Το πρόβλημα της μεροληψίας στην αναγνώριση προσώπου

Source: https://lab.imedd.org/to-provlima-tis-merolipsias-stin-anagn/

Το iMEdD Lab αναδημοσιεύει, σε ελληνική μετάφραση, το άρθρο του Γουίλιαμ Κράμπλερ από το Technology Policy Blog του Κέντρου Στρατηγικών και Διεθνών Σπουδών στην Ουάσινγκτον.

Ερευνητές έχουν διαπιστώσει ότι κορυφαίοι αλγόριθμοι αναγνώρισης προσώπου έχουν διαφορετικά ποσοστά ακρίβειας για διαφορετικές δημογραφικές κατηγορίες. Η πρώτη μελέτη που κατέγραψε αυτό το αποτέλεσμα ήταν μια έκθεση του 2003 από το Εθνικό Ινστιτούτο Προτύπων και Τεχνολογίας των ΗΠΑ (National Institute of Standards and Technology, NIST), το οποίο διαπίστωσε ότι οι αλγόριθμοι δυσκολεύονταν περισσότερο στην αναγνώριση γυναικών από όσο ανδρών, όπως και στην αναγνώριση νεαρών ατόμων σε σχέση με αυτά μεγαλύτερης ηλικίας. Το 2018, ερευνητές από το MIT και τη Microsoft προκάλεσαν αίσθηση με μια έκθεση που δείχνει ότι οι αλγόριθμοι ταξινόμησης φύλου –οι οποίοι σχετίζονται, αν και δεν ταυτίζονται με τους αλγόριθμους αναγνώρισης προσώπου– είχαν ποσοστό σφάλματος μόλις 1% για τους λευκούς άνδρες αλλά σχεδόν 35% για σκουρόχρωμες γυναίκες. Η πιο ενδελεχής μελέτη πάνω σε αυτήν την ανισότητα ολοκληρώθηκε από το NIST το 2019. Μέσα από τις δοκιμές τους, οι ερευνητές του NIST επιβεβαίωσαν ότι η πλειονότητα των αλγορίθμων εμφανίζει δημογραφικές διαφορές στα ποσοστά τόσο των ψευδώς αρνητικών (απόρριψη σωστής αντιστοίχισης) όσο των ψευδώς θετικών (αντιστοίχιση με λάθος άτομο) αποτελεσμάτων.

Το NIST βρήκε ότι δημογραφικοί παράγοντες είχαν πολύ μεγαλύτερο αντίκτυπο στα ποσοστά των ψευδώς θετικών αποτελεσμάτων –όπου οι διαφορές στο ποσοστό σφάλματος μπορούσαν να είναι από δεκαπλάσιες ως και εκατονταπλάσιες ανάμεσα στις διάφορες δημογραφικές κατηγορίες. Ως προς τα ψευδώς αρνητικά, σε γενικές γραμμές, σημειώνονταν διαφορές στο ποσοστό σφάλματος περίπου κατά συντελεστή τρία. Οι διαφορές στα ποσοστά των ψευδώς θετικών αποτελεσμάτων προκαλούν κατά κανόνα μεγαλύτερη ανησυχία, καθώς η εσφαλμένη ταυτοποίηση κάποιου ατόμου συνήθως ενέχει μεγαλύτερους κινδύνους από το να απορρίπτεται εσφαλμένα κάποιος από ένα σύστημα αναγνώρισης προσώπου (για παράδειγμα, όταν το iPhone σας δεν σας συνδέει με την πρώτη προσπάθεια). Το NIST διαπίστωσε ότι Ασιάτες, Αφροαμερικανοί και Αμερικανοί Ινδιάνοι είχαν γενικά υψηλότερα ποσοστά σφάλματος ψευδώς θετικών αποτελεσμάτων από τα λευκά άτομα, οι γυναίκες είχαν υψηλότερα ποσοστά ψευδώς θετικών αποτελεσμάτων από τους άνδρες, και τα παιδιά και οι ηλικιωμένοι είχαν υψηλότερα ποσοστά ψευδώς θετικών αποτελεσμάτων από τους ενήλικες μέσης ηλικίας.

Ο πιο σημαντικός παράγοντας για τη μείωση της μεροληψίας φαίνεται να είναι η επιλογή των δεδομένων, με τα οποία οι αλγόριθμοι εκπαιδεύονται για τη δημιουργία μοντέλων. Ωστόσο, ένα διευρυμένο καθεστώς ελέγχου των «training data» θα μπορούσε να βρει αντίσταση από προγραμματιστές.

Ωστόσο, το NIST, επίσης, κατέληξε σε πολλά ενθαρρυντικά συμπεράσματα. Το πρώτο είναι ότι οι διαφορές ανάμεσα στις δημογραφικές κατηγορίες ήταν πολύ χαμηλότερες σε αλγόριθμους που ήταν πιο ακριβείς συνολικά. Αυτό σημαίνει ότι, καθώς τα συστήματα αναγνώρισης προσώπου συνεχίζουν να βελτιώνονται, ο αντίκτυπος της μεροληψίας θα μειωθεί. Ακόμα πιο ευοίωνο ήταν το γεγονός ότι ορισμένοι αλγόριθμοι δεν έδειξαν απολύτως καμία διακριτή μεροληψία, υποδηλώνοντας ότι η μεροληψία μπορεί να εξαλειφθεί εξ ολοκλήρου με τους σωστούς αλγόριθμους και διαδικασίες ανάπτυξης. Ο πιο σημαντικός παράγοντας για τη μείωση της μεροληψίας φαίνεται να είναι η επιλογή των «δεομένων εκπαίδευσης» («training data»), των δεδομένων με τα οποία οι αλγόριθμοι εκπαιδεύονται για τη δημιουργία μοντέλων. Εάν οι αλγόριθμοι εκπαιδεύονται σε σύνολα δεδομένων που περιέχουν πολύ λίγα παραδείγματα μιας συγκεκριμένης δημογραφικής κατηγορίας, το μοντέλο που θα προκύψει θα είναι χειρότερο στην ακριβή αναγνώριση μελών αυτή της κατηγορίας σε πραγματικές εφαρμογές. Οι ερευνητές του NIST ανέπτυξαν τη θεωρία ότι αυτός ενδέχεται να είναι ο λόγος για τον οποίο πολλοί από τους αλγόριθμους που αναπτύχθηκαν στις Ηνωμένες Πολιτείες είχαν χειρότερη απόδοση σε πρόσωπα Ασιατών από αλγόριθμους που αναπτύχθηκαν στην Κίνα. Οι κινεζικές ομάδες πιθανώς χρησιμοποίησαν, για την εκπαίδευση των αλγορίθμων, σύνολα δεδομένων με μεγαλύτερη αντιπροσώπευση Ασιατών, βελτιώνοντας την απόδοσή τους σε αυτήν την κατηγορία.

Λόγω της σημασίας της επιλογής των «training data» για την απόδοση και μεροληψία των αλγορίθμων αναγνώρισης προσώπου, αυτά τα σύνολα δεδομένων γίνονται ολοένα και πιο δημοφιλής στόχος, στο πλαίσιο κανονιστικών προτάσεων. Η Ευρωπαϊκή Ένωση, για παράδειγμα, εισηγήθηκε πρόσφατα ότι ένα κανονιστικό πλαίσιο για συστήματα τεχνητής νοημοσύνης υψηλού κινδύνου, όπως η αναγνώριση προσώπου, περιλαμβάνει την προϋπόθεση πως τα «δεδομένα εκπαίδευσης» οφείλουν να «έχουν επαρκή ευρύτητα» και «να διασφαλίζεται ότι όλες οι σχετικές διαστάσεις του φύλου, της εθνότητας και άλλοι πιθανοί λόγοι απαγορευμένων διακρίσεων αντικατοπτρίζονται δεόντως στα εν λόγω σύνολα δεδομένων». Οι έλεγχοι για την επιβεβαίωση της ποιότητας των «training data» θα μπορούσαν να αποτελέσουν σημαντικό εργαλείο για την αντιμετώπιση των κινδύνων της μεροληψίας στην αναγνώριση προσώπου. Ωστόσο, ένα διευρυμένο καθεστώς ελέγχου θα μπορούσε να βρει αντίσταση από προγραμματιστές που ενδέχεται να αντιταχθούν στην αύξηση του χρόνου ή του κόστους της διαδικασίας ανάπτυξης ή στο να διαθέσουν οποιοδήποτε τμήμα του αλγορίθμου τους σε έρευνα τρίτων.

Οι Αφροαμερικανοί άνδρες, για παράδειγμα, αντιπροσωπεύονται δυσανάλογα στις φωτογραφικές βάσεις δεδομένων που χρησιμοποιούν πολλά συστήματα των αρχών επιβολής του νόμου. Αυτό είναι αποτέλεσμα ευρύτερων κοινωνικών τάσεων, αλλά θα μπορούσε να σημαίνει ότι οι Αφροαμερικανοί άνδρες θα αναγνωρίζονται και θα παρακολουθούνται συχνότερα.

Η κυβερνητική δράση θα είναι απαραίτητη για να ενθαρρυνθεί η υιοθέτηση πρακτικών ελέγχου των δεδομένων με τα οποία εκπαιδεύονται οι αλγόριθμοι. Το ευκολότερο πρώτο βήμα θα ήταν η ενημέρωση των πολιτικών προμηθειών σε πολιτειακό, τοπικό και ομοσπονδιακό επίπεδο, ώστε να απαγορευτούν οι κυβερνητικές αγορές από προμηθευτές λογισμικού αναγνώρισης προσώπου που δεν έχουν περάσει έναν αλγοριθμικό έλεγχο ο οποίος να ενσωματώνει την αξιολόγηση των «training data» ως προς τη μεροληψία. Αυτοί οι έλεγχοι θα μπορούσαν να

διενεργηθούν από μια ρυθμιστική αρχή ή από ανεξάρτητους αξιολογητές διαπιστευμένους από την εκάστοτε κυβέρνηση. Τουλάχιστον, αυτό θα έπρεπε να απαιτείται από τον νόμο ή τις πολιτικές για χρήσεις υψηλού κινδύνου, όπως η χρήση από υπηρεσίες επιβολής του νόμου. Οι ομοσπονδιακοί, υπεύθυνοι για τη χάραξης πολιτικής, φορείς, επίσης, θα μπορούσαν να βοηθήσουν στη μείωση των κινδύνων μεροληψίας, εξουσιοδοτώντας το NIST με την επίβλεψη της δημιουργίας δημόσιων, δημογραφικά αντιπροσωπευτικών συνόλων δεδομένων, που οποιαδήποτε εταιρεία λογισμικού αναγνώρισης προσώπου θα μπορούσε να χρησιμοποιήσει για εκπαίδευση αλγορίθμων.

Ωστόσο, η μεροληψία μπορεί να εκδηλωθεί όχι μόνο ως προς τους αλγόριθμους που χρησιμοποιούνται, αλλά και ως προς τους καταλόγους επιτήρησης με τους οποίους αντιστοιχίζονται αυτά τα συστήματα. Ακόμη και αν ένας αλγόριθμος δεν εμφανίζει καμία διαφορά στην ακρίβεια ανάμεσα σε δημογραφικές ομάδες, και πάλι η χρήση του θα μπορούσε να έχει κάποια ανισότητα ως αντίκτυπο, εάν ορισμένες ομάδες υπερεκπροσωπούνται στις βάσεις δεδομένων. Οι Αφροαμερικανοί άνδρες, για παράδειγμα, αντιπροσωπεύονται δυσανάλογα στις φωτογραφικές βάσεις δεδομένων που χρησιμοποιούν πολλά συστήματα αναγνώρισης προσώπου των αρχών επιβολής του νόμου για την αντιστοίχιση ατόμων. Αυτό είναι αποτέλεσμα ευρύτερων κοινωνικών τάσεων, αλλά εάν η αναγνώριση προσώπου γίνει κοινό εργαλείο αστυνόμευσης, αυτό θα μπορούσε να σημαίνει ότι οι Αφροαμερικανοί άνδρες θα αναγνωρίζονται και θα παρακολουθούνται συχνότερα, καθώς πολλοί έχουν ήδη καταγραφεί σε βάσεις δεδομένων των αρχών επιβολής του νόμου. Σε αντίθεση με το ζήτημα της διαφοροποιημένης ακρίβειας, αυτό δεν είναι ένα πρόβλημα που μπορεί να λυθεί με τη βελτίωση της τεχνολογίας.

Με αυτό τονίζεται πόσο σημαντικό είναι να μετατοπιστεί η συζήτηση γύρω από τους κινδύνους της αναγνώρισης προσώπου. Όλο και περισσότερο, οι πρωταρχικοί κίνδυνοι δεν θα προέρχονται από περιπτώσεις όπου η τεχνολογία αποτυγχάνει, αλλά μάλλον από περιπτώσεις όπου η τεχνολογία λειτουργεί ακριβώς όπως πρέπει. Οι συνεχείς βελτιώσεις στα δεδομένα τεχνολογίας και εκπαίδευσης θα εξαλείψουν σιγά σιγά τις υπάρχουσες μεροληψίες των αλγορίθμων, μειώνοντας πολλούς από τους τρέχοντες κινδύνους της τεχνολογίας και διευρύνοντας τα οφέλη που μπορούν να αποκτηθούν από την υπεύθυνη χρήση. Από την άλλη, αυτό, επίσης, θα καταστήσει τις εφαρμογές πιο ελκυστικές για τους χειριστές τους, δημιουργώντας νέους προβληματισμούς. Καθώς οι υπεύθυνοι φορείς για τη χάραξη πολιτικής μελετούν τον καλύτερο τρόπο για την κατασκευή συστημάτων διακυβέρνησης που θα διαχειρίζονται την αναγνώριση προσώπου, θα πρέπει να διασφαλίσουν ότι οι λύσεις τους είναι προσαρμοσμένες στον κατεύθυνση που παίρνει η τεχνολογία, και όχι στο σημείο όπου βρίσκεται σήμερα. Η μεροληψία στους αλγόριθμους αναγνώρισης προσώπου είναι ένα πρόβλημα με περισσότερες από μία διαστάσεις. Οι τεχνικές βελτιώσεις ήδη προσπαθούν να συμβάλλουν στη λύση, αλλά πολλά θα συνεχίσουν να εξαρτώνται από τις αποφάσεις που λαμβάνουμε σχετικά με τον τρόπο χρήσης και διακυβέρνησης της τεχνολογίας.

Μετάφραση: Εβίτα Λύκου

Πηγή: Γουίλιαμ Κράμπλερ, *The Problem of Bias in Facial Recognition*, Technology Policy Blog, CSIS: Technology Policy Program, 1 *Μαΐου* 2020

[*]*Ο συγγραφέας του άρθρου, Γουίλιαμ Κράμπλερ, είναι βοηθός έρευνας στο Πρόγραμμα Τεχνολογικής Πολιτικής (Technology Policy Program) στο Κέντρο Στρατηγικών & Διεθνών Σπουδών (Center for Strategic & International Studies, CSIS) στην Ουάσιγκτον.*

Το παρόν αποτελεί αναδημοσίευση, σε ελληνική μετάφραση. Το άρθρο πρωτογενώς δημοσιεύθηκε στο Technology Policy Blog, το οποίο παράγεται από το Technology Policy Program στο CSIS, έναν ιδιωτικό, φοροαπαλλασσόμενο οργανισμό που εστιάζει σε διεθνή θέματα δημόσιας πολιτικής. Η έρευνά του είναι μη κομματική και μη ιδιοκτησιακή. Το CSIS δεν λαμβάνει συγκεκριμένες πολιτικές θέσεις. Κατά συνέπεια, όλες οι απόψεις, οι θέσεις και τα συμπεράσματα που εκφράζονται σε αυτή τη δημοσίευση θα πρέπει να θεωρείται ότι ανήκουν αποκλειστικά στον συγγραφέα.

The Problem of Bias in Facial Recognition

May 1, 2020

By: William Crumpler

Source: https://www.csis.org/blogs/technology-policy-blog/problem-bias-facial-recognition

Researchers have found that leading facial recognition algorithms have different accuracy rates for different demographic groups. The first study to demonstrate this result was a 2003 report by the National Institute of Standards and Technology (NIST), which found that female subjects were more difficult for algorithms to recognize than male subjects, and young subjects more difficult to recognize than older subjects. In 2018, researchers from MIT and Microsoft generated news with a report showing that gender classification algorithms—which are related, though distinct from face identification algorithms—had error rates of just 1% for white men, but almost 35% for dark-skinned women. The most thorough investigation of this disparity was completed by NIST in 2019. Through their testing, NIST confirmed that a majority of algorithms exhibit demographic differences in both false negative rates (rejecting a correct match) and false positive rates (matching to the wrong person).

NIST found that demographic factors had a much larger effect on false positive rates—where differences in the error rate between demographic groups could vary by

a factor of ten or even one hundred—than false negative rates—where differences were generally within a factor of three. Differences in false positive rates are generally of greater concern, as there is usually greater risk in misidentifying someone than in having someone be incorrectly rejected by a facial recognition system (as when your iPhone doesn't log you in on the first try). NIST found that Asians, African Americans, and American Indians generally had higher false positive error rates than white individuals, women had higher false positive rates than men, and children and the elderly had higher false positive rates than middle aged adults.

However, NIST also came to several encouraging conclusions. The first is that differences between demographic groups were far lower in algorithms that were more accurate overall. This means that as facial recognition systems continue to improve, the effects of bias will be reduced. Even more promising was that some algorithms demonstrated no discernible bias whatsoever, indicating that bias can be eliminated entirely with the right algorithms and development processes. The most important factor in reducing bias appears to be the selection of training data used to build algorithmic models. If algorithms are trained on datasets that contain very few examples of a particular demographic group, the resulting model will be worse at accurately recognizing members of that group in real world deployments. NIST's researchers theorized that this may be the reason many algorithms developed in the United States performed worse on Asian faces than algorithms developed in China. Chinese teams likely used training datasets with greater representation of Asian faces, improving their performance on that group.

Because of the importance of training data selection on the performance and bias of facial recognition algorithms, these datasets have become an increasingly popular target for regulatory proposals. The EU, for example, recently proposed that a regulatory framework for high-risk AI systems like facial recognition include requirements that training data be "sufficiently broad," and reflect "all relevant dimensions of gender, ethnicity and other possible grounds of prohibited discrimination." Training data audits to confirm the quality of training datasets could become an important tool for addressing the risks of bias in facial recognition. However, an expanded audit regime could face resistance from developers who will oppose adding additional time or cost to the development process, or opening any part of their algorithm to third party investigation.

Government action will be necessary to encourage the adoption of training data audit practices. The easiest first step would be to update procurement policies at the state, local, and federal level to ban government purchases from facial recognition vendors that have not passed an algorithmic audit incorporating the evaluation of training data for bias. These audits could be undertaken by a regulator or by independent assessors accredited by a government. At a minimum, this should be required by law or policy

for high-risk uses like law enforcement deployments. Federal policymakers could also help to reduce bias risks by empowering NIST to oversee the construction of public, demographically representative datasets that any facial recognition company could use for training.

However, bias can manifest not only in the algorithms being used, but also in the watchlists these systems are matching against. Even if an algorithm shows no difference in its accuracy between demographics, its use could still result in a disparate impact if certain groups are over-represented in databases. African American males, for example, are disproportionately represented in the mugshot databases many law enforcement facial recognition systems use for matching. This is the result of larger social trends, but if facial recognition becomes a common policing tool, this could mean that African American males will be more frequently identified and tracked since many are already enrolled in law enforcement databases. Unlike the question of differential accuracy, this is not a problem that can be solved with better technology.

This highlights the importance of shifting the conversation around the risks of facial recognition. Increasingly, the primary risks will not come from instances where the technology fails, but rather from instances where the technology works exactly as it is meant to. Continued improvements to technology and training data will slowly eliminate the existing biases of algorithms, reducing many of the technology's current risks and expanding the benefits that can be gained from responsible use. But this will also make deployments more attractive to operators, creating new sets of concerns. As policymakers consider how best to construct governance systems to manage facial recognition, they should ensure their solutions are tailored to where the technology is heading, not where it is at today. Bias in facial recognition algorithms is a problem with more than one dimension. Technical improvements are already helping contribute to the solution, but much will continue to depend on the decisions we make about how the technology is used and governed.

William Crumpler is a research assistant with the Technology Policy Program at the Center for Strategic and International Studies in Washington, DC.