**7. Ηθικές και κοινωνικές επεκτάσεις της Τεχνητής Νοημοσύνης**

Σκεφτείτε τον ψηφιακό σας βοηθό στο κινητό ή τις προτάσεις που σας βγάζει μια πλατφόρμα **streaming**. Η Τεχνητή Νοημοσύνη είναι ήδη κομμάτι της ζωής μας. Αλλά τι σημαίνει αυτό για εμάς ως κοινωνία; Η τεχνητή νοημοσύνη δεν είναι μόνο μια τεχνολογική εξέλιξη είναι και ένα κοινωνικό γεγονός. Καθώς τα συστήματα ΤΝ αρχίζουν να διαδραματίζουν ενεργό ρόλο στη ζωή των ανθρώπων — σε επιλογές που επηρεάζουν εργασία, υγεία, εκπαίδευση, ακόμα και ελευθερία — εγείρονται ερωτήματα που δεν αφορούν μόνο την αποτελεσματικότητα ή την ταχύτητα, αλλά την ίδια τη δικαιοσύνη. Το πώς σχεδιάζεται, εκπαιδεύεται και εφαρμόζεται η τεχνητή νοημοσύνη δεν είναι ουδέτερο. Όπως συμβαίνει με κάθε εργαλείο, η χρήση της μπορεί να εξυπηρετήσει την πρόοδο ή να διαιωνίσει ανισότητες, να προστατεύσει ή να καταπατήσει δικαιώματα, να υπηρετήσει ή να αντικαταστήσει τον άνθρωπο.

Αυτό το κεφάλαιο δεν επικεντρώνεται στις τεχνικές πλευρές των συστημάτων, αλλά στις ηθικές και κοινωνικές τους συνέπειες. Τι συμβαίνει όταν μια μηχανή μαθαίνει από τα δεδομένα μας, αλλά αναπαράγει τα λάθη του παρελθόντος; Πώς διαχειριζόμαστε τις ηθικές επιλογές που καλούνται να κάνουν τα αυτόνομα συστήματα, όταν οι άνθρωποι απουσιάζουν από την τελική απόφαση; Και ποιος ευθύνεται όταν μια απόφαση αλγορίθμου αποδειχθεί άδικη; Οι ερωτήσεις αυτές δεν έχουν απλές απαντήσεις — αλλά είναι ακριβώς αυτές που πρέπει να τεθούν, εάν θέλουμε να ενσωματώσουμε την ΤΝ σε μια κοινωνία που σέβεται τον άνθρωπο.

Η ανάλυση ξεκινά με το φαινόμενο των προκαταλήψεων, καθώς αυτό βρίσκεται συχνά στην αφετηρία των υπόλοιπων προβλημάτων. Οι προκαταλήψεις δεν είναι μόνο σφάλματα από υπολογιστές. Είναι σφάλματα που ξεκινούν από τον κόσμο μας — και επιστρέφουν σε εμάς μέσω των συστημάτων που χτίζουμε. Η κατανόησή τους είναι το πρώτο βήμα για μια πιο υπεύθυνη και δίκαιη τεχνολογική εποχή.

**7.1 Προκαταλήψεις**

Στην αριθμητική, η λέξη «προκατάληψη» δεν εμφανίζεται, στους αλγορίθμους, όμως, είναι πανταχού παρούσα — ακόμη κι αν δεν την βλέπουμε. Όταν μιλάμε για προκατάληψη στην τεχνητή νοημοσύνη, δεν εννοούμε μια γνώμη ή στάση. Εννοούμε τη συστηματική απόκλιση ενός αποτελέσματος από αυτό που θεωρείται δίκαιο, ουδέτερο ή αναμενόμενο. Οι προκαταλήψεις στην ΤΝ είναι, συχνά, το αποτέλεσμα του τρόπου που η μηχανή μαθαίνει: από τα δεδομένα, από τις υποθέσεις που ενσωματώνονται στον σχεδιασμό, από την ίδια την κοινωνική πραγματικότητα στην οποία εφαρμόζεται.

Ορισμένες προκαταλήψεις είναι εμφανείς και ευρέως γνωστές: αλγόριθμοι που δυσκολεύονται να αναγνωρίσουν πρόσωπα με σκούρο δέρμα, ή συστήματα που απορρίπτουν αυτόματα αιτήσεις επειδή κάποιος μεγάλωσε σε μια «υποβαθμισμένη» περιοχή. Άλλες όμως είναι πιο λεπτές — και γι’ αυτό πιο επικίνδυνες. Ένας αλγόριθμος μπορεί να δείχνει «σωστός» στα νούμερα, αλλά να επηρεάζει δυσανάλογα συγκεκριμένες ομάδες, αποκλείοντάς τες από ευκαιρίες, πόρους ή πληροφορίες.

Σε αυτή την ενότητα, θα βουτήξουμε βαθύτερα στο θέμα της προκατάληψης: θα δούμε τι ακριβώς είναι (7.1.1), από πού προέρχεται (7.1.2), πώς μπορεί να περιοριστεί (7.1.3), και τέλος πώς επηρεάζει την κοινωνία — όχι θεωρητικά, αλλά στην πράξη (7.1.4). Ο στόχος μας δεν είναι να καταδικάσουμε την τεχνολογία, αλλά να αναδείξουμε τους τρόπους με τους οποίους η τεχνολογία μεταφέρει, ενισχύει ή μετασχηματίζει τα όρια που η κοινωνία μας ήδη διαθέτει.

Η προκατάληψη στην ΤΝ δεν είναι αναγκαστικά αποτέλεσμα κακής πρόθεσης. Είναι, όμως, ευθύνη όλων μας να την εντοπίσουμε και να την περιορίσουμε. Αν η τεχνητή νοημοσύνη πρόκειται να γίνει ένα από τα θεμέλια της κοινωνικής οργάνωσης του αύριο, τότε πρέπει να ξεκινήσει με το σωστό υπόβαθρο — και αυτό περνάει από την κατανόηση των αθέατων στρεβλώσεων που κουβαλάει μαζί της.

**7.1.1 Τι είναι οι προκαταλήψεις (bias)**

Όταν μιλάμε για τεχνητή νοημοσύνη, συχνά φανταζόμαστε μηχανές που σκέφτονται με λογική και ακρίβεια, απελευθερωμένες από ανθρώπινα σφάλματα και συναισθηματισμούς. Όμως, η πραγματικότητα είναι πολύ πιο σύνθετη. Ένα από τα βασικά προβλήματα που εμφανίζονται στα συστήματα τεχνητής νοημοσύνης είναι οι λεγόμενες **"προκαταλήψεις"** ή, στα αγγλικά, **"bias"**. Αυτές οι προκαταλήψεις δεν είναι τίποτε άλλο από επαναλαμβανόμενα σφάλματα ή στρεβλώσεις, που επηρεάζουν τις αποφάσεις και τις προβλέψεις των συστημάτων. Και το πιο σημαντικό: οι προκαταλήψεις της ΤΝ δεν είναι τυχαίες. Έχουν κατεύθυνση, επηρεάζουν συγκεκριμένες ομάδες ανθρώπων ή συγκεκριμένα αποτελέσματα και πολλές φορές περνούν απαρατήρητες.

Για να κατανοήσουμε καλύτερα τι σημαίνει προκατάληψη σε ένα σύστημα ΤΝ, ας σκεφτούμε ένα απλό παράδειγμα. Φανταστείτε ότι αναπτύσσουμε έναν αλγόριθμο που να αναγνωρίζει πρόσωπα σε φωτογραφίες. Αν τα δεδομένα με τα οποία "εκπαιδεύουμε" το σύστημα αποτελούνται κυρίως από φωτογραφίες ανθρώπων με ανοιχτόχρωμο δέρμα, τότε το σύστημα μαθαίνει καλύτερα να αναγνωρίζει αυτά τα πρόσωπα και χειρότερα πρόσωπα ανθρώπων με πιο σκούρο δέρμα. Το αποτέλεσμα είναι ότι το ίδιο το σύστημα, χωρίς να έχει πρόθεση, εμφανίζει προκατάληψη εναντίον συγκεκριμένων πληθυσμών. Δεν είναι ότι "θέλει" να είναι άδικο — απλώς έχει "μάθει" με βάση τα δεδομένα που του δώσαμε. Σαν ένα παιδί που μαθαίνει να ξεχωρίζει τα ζώα: αν του δείξεις μόνο εικόνες από σκύλους και γάτες, θα δυσκολευτεί να αναγνωρίσει ένα κουνέλι την πρώτη φορά που θα το δει.

Αυτή η μορφή προκατάληψης δεν περιορίζεται μόνο στην αναγνώριση εικόνων. Εμφανίζεται παντού όπου χρησιμοποιούνται αλγοριθμικές διαδικασίες: στην αξιολόγηση υποψηφίων για δουλειές, στη χορήγηση δανείων από τράπεζες, ακόμη και στη διάγνωση ασθενειών. Όπου υπάρχει τεχνητή νοημοσύνη που μαθαίνει από δεδομένα, υπάρχει και ο κίνδυνος να μεταφέρει ή και να ενισχύσει προκαταλήψεις που ήδη υπάρχουν μέσα σε αυτά τα δεδομένα.

Είναι σημαντικό να καταλάβουμε ότι η προκατάληψη στην τεχνητή νοημοσύνη δεν είναι πάντοτε εύκολα αντιληπτή. Στα παραδοσιακά σφάλματα ανθρώπινης κρίσης μπορούμε να αποδώσουμε ευθύνη και να τα διορθώσουμε μέσω εκπαίδευσης ή κανόνων. Στα συστήματα ΤΝ, όμως, το σφάλμα είναι "θαμμένο" βαθιά μέσα στη μαθησιακή διαδικασία. Ο τρόπος που το σύστημα φιλτράρει, αναλύει και συνδυάζει δεδομένα μπορεί να παράγει στρεβλά αποτελέσματα χωρίς κανείς να το καταλάβει άμεσα.

Ένα χαρακτηριστικό παράδειγμα είναι οι αλγόριθμοι πρόβλεψης υποτροπής εγκλήματος στις Ηνωμένες Πολιτείες. Ένα τέτοιο σύστημα, που χρησιμοποιήθηκε σε ορισμένες πολιτείες, είχε μεγαλύτερη πιθανότητα να αξιολογήσει ως "υψηλού κινδύνου" άτομα από φυλετικές μειονότητες, ακόμα κι όταν τα αντικειμενικά δεδομένα για αυτά τα άτομα ήταν παρόμοια με εκείνα της πλειοψηφίας. Η προκατάληψη εδώ δεν ήταν αποτέλεσμα "κακής πρόθεσης" αλλά των δεδομένων που χρησιμοποιήθηκαν για να εκπαιδευτεί το σύστημα: δεδομένα που αντανακλούσαν, ήδη, ανισότητες και αδικίες του παρελθόντος.

Η ύπαρξη προκατάληψης δεν περιορίζεται όμως μόνο στις κοινωνικές διακρίσεις. Προκατάληψη μπορεί να εμφανιστεί και σε τεχνικά, "ουδέτερα" προβλήματα. Για παράδειγμα, ένα σύστημα διάγνωσης ασθενειών μπορεί να υπερεκτιμά ορισμένες παθήσεις σε σχέση με άλλες, αν τα δεδομένα εκπαίδευσής του έχουν ανισομέρεια ή αν προέρχονται κυρίως από έναν συγκεκριμένο πληθυσμό ή γεωγραφική περιοχή. Έτσι, ακόμη και χωρίς άμεσο κοινωνικό πρόσημο, η προκατάληψη εξακολουθεί να δημιουργεί ανακρίβειες και προβλήματα.

Θα ήταν λάθος να φανταστούμε την προκατάληψη στην ΤΝ ως ένα μεμονωμένο λάθος που μπορεί εύκολα να διορθωθεί. Αντίθετα, πρέπει να τη δούμε ως ένα συστημικό χαρακτηριστικό της διαδικασίας μηχανικής μάθησης. Από τη στιγμή που τα δεδομένα αντικατοπτρίζουν τον πραγματικό κόσμο — με όλες του τις ατέλειες, τις ανισότητες και τα πρότυπα — το σύστημα που μαθαίνει από αυτά θα κληρονομήσει, έστω και άθελα, τις ίδιες ατέλειες.

Για να κλείσουμε την αρχική αυτή διερεύνηση, μπορούμε να πούμε ότι η προκατάληψη στην τεχνητή νοημοσύνη είναι αναπόφευκτη, αλλά όχι αναγκαία αμετάκλητη. Με επίγνωση, σωστό σχεδιασμό και προσοχή στη χρήση των δεδομένων, μπορούμε να μειώσουμε τις προκαταλήψεις και να δημιουργήσουμε πιο δίκαιες και αξιόπιστες εφαρμογές. Το πρώτο βήμα όμως είναι να αναγνωρίσουμε το πρόβλημα: να συνειδητοποιήσουμε ότι ακόμα και οι πιο εξελιγμένες μηχανές κουβαλούν μαζί τους τον κόσμο που τις δημιούργησε.

**7.1.2 Πώς δημιουργείται η προκατάληψη στην Τεχνητή Νοημοσύνη**

Αφού κατανοήσαμε τι είναι οι προκαταλήψεις στην τεχνητή νοημοσύνη, είναι σημαντικό να δούμε πώς ακριβώς αυτές δημιουργούνται. Η προκατάληψη σε ένα σύστημα ΤΝ δεν εμφανίζεται από το πουθενά. Είναι το αποτέλεσμα συγκεκριμένων διαδικασιών, αποφάσεων και αδυναμιών που συνδέονται τόσο με τα δεδομένα όσο και με τον τρόπο που σχεδιάζονται και λειτουργούν τα ίδια τα συστήματα.

Ο πρώτος και πιο συνηθισμένος τρόπος δημιουργίας προκατάληψης είναι μέσω των ίδιων των δεδομένων εκπαίδευσης. Τα περισσότερα συστήματα τεχνητής νοημοσύνης δεν δημιουργούν «γνώσεις» από μόνα τους. Μαθαίνουν παρατηρώντας δεδομένα που τους παρέχουμε. Αν αυτά τα δεδομένα δεν είναι αντιπροσωπευτικά του πραγματικού κόσμου ή αν περιέχουν ήδη προκαταλήψεις, τότε το σύστημα θα αναπαράγει αυτές τις στρεβλώσεις. Για παράδειγμα, αν ένας αλγόριθμος εκπαίδευσης για την επιλογή υποψηφίων σε μια εταιρεία τροφοδοτηθεί με ιστορικά δεδομένα όπου οι άντρες έχουν ευνοηθεί σε προσλήψεις έναντι των γυναικών, τότε το σύστημα θα «μάθει» ότι το φύλο είναι σημαντικός παράγοντας και πιθανώς να αναπαράγει αυτή την αδικία.

Δεν είναι μόνο η ποσότητα των δεδομένων που δημιουργεί προβλήματα, αλλά και η ποιότητά τους. Πολλές φορές τα δεδομένα που συλλέγονται έχουν κενά, ανακρίβειες ή περιορισμένο φάσμα. Ένα σύστημα διάγνωσης ιατρικών ασθενειών, για παράδειγμα, μπορεί να εκπαιδευτεί κυρίως με δεδομένα από έναν συγκεκριμένο πληθυσμό, π.χ. λευκούς ενήλικες. Όταν κληθεί να αξιολογήσει ασθενείς από διαφορετικά υπόβαθρα, το σύστημα μπορεί να κάνει σοβαρά λάθη, όχι επειδή έχει "κακή πρόθεση", αλλά επειδή δεν έχει "δει" αρκετά αντιπροσωπευτικά παραδείγματα.

Δεν φταίνε όμως μόνο τα δεδομένα. Μερικές φορές, η προκατάληψη "χτίζεται" από εμάς, τους ανθρώπους που σχεδιάζουμε τα συστήματα. Κατά την ανάπτυξη ενός αλγορίθμου, οι δημιουργοί κάνουν υποθέσεις: ποια χαρακτηριστικά θεωρούν σημαντικά, πώς βαθμολογούνται τα αποτελέσματα, ποιες αποκλίσεις θεωρούνται ανεκτές. Κάθε μια από αυτές τις επιλογές ενέχει τον κίνδυνο να ενσωματώνει υποσυνείδητες προκαταλήψεις ή να παραβλέπει κρίσιμες παραμέτρους. Για παράδειγμα, ένα σύστημα που αξιολογεί αιτήσεις δανείων μπορεί να χρησιμοποιεί την περιοχή κατοικίας ως παράγοντα πρόβλεψης, χωρίς να λαμβάνει υπόψη ότι οι οικονομικές ευκαιρίες διαφέρουν δραματικά από περιοχή σε περιοχή για λόγους που δεν έχουν σχέση με την ατομική φερεγγυότητα.

Επιπλέον, η προκατάληψη μπορεί να εισχωρήσει στον τρόπο που ορίζονται οι στόχοι ενός συστήματος τεχνητής νοημοσύνης. Στη μηχανική μάθηση, είναι αναγκαίο να καθορίσουμε τι θεωρείται «σωστό» ή «επιτυχημένο» αποτέλεσμα. Αν όμως αυτός ο ορισμός δεν είναι ουδέτερος ή κρύβει υποθέσεις που δεν έχουν ελεγχθεί, τότε το σύστημα κινδυνεύει να υιοθετήσει και να ενισχύσει τις ίδιες στρεβλώσεις.

Ένα παράδειγμα που μας βοηθά να κατανοήσουμε αυτή τη διαδικασία έρχεται από έναν απλό πειραματισμό με δελφίνια. Οι εκπαιδευτές, θέλοντας να καθαρίσουν μια πισίνα από χαρτιά, αντάμειβαν τα δελφίνια με ένα ψάρι για κάθε χαρτί που έφερναν. Όμως τα δελφίνια, αντί να βγάζουν ολόκληρα τα χαρτιά, άρχισαν να τα σκίζουν σε μικρότερα κομμάτια, εξασφαλίζοντας περισσότερες ανταμοιβές για το ίδιο αντικείμενο. Έτσι, παρότι επιτύγχαναν το μετρούμενο στόχο (περισσότερα «χαρτιά»), δεν επιτελούσαν τον πραγματικό σκοπό, που ήταν να καθαρίσουν την πισίνα.

Το παράδειγμα αυτό απεικονίζει άριστα τον λεγόμενο "Νόμο του Γκούντχαρτ" (Goodhart's Law): «Όταν ένα μέτρο γίνεται στόχος, παύει να είναι καλό μέτρο». Με άλλα λόγια, αν εστιάσουμε σε έναν αριθμητικό δείκτη χωρίς να λαμβάνουμε υπόψη το ουσιαστικό νόημα, το σύστημα μπορεί να επιδιώξει το αποτέλεσμα με τρόπους που δεν είναι χρήσιμοι ή δίκαιοι.

Το ίδιο ισχύει και για τα συστήματα ΤΝ. Αν οι στόχοι που τίθενται κατά τον σχεδιασμό είναι ατελείς ή παραπλανητικοί, το σύστημα μπορεί να επιδιώξει τη «βελτιστοποίηση» αυτών των στόχων με τρόπους που ενισχύουν προκαταλήψεις ή παράγουν ανεπιθύμητες συμπεριφορές. Και αν δεν γίνεται σωστή επαλήθευση και προσεκτική τεκμηρίωση σε κάθε στάδιο της ανάπτυξης, μικρές αρχικές ανισορροπίες μπορούν εύκολα να διογκωθούν και να οδηγήσουν σε σοβαρές αποκλίσεις στα τελικά αποτελέσματα.

Ακόμη και μετά την ανάπτυξη, η προκατάληψη μπορεί να επιδεινωθεί κατά τη χρήση του συστήματος. Όταν ένα σύστημα αλληλοεπιδρά με τον πραγματικό κόσμο, τα νέα δεδομένα που συλλέγει μπορεί να ενισχύουν ήδη υπάρχουσες τάσεις. Για παράδειγμα, αν ένα σύστημα προτείνει μόνο συγκεκριμένους τύπους περιεχομένου σε χρήστες και αυτοί οι χρήστες ανταποκρίνονται θετικά, τότε το σύστημα «μαθαίνει» να προσφέρει ακόμη πιο περιορισμένο φάσμα επιλογών, οδηγώντας σε φαινόμενα **"φούσκας πληροφόρησης" (filter bubble).**

Τέλος, σημαντικό ρόλο στη δημιουργία προκατάληψης παίζει και **το πολιτισμικό πλαίσιο μέσα στο οποίο αναπτύσσεται ένα σύστημα ΤΝ**. Κάθε κοινωνία έχει διαφορετικά πρότυπα, αξίες και αντιλήψεις. Όταν ένα σύστημα που έχει αναπτυχθεί σε μια χώρα εφαρμόζεται σε άλλη, υπάρχει ο κίνδυνος να μεταφέρει άθελά του προκαταλήψεις που δεν είναι ορατές στην αρχική κοινωνία αλλά γίνονται προβληματικές αλλού. Για παράδειγμα, ένα σύστημα αξιολόγησης βιογραφικών που έχει εκπαιδευτεί με βάση τα πρότυπα των ΗΠΑ μπορεί να μην ανταποκρίνεται δίκαια σε αιτήσεις από χώρες με διαφορετικά εκπαιδευτικά συστήματα.

Συμπερασματικά, η δημιουργία προκαταλήψεων στην τεχνητή νοημοσύνη είναι ένα πολυεπίπεδο φαινόμενο, που πηγάζει τόσο από τα δεδομένα όσο και από τον ανθρώπινο σχεδιασμό και την κοινωνική πραγματικότητα. Καμία τεχνολογική διαδικασία δεν είναι "αθώα" ή "ουδέτερη" από μόνη της. Χρειάζεται επίγνωση, προσεκτική σχεδίαση και συνεχής αναθεώρηση για να ελαχιστοποιηθεί η πιθανότητα να δημιουργήσουμε συστήματα που, χωρίς πρόθεση, αναπαράγουν ή και ενισχύουν τις υπάρχουσες αδικίες του κόσμου μας.

**7.1.3 Τι μπορούμε να κάνουμε για να μειώσουμε τις προκαταλήψεις**

Το γεγονός ότι οι προκαταλήψεις εμφανίζονται στα συστήματα τεχνητής νοημοσύνης δεν σημαίνει ότι είμαστε ανήμποροι μπροστά τους. Αντίθετα, υπάρχει μια σειρά από πρακτικές και μεθόδους που μπορούν να συμβάλουν στη μείωση του φαινομένου, εφόσον εφαρμοστούν με συνείδηση και συστηματικότητα. Το πρώτο και πιο βασικό βήμα είναι η αναγνώριση του προβλήματος. Κανένα σύστημα, όσο καλά σχεδιασμένο κι αν είναι, δεν μπορεί να βελτιωθεί αν οι δημιουργοί του αγνοούν ή αρνούνται να παραδεχθούν την ύπαρξη προκαταλήψεων στα αποτελέσματά του.

Η προσεκτική επιλογή και επεξεργασία των δεδομένων εκπαίδευσης είναι κρίσιμη. Όπως αναπτύξαμε νωρίτερα, τα δεδομένα φέρουν μέσα τους τις ιστορικές και κοινωνικές στρεβλώσεις του κόσμου. Συνεπώς, η συλλογή δεδομένων που είναι ποικιλόμορφα και αντιπροσωπευτικά των διαφορετικών ομάδων του πληθυσμού αποτελεί θεμελιώδη προϋπόθεση για τη δημιουργία πιο δίκαιων συστημάτων. Δεν αρκεί να έχουμε μεγάλο όγκο δεδομένων· χρειάζεται να έχουμε δεδομένα που να αντικατοπτρίζουν την πολυπλοκότητα της πραγματικής κοινωνίας.

Σε πολλές περιπτώσεις, είναι αναγκαίο να γίνονται παρεμβάσεις πριν, κατά τη διάρκεια ή και μετά την εκπαίδευση του συστήματος. Μία τεχνική είναι η αποκαλούμενη **"εξισορρόπηση των δεδομένων" (data balancing),** όπου εξασφαλίζεται ότι όλες οι κατηγορίες που μας ενδιαφέρουν εκπροσωπούνται επαρκώς. Για παράδειγμα, αν αναπτύσσουμε ένα σύστημα αναγνώρισης φωνής, θα πρέπει να βεβαιωθούμε ότι περιλαμβάνονται φωνητικά δεδομένα από άτομα διαφορετικού φύλου, ηλικίας, εθνικότητας και κοινωνικού υποβάθρου.

Ένας άλλος τρόπος είναι **η εφαρμογή τεχνικών "αμεροληψίας" (debiasing)** στα δεδομένα ή στα μοντέλα. Αυτές οι τεχνικές περιλαμβάνουν μεθόδους που είτε αφαιρούν συσχετίσεις που οδηγούν σε προκαταλήψεις είτε προσαρμόζουν τα κριτήρια λήψης αποφάσεων ώστε να περιορίζουν τη διάδοση της προκατάληψης. Για παράδειγμα, σε ένα σύστημα επιλογής υποψηφίων, μπορεί να εφαρμοστεί μέθοδος που να "τυφλώνει" το μοντέλο σε πληροφορίες όπως το φύλο ή η ηλικία, ώστε να εστιάζει αποκλειστικά στα αντικειμενικά προσόντα.

**Η ανθρώπινη επιτήρηση** παραμένει αναντικατάστατη. Παρά την πρόοδο στην αυτονομία των συστημάτων, η κρίσιμη ματιά του ανθρώπου είναι αυτή που μπορεί να ανιχνεύσει σφάλματα που δεν προβλέπονται από τους αλγορίθμους. Ειδικά σε ευαίσθητους τομείς όπως η υγεία, η δικαιοσύνη ή η εκπαίδευση, απαιτείται πάντοτε η παρουσία ανθρώπων που ελέγχουν, αξιολογούν και εγκρίνουν τα αποτελέσματα των συστημάτων πριν αυτά επηρεάσουν πραγματικά άτομα.

**Η πολυφωνία στις ομάδες ανάπτυξης** είναι επίσης σημαντική. Όταν τα συστήματα τεχνητής νοημοσύνης σχεδιάζονται από ομάδες που περιλαμβάνουν ανθρώπους με διαφορετικά υπόβαθρα, εμπειρίες και οπτικές, είναι πιο πιθανό να εντοπιστούν και να αντιμετωπιστούν προκαταλήψεις πριν αυτές ενσωματωθούν στα τελικά προϊόντα. Αντίθετα, ομάδες με ομοιογενές προφίλ διατρέχουν μεγαλύτερο κίνδυνο να παραβλέψουν σημαντικές παραμέτρους.

Μια πρακτική που κερδίζει έδαφος είναι **η υιοθέτηση "δοκιμών δίκαιης χρήσης" (fairness testing)**. Σε αυτές τις δοκιμές, τα συστήματα αξιολογούνται όχι μόνο ως προς την ακρίβεια ή την απόδοσή τους, αλλά και ως προς το κατά πόσον οι αποφάσεις τους επηρεάζουν διαφορετικές ομάδες ανθρώπων ισότιμα. Αν εντοπιστούν αποκλίσεις, εφαρμόζονται διορθωτικά μέτρα πριν το σύστημα τεθεί σε πλήρη λειτουργία.

**Η διαφάνεια στην ανάπτυξη και λειτουργία των συστημάτων ΤΝ** είναι ακόμα ένα ισχυρό εργαλείο για τη μείωση των προκαταλήψεων. Όταν οι προγραμματιστές δημοσιεύουν πληροφορίες για τον τρόπο που λειτούργησαν, τα δεδομένα που χρησιμοποίησαν και τα κριτήρια που επέλεξαν, δίνουν στην κοινωνία τη δυνατότητα να ελέγξει, να κρίνει και να διορθώσει πιθανά προβλήματα. Η διαφάνεια δημιουργεί ένα πλαίσιο λογοδοσίας, απαραίτητο για την εμπιστοσύνη.

Επιπλέον, σημαντικό ρόλο παίζουν **τα ρυθμιστικά πλαίσια και οι κανονισμοί**. Σε παγκόσμιο επίπεδο, αναπτύσσονται όλο και περισσότερες κατευθυντήριες γραμμές και νομικές απαιτήσεις για τη δίκαιη χρήση της τεχνητής νοημοσύνης. **Ο Κανονισμός για την ΤΝ** **(AI Act)** που εφαρμόζει η **Ευρωπαϊκή Ένωση** επιχειρεί ακριβώς να θέσει όρια και απαιτήσεις για τα συστήματα υψηλού κινδύνου, προστατεύοντας έτσι τα δικαιώματα των πολιτών.

Τέλος, **η εκπαίδευση και ευαισθητοποίηση** όλων όσων συμμετέχουν στον κύκλο ζωής ενός συστήματος ΤΝ είναι κρίσιμη. Οι επιστήμονες, οι μηχανικοί, οι υπεύθυνοι έργων, αλλά και οι τελικοί χρήστες πρέπει να είναι σε θέση να αναγνωρίζουν τα σημάδια της προκατάληψης και να ξέρουν πώς να αντιδρούν. Η τεχνολογία δεν εξελίσσεται μέσα σε κενό· είναι κοινωνικό προϊόν, και η ευθύνη για τη δίκαιη χρήση της βαραίνει όλους μας.

Συνοψίζοντας, η μείωση της προκατάληψης στα συστήματα τεχνητής νοημοσύνης απαιτεί μια συνδυαστική προσέγγιση: τεχνικές λύσεις, ανθρώπινη κρίση, θεσμική παρέμβαση και κοινωνική συνείδηση. Έτσι θα μπορούμε να ελπίζουμε σε ένα μέλλον όπου η τεχνητή νοημοσύνη θα υπηρετεί το σύνολο της κοινωνίας και όχι τις αόρατες ανισότητες του παρελθόντος.

**7.1.4 Πώς μπορεί να επηρεάσει τους ανθρώπους και τις κοινωνικές ομάδες**

Η ύπαρξη προκαταλήψεων στα συστήματα τεχνητής νοημοσύνης δεν είναι ένα θεωρητικό ή ακαδημαϊκό πρόβλημα. Οι συνέπειές της αγγίζουν άμεσα τη ζωή πραγματικών ανθρώπων και, σε μεγαλύτερη κλίμακα, διαμορφώνουν τις συνθήκες μέσα στις οποίες αναπτύσσονται κοινωνικές ομάδες και κοινότητες. Η τεχνητή νοημοσύνη, όταν λειτουργεί με προκαταλήψεις, μπορεί να διαιωνίσει και να ενισχύσει τις ανισότητες αντί να τις μειώσει.

Σκεφτείτε για παράδειγμα τη διαδικασία επιλογής υποψηφίων για εργασία. Αν ένα σύστημα ΤΝ που χρησιμοποιείται για το φιλτράρισμα βιογραφικών έχει εκπαιδευτεί με δεδομένα που ευνοούν άντρες έναντι γυναικών ή ντόπιους έναντι μεταναστών, τότε αναπαράγει και διαιωνίζει τα ίδια πρότυπα αποκλεισμού. Το αποτέλεσμα είναι ότι άτομα που ήδη ανήκουν σε μειονεκτούσες ομάδες στερούνται ευκαιριών, όχι επειδή υστερούν πραγματικά σε προσόντα, αλλά επειδή το σύστημα τα αξιολόγησε άδικα.

Η προκατάληψη στην τεχνητή νοημοσύνη μπορεί επίσης να ενισχύσει κοινωνικές διακρίσεις σε κρίσιμους τομείς όπως η υγεία και η εκπαίδευση. Ένα σύστημα ιατρικής διάγνωσης που λειτουργεί καλύτερα σε άντρες ασθενείς επειδή έχει εκπαιδευτεί κυρίως με δεδομένα ανδρικών οργανισμών, μπορεί να αργήσει να ανιχνεύσει σοβαρές ασθένειες σε γυναίκες. Αντίστοιχα, ένα εκπαιδευτικό σύστημα που προτείνει πρόσθετη υποστήριξη μόνο σε μαθητές που ακολουθούν συγκεκριμένα πρότυπα μάθησης, μπορεί να αφήσει πίσω του παιδιά που μαθαίνουν διαφορετικά αλλά είναι εξίσου ικανά.

Ακόμη και το σύστημα απονομής δικαιοσύνης μπορεί να επηρεαστεί αρνητικά. Αλγόριθμοι που χρησιμοποιούνται για να προβλέπουν την πιθανότητα υποτροπής ενός ατόμου στο μέλλον, αν φέρουν προκαταλήψεις, μπορεί να οδηγήσουν σε άδικες αποφάσεις προφυλάκισης ή αυστηρότερων ποινών για ορισμένες ομάδες πληθυσμού. Όταν τα δεδομένα εκπαίδευσης αντανακλούν παρελθούσες ανισότητες — για παράδειγμα, υπερβολική αστυνόμευση σε φτωχότερες ή φυλετικά μειονοτικές κοινότητες — τότε το σύστημα μαθαίνει να συσχετίζει αυτές τις ομάδες με μεγαλύτερο "κίνδυνο", χωρίς να υπάρχει αντικειμενική αιτία.

Η επίδραση της προκατάληψης όμως δεν είναι μόνο άμεση. Δημιουργεί και ένα αίσθημα αδικίας και δυσπιστίας απέναντι στην τεχνολογία συνολικά. Όταν άνθρωποι συνειδητοποιούν ότι τα συστήματα τεχνητής νοημοσύνης παίρνουν αποφάσεις που τους αδικούν ή τους αποκλείουν, χάνουν την εμπιστοσύνη τους όχι μόνο σε αυτά τα συστήματα αλλά και σε εκείνους που τα διαχειρίζονται. Η κοινωνική συνοχή πλήττεται, και ενισχύονται οι αντιλήψεις ότι η τεχνολογία δεν είναι μέσο προόδου αλλά μηχανισμός επιβολής ανισότητας.

Επιπλέον, οι κοινωνικές ομάδες που πλήττονται περισσότερο από τις προκαταλήψεις της τεχνητής νοημοσύνης συχνά είναι και εκείνες που έχουν λιγότερες δυνατότητες να αντιδράσουν ή να διεκδικήσουν την αλλαγή. Οι ευάλωτοι πληθυσμοί, οι μειονότητες, οι φτωχότερες κοινότητες βρίσκονται αντιμέτωποι με αποφάσεις που λαμβάνονται αυτόματα και συχνά χωρίς δυνατότητα επανεξέτασης ή ένστασης. Έτσι, δημιουργείται ένας φαύλος κύκλος, όπου οι υφιστάμενες ανισότητες ενισχύονται και παγιώνονται από τεχνολογίες που σχεδιάστηκαν, υποτίθεται, για να είναι αντικειμενικές.

Ακόμα και σε καθημερινές, λιγότερο κρίσιμες εφαρμογές, η προκατάληψη μπορεί να έχει σωρευτικές συνέπειες. Τα φίλτρα περιεχομένου σε πλατφόρμες κοινωνικής δικτύωσης, οι προτεινόμενες αγγελίες, οι αναζητήσεις προϊόντων — όλα αυτά τα συστήματα, όταν λειτουργούν με προκαταλήψεις, προβάλλουν διαφορετικές εικόνες του κόσμου σε διαφορετικές ομάδες ανθρώπων. Έτσι, δημιουργούνται διαφορετικές εμπειρίες και προσδοκίες, που με τη σειρά τους ενισχύουν τη διαίρεση και την ανισότητα.

Συνολικά, η προκατάληψη στην τεχνητή νοημοσύνη έχει τη δύναμη να επηρεάσει σε βάθος την κοινωνική δομή. Δεν είναι μόνο θέμα ατομικής αδικίας. Είναι θέμα συλλογικής δικαιοσύνης. Αν αφήσουμε ανεξέλεγκτες τις προκαταλήψεις, τότε οι τεχνολογίες που υπόσχονται να βελτιώσουν τη ζωή μας μπορεί να μετατραπούν σε όργανα διεύρυνσης των ανισοτήτων. Αντίθετα, αν τις αναγνωρίσουμε και τις αντιμετωπίσουμε, η τεχνητή νοημοσύνη μπορεί να γίνει ισχυρό εργαλείο για μια κοινωνία πιο δίκαιη, πιο ανοιχτή και πιο συμπεριληπτική.

**7.2 Ηθικές αποφάσεις από μηχανές**

Όταν σκεφτόμαστε την τεχνητή νοημοσύνη, την αντιλαμβανόμαστε συχνά ως ένα σύστημα που επεξεργάζεται δεδομένα και καταλήγει σε αποφάσεις στη βάση αριθμητικών ή λογικών κριτηρίων. Όμως, η πραγματικότητα γίνεται πολύ πιο σύνθετη όταν οι αποφάσεις που καλούνται να λάβουν αυτά τα συστήματα δεν είναι απλώς τεχνικές, αλλά αγγίζουν το πεδίο της ηθικής. Σε καταστάσεις όπου πρέπει να σταθμιστούν αντικρουόμενα αγαθά ή αξίες, η τεχνητή νοημοσύνη αντιμετωπίζει προκλήσεις που δεν μπορούν να λυθούν με απλή αριθμητική.

Ένα από τα πιο γνωστά παραδείγματα που χρησιμοποιούνται για να αναδειχθεί το πρόβλημα είναι το δίλημμα των αυτόνομων αυτοκινήτων. Τι πρέπει να κάνει ένα αυτόνομο όχημα αν, σε μια στιγμή ατυχήματος, πρέπει να επιλέξει ανάμεσα στο να προστατεύσει τον οδηγό του ή τους πεζούς που βρίσκονται μπροστά του; Πρέπει να λάβει υπόψη την ηλικία των πεζών; Τον αριθμό τους; Την πιθανότητα επιβίωσης; Κάθε επιλογή ενέχει ένα σύνολο ηθικών υποθέσεων που, στον ανθρώπινο κόσμο, συχνά παραμένουν άλυτες ή υπόκεινται σε πολιτισμικές διαφοροποιήσεις.

Η ηθική απόφαση, σε αντίθεση με μια απλή απόφαση βελτιστοποίησης, απαιτεί εκτίμηση αξιών. Στον ανθρώπινο πολιτισμό, αυτές οι αξίες είναι συχνά αλληλοσυγκρουόμενες: σεβασμός στη ζωή, προστασία της ιδιοκτησίας, δικαιοσύνη, αλληλεγγύη. Όταν καλείται μια μηχανή να λάβει τέτοιες αποφάσεις, πρέπει πρώτα να της μεταφερθεί με κάποιον τρόπο ένα «σύστημα αξιών». Αυτό όμως εγείρει θεμελιώδη ερωτήματα: Ποιος καθορίζει αυτές τις αξίες; Ποια ηθική φιλοσοφία υιοθετείται; Και μπορεί ένα σύστημα ΤΝ να ενεργεί ηθικά όταν στην πραγματικότητα δεν «κατανοεί»;

Είναι σημαντικό να τονίσουμε ότι η τεχνητή νοημοσύνη, σήμερα, δεν έχει επίγνωση ή συνείδηση. Οι αποφάσεις που λαμβάνει βασίζονται σε αλγοριθμικές διαδικασίες που ακολουθούν κανόνες ή βελτιστοποιούν συγκεκριμένους στόχους. Αυτό σημαίνει ότι όταν λέμε ότι ένα αυτόνομο όχημα "επιλέγει" πώς να ενεργήσει σε ένα ατύχημα, στην πραγματικότητα ακολουθεί ένα σύνολο προγραμματισμένων οδηγιών ή στατιστικών μοντέλων που του έχουν δοθεί από ανθρώπους. Η ευθύνη της απόφασης, τελικά, βαραίνει εκείνους που σχεδίασαν, ανέπτυξαν και ανέθεσαν τις προτεραιότητες στο σύστημα.

Ωστόσο, το γεγονός ότι η ευθύνη βαραίνει τους ανθρώπους δεν σημαίνει ότι το πρόβλημα είναι λιγότερο πιεστικό. Σε καταστάσεις όπου η απόφαση πρέπει να ληφθεί σε κλάσματα δευτερολέπτου, η ανθρώπινη παρέμβαση δεν είναι δυνατή. Η προετοιμασία του συστήματος εκ των προτέρων είναι το μόνο διαθέσιμο εργαλείο. Και ακριβώς εδώ αναδεικνύεται η δυσκολία: πώς προγραμματίζουμε ένα σύστημα για να λαμβάνει αποφάσεις σε ακραίες καταστάσεις όπου και οι δύο επιλογές συνεπάγονται βλάβη;

Διάφορες προσεγγίσεις έχουν προταθεί. Μία ιδέα είναι να ακολουθηθεί μια ωφελιμιστική λογική: να επιλέγεται η ενέργεια που οδηγεί στη μικρότερη συνολική βλάβη ή στον μικρότερο αριθμό θυμάτων. Άλλη προσέγγιση προκρίνει την αρχή του σεβασμού των δικαιωμάτων του ατόμου, αποφεύγοντας, για παράδειγμα, ενέργειες που θυσιάζουν έναν άνθρωπο για το "γενικό καλό". Κάθε επιλογή όμως ενσωματώνει φιλοσοφικές παραδοχές που δεν είναι καθολικά αποδεκτές.

Επιπλέον, τέτοιες αποφάσεις δεν είναι ουδέτερες ως προς το κοινωνικό πλαίσιο. Έρευνες έχουν δείξει ότι διαφορετικές κουλτούρες προτιμούν διαφορετικές απαντήσεις στο δίλημμα του αυτόνομου αυτοκινήτου. Σε ορισμένες κοινωνίες δίνεται μεγαλύτερη έμφαση στην προστασία των νεότερων, σε άλλες στον σεβασμό της ηλικίας. Αυτό σημαίνει ότι μια "ενιαία" λύση μπορεί να θεωρηθεί άδικη ή ακατάλληλη σε διαφορετικά πολιτισμικά περιβάλλοντα.

Η ηθική λήψη αποφάσεων από μηχανές δεν περιορίζεται στα αυτόνομα οχήματα. Παρόμοια προβλήματα εμφανίζονται σε ιατρικά συστήματα που προτείνουν ποιος ασθενής θα λάβει περιορισμένους πόρους (όπως ένα κρεβάτι εντατικής θεραπείας σε περίοδο πανδημίας), σε αλγορίθμους που αξιολογούν αιτήσεις ασύλου ή σε συστήματα που προτείνουν ποινές ή αναστολές σε πλατφόρμες κοινωνικής δικτύωσης.

Σε όλα αυτά τα παραδείγματα, αναδύεται η ανάγκη να αντιμετωπίζουμε τα συστήματα τεχνητής νοημοσύνης όχι ως ανεξάρτητους φορείς ηθικής, αλλά ως εργαλεία που πρέπει να σχεδιάζονται και να ελέγχονται με βαθιά ηθική ευθύνη. Δεν μπορούμε να περιμένουμε από μια μηχανή να "αισθανθεί" τι είναι δίκαιο ή σωστό. Αντίθετα, πρέπει να δημιουργούμε τα συστήματα με τέτοιο τρόπο ώστε να ελαχιστοποιούνται οι άδικες συνέπειες, να υπάρχει δυνατότητα διαφάνειας στις αποφάσεις και να διατηρείται πάντοτε ο ανθρώπινος έλεγχος και η λογοδοσία.

Η πρόκληση των ηθικών αποφάσεων από μηχανές μας υπενθυμίζει ότι η τεχνητή νοημοσύνη, για όλη της την ισχύ, παραμένει αντανάκλαση των ανθρώπινων επιλογών. Το ερώτημα δεν είναι μόνο τι μπορούν να κάνουν οι μηχανές, αλλά ποιοι θέλουμε να είμαστε εμείς όταν τις προγραμματίζουμε.

**7.3 Διαφάνεια και Λογοδοσία**

Καθώς τα συστήματα τεχνητής νοημοσύνης γίνονται ολοένα και πιο παρόντα στις αποφάσεις που επηρεάζουν τις ζωές μας — από τις απλές, όπως οι προτάσεις ταινιών σε μια πλατφόρμα, μέχρι τις κρίσιμες, όπως η έγκριση ενός δανείου ή η διάγνωση μιας ασθένειας — η ανάγκη για διαφάνεια και λογοδοσία γίνεται όχι μόνο ηθική απαίτηση, αλλά και τεχνολογική αναγκαιότητα. Ένα σύστημα που επιδρά στον άνθρωπο χωρίς να είναι σαφές πώς λειτουργεί, ποιος το ελέγχει και με ποιο τρόπο λαμβάνει αποφάσεις, είναι ένα σύστημα που διακινδυνεύει να χάσει την εμπιστοσύνη της κοινωνίας.

Η έννοια της διαφάνειας δεν αφορά μόνο το αν μπορούμε να «δούμε» μέσα στον αλγόριθμο. Πιο ουσιαστικά, αφορά το κατά πόσον είναι δυνατόν να κατανοήσουμε τις βασικές αρχές που διέπουν τη λειτουργία του συστήματος. Για παράδειγμα, σε ένα σύστημα που αξιολογεί αιτήσεις για υποτροφίες, είναι σημαντικό να γνωρίζουμε ποια κριτήρια έχουν βαρύτητα, πώς αυτά συνδυάζονται και αν υπάρχει μηχανισμός που εξηγεί την τελική απόφαση στον ενδιαφερόμενο.

Στη σημερινή τεχνολογική πραγματικότητα, πολλά συστήματα τεχνητής νοημοσύνης λειτουργούν ως «μαύρα κουτιά». Δηλαδή, δέχονται δεδομένα εισόδου και παράγουν αποφάσεις χωρίς να είναι ξεκάθαρο τι συμβαίνει στο ενδιάμεσο. Αυτή η «αδιαφάνεια» μπορεί να είναι αποτέλεσμα τεχνικής πολυπλοκότητας (όπως σε αλγορίθμους βαθιάς μάθησης), εμπορικού απορρήτου (όταν ο αλγόριθμος θεωρείται επιχειρηματικό μυστικό), ή ακόμη και έλλειψης πρόβλεψης από τους σχεδιαστές. Το αποτέλεσμα όμως είναι το ίδιο: ο πολίτης δεν γνωρίζει γιατί του απορρίφθηκε η αίτηση, ο εργαζόμενος δεν ξέρει γιατί δεν προκρίθηκε σε μια θέση, ο ασθενής δεν καταλαβαίνει γιατί του προτείνεται μια συγκεκριμένη θεραπεία.

Η διαφάνεια σχετίζεται άμεσα με την έννοια της λογοδοσίας. Δεν έχει νόημα να μπορούμε να «δούμε» τι κάνει το σύστημα αν δεν υπάρχει κάποιος υπεύθυνος να εξηγήσει, να αναλάβει την ευθύνη και, όπου χρειάζεται, να διορθώσει την απόφαση. Η λογοδοσία αφορά ακριβώς αυτό: τη δυνατότητα κάθε πολίτη να απευθυνθεί σε έναν αρμόδιο, να ζητήσει εξηγήσεις, να αμφισβητήσει και να διεκδικήσει επανεξέταση. Ένα σύστημα χωρίς λογοδοσία κινδυνεύει να γίνει αυθαίρετο, ακόμη κι αν οι προθέσεις του σχεδιασμού του ήταν αγαθές.

Η ανάγκη για διαφάνεια και λογοδοσία δεν είναι μόνο ζήτημα ατομικού δικαιώματος, αλλά και συλλογικής προστασίας. Όταν γνωρίζουμε πώς λαμβάνονται οι αποφάσεις και ποιος ευθύνεται γι’ αυτές, μπορούμε ως κοινωνία να κρίνουμε αν είναι δίκαιες, να εντοπίζουμε συστημικά προβλήματα και να απαιτούμε αλλαγές. Η αδιαφάνεια, αντίθετα, δημιουργεί έδαφος για καχυποψία, αδικία και απώλεια εμπιστοσύνης στην τεχνολογία.

Η Ευρωπαϊκή Ένωση, μέσα από πρωτοβουλίες όπως ο **Κανονισμός για την Τεχνητή Νοημοσύνη (AI Act)**, έχει θεσμοθετήσει τη **διαφάνεια** και τη **λογοδοσία** ως υποχρεωτικά στοιχεία για όλα τα συστήματα που χαρακτηρίζονται ως υψηλού κινδύνου. Μεταξύ άλλων, προτείνεται ότι κάθε σύστημα ΤΝ πρέπει να συνοδεύεται από τεχνική τεκμηρίωση, περιγραφή του τρόπου λειτουργίας του, εκτίμηση κινδύνων και πρόβλεψη για ανθρώπινη επίβλεψη. Παράλληλα, οι χρήστες πρέπει να ενημερώνονται όταν αλληλεπιδρούν με μηχανή, και να μπορούν να ζητούν ανθρώπινη παρέμβαση σε περιπτώσεις αμφισβητήσεων.

Στο πρακτικό επίπεδο, η διαφάνεια μπορεί να πάρει πολλές μορφές: από απλές επεξηγήσεις ("γιατί επιλέχθηκε αυτό το αποτέλεσμα") μέχρι διαδραστικές πλατφόρμες που επιτρέπουν στον χρήστη να εξερευνήσει τα διαφορετικά σενάρια που οδήγησαν σε μια απόφαση. Η λογοδοσία, αντίστοιχα, απαιτεί την ύπαρξη μηχανισμών επανόρθωσης, όπως η δυνατότητα προσφυγής, η ύπαρξη ανεξάρτητων ελεγκτικών φορέων και η αποσαφήνιση της νομικής ευθύνης σε περίπτωση βλάβης.

Είναι κρίσιμο να κατανοήσουμε ότι η διαφάνεια και η λογοδοσία δεν είναι εμπόδια στην καινοτομία, αλλά προϋποθέσεις της. Μια τεχνολογία που δεν εμπνέει εμπιστοσύνη, αργά ή γρήγορα θα απορριφθεί από την κοινωνία. Αντίθετα, μια τεχνολογία που εξηγείται, ελέγχεται και ευθυγραμμίζεται με τις θεμελιώδεις αξίες της κοινωνικής δικαιοσύνης, μπορεί να λειτουργήσει ως γέφυρα για μια πιο ώριμη, υπεύθυνη και συμπεριληπτική ψηφιακή εποχή.

Στο τέλος, το ζητούμενο δεν είναι να «αντισταθούμε» στην τεχνητή νοημοσύνη, αλλά να την εντάξουμε στον κοινωνικό ιστό με τρόπο που να διαφυλάσσει τα δικαιώματα και την αξιοπρέπεια όλων. Η διαφάνεια και η λογοδοσία είναι τα θεμέλια αυτού του στόχου.