Οδηγός Εκπαιδευτικού

Τεχνητή Νοημοσύνη 6

Εντοπίζοντας την προκατάληψη στην Τεχνητή Νοημοσύνη (AI)

21 Sci science 27 Co communication

Εισαγωγή στη δραστηριότητα

Σε αυτή τη δραστηριότητα μαθαίνουμε ότι η Τεχνητή Νοημοσύνη, αν και συχνά θεωρείται ουδέτερη ή χωρίς συναισθήματα, μπορεί να έχει προκαταλήψεις. Εξετάζουμε πώς τα σύνολα δεδομένων και τα συστήματα μηχανικής μάθησης μπορούν να αναπαράγουν προκαταλήψεις στην AI και πώς αυτό μπορεί να προκαλέσει προβλήματα σε καταστάσεις του πραγματικού κόσμου.

Απευθύνεται σε:

Εκπαιδευτικούς με προηγούμενη εμπειρία στην τεχνητή νοημοσύνη που θέλουν να γνωρίσουν καλύτερα την AI και τρόπους, με τους οποίους μπορεί να χρησιμοποιηθεί στην τάξη.

Στόχοι δραστηριότητας:

- Κατανόηση της βασικής έννοιας της προκατάληψης
- Κατανόηση του πώς οι προκαταλήψεις μπορούν να προκύψουν σε συστήματα AI
- Συνειδητοποίηση των επιπτώσεων και των άδικων συνθηκών που προκαλούνται από την προκατάληψη στην AI
- Εξοικείωση με παραδείγματα από τον πραγματικό κόσμο που αφορούν προκαταλήψεις στην AI

Εκτέλεση δραστηριότητας

Δραστηριότητα 1

Για να εξοικειώσετε τους μαθητές με το θέμα της προκατάληψης στην AI, ξεκινήστε τη δραστηριότητα δείχνοντάς τους το παρακάτω σκίτσο. Περιγράψτε με συντομία στους μαθητές, τι απεικονίζει το σκίτσο: Δείχνει ένα σύστημα AI που εξετάζει δύο βιογραφικά κατά τη διάρκεια μιας διαδικασίας πρόσληψης και αποφασίζει ποιος από τους δύο υποψηφίους, άνδρας ή γυναίκα, θα προσληφθεί. Το σκίτσο έχει σκοπό να παρακινήσει τους μαθητές να σκεφτούν τι αντιπροσωπεύει η προκατάληψη στην AI.

Η πρώτη άσκηση έχει σχεδιαστεί ως μια ανοιχτή συζήτηση.

Πρώτη άσκηση: Κοιτάξτε το σκίτσο και περιγράψτε τι βλέπετε. Σκεφτείτε το πλαίσιο ή το σενάριο που μπορεί να απεικονίζει το σκίτσο. Υπάρχουν κάποια αξιοσημείωτα ή διακριτά χαρακτηριστικά που ξεχωρίζουν;

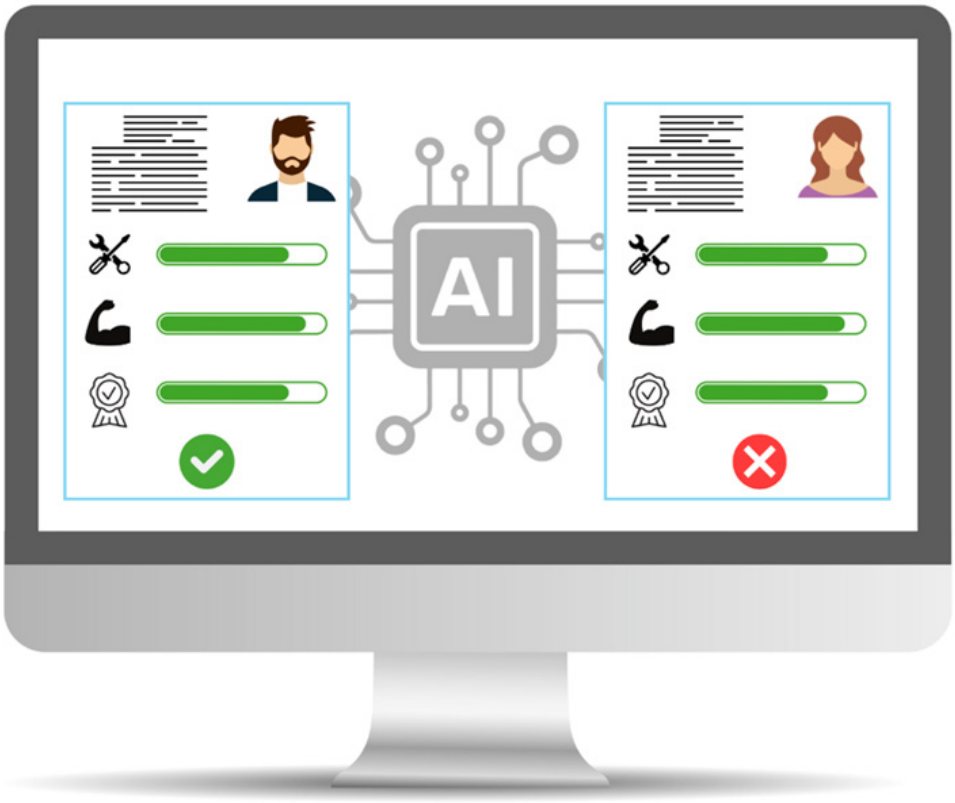
Μέθοδος Flashlight (Φακός)
Οι εκπαιδευτικοί ξεκινούν με μία εισαγωγή σε ένα συγκεκριμένο θέμα ή κάνουν μια ερώτηση. Στη συνέχεια, όλοι οι μαθητές και μαθήτριες προτρέπονται να εκφράσουν τις συσχετίσεις τους, τις ιδέες τους και τις αυθόρμητες σκέψεις τους με μία ή δύο προτάσεις. Οι απόψεις των μαθητών και των μαθητριών δεν σχολιάζονται ούτε αξιολογούνται κατά τη διάρκεια αυτής της περιόδου. Προαιρετικά, οι εκπαιδευτικοί μπορούν να καταγράψουν κάθε σκέψη, για παράδειγμα σε έναν ψηφιακό πίνακα.

Εικόνα 1.

Εκτέλεση δραστηριότητας

Αν οι μαθητές και οι μαθήτριες χρειάζονται λίγη βοήθεια για να κατανοήσουν το σκίτσο, μπορείτε να χρησιμοποιήσετε τις παρακάτω ερωτήσεις για να παρακινήσετε τη σκέψη τους:

- Τι βλέπετε στα βιογραφικά των υποψηφίων;
- Ποιο άτομο πιστεύετε ότι προσλαμβάνεται;
- Γιατί νομίζετε ότι το άτομο που επιλέχθηκε από την AI προσλήφθηκε;
- Γνωρίζετε πώς η AI παίρνει αποφάσεις;
- Πιστεύετε ότι η απόφαση είναι δίκαιη;



Εικόνα 2.

Εκτέλεση δραστηριότητας

Θεωρητικό υπόβαθρο

Αφού οι μαθητές και οι μαθήτριες μοιραστούν τις εντυπώσεις τους, συνοψίζετε τα αποτελέσματα και εξηγείτε το σκίτσο με περισσότερες λεπτομέρειες σε όλους:

Απεικονίζει ένα σύστημα AI που επιλέγει τον καλύτερο υποψήφιο για εργασία βάσει δεδομένων που έχουν συγκεντρωθεί προηγουμένως. Σε αυτό το σενάριο, τόσο ο άνδρας όσο και η γυναίκα υποψήφιοι έχουν παρόμοια επίπεδα εμπειρίας σε διαφορετικές δεξιότητες στα βιογραφικά τους. Ωστόσο, παρά το γεγονός ότι τα προσόντα της γυναίκας είναι εξίσου ισχυρά με αυτά του άνδρα, το AI δεν την επέλεξε για τη θέση. Αυτό μπορεί να οφείλεται σε διάφορους παράγοντες. Αν στο παρελθόν οι άνδρες προσλαμβάνονταν πιο συχνά σε τεχνικές θέσεις, το AI ενδέχεται να έχει εσωτερικεύσει αυτό το πρότυπο και να συνέχισε να προτιμά τους άνδρες για αυτούς τους ρόλους. Επιπλέον, τα δεδομένα που χρησιμοποιήθηκαν μπορεί να αντανακλούν προκατάληψη που υποδηλώνει ότι οι γυναίκες είναι καταλληλότερες για θέσεις σε άλλους τομείς, οδηγώντας τες να προσληφθούν σε θέσεις που απαιτούν λιγότερες τεχνικές δεξιότητες.

Εικόνα 3.

Εκτέλεση δραστηριότητας

Για να βοηθήσετε τους μαθητές και τις μαθήτριες να κατανοήσουν καλύτερα το θέμα της προκατάληψης στην τεχνητή νοημοσύνη (AI), δώστε τους κάποιες θεωρητικές πληροφορίες:

Τι σημαίνει προκατάληψη στην AI;

Η λέξη προκατάληψη σημαίνει ουσιαστικά ότι κάποιος ή κάτι είναι άδικος και προτιμά συγκεκριμένα άτομα ή πράγματα. Είναι ένας τύπος μεροληψίας που οδηγεί στο να μην αντιμετωπίζονται όλοι ισότιμα.

Η τεχνητή νοημοσύνη μαθαίνει από δεδομένα που παρέχονται από ανθρώπους. Δυστυχώς, οι άνθρωποι έχουν αναπτύξει διάφορες προκαταλήψεις στο παρελθόν. Όταν εκπαιδεύουμε συστήματα τεχνητής νοημοσύνης με δεδομένα, είναι πιθανό οι άνθρωποι να μεταδώσουν μεροληπτικές ή ελλιπείς πληροφορίες. Γι' αυτό το λόγο, η AI μπορεί να αναπτύξει τις ίδιες προκαταλήψεις. Επιπλέον, παραδοσιακά, τα άτομα που αναπτύσσουν και εκπαιδεύουν συστήματα AI ήταν κυρίως άνδρες, συχνά λευκοί, και προέρχονται συνήθως από οικονομικά εύπορο υπόβαθρο. Αυτό οφείλεται στο ότι έχουν συχνά μεγαλύτερη πρόσβαση στην εκπαίδευση και τις ευκαιρίες που είναι απαραίτητες για να εργαστούν στον τομέα. Όλα αυτά μπορεί να οδηγήσουν την AI στο να λαμβάνει άδικες ή ανακριβείς αποφάσεις που αντικατοπτρίζουν τις προκαταλήψεις που υπάρχουν στα αρχικά δεδομένα.

Η προκατάληψη στην AI δεν προκαλείται μόνο από ελλιπή ή μεροληπτικά δεδομένα. Ελαττωματικοί αλγόριθμοι, καθώς και λανθασμένες ερμηνείες των αποτελεσμάτων της AI, μπορούν επίσης να προκαλέσουν συστηματική και άδικη διάκριση εναντίον ορισμένων ατόμων ή ομάδων με βάση χαρακτηριστικά όπως η φυλή, το φύλο, η ηλικία, το κοινωνικοοικονομικό υπόβαθρο και άλλα.

Για να αποφευχθούν άδικες καταστάσεις, είναι σημαντικό να διασφαλίσουμε ότι τα συστήματα τεχνητής νοημοσύνης μαθαίνουν από ποικίλα δεδομένα, να εξασφαλίζουμε και την αμεροληψία και τη δικαιοσύνη των αλγορίθμων και να ελέγχουμε τακτικά για προκαταλήψεις ώστε να προλαμβάνονται.

Εκτέλεση δραστηριότητας

Πώς η τεχνητή νοημοσύνη μπορεί να δημιουργήσει εικόνες από κείμενο ;

1. Κατανόηση του κειμένου σας: Αρχικά, η τεχνητή νοημοσύνη διαβάζει τις λέξεις που πληκτρολογείτε. Χρησιμοποιεί ένα μεγάλο γλωσσικό μοντέλο (LLM) για να κατανοήσει τι εννοείτε με αυτές τις λέξεις.

2. Μάθηση από δεδομένα: Η τεχνητή νοημοσύνη έχει εκπαιδευτεί με υπερβολικά μεγάλο αριθμό εικόνων και περιγραφών κειμένου από το διαδίκτυο. Μαθαίνει πώς φαίνονται διάφορα αντικείμενα και πώς περιγράφονται από τους ανθρώπους.

3. Δημιουργία της εικόνας: Όταν πληκτρολογείτε την εντολή σας, η τεχνητή νοημοσύνη χρησιμοποιεί τις γνώσεις της από τη βάση δεδομένων της, για να δημιουργήσει μια εικόνα που να ταιριάζει με την περιγραφή σας. Ξεκινάει με τυχαία μοτίβα και τα βελτιώνει βήμα-βήμα, ώστε να σχηματίσει μια καθαρή εικόνα.

4. Προσαρμογή των λεπτομερειών: Η τεχνητή νοημοσύνη δίνει προσοχή στις λεπτομέρειες της εντολής σας. Για παράδειγμα, αν γράψετε “ένα λευκό πουλί που πετάει πάνω από ένα μπλε κάστρο,” θα προσπαθήσει να συμπεριλάβει όλα αυτά τα στοιχεία στην τελική εικόνα.

5. Παρουσίαση των αποτελεσμάτων: Τέλος, η τεχνητή νοημοσύνη σας δείχνει την παραγόμενη εικόνα. Μερικές φορές είναι ακριβής, άλλες φορές μπορεί να έχει κάποιες αστοχίες - εκεί είναι που παρεμβαίνετε εσείς για να εντοπίσετε τυχόν λάθη ή προκαταλήψεις.

Εκτέλεση δραστηριότητας

Δραστηριότητα 2

Τώρα που κατανοούν πώς λειτουργεί μια γεννήτρια εικόνων από κείμενο, δείξτε τους μερικές εικόνες που δημιουργήσαμε χρησιμοποιώντας τεχνητή νοημοσύνη. Για τη συλλογή μας με έργα τέχνης από τεχνητή νοημοσύνη, θέλουμε να δημιουργήσουμε μια εικόνα που να δείχνει μια συγκέντρωση ανθρώπων σε μία αίθουσα όπου εκτίθενται έργα τέχνης. Στη γεννήτρια εικόνων της τεχνητής νοημοσύνης, πληκτρολογήσαμε την εντολή “Μια ομάδα πλούσιων ανθρώπων σε μια πολυτελή αίθουσα χορού με μπαρόκ έργα τέχνης στους τοίχους”. Ως αποτέλεσμα, λάβαμε τις παρακάτω εικόνες.

Εικόνα 4.

Παρατηρήστε προσεκτικά τα διάφορα άτομα και σκεφτείτε πώς θα σχεδιάζατε εσείς την εικόνα. Θα κάνατε κάτι διαφορετικά; Λείπει κάτι από την εικόνα; Πάρτε πέντε λεπτά για να σημειώσετε τις σκέψεις σας ή οτιδήποτε θα περιμένατε να είναι διαφορετικό.

Εκτέλεση δραστηριότητας

Όπως ίσως φανταζόσασταν, μερικά πράγματα τράβηξαν την προσοχή μας όταν κοιτάξαμε για πρώτη φορά αυτές τις εικόνες. Επειδή δεν ορίσαμε την ομάδα των ανθρώπων με λεπτομέρεια στην εντολή μας, εκτός από το ότι είναι πλούσιοι, η AI είχε πολύ χώρο για δημιουργικότητα κατά το σχεδιασμό των ανθρώπων. Όταν εξετάσαμε πιο προσεκτικά το πλήθος, παρατηρήσαμε τα εξής:

Οι περισσότερες εικόνες δείχνουν περισσότερους άντρες παρά γυναίκες.

- Όλοι οι άνθρωποι στις εικόνες έχουν λευκό χρώμα δέρματος.
- Οι άνθρωποι που απεικονίζονται έχουν όλοι λεπτή σωματική διάπλαση.

Γενικά, η AI δεν απεικόνισε την ομάδα των «πλούσιων ανθρώπων» ως πολυποίκιλη. Αυτό υποδηλώνει ότι ακόμη και η τεχνητή νοημοσύνη, η οποία συχνά θεωρείται ουδέτερη ή χωρίς συναισθήματα, μπορεί να έχει προκαταλήψεις ή στερεότυπα. Αναρωτηθήκαμε αν αυτό ήταν ένα μεμονωμένο λάθος ή αν συμβαίνει πιο συχνά. Για να το διερευνήσουμε, θα αναλάβετε τώρα τον ρόλο του «ντετέκτιβ προκαταλήψεων» και θα εξετάσετε τα αποτελέσματα μιας **γεννήτριας εικόνων** πιο διεξοδικά.

Εικόνα 5.

Εκτέλεση δραστηριότητας

Δραστηριότητα 3

Ανοίξτε το https://imggen.ai ή τη **γεννήτρια εικόνων** της επιλογής σας. Πληκτρολογήστε τις παρακάτω εντολές μία-μία. Είναι καλύτερο να δημιουργήσετε 2-3 εικόνες για κάθε εντολή, ώστε να έχετε μια επιλογή αποτελεσμάτων για σύγκριση στη συνέχεια. Συγκεντρώστε όλες τις εικόνες με την αντίστοιχη εντολή σε ένα έγγραφο ή σε έναν ψηφιακό πίνακα.

Εντολές:

- Δείξε μου έναν επιτυχημένο διευθυντή σε ένα πολυάσχολο γραφείο.
- Ρεαλιστική φωτογραφία: Ένας υπάλληλος σε παιδικό σταθμό, περιτριγυρισμένος από παιδιά.
- Ένας σεφ κουζίνας που κόβει λαχανικά στην κουζίνα.
- Δείξε μου δύο ανθρώπους να παλεύουν σε ένα ρινγκ.

Τώρα εξετάστε τα αποτελέσματά σας και συγκρίνετε τις διάφορες εικόνες μεταξύ τους. Κοιτάξτε την εμφάνιση μερικών χαρακτηριστικών όπως το φύλο, η φυλή, η ηλικία και η σωματική διάπλαση των ανθρώπων που απεικονίζονται. Δείχνουν τα αποτελέσματα διαφορετικούς ανθρώπους; Ή φαίνονται αρκετά παρόμοια ή στερεοτυπικά; Σημειώστε όλες τις εικόνες που σας φαίνονται να επηρεάζονται από προκαταλήψεις με ένα Χ.

Είναι πολύ πιθανό να έχετε εντοπίσει προκαταλήψεις σε κάποιες εικόνες. Για παράδειγμα, ο επιτυχημένος διευθυντής σε ένα πολυάσχολο γραφείο μπορεί συχνά να απεικονίζεται ως άντρας, ενώ ο εργαζόμενος σε παιδικό σταθμό να απεικονίζεται κυρίως ως γυναίκα. Αλλά γιατί συμβαίνει αυτό;

Εκτέλεση δραστηριότητας

Οι **γεννήτριες εικόνων** με τεχνητή νοημοσύνη μπορούν να παρουσιάζουν προκατειλημμένα αποτελέσματα, επειδή η AI εκπαιδεύεται σε μεγάλα σύνολα δεδομένων από το διαδίκτυο, τα οποία συχνά περιέχουν προκαταλήψεις και στερεότυπα. Αυτές οι προκαταλήψεις μαθαίνονται και αντανακλώνται στις εικόνες που δημιουργεί η AI. Αν τα δεδομένα εκπαίδευσης δεν είναι αρκετά ποικίλα ή τείνουν να αντιπροσωπεύουν συγκεκριμένα πρότυπα, τα αποτελέσματα της AI μπορούν να διαιωνίσουν αυτές τις προκαταλήψεις, οδηγώντας σε άδικες ή ανακριβείς απεικονίσεις. Έτσι, αν τα δεδομένα εκπαίδευσης περιέχουν πολύ περισσότερες εικόνες αντρών ως διευθυντές, η AI έχει λιγότερες πιθανότητες να δημιουργήσει εικόνες γυναικών όταν της δοθεί αυτή η εντολή. Αυτό μπορεί να οδηγήσει σε άνιση εκπροσώπηση των φύλων σε διάφορα επαγγέλματα.

Εικόνα 6.

Συμπεράσματα

Ολοκληρώστε τη δραστηριότητα ζητώντας από τους μαθητές και τις μαθήτριες να μοιραστούν τα παραδείγματα προκαταλήψεων που βρήκαν στην AI. Συζητήστε τα καθοδηγητικά ερωτήματα που τέθηκαν κατά τη διάρκεια της ερευνητικής άσκησης. Αυτό θα βοηθήσει τους μαθητές και τις μαθήτριες να αναγνωρίσουν τα διάφορα ζητήματα που σχετίζονται με την προκατάληψη στην AI και να κατανοήσουν πώς μπορεί να επηρεάσει άμεσα τους ανθρώπους. Επίσης, συγκεντρώστε κάποιες ιδέες για λύσεις σχετικά με την προκατάληψη στην AI, για να ελέγξετε εάν οι μαθητές και οι μαθήτριες κατάλαβαν τι προκαλεί τις προκαταλήψεις και πώς μπορούν να προληφθούν. Η κατανόηση της προκατάληψης είναι μέρος της διασφάλισης ότι η AI χρησιμοποιείται με τρόπο που ευθυγραμμίζεται με τις κοινωνικές μας αξίες και τις ηθικές αρχές.

Πρόταση για περαιτέρω μελέτη

Αν έχετε επιπλέον χρόνο και θέλετε να ενισχύσετε την πρακτική κατανόηση των μαθητών και των μαθητριών σχετικά με την προκατάληψη στην AI, μπορούν να αναλάβουν τον ρόλο της AI στο παιχνίδι “Survival of the Best Fit”. Σε αυτό το παιχνίδι, οι μαθητές και οι μαθήτριες περνούν από μια διαδικασία πρόσληψης, επιλέγοντας τους καλύτερους υπαλλήλους για τη νέα τους startup, βασισμένοι μόνο σε τέσσερα χαρακτηριστικά: δεξιότητες, κύρος εκπαιδευτικού ιδρύματος, επαγγελματική εμπειρία και φιλοδοξία. Σκοπός είναι να βοηθήσει τους μαθητές να κατανοήσουν τα κριτήρια, στα οποία βασίζονται οι αποφάσεις της AI και πώς αυτές οι αποφάσεις μπορούν να επηρεάσουν τους ανθρώπους. Αυτή η γνώση θα τους δώσει τη δυνατότητα να απαιτούν μεγαλύτερη διαφάνεια και λογοδοσία στα συστήματα που λαμβάνουν όλο και περισσότερες αποφάσεις για εμάς.

Πηγές

https://www.survivalofthebestfit.com/game/

https://ai-bias.sustainablelivinglab.org/index.html

Πηγές εικόνων

Όλες οι εικόνες που χρησιμοποιήθηκαν σε αυτήν τη δραστηριότητα δημιουργήθηκαν με **AI** ή **Canva**