεύθυνοι για τις εκφραζόμενες απόψεις. Κωδικός έργου: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

δικαιώματα, όπως εκείνες που εμπνέονται από τη μεθοδολογία HUDERIA του Συμβουλίου της Ευρώπης - Μεθοδολογία για την Εκτίμηση Κινδύνου και Επιπτώσεων Συστημάτων Τεχνητής Νοημοσύνης από την άποψη των Ανθρωπίνων Δικαιωμάτων, της Δημοκρατίας και του Κράτους Δικαίου- θα πρέπει επομένως να είναι υποχρεωτικές για την ιατρική ΤΝ υψηλού κινδύνου, εξετάζοντας ρητά τη διαφοροποιημένη απόδοση και τα αποτελέσματα μεταξύ προστατευόμενων ομάδων. Οι εκτιμήσεις αυτές πρέπει να περιλαμβάνουν ουσιαστική συμμετοχή των ενδιαφερόμενων μερών, συμπεριλαμβανομένων των οργανώσεων της κοινωνίας των πολιτών, εκπροσώπων ασθενών και φορέων ισότητας, προκειμένου να αναδεικνύονται βλάβες που ενδέχεται να παραμένουν αόρατες από μια αμιγώς τεχνική ή κλινική οπτική.

Θα πρέπει να απαιτούνται περιοδικοί έλεγχοι (*audits*) των βιοϊατρικών συστημάτων ΤΝ, προκειμένου να επαληθεύεται η διαρκής συμμόρφωσή τους με τα πρότυπα θεμελιωδών δικαιωμάτων, με ιδιαίτερη προσοχή στη μετατόπιση μεροληψίας (*bias drift*), στις μεταβολές των συνόλων δεδομένων (*dataset shifts*) και στις αλλαγές της κλινικής χρήσης με την πάροδο του χρόνου. Όταν οι έλεγχοι αποκαλύπτουν επίμονες ή μη αντιμετωπίσιμες διακριτικές επιπτώσεις, πρέπει να υπάρχουν σαφείς νομικές και θεσμικές διαδικασίες για τον περιορισμό, την αναστολή ή τον τερματισμό της χρήσης του συστήματος. Το δικαίωμα στην υγεία δεν μπορεί να δικαιολογεί τη συνεχιζόμενη ανάπτυξη και χρήση εργαλείων ΤΝ που συστηματικά θέτουν ορισμένες ομάδες σε μειονεκτική θέση, ακόμη και εάν οι συνολικοί δείκτες απόδοσης εμφανίζονται ευνοϊκοί.

Τρίτον, οι αρχές της ΕΕ και των κρατών μελών οφείλουν να αντιμετωπίσουν τον κίνδυνο κακής χρήσης και δευτερογενών βλαβών που συνδέονται με τη βιοϊατρική ΤΝ. Αυτό περιλαμβάνει ευπάθειες κυβερνοασφάλειας που θα μπορούσαν να θέσουν σε κίνδυνο την ακεραιότητα των συστημάτων ή να επιτρέψουν κακόβουλη χειραγώγηση των κλινικών αποτελεσμάτων, καθώς και την επαναχρησιμοποίηση συστημάτων ΤΝ υγείας για σκοπούς επιτήρησης, σκιαγράφησης προφίλ ή πρακτικές που οδηγούν σε

Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης. Οι απόψεις και οι γνώμες που διατυπώνονται εκφράζουν αποκλειστικά τις απόψεις των συντακτών και δεν αντιπροσωπεύουν κατ'ανάγκη τις απόψεις της Ευρωπαϊκής Ένωσης ή του Ευρωπαϊκού Εκτελεστικού Οργανισμού Εκπαίδευσης και Πολιτισμού (EACEA). Η Ευρωπαϊκή Ένωση και ο EACEA δεν μπορούν να θεωρηθούν υπεύθυνοι για τις εκφραζόμενες απόψεις. Κωδικός έργου: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

αποκλεισμούς. Τα βιοϊατρικά συστήματα ΤΝ θα πρέπει να υπόκεινται σε τακτικές αξιολογήσεις ασφάλειας και σε αυστηρές υποχρεώσεις αναφοράς περιστατικών, με σαφείς μηχανισμούς λογοδοσίας σε περιπτώσεις όπου μεροληπτικά ή παραβιασμένα συστήματα οδηγούν σε παραβιάσεις δικαιωμάτων. Τα πλαίσια ευθύνης πρέπει να διασφαλίζουν ότι η ευθύνη δεν μπορεί να μετακυλίεται αποκλειστικά σε μεμονωμένους κλινικούς ιατρούς, όταν οι βλάβες είναι δομικά ενσωματωμένες στον σχεδιασμό ή στις αποφάσεις ανάπτυξης και εφαρμογής των συστημάτων ΤΝ.

Τέταρτον, η προώθηση ηθικών και υπεύθυνων πρακτικών πρέπει να ενσωματωθεί σε ολόκληρη την αξιακή αλυσίδα της βιοϊατρικής ΤΝ. Οι προγραμματιστές και οι πάροχοι συστημάτων θα πρέπει να υποχρεούνται να αντιμετωπίζουν προληπτικά τους κινδύνους μεροληψίας μέσω συλλογής αντιπροσωπευτικών δεδομένων, προσεκτικής επιλογής στόχων και υποκατάστατων δεικτών (proxies), επικύρωσης ανά υποομάδα και διαφανούς αναφοράς της απόδοσης σε σχέση με το φύλο και τις φυλετικές ομάδες. Σημαντικό είναι ότι τα στοιχεία που εξετάζονται στο παραδοτέο D2.1 δείχνουν πως η προσέγγιση της «δικαιοσύνης μέσω άγνοιας» (“fairness through unawareness”[1]), καθώς και οι αμιγώς τεχνικές στρατηγικές για την εξάλειψη των μεροληψιών, συχνά δεν επαρκούν στο πλαίσιο της υγειονομικής περίθαλψης. Κατά συνέπεια, η κανονιστική καθοδήγηση και τα σχετικά πρότυπα θα πρέπει να υπερβαίνουν αφηρημένους δείκτες δικαιοσύνης και να απαιτούν από τους προγραμματιστές να αποδεικνύουν κλινικά ουσιώδη αποτελέσματα ισότητας, αξιολογούμενα σε συνάρτηση με τις πραγματικές διαδρομές και τα πρότυπα πρόσβασης στις υπηρεσίες υγείας.

Οι πολιτικές δημόσιων προμηθειών και χρηματοδότησης διαδραματίζουν καθοριστικό ρόλο στη διαμόρφωση των κινήτρων προγραμματισμού/ ανάπτυξης. Οι αρχές υγείας και τα δημόσια νοσοκομεία θα πρέπει να ενσωματώνουν κριτήρια θεμελιωδών

[1] Στο πλαίσιο της προσέγγισης “fairness through awareness” ορίζεται ένας αλγόριθμος ως fair αν αυτός δίνει παρόμοια αποτελέσματα για άτομα με παρόμοια χαρακτηριστικά.

Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης. Οι απόψεις και οι γνώμες που διατυπώνονται εκφράζουν αποκλειστικά τις απόψεις των συντακτών και δεν αντιπροσωπεύουν κατ'ανάγκη τις απόψεις της Ευρωπαϊκής Ένωσης ή του Ευρωπαϊκού Εκτελεστικού Οργανισμού Εκπαίδευσης και Πολιτισμού (EACEA). Η Ευρωπαϊκή Ένωση και ο EACEA δεν μπορούν να θεωρηθούν υπεύθυνοι για τις εκφραζόμενες απόψεις. Κωδικός έργου: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

δικαιωμάτων και μεροληψίας στις αποφάσεις προμήθειας συστημάτων ΤΝ, ευνοώντας λύσεις που αποδεικνύουν ισχυρές, διαφανείς και ανεξάρτητα επαληθευμένες πρακτικές μετριασμού της μεροληψίας. Τα χρηματοδοτικά εργαλεία της ΕΕ, συμπεριλαμβανομένων των μελλοντικών προγραμμάτων έρευνας και καινοτομίας, θα πρέπει να συνεχίσουν να δίνουν προτεραιότητα σε έργα που συνδυάζουν την τεχνική καινοτομία με διακυβέρνηση βασισμένη στα δικαιώματα, στη συμμετοχή των ενδιαφερόμενων μερών και στην ενίσχυση ικανοτήτων, σε ευθυγράμμιση με το μοντέλο AEQUITAS.

Τέλος, η ενίσχυση της κοινωνικής ανθεκτικότητας απέναντι σε μεροληπτικά βιοϊατρικά συστήματα ΤΝ απαιτεί διαρκή επένδυση στην ενημέρωση του κοινού, στη συμμετοχή της κοινωνίας των πολιτών και στη διατομεακή συνεργασία. Τα άτομα πρέπει να ενδυναμώνονται ώστε να κατανοούν τα δικαιώματά τους στο πλαίσιο της υγειονομικής περίθαλψης που διαμεσολαβείται από ΤΝ, καθώς και τους διαθέσιμους μηχανισμούς προστασίας τους. Οι οργανώσεις της κοινωνίας των πολιτών, οι φορείς ισότητας και οι ομάδες ασθενών θα πρέπει να αναγνωρίζονται ως ουσιώδεις παράγοντες για την παρακολούθηση των επιπτώσεων της ΤΝ, την υποστήριξη των θιγόμενων ατόμων και τη διαμόρφωση πολιτικών. Η συνεργασία μεταξύ κυβερνήσεων, παρόχων υγείας, ερευνητών/ερευνητριών, βιομηχανίας και κοινωνίας των πολιτών είναι αναγκαία ώστε να διασφαλιστεί ότι τα οφέλη της βιοϊατρικής ΤΝ κατανέμονται δίκαια και δεν αναπαράγουν τις υφιστάμενες ανισότητες στην υγεία.

Συνολικά, τα ευρήματα του Παραδοτέου D2.1 στηρίζουν ένα σαφές συμπέρασμα πολιτικής: η βιοϊατρική ΤΝ μπορεί να θεωρηθεί αξιόπιστη και θεμιτή στην ΕΕ μόνο όταν ο σχεδιασμός, η ανάπτυξη και η διακυβέρνησή της εδράζονται σταθερά στην προστασία των θεμελιωδών δικαιωμάτων. Ο Κανονισμός της ΕΕ για την Τεχνητή Νοημοσύνη, ερμηνευόμενος υπό το πρίσμα του Χάρτη Θεμελιωδών Δικαιωμάτων της Ευρωπαϊκής Ένωσης και υλοποιούμενος μέσω συγκεκριμένων μηχανισμών εποπτείας, εκτίμησης επιπτώσεων και λογοδοσίας, προσφέρει μια κρίσιμη ευκαιρία ώστε να

Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης. Οι απόψεις και οι γνώμες που διατυπώνονται εκφράζουν αποκλειστικά τις απόψεις των συντακτών και δεν αντιπροσωπεύουν κατ'ανάγκη τις απόψεις της Ευρωπαϊκής Ένωσης ή του Ευρωπαϊκού Εκτελεστικού Οργανισμού Εκπαίδευσης και Πολιτισμού (EACEA). Η Ευρωπαϊκή Ένωση και ο EACEA δεν μπορούν να θεωρηθούν υπεύθυνοι για τις εκφραζόμενες απόψεις. Κωδικός έργου: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

διασφαλιστεί ότι η καινοτομία στον τομέα της υγείας προάγει την ισότητα και δεν αναπαράγει ιστορικά πρότυπα διακρίσεων.

Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης. Οι απόψεις και οι γνώμες που διατυπώνονται εκφράζουν αποκλειστικά τις απόψεις των συντακτών και δεν αντιπροσωπεύουν κατ'ανάγκη τις απόψεις της Ευρωπαϊκής Ένωσης ή του Ευρωπαϊκού Εκτελεστικού Οργανισμού Εκπαίδευσης και Πολιτισμού (EACEA). Η Ευρωπαϊκή Ένωση και ο EACEA δεν μπορούν να θεωρηθούν υπεύθυνοι για τις εκφραζόμενες απόψεις. Κωδικός έργου: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Με τη συγχρηματοδότηση της Ευρωπαϊκής Ένωσης

AEQUITAS