Δήμου Παντελής

Α.Μ.: 71447092

Τρέχον έτος σπουδών: 12ο

Ηθική και Τεχνητή Νοημοσύνη

Μέλη τριμελούς επιτροπής:

1. Νικολάου Γρηγόριος GRIGORIOS NIKOLAOU Digitally signed by GRIGORIOS NIKOLAOU Date: 2022.11.08 12:57:45 +02'00'

2. Βασιλειάδου Σουλτάνα Soultana Vasileiadou Digitally signed by Soultana Vasileiadou Date: 2022.11.11 20:10:53 +02'00'

3. Δρόσος Χρήστος Christos Drosos Digitally signed by Christos Drosos Date: 2022.11.08 21:30:00 +02'00'

ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΠΤΥΧΙΑΚΗΣ/ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

Ο/η κάτωθι υπογεγραμμένος/η Δήμου Παντελής του Γεωργίου, με αριθμό μητρώου 71447092 φοιτητής/τρια του Πανεπιστημίου Δυτικής Αττικής της Σχολής Μηχανικών του Τμήματος Βιομηχανικής Σχεδίασης και Παραγωγής, δηλώνω υπεύθυνα ότι:

«Είμαι συγγραφέας αυτής της πτυχιακής/διπλωματικής εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, οι όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε ακριβώς είτε παραφρασμένες, αναφέρονται στο σύνολό τους, με πλήρη αναφορά στους συγγραφείς, τον εκδοτικό οίκο ή το περιοδικό, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το διαδίκτυο. Επίσης, βεβαιώνω ότι αυτή η εργασία έχει συγγραφεί από μένα αποκλειστικά και αποτελεί προϊόν πνευματικής ιδιοκτησίας τόσο δικής μου, όσο και του Ιδρύματος.

Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του πτυχίου μου».

Ο/Η Δηλών/ούσα

Πίνακας περιεχομένων

Περιεχόμενα

Περίληψη

Η τεχνητή νοημοσύνη είναι ένα αντικείμενο το οποίο είναι ευρέως διαδεδομένα σε πολλούς τομείς της καθημερινότητας μας. Κάτι το οποίο δημιουργεί πολλές αμφιβολίες όσο αφορά το θέμα της ηθικής, καθώς το AI δεν είναι ακόμα εντελώς έτοιμο για χρήση σε όλους τους τομείς στη ζωή μας. Συνεπώς στη παρούσα διπλωματική εργασία θα αναλύσουμε κάποιους πολύ σημαντικούς όρους όπως είναι το FAT το οποίο είναι η δικαιοσύνη, η ευθύνη και η διαφάνεια, όροι οι οποίοι πρέπει να λαμβάνονται υπόψη σε κάθε εφαρμογή τεχνητής νοημοσύνης. Βέβαια όπως θα αναφέρουμε παρακάτω, αυτές οι αρχές δεν λαμβάνονται πάντα υπόψη, με απόρροια να δημιουργούνται συστήματα τα οποία έχουν πολλές προκαταλήψεις στον κώδικα τους, κάτι το οποίο με τη σειρά του οδηγεί σε πολλά προβλήματα. Επιπρόσθετα θα αναφερθούμε σε ηθικά διλήμματα που υπάρχουν στον τομέα της ιατρικής, διότι είναι από τα πιο σημαντικά στοιχεία της ζωής μας, και ένας τομέας ο οποίος επηρεάζει τις ζωές όλων των ανθρώπων άμεσα Επιπρόσθετα θα αναλύσουμε πόσο ασφαλή είναι η μηχανική μάθηση καθώς είναι μια πολύ σημαντική προέκταση της τεχνητής νοημοσύνης και θα μπορούσε κανείς να πει ότι είναι ο πυλώνας της σωστής λειτουργίας της. Τέλος θα παραθέσω τις απόψεις μου σχετικά με την ηθική της τεχνητής νοημοσύνης και τις εφαρμογές της στον τομέα της υγείας καθώς και την ασφάλεια χρήσης της όσο αφορά την εκπαίδευση που μπορεί να δεχτεί.

Κεφάλαιο 1 : Εισαγωγή

Η παρούσα διπλωματική εργασία αφορά την ηθική στο θέμα της Τεχνητής Νοημοσύνης . Για να αναφερθούμε στην ηθική πρέπει πρώτα να γνωρίζουμε τι είναι η Τεχνητή Νοημοσύνη ή αλλιώς γνωστή και ως "A.I."(Artificial Intelligence). Ως Τεχνητή Νοημοσύνη ορίζουμε την επιστήμη και μηχανική της κατασκευής έξυπνων μηχανών, πιο συγκεκριμένα ευφυών προγραμμάτων. Λέγοντας ευφυή προγράμματα αναφερόμαστε σε προγράμματα τα οποία έχουν την ικανότητα να βελτιώνονται από μόνα τους και να "μαθαίνουν" από τα λάθη τους για την επίτευξη καλύτερων αποτελεσμάτων, το οποίο αυτό ονομάζεται μηχανική μάθηση. Η μηχανική μάθηση έχει γίνει ένα θεμελιώδες εργαλείο για τον κόσμο του προγραμματισμού, καθώς μπορεί να εξελίσσεται ταχύτατα και να προσαρμόζεται σε διάφορες καταστάσεις , κάτι το οποίο είναι ταυτόχρονα ένα ελάττωμα καθώς πολλά άτομα που θέλουν να κάνουν "επίθεση" μπορούν να εκμεταλλευτούν τα συστήματα της . Για αυτόν ακριβώς τον λόγω στα κεφάλαια που θα ακολουθήσουν θα αναλύσουμε το F.A.T.(fairness , accountability ,transparency) στη μηχανική μάθηση τα οποία είναι βασικές αρχές για τη διατήρηση της ηθικής στο τομέα της τεχνητής νοημοσύνης, καθώς και κάποιες άλλες σημαντικές αρχές οι οποίες είναι εξίσου σημαντικές.

Το FAT είναι από τα πιο σημαντικά στοιχεία που πρέπει να λαμβάνουν υπόψη τους οι ερευνητές κατά τη δημιουργία ενός αλγορίθμου AI, καθώς με την βοήθεια των αρχών αυτών μπορούμε να φτιάχνουμε συστήματα τα οποία είναι πιο κοντά στην ανθρώπινη φύση. Κάτι το οποίο είναι εξαιρετικά σημαντικό όταν ο σκοπός του AI είναι να δημιουργηθούν αυτοματοποιημένα συστήματα, τα οποία θα μιμούνται την ανθρώπινη συμπεριφορά και θα μπορούν να λαμβάνουν πολλά πράγματα υπόψιν τους όταν παίρνουν αποφάσεις. Οπότε το FAT είναι ένα πολύ σημαντικό στοιχείο το οποίο θα αναλύσουμε σε αυτή τη διπλωματική εργασία.

Επιπρόσθετα στη παρούσα διπλωματική εργασία, θα παραθέσουμε παραδείγματα στα οποία δεν εφαρμόστηκαν αρκετά μέτρα για τη προστασία της ηθικής στον τομέα της ιατρικής, όπως και αντίστοιχα τι θα έπρεπε να γίνει ώστε να αποφευχθεί κάτι αντίστοιχο στο μέλλον. Όσο αφορά τον τομέα της υγείας, είναι ένα θέμα το οποίο είναι πολύ ενδιαφέρον σχετικά με την ηθική του AI, καθώς δημιουργούνται πολλά διλήμματα για το αν είναι σωστό να χρησιμοποιείται ένα αυτοματοποιημένο σύστημα για τη διάγνωση και θεραπεία των ασθενών. Ειδικότερα, στον τομέα της υγείας τα λάθη που μπορεί να προκύψουν από τη κακή χρήση της τεχνητής νοημοσύνης μπορεί να οδηγήσουν σε ατυχήματα τα οποία είναι ανεπανόρθωτα ή ακόμα και σε θάνατο κάποιων ατόμων. Αντίστοιχα μπορούν να δημιουργηθούν πολύ σοβαρά προβλήματα από τη χρήση τεχνητής νοημοσύνης και σε άλλους τομείς, όπως είναι ο τομέας της ασφάλειας ή ακόμα και ο τομέας των επιχειρήσεων.

Τέλος θα γίνει μια κριτική πάνω στις εφαρμογές της τεχνητής νοημοσύνης και των ηθικών αρχών που πρέπει να ακολουθεί. Διότι χωρίς αυτές τις ηθικές αρχές όπως θα δούμε σε κάποια παραδείγματα στη παθολογία και σε άλλες περιπτώσεις της ιατρικής, μπορεί να δημιουργηθούν

πολλά προβλήματα τα οποία κιόλας σε μερικές περιπτώσεις μπορεί να μην είναι εφικτό να διορθωθούν.

Κεφάλαιο 2 : Ανάλυση του F.A.T.

2.1 F.A.T. Fairness, Accountability, Transparency

Στην εποχή μας η τεχνητή νοημοσύνη όπως θα αναφερθεί περαιτέρω σε αυτήν την εργασία είναι ένα αναπόσπαστο κομμάτι της καθημερινότητας μας. Αυτό είναι περισσότερο εμφανές αν λάβουμε υπόψη την άνοδο που έχει τα τελευταία χρόνια η Τεχνητή Νοημοσύνη, καθώς έχει καταστήσει την ενσωμάτωση αυτοματοποιημένων συστημάτων λήψης αποφάσεων σε πολλές εφαρμογές. Βέβαια όμως, με την άνοδο της, η τεχνητή νοημοσύνη έχει καταφέρει να θέσει πολύ κόσμο σε προβληματισμό και να τους δημιουργήσει αυξανόμενες ανησυχίες όσο αφορά τις ηθικές συνέπειες αυτών των συστημάτων, καθώς όπως έχουμε διακρίνει πολλές φορές στο παρελθόν, τέτοια συστήματα έχουν βάλει αρκετό κόσμο σε κίνδυνο. Συνεπώς, για αυτόν τον λόγο έχουν πραγματοποιηθεί έρευνες σχετικά με τις ηθικές πτυχές των συστημάτων τεχνητής νοημοσύνης, οι οποίες έρευνες είναι γνωστές και ως ηθικό AI, FAT, FAT/ML, FAT* ή FAccT. Το FAT αναφέρεται σε τρείς πολύ βασικές έννοιες της ηθικής της τεχνητής νοημοσύνης: FAccT. Το ML είναι η μηχανική μάθηση (Machine Learning) και αναφέρεται στις εφαρμογές της ενώ ο "*" είναι για να δείξει ότι υπάρχουν και άλλα ηθικά διλήμματα όπως είναι όπως είναι η εξουσία και η δικαιοσύνη. Το FAccT είναι η τελευταία ονομασία που διαδέχθηκε από ένα από τα συνέδρια με τη μεγαλύτερη επιρροή στο τομέα της ηθικής του AI. Όπου οι τρείς πιο σημαντικές ηθικές αξίες της τεχνητής νοημοσύνης είναι η Δικαιοσύνη, η Ευθύνη και η Διαφάνεια. Βέβαια η έρευνα στο τομέα της ηθικής στο AI είναι "ιδιαίτερη" καθώς υπάρχουν πολλές έννοιες οι οποίες έχουν πολλαπλές ερμηνείες, ενώ ταυτόχρονα προσπαθεί να βρει τεχνητές λύσεις για τον έλεγχο και την διασφάλιση των ηθικών αυτών αξιών. Λόγου χάρη η δικαιοσύνη ως γενικότερη ερμηνεία αναφέρεται στην έλλειψη προκαταλήψεων σε ένα σύστημα AI προς κάποιες ομάδες ατόμων, αλλά ταυτόχρονα στο FAccT υπάρχουν και άλλοι ορισμοί της έννοιας δικαιοσύνης στη βιβλιογραφία τους. Κάποιοι από αυτούς τους ορισμούς έχουν μια πιο στατιστική λογική καθώς προσπαθούν να δώσουν την ίδια απόδοση στους πάντες σε μετρήσεις ταξινόμησης, δηλαδή προσπαθούν να "υπερνικήσουν" τις προκαταλήψεις έτσι ώστε να μπορούν όλοι από οποιαδήποτε πολιτισμική ομάδα να έχουν τις ίδιες πιθανότητες να επιτύχουν σε μια προκαθορισμένη αλγοριθμική έξοδο. Αντίστοιχα υπάρχουν και άλλοι ορισμοί όπως αυτός που ελέγχει αν μια προκαθορισμένη μεταβλητή μπορεί να επηρεάσει τις εξόδους ενός αλγορίθμου. Βέβαια πολλές από τις ερμηνείες της δικαιοσύνης δεν επιτυγχάνουν τα ίδια αποτελέσματα με τους προηγούμενους ορισμούς καθώς είναι πολύ ευαίσθητες και τα δεδομένα που χρησιμοποιούμε στους αλγόριθμους μπορούν να τις επηρεάσουν πολύ εύκολα και με αρνητικό αντίκτυπο. Κάτι το οποίο τις καθιστά μη συμβατές για χρήση στο τομέα της ηθικής του AI. Συνεπώς δεν είναι όλοι οι ορισμοί κατάλληλοι για χρήση στα αυτοματοποιημένα συστήματα. Οπότε όποιος έχει σκοπό να χρησιμοποιήσει αλγόριθμους AI θα

πρέπει να συμβουλευτεί εξαιρετικές κριτικές πάνω στη δικαιοσύνη, ευθύνη και διαφάνεια οι οποίες βρίσκονται στην βιβλιογραφία της FAccT. [5]

Στο σημείο που βρισκόμαστε κρίνεται απαραίτητο να αναφέρουμε γιατί η έρευνα στο τομέα του FAccT απαιτεί μια πολύ στενή συνεργασία μεταξύ της επιστήμης των δεδομένων, όπως και αντίστοιχα και των ανθρωπιστικών επιστημών. Βέβαια για να λειτουργήσει σωστά ένας αλγόριθμος, εκτός από αυτές τις 2 γνώσεις, είναι ύψιστης σημασίας ο ερευνητής να έχει γνώσης σχετικά με τον τομέα στον οποίο θα χρησιμοποιήσει τον αλγόριθμο. Για παράδειγμα, ένας AI αλγόριθμος στη παθολογία δεν μπορεί να είναι σωστά εκπαιδευμένος και να λειτουργεί άριστα αν δεν συμβουλευτεί ο ερευνητής πρώτα αρκετούς παθολόγους. Καθώς οι γνώστες του θέματος έχουν τη δυνατότητα να ορίσουν τις σωστές παραμέτρους που θα έχουν τα δεδομένα εκπαίδευσης έτσι ώστε να μην μπουν σε κίνδυνο ζωές ή επιχειρήσεις αντίστοιχα. Για αυτόν ακριβώς τον λόγο η κοινότητα του FAccT αποτελείται από ερευνητές από πολλούς κλάδους εργασίας, όπως είναι η μηχανική μάθηση, η στατιστική, η επιστήμη δεδομένων, η νομοθεσία και οι κοινωνικές επιστήμες, όπως και αντίστοιχα από πολλούς ενδιαφερόμενους φορείς του κλάδου όπως είναι η Google, η IBM, η Microsoft και άλλες πολυεθνικές οι οποίες θέλουν να διατηρηθούν οι ηθικές αξίες στον χώρο του AI. Ωστόσο, μια άποψη η οποία υπάρχει σχετικά με το FAccT είναι ότι, ακόμα και αν χρησιμοποιηθούν τέλειες τεχνικές λύσεις για τις αξίες του FAccT, οι οποίες θα μπορούσαν να ενσωματωθούν στα συστήματα AI, αυτό δεν θα έλυνε ένα μεγάλο πρόβλημα κατανόησης που υπάρχει ως προς τα κοινωνικά, πολιτικά και πολιτιστικά περιβάλλοντα στα οποία αυτά τα συστήματα αναπτύσσονται. Μια γενική αλήθεια είναι ότι τα συνέδρια FAccT έχουν κατά κύριο λόγο συμμετέχοντες από Αμερικάνικα ινστιτούτα και πολλούς λευκούς συγγραφείς κάτι το οποίο δεν θέτει ένα σωστό παράδειγμα ως προς τη δικαιοσύνη η οποία θα έπρεπε να υπάρχει. Συνεπώς είναι ένα θέμα ύψιστης σημασίας μεταξύ των ατόμων του FAccT να αποκτήσουν πιο πολύ ποικιλομορφία, και συνεπώς να γίνουν πιο αποτελεσματικοί ως προς τον τρόπο που πράττουν.[5]

2.1.1 FAIRNESS (Δικαιοσύνη στη Μηχανική Μάθηση)

Αυτή η έννοια βασίζεται στο ηθικό ιδεώδες της ανθρώπινης ισότητας. Η δικαιοσύνη είναι ένα θέμα που αμφισβητείται έντονα στις συζητήσεις οι οποίες αφορούν το δίκαιο. Κάτι το οποίο καθιστά την υιοθέτηση των ορισμών δίκαιου και δικαιοσύνης αρκετά δύσκολο με συνέπεια να υπάρχουν πολλές απόκλειόμενες θεωρίες της δικαιοσύνης όπως είναι διορθωτικές, διανεμητικές, διαδικαστικές, ουσιαστικές και άλλες θεωρίες. Τίθεται επίσης το ζήτημα της εμβέλειας ή της αποστολής εντός της οποίας συζητούνται οι έννοιες της δικαιοσύνης/δικαιοσύνης, δηλαδή η αμεροληψία στο πλαίσιο των πολιτικών κοινοτήτων (δικαιώματα του πολίτη) ή/και οικουμενικές ανθρώπινες ανησυχίες και, εάν ισχύει, πώς να οριστούν τα δημογραφικά στοιχεία , δηλαδή φύλο, εθνικότητα, φυλή, κοινωνικοοικονομικό υπόβαθρο κ.λπ. Κάποιες πολύ βασικές έννοιες που χρησιμοποιούνται όταν μιλάμε για δικαιοσύνη στη τεχνητή νοημοσύνη είναι η μεροληψία (bias), η προσβασιμότητα (accessibility) και η συμμετοχή (participation). Ο όρος **μεροληψία** σε αυτό το

πλαίσιο αναφέρεται σε προνομιακή ή μεροληπτική μεταχείριση ατόμων ή ομάδων. Οι ανησυχίες περιλαμβάνουν μεροληψία σε (ιστορικά) σύνολα δεδομένων, σκόπιμη εκμετάλλευση ατόμων όπως είναι η τοπική τιμολόγηση κάποιων συγκεκριμένων ατόμων, και την ποιότητα της παροχής υπηρεσιών. Περιλαμβάνεται επίσης μια διάκριση μεταξύ δικαιοσύνης στη θεραπεία και δικαιοσύνης επιπτώσεων. Πολλά παραδείγματα σχετικά με αυτόν τον όρο έχουν να κάνουν με τη διάκριση που υπάρχει σε ορισμένους τομείς ως προς τα άτομα από διαφορετικούς πολιτισμούς. Κάτι το οποίο στον τομέα της υγείας για παράδειγμα βάζει πολλές ζωές σε κίνδυνο, κάτι το οποίο είναι ηθικά κατακριτέο. Αν και το μεγαλύτερο μέρος της συζήτησης στον τομέα της ηθικής της τεχνητής νοημοσύνης επικεντρώνεται στην πρόληψη της βλάβης, είναι επίσης προφανές ότι τα άτομα και οι κοινότητες πρόκειται να κερδίσουν πολλά από τα συστήματα τεχνητής νοημοσύνης, εάν αυτά τεθούν στη διάθεσή τους. Επομένως, είναι ζωτικής σημασίας όσο το δυνατόν περισσότεροι άνθρωποι να έχουν ίση **πρόσβαση** σε αυτές τις τεχνολογίες. όχι μόνο θα πρέπει να είναι φθηνά, αλλά θα πρέπει επίσης να σχεδιαστούν έτσι ώστε να είναι φιλικά προς τον χρήστη, έτσι ώστε να μπορούν όλοι να κάνουν χρήση του τελικού προϊόντος, χωρίς να υπάρχουν διακρίσεις, ειδικότερα να μπορούν άτομα από διάφορες φυλετικές ομάδες να έχουν μια ευχάριστη "εμπειρία" κατά τη χρήση του εκάστοτε αλγορίθμου AI και ακόμα πιο σημαντικό είναι να μπορούν να κάνουν χρήση του άτομα τα οποία έχουν αναπηρίες και γενικά δυσκολίες. Όταν αναφερόμαστε στη **συμμετοχή**, αναφερόμαστε στην ουσιαστική συμμετοχή του κοινού στα συστήματα τεχνητής νοημοσύνης, καθώς αυτή η συμμετοχή καθίσταται δυνατή χάρη στη σαφή και κατανοητή επικοινωνία των χρηστών. Αυτό κατορθώνει να ενθαρρύνει επίσης τη μάθηση και να καθιερώσει μια πιο ολιστική προσέγγιση για τη συναίνεση. Η συμμετοχή περιλαμβάνει επίσης, την αναζήτηση απόψεων των ενδιαφερομένων κατά τη δημιουργία του συστήματος. Αυτό επεκτείνεται στην ποικιλομορφία στη στρατολόγηση και στις διεπιστημονικές ομάδες που συμμετέχουν στη διακυβέρνηση και την ανάπτυξη. **[6]**

2.1.2 Accountability (Ευθύνη)

Το Accountability είναι μια από τις 3 πιο σημαντικές ηθικές αξίες στη τεχνητή νοημοσύνη. Ως ένας από του τομείς της εφαρμοσμένης ηθικής, η ηθική τεχνητή νοημοσύνη εξετάζει τις επιπτώσεις που έχει η τεχνολογία AI στους ανθρώπους. Από μόνη της η λογοδοσία (accountability) είναι μια περίπλοκη και παραδοσιακή έννοια η οποία έχει διαφορετικές ερμηνείες ανάλογα τον τομέα στον οποίο χρησιμοποιείται, όπως είναι οι χρηματοοικονομικοί τομείς, διαχείριση πληροφορικής, κοινωνικές επιστήμες, ακόμα και επιστήμη των υπολογιστών. Μια γενικά αποδεκτή έννοια όμως από τους περισσότερους είναι ότι η λογοδοσία έχει να κάνει με το να γνωρίζουμε ποιος έλαβε τις αποφάσεις, πώς ελήφθησαν και ποιες διαδικασίες ή όργανα τέθηκαν σε εφαρμογή για την αξιολόγηση και την παρακολούθηση, δηλαδή τη διακυβέρνηση. Μια πολύ σημαντική πτυχή της λογοδοσίας είναι η ανθρώπινη επίβλεψη που χρειάζονται οι εφαρμογές AI, καθώς τα μέτρα αυτά είναι ζωτικής σημασίας, διότι σύμφωνα και με το νομικό

καθεστώς, οι άνθρωποι είναι εντελώς υπεύθυνοι για τραυματισμούς και προβλήματα που μπορεί να προκύψουν από τα συστήματα τεχνητής νοημοσύνης. [7]

Υπάρχει ένας αυξανόμενος όγκος βιβλιογραφίας για το θέμα της διατήρησης του "human-in-the-loop", το οποίο συζητείται με δύο τρόπους: πρώτον, ως μέσο διασφάλισης της ανθρώπινης συμμετοχής στις διαδικασίες λήψης αποφάσεων των αυτοματοποιημένων συστημάτων, και δεύτερον, ως μέσο παροχής ανθρώπινης εποπτείας των αυτοματοποιημένων αποφάσεων, ειδικά στο πλαίσιο των «συστημάτων υποστήριξης αποφάσεων». Στην περίπτωση του τελευταίου, μπορεί κανείς να φανταστεί ένα σύστημα «ημι-αυτοματοποιημένης απόφασης» στο οποίο ένα σύστημα υπολογιστή παράγει αποτελέσματα, οδηγίες και συστάσεις, και στη συνέχεια αυτά εξετάζονται από έναν άνθρωπο που είτε εγκρίνει είτε αποδοκιμάζει την αρχική απόφαση. Πρέπει να υπάρχει κάτι το οποίο θα εμποδίσει τον άνθρωπο από το να συμφωνεί χωρίς να ελέγχει όντως αν η απόφαση του αλγορίθμου AI ήταν σωστή, καθώς αν συμβεί κάτι τέτοιο η απόφαση θα είναι πρακτικά αυτοματοποιημένη και η συνεισφορά του ανθρώπου στη προκειμένη περίπτωση θα είναι απλά εικονική και τυπική, κάτι το οποίο θα έκανε την διαδικασία της επανεξέτασης αχρείαστη. **[6]**

Βέβαια όσο αφορά τα συστήματα AI, η λογοδοσία δεν σχετίζεται μόνο με την ευθύνη που έχουν οι ενδιαφερόμενοι σε κάθε στάδιο της δημιουργίας του αλγορίθμου AI, αλλά και με άλλες κοινωνικές και τεχνολογικές πτυχές. Κάποιες από αυτές τις πτυχές είναι ο ορισμός νέων κανονισμών και νόμων οι οποίοι θα είναι για συγκεκριμένες εφαρμογές και δεν θα εφαρμόζονται παντού, εκτιμήσεις επιπτώσεων ώστε να υπάρχει μια πρόβλεψη των προβλημάτων που θα αντιμετωπίσουν αργότερα οι ερευνητές, επαληθευσιμότητα και δυνατότητα αναπαραγωγής του συστήματος, και ένα από τα σημαντικότερα από αυτά το οποίο είναι τα διορθωτικά μέτρα για αυτοματοποιημένες αποφάσεις, κάτι το οποίο είναι ύψιστης σημασίας σε εφαρμογές οι οποίες μπορούν να βάλουν σε κίνδυνο την ανθρώπινη ζωή. Επιπρόσθετα αν λάβουμε υπόψη τον ορισμό που έχει δοθεί στο Accountability από την ομάδα εμπειρογνωμόνων Υψηλού Επιπέδου της Ευρωπαϊκής Ένωσης για την τεχνητή νοημοσύνη, καταλαβαίνει κανείς ότι η λογοδοσία έχει πολλές δυνατότητες από τις οποίες κάποιες είναι η δυνατότητα ελέγχου, ελαχιστοποίηση και αναφορά αρνητικών επιπτώσεων, συμβιβασμών και η επανόρθωση.

Η λογοδοσία μπορεί να συσχετισθεί με 6 διαφορετικές διαστάσεις σύμφωνα με τους στόχους που έχουν οι ενδιαφερόμενοι. Αυτές οι διαστάσεις απαρτίζονται από τα εξής: ευθύνη, αιτιολόγηση, έλεγχος, αναφορά, επανόρθωση και ιχνηλασιμότητα, τα οποία έχουν πολλές εξαιρετικές εργασίες οι οποίες ασχολούνται με αυτά. Επιπρόσθετα υπάρχουν δύο κύριοι τύποι λογοδοσίας τους οποίους θα αναλύσουμε περαιτέρω: την εσωτερική λογοδοσία και την εξωτερική.

Η εξωτερική λογοδοσία, όπως είναι αντιληπτό και από την ονομασία της είναι η λογοδοσία η οποία είναι επικεντρωμένη στην αξιολόγηση και επαλήθευση ενός συστήματος AI αλλά από έναν εξωτερικό φορέα του οργανισμού ο οποίος πραγματοποίησε την ανάπτυξη του συστήματος αυτού. Αυτό συνεπάγεται ότι μία εξωτερική ομάδα είτε είναι ρυθμιστικές αρχές, είτε χρήστες, ή ακόμα και ανεξάρτητοι ερευνητές, πραγματοποιεί τη διαδικασία της αξιολόγησης και επικύρωσης αποκτώντας πρόσβαση σε ορισμένα σημεία του συστήματος όπως είναι οι έξοδοι μέσω API. Υπάρχουν διάφορες τεχνικές οι οποίες μπορούν να χρησιμοποιηθούν για την “δοκιμή” της

λογοδοσίας όπως είναι η ποιοτική έρευνα, η δοκιμή μαύρου κουτιού για την απόκτηση αποτελεσμάτων κ.λπ. Κρίνεται απαραίτητο να αναφερθεί ότι στο τομέα της ηθικής του AI υπάρχει μια ανοιχτή συζήτηση - διαμάχη όσο αφορά τον τρόπο με τον οποίο θα έπρεπε να εφαρμόζεται η εξωτερική λογοδοσία, και τη πρόσβαση της σε πιο ευαίσθητα στοιχεία όπως είναι τα δεδομένα εκπαίδευσης και ο αλγόριθμος, καθώς και το πως θα μπορούν να προστατευτούν αυτά τα στοιχεία. Από την άλλη πλευρά, η εσωτερική λογοδοσία πραγματοποιείται στο πλαίσιο του ίδιου του συστήματος AI, δηλαδή μέσω προγραμματιστών, εσωτερικών ελεγκτών και άλλων ειδικών στον τομέα. Η δουλειά τους είναι να ελέγχουν εάν όλες οι διαδικασίες του αγωγού τεχνητής νοημοσύνης συμμορφώνονται με τα ισχύοντα πρότυπα, τη νομοθεσία και τις οργανωτικές αρχές και απαιτήσεις τεχνητής νοημοσύνης, συμπεριλαμβανομένων των ηθικών προσδοκιών.

Η συντριπτική πλειονότητα των εργασιών που έχουν γίνει στο θέμα της λογοδοσίας της τεχνητής νοημοσύνης διερευνούν τις διάφορες πτυχές της λογοδοσίας μεμονωμένα τη μία από την άλλη, χωρίς να διεξάγουν κανενός είδους έρευνα για τη μεταξύ τους διασύνδεση. Για να βελτιωθεί η συνολική υπευθυνότητα ενός συστήματος τεχνητής νοημοσύνης και για να ανταποκριθεί καλύτερα στις απαιτήσεις των πολλών ενδιαφερομένων του, είναι χρήσιμο να έχουμε μια σταθερή κατανόηση των αλληλεξαρτήσεων που υπάρχουν μεταξύ των πολλών στοιχείων του, όπως είναι η ευθύνη, η δικαιολόγηση, η εμπιστοσύνη, η ασφάλεια και η αποκατάσταση . Η ανάπτυξη της εν συναίσθησης είναι ένα από τα πιο σημαντικά εργαλεία για την επίτευξη αυτού του στόχου. **[7]**

2.1.3 Transparency (Διαφάνεια της Μηχανικής μάθησης)

Ως διαφάνεια στη Μηχανική Μάθηση ορίζουμε τις συνθήκες και τις διαδικασίες οι οποίες δίνουν τη δυνατότητα στους χρήστες να γνωρίζουν το τρόπο διαχείρισης των δεδομένων τους. Η διαφάνεια, ως βασική έννοια αυτού του ζητήματος, είναι θεμελιώδης για την οικοδόμηση αξιοπιστίας και την ανάληψη ευθύνης. Τόσο οι αποφάσεις που λαμβάνονται όσο και το σκεπτικό πίσω από αυτές τις αποφάσεις μπορούν να θεωρηθούν πτυχές διαφάνειας όταν πρόκειται για τη χρήση συστημάτων τεχνητής νοημοσύνης. Πιο συγκεκριμένα όταν αναφερόμαστε στη διαφάνεια αναφερόμαστε στην ικανότητα ενός συστήματος να μπορεί να εξηγηθεί σε διαφορετικά άτομα, ανάλογα με το τεχνικό επίπεδο της τεχνικής τους εξειδίκευσης, τη συμμετοχή τους στη δημιουργία και τη συντήρηση του συστήματος. [6]

Ένας σύνηθες τρόπος με τον οποίο οι ιστοσελίδες χρησιμοποιούν τα δεδομένα μας είναι για να μας προβάλλουν αντικείμενα που θα μας ‘’ενδιαφέρουν’’ . Αυτό γίνεται με την βοήθεια των cookies. Τα Cookies είναι δεδομένα από έναν ιστότοπο που είναι αποθηκευμένα σε ένα πρόγραμμα περιήγησης από το οποίο ο ιστότοπος μπορεί να τα ανακτήσει αργότερα. Έτσι η ιστοσελίδα μπορεί να προβάλει αντικείμενα που άρεσαν σε όμοια άτομα με τον χρήστη. Πολλές ιστοσελίδες και εταιρίες χρησιμοποιούν αυτή τη τεχνική όπως είναι η Amazon , CDNow , Ebay

και άλλες πολλές. Συνεπώς αυτές οι εταιρίες έχουν έναν αλγόριθμο ο οποίος προτείνει αντικείμενα τα οποία θα φανούν ενδιαφέρον στο χρήστη και συνεπώς έχουν μεγαλύτερη πέραση προς αυτόν. **[2]**

Η συντριπτική πλειονότητα των συστημάτων συστάσεων στο Διαδίκτυο λειτουργεί με μυστηριώδη τρόπο, χωρίς να παρέχουν στον χρήστη κάποια εικόνα για τη λογική του συστήματος ή εξήγηση για τις συστάσεις που έγιναν. Ο τυπικός τρόπος αλληλεπίδρασης συνεπάγεται το να ζητάτε από τον χρήστη κάποια στοιχεία (όπως αξιολογήσεις πραγμάτων ή μια λίστα με τους μουσικούς ή τους συγγραφείς που προτιμούν), την επεξεργασία των δεδομένων και, στη συνέχεια, την παροχή στον χρήστη με κάποια έξοδο με τη μορφή προτάσεων. Αυτό πραγματοποιείται με αλγορίθμους γνωστού και ως αλγόριθμοι CF οι οποίοι συλλέγουν δεδομένα από τους χρήστες και προβάλουν πράγματα που θα "αρέσουν" στον χρήστη.

Στην πραγματικότητα, οι αλγόριθμοι CF (Collaborating Filtering) είναι επίσης γνωστοί ως αλγόριθμοι κοινωνικού φιλτραρίσματος. Αυτό οφείλεται στο γεγονός ότι έχουν διαμορφωθεί σύμφωνα με τη δοκιμασμένη και αληθινή κοινωνική διαδικασία λήψης συστάσεων, ζητώντας από φίλους με παρόμοια ενδιαφέροντα να προτείνουν ταινίες, βιβλία ή μουσική που τους αρέσει. Με αυτόν τον τρόπο, οι αλγόριθμοι CF είναι σε θέση να αναπαράγουν τη δοκιμασμένη στο χρόνο κοινωνική διαδικασία λήψης συστάσεων. Ο παραλήπτης μιας σύστασης έχει στη διάθεσή του έναν αριθμό επιλογών όταν αποφασίζει εάν θα εμπιστευτεί ή όχι τη σύσταση. Αυτές οι επιλογές περιλαμβάνουν:

(α) ανάλυση του βαθμού στον οποίο τα γούστα του παραλήπτη και του συστάτη είναι παρόμοια·

(β) ανάλυση του βαθμού στον οποίο οι προηγούμενες προτάσεις του εισηγητή ήταν επιτυχείς· και

(γ) να ζητήσει από τον συντάκτη περισσότερες πληροφορίες σχετικά με τους λόγους για τους οποίους έγινε η σύσταση.

Με παρόμοιο τρόπο, τα συστήματα συστάσεων πρέπει να παρέχουν στους χρήστες κάποιο έλεγχο σχετικά με τον τρόπο με τον οποίο αξιολογούν την καταλληλόλητα των προτάσεων που λαμβάνουν.

Προηγούμενες μελέτες έχουν δείξει ότι τα έμπειρα συστήματα που λειτουργούν ως οδηγοί αποφάσεων απαιτείται να προσφέρουν αιτιολογήσεις και εξηγήσεις για τις συστάσεις που προσφέρουν. Η έρευνα που διεξάγεται με τις μηχανές αναζήτησης υπογραμμίζει επίσης την ανάγκη διαφάνειας. Σύμφωνα με τους Johnson και Johnson (1993), οι εξηγήσεις είναι ένα εξαιρετικά σημαντικό συστατικό της δυναμικής σχέσης που υπάρχει μεταξύ των χρηστών και των πολύπλοκων συστημάτων. Σύμφωνα με τα ευρήματα της έρευνάς τους, μια λειτουργία της εξήγησης είναι να παρέχει μια απεικόνιση της σύνδεσης μεταξύ του προηγουμένου και του επακόλουθου(δηλαδή, μεταξύ αιτίας και αποτελέσματος).

Όταν πρόκειται για συστήματα συστάσεων, η κατανόηση της σχέσης μεταξύ της εισόδου στο σύστημα δηλαδή τις αξιολογήσεις που παρέχονται από τον χρήστη και της εξόδου όπως είναι οι συστάσεις, επιτρέπει στον χρήστη να συμμετάσχει σε μια δέσμευση που είναι προβλέψιμη και

αποτελεσματική με το Σύστημα.

Αντί να τραβούν «στιγμιότυπα στο σκοτάδι», δίνεται στους χρήστες η δυνατότητα, όταν υπάρχει διαφάνεια, να αλλάζουν ουσιαστικά τα δεδομένα για να βελτιώσουν τις συστάσεις. Ο Herlocker και οι συνεργάτες του πιστεύουν ότι η έλλειψη διαφάνειας είναι ο λόγος για τον οποίο τα συστήματα συστάσεων δεν έχουν εφαρμοστεί στη λήψη αποφάσεων με υψηλά επίπεδα κινδύνου. Ενώ οι πελάτες θα μπορούσαν να είναι έτοιμοι να ρισκάρουν σε μια ασαφή πρόταση ταινίας, μπορεί να μην είναι διατεθειμένοι να δεσμευτούν σε έναν ιστότοπο διακοπών χωρίς πρώτα να κατανοήσουν τους λόγους πίσω από μια τέτοια σύσταση. Συνεπώς είναι απαραίτητο να υπάρχει διαφάνεια ώστε να γνωρίζουν οι χρήστες γιατί βλέπουν τις προτάσεις του συστήματος AI, έτσι θα είναι πιο κατανοητό από τους πάντες και θα μπορέσουν μακροπρόθεσμα να χρησιμοποιήσουν τέτοιες προτάσεις και σε πιο "σοβαρές" εφαρμογές. **[2]**

2.1.4 Safety (**ασφάλεια**)

Αυτή η έννοια βασίζεται στην ηθική αρχή της μη κακίας, η οποία δηλώνει ότι καμία ενέργεια δεν πρέπει να προκαλεί πόνο ή ταλαιπωρία σε άλλο ον, όπου η ταλαιπωρία θεωρείται ότι περιλαμβάνει αρνητικές επιπτώσεις στη σωματική, ψυχική και κοινωνική υγεία ενός ατόμου και στο περιβάλλον του. **[6]**

Εφόσον δεν είναι εφικτό με κάποιο τρόπο να προστατεύσουμε τους χρήστες αφού γίνει κάτι "κακό", προσπαθούμε να έχουμε μια προληπτική συμπεριφορά ή γενικά να γίνει εντοπισμός των ρίσκων που μπορεί να προκύψουν. Συνεπώς υπάρχουν κάποιες βασικές αρχές που λαμβάνονται υπόψη κατά τη διαδικασία αυτή. Η δύναμη ενός συστήματος απέναντι σε εχθρικές ή κακόβουλες ενέργειες είναι κάτι το οποίο διαφοροποιεί τα καλά με τα "κακοφτιαγμένα "συστήματα, και το πόσο ανθεκτικό είναι ένα σύστημα στο hacking και σε άλλες μορφές κυβερνοεπίθεσης, φαινόμενο το οποίο είναι σύνηθες στις μέρες μας. Επιπρόσθετα το να υπάρχουν διασφαλίσεις για την αποφυγή εκμετάλλευσης του συστήματος είναι ύψιστης σημασίας, ώστε να επιτευχθεί η ευρωστία του συστήματος και να μην υπάρχουν κακά περιστατικά όπως είναι η διαρροή του μοντέλου. Ένας άλλος τρόπος με τον οποίο πολλά συστήματα χρησιμοποιούνται λάθος, είναι όταν χρησιμοποιούνται για ένας καλό σκοπό, και ταυτόχρονα και για έναν κακόβουλο. Αυτό συμβαίνει πολλές φορές καθώς το αρχικό σύστημα μετατρέπεται από τρίτους σε κάτι το οποίο έχει σκοπό να βλάψει κόσμο. Κάτι το οποίο συμβαίνει πολλές φορές σε εφαρμογές που έχουν να κάνουν με τον στρατό, όπως είναι η χρήση οπλισμένων drones και άλλα συστήματα τα οποία καταλήγουν να χρησιμοποιούνται ενάντια σε ανθρώπους. Άλλες 2 πολύ βασικές αρχές της ασφάλειας είναι η αξιοπιστία και η αναπάραγωγιμότητα, τα οποία παίζουν καθοριστικό ρόλο όσο αφορά την ευρωστία του συστήματος. Όταν αναφερόμαστε στην αξιοπιστία, αναφερόμαστε στο βαθμό στον οποίο το σύστημα πληροί τις απαραίτητες προϋποθέσεις και απαιτήσεις για τις οποίες σχεδιάστηκε και εφαρμόστηκε. Αντίστοιχα όταν θέλουμε να αναφερθούμε στην αναπάραγωγιμότητα, αναφερόμαστε στο πόσο καλά αποδίδει ένα σύστημα όταν υπόκειται στις ίδιες εισροές και

συνθήκες. Αυτές οι δύο αρχές είναι πολύ σημαντικές σχετικά με την ευρωστία καθώς, αν δεν υπάρχει αξιοπιστία και δεν είναι εφικτή η αναπαραγωγή αποτελεσμάτων, αυτό υπονομεύει την εμπιστοσύνη που υπάρχει στο σύστημα. Κάτι το οποίο είναι πολύ σημαντικό ειδικά σε εφαρμογές μεγάλης κλίμακας, και ειδικότερα σε εφαρμογές που σχετίζονται με τον τομέα της υγείας. Λαμβάνοντας υπόψη όλα τα προαναφερθέντα καταλαβαίνει κανείς ότι η ευρωστία είναι ένα αναπόσπαστο κομμάτι του συστήματος, το οποίο πρέπει να διασφαλίζετε σε κάθε περίπτωση. Συνεπώς οι ερευνητές οφείλουν να προσέχουν τα ρίσκα που προαναφέραμε αλλά και να προσπαθούν να προνοήσουν για αυτά που δεν έχουν αντιμετωπίσει ακόμα. Οπότε είναι προτεραιότητα για τους προγραμματιστές και τα άτομα τα οποία ασχολούνται περαιτέρω με τον αλγόριθμο, να έχουν κάποια συστήματα ελέγχου και παρακολούθησης ώστε να μπορέσουν να αποφευχθούν τυχόν "ατυχήματα". Βέβαια χωρίς να θέλουμε να υποτιμήσουμε ότι είπαμε προηγουμένως, είναι αδύνατη η ολική αποφυγή των ρίσκων καθώς δεν υπάρχει κάποιος μηχανισμός ή πρόγραμμα που να μπορεί να υπολογίσει όλα τα ρίσκα και να βρει τρόπους να αποφευχθούν, παρόλα αυτά είναι πολύ σημαντικό να υπάρχει ένα σύστημα παρακολούθησης, καθώς αυτό είναι προϋπόθεση για την διασφάλιση της αποτελεσματικής λειτουργίας του συστήματος. **[6]**

2.1.5 Privacy (Απόρρητο)

Στη σήμερον εποχή το ζήτημα του απορρήτου είναι κάτι το οποίο περιλαμβάνεται σε πολλές ηθικές συζητήσεις, είτε είναι σχετικά με τα μέσα κοινωνικής δικτύωσης, είτε ακόμα και σχετικά με τα δεδομένα που συλλέγονται από πολλές ιστοσελίδες. Συνεπώς είναι λογικό να είναι και ένα ηθικό θέμα και στον τομέα της ηθικής του AI καθώς σχετίζεται άμεσα και με τα 2 παραδείγματα που αναφέραμε μόλις. Όταν μιλάμε για απόρρητο γενικά, αναφερόμαστε στη δημόσια και πολιτική ανάγκη για σεβασμό των προσωπικών δεδομένων των ανθρώπων, κάτι το οποίο θα έπρεπε να είναι μεγαλύτερη προτεραιότητα όλων. Βέβαια δεν υπάρχει ισότητα στο απόρρητο σε όλους τους τομείς καθώς το απόρρητο στη προσωπική μας ζωή έχει παρεμβληθεί και ο κόσμος έχει αφοσιωθεί πιο πολύ στα πολιτικό-οικονομικά ζητήματα του απορρήτου. Αυτό δημιουργεί πολλά προβλήματα καθώς πολλά δεδομένα από τη καθημερινότητα μας αντλούνται από εταιρίες, έθνη και πολιτικούς χωρίς να το γνωρίζουμε. Για αυτό το λόγο είναι απαραίτητο οι άνθρωποι να κατανοούν σε τι "συμφωνούν" όταν μπαίνουν σε μια σελίδα, καθώς αυτή η σελίδα είναι πολύ πιθανό να έχει πρόσβαση σε πολλά δεδομένα τους. Το μεγαλύτερο πρόβλημα σε αυτό το ενδεχόμενο είναι ότι η σελίδα μπορεί να κάνει κακόβουλη χρήση των δεδομένων αυτών ή ακόμα και να τα πουλήσει για να βγάλει κέρδη. Επιπρόσθετα ένα ζήτημα στο πολιτικό τομέα είναι η παρακολούθηση που υπάρχει από το κράτος είτε για πολιτικούς , είτε και για οικονομικούς λόγους, χωρίς να το γνωρίζουμε, κάτι το οποίο δεν είναι ηθικά σωστό και γίνεται δυστυχώς παντού. Συνεπώς για να διατηρείται το απόρρητο στις εφαρμογές AI πρέπει να ακολουθούνται

κάποιες αρχές. Μία από αυτές είναι η διαχείριση δεδομένων η οποία είναι από τις πιο σημαντικές διαδικασίες. Όλες οι διαδικασίες που πραγματοποιούνται κατά τη συλλογή δεδομένων, όπως είναι η αρχική συλλογή, ο καθαρισμός, η ανάλυση, η επαναχρησιμοποίηση και η ανακύκλωση είναι κατεργασίες οι οποίες πρέπει να πραγματοποιούνται χωρίς να διακυβεύεται η ασφάλεια ή το απόρρητο των δεδομένων. Βέβαια δεν είναι πάντα διασφαλισμένη η ασφάλεια των δεδομένων, οπότε είναι μια πολύ σημαντική υποχρέωση που θα έπρεπε να τηρείται από όλους. Καθώς με την ασφαλή διαχείριση των δεδομένων και την διασφάλιση της προστασίας μπορεί και υπάρχει η διατήρηση της εμπιστευτικότητας στο σύστημα. Στο πλαίσιο της προστασίας της ιδιωτικής ζωής και των δεδομένων, η πρακτική της συλλογής και αποθήκευσης μόνο της ελάχιστης ποσότητας πληροφοριών που είναι απαραίτητη είναι γνωστή ως «ελαχιστοποίηση δεδομένων». Καθορίζεται ότι υπάρχουν τρεις παράγοντες που καθορίζουν την ελαχιστοποίηση δεδομένων: Αρχικά είναι η επάρκεια, όπου τα δεδομένα είναι επαρκή για να εξυπηρετήσουν έναν συγκεκριμένο σκοπό. Έπειτα είναι η συνάφεια, η οποία εφαρμόζεται όταν τα δεδομένα έχουν αιτιολογημένη σύνδεση με τον δηλωμένο σκοπό· και αναγκαιότητα, όπου τα δεδομένα είναι περιορισμένα και φυλάσσονται μόνο όταν χρειάζεται και στη συγκεκριμένη περίπτωση διαγράφονται όσα δεδομένα δεν χρησιμοποιούνται. Με αυτούς τους τρόπους διασφαλίζεται η ασφάλεια των προσωπικών δεδομένων με αποτέλεσμα οι άνθρωποι να έχουν περισσότερη εμπιστοσύνη στα συστήματα AI κάτι το οποίο με τη σειρά του θα οδηγήσει σε περισσότερες εφαρμογές που θα έχουν ως βάση αλγορίθμους AI. **[6]**

Κεφάλαιο 3 : Ηθικά διλήμματα Τεχνητής Νοημοσύνης

3.1 Μαύρο Κουτί και Ηθικοί Φραγμοί

Το μαύρο κουτί και η γενική καταγραφή πληροφοριών κατά την διάρκεια της πτήσης. Ξεκίνησαν να εφαρμόζονται από το 1958, αρχικά για να καταγράφονται βασικοί παράμετροι , όπως η ταχύτητα και η κατεύθυνση. Τα πρώτα μαύρα κουτιά ήταν σε θέση να καταγράφουν μέχρι πέντε παραμέτρους, σε αντίθεση με τα πιο εξελιγμένα μοντέρνα μαύρα κουτιά, τα οποία έχουν την ικανότητα να καταγράφουν παραπάνω από 1000 παραμέτρους. Σε αυτές τις παραμέτρους συγκαταλέγεται και η καταγραφή της φωνής στον χώρο του πιλοτηρίου, το οποίο δίνει την δυνατότητα στους ερευνητές να μπορούν να ελέγξουν ακόμη καλύτερα σε περίπτωση ατυχήματος τις αιτίες του από την επικοινωνία των πιλότων σε συνδυασμό με τις υπόλοιπες παραμέτρους. Αν και η καταγραφή αυτή αρχικά έφερε πολλούς ενδοιασμούς από το ευρύ κοινό, έχει αποδειχθεί ότι είναι πολύτιμη κατά την προσομοίωση του ατυχήματος, διότι παρέχει την δυνατότητα στους ερευνητές να κάνουν μια πολύ ακριβή αναπαράσταση του δυστυχήματος, οδηγώντας τους έτσι στην εύρεση του αιτίου. **[55]**

Μπορούμε να κάνουμε συγκρίσεις μεταξύ των τρεχουσών αιτιών αεροπορικών καταστροφών και των πιθανών μελλοντικών ατυχιών που θα προκληθούν από την αυξανόμενη εξάρτησή μας από τα ρομπότ. Συνήθως οι απροσδόκητες αλληλεπιδράσεις μεταξύ των συμπεριφορών και των ενεργειών του πληρώματος του αεροσκάφους, της ανταπόκρισης και της συμπεριφοράς των συστημάτων του αεροσκάφους και των συνθηκών περιβάλλοντος της πτήσης οδηγούν σε αεροπορικές ατυχίες. Με παρόμοιο τρόπο, προβλέπουμε ότι ένας παρόμοιος συνδυασμός περιστάσεων θα επηρεάσει τα ρομπότ στο μέλλον. Τα ρομπότ αναπόφευκτα θα προοδεύσουν στην πολυπλοκότητα, θα λειτουργούν όλο και πιο αυτόνομα και θα το κάνουν σε περιβάλλοντα ελεύθερης ροής όπως τα θέλουν οι άνθρωποι. Αυτές οι βελτιωμένες ελευθερίες συνοδεύονται από περισσότερες πιθανότητες οι απροσδόκητοι συνδυασμοί απρόβλεπτων στοιχείων να οδηγήσουν σε επικίνδυνες καταστάσεις ή ακόμα και σε σωματική βλάβη. Αυτό δεν σημαίνει ότι τέτοια περιστατικά θα συμβαίνουν τακτικά. Όπως και με τις αεροπορικά δυστυχήματα, ο σοβαρός τραυματισμός μπορεί να είναι σπάνιος. Ωστόσο, πρέπει να αποδεχτούμε ότι οι επικίνδυνες καταστάσεις είναι αναπόφευκτες. Έτσι, η επιθυμία μας τα ρομπότ να συνεισφέρουν στις ανθρώπινες δραστηριότητες με νέους τρόπους έχει ως αποτέλεσμα την παροχή νέων δυνάμεων και ελευθεριών, γεγονός που συνδράμει στην εμπλοκή τους σε σημαντικό κίνδυνο και συνεπώς σε τραυματισμό. Αυτό μπορεί να αποτελέσει κίνητρο για μια συνολική εξέταση των πλεονεκτημάτων έναντι των κινδύνων της αύξησης της εξάρτησής μας από τα ρομπότ, αλλά πιστεύουμε ότι απαιτούνται περισσότερα κριτήρια για την εισαγωγή των ρομπότ στον χώρο της αεροπορίας. Υποστηρίζουμε ότι η αποδοχή των αεροπορικών ταξιδιών, παρά τις καταστροφές

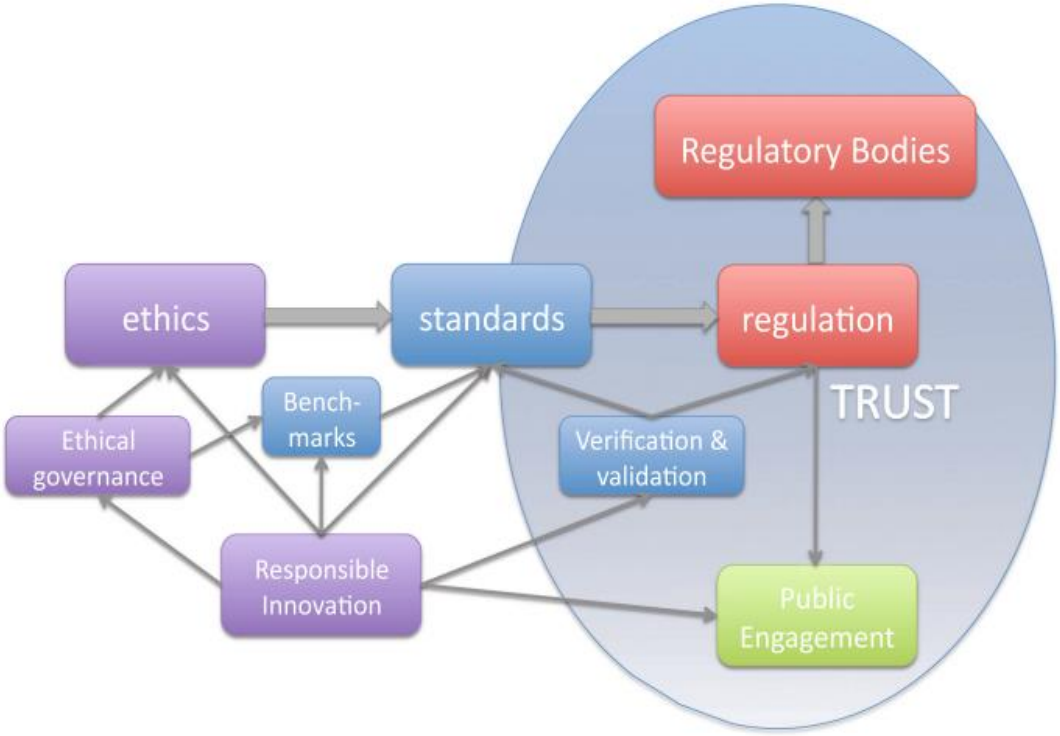
τους, συνδέεται εν μέρει με τη διακυβέρνηση της αεροπορίας, η οποία έχει πολιτιστικές, συμβολικές καθώς και πρακτικές επιπτώσεις. Το να γίνει κατανοητός ο τρόμος της καταστροφής μέσω της διαδικασίας έρευνας και ανοικοδόμησης είναι ένα κρίσιμο μέρος στην εύρεση αποτελεσματικών λύσεων για την αποφυγή τους. **[55]**

Λόγω αυτής της αναγκαιότητας, όπως επίσης και ευχρηστίας, του μαύρου κουτιού έχουν γίνει προσπάθειες ένταξης της τεχνολογίας αυτής και σε άλλους τομείς που συμπεριλαμβάνουν προγραμματισμό και μικροελεγκτές. Ωστόσο ο τομέας που έχει επωφεληθεί σε μεγάλο βαθμό είναι αυτός της αυτοκινητοβιομηχανίας, όπου χρησιμοποιείται κυρίως για καταγραφή δεδομένων αλλά γίνεται και προσπάθεια για την εκμετάλλευση της τεχνολογίας αυτής για την καταγραφή οδηγιών συμπεριφορών. Υπάρχει μια σχέση μεταξύ ηθικής, κανόνων και προτύπων. **[55]**

Οι ηθικές αρχές επισημοποιούνται σε ένα πλαίσιο από πρότυπα, τα οποία μπορούν να χρησιμοποιηθούν για τη δοκιμή συμμόρφωσης ή, πιο πρακτικά για τα ηθικά πρότυπα, για να κατευθύνουν τους σχεδιαστές μέσω μιας αξιολόγησης ηθικού κινδύνου για ένα συγκεκριμένο ρομπότ και να τους βοηθήσουν να εξαλείψουν τυχόν κινδύνους που εντοπίστηκαν. Ωστόσο, υπάρχουν περιπτώσεις όπου τα πρότυπα πρέπει να επιβληθούν μέσω νομοθεσίας που απαιτεί τα συστήματα να πιστοποιούνται ως συμβατά με πρότυπα ή τμήματα προτύπων. Ως αποτέλεσμα, οι κανόνες και οι κανονισμοί συνδέονται άρρηκτα με την ηθική ή τις ηθικές έννοιες. **[55]**

Δεν υπάρχει αμφιβολία ότι τεχνολογίες με μεγάλη αναστάτωση όπως τα drones, τα αυτοκίνητα χωρίς οδηγό και η υποστηρικτική ρομποτική απαιτούν αν όχι νέο νόμο, τότε κανονιστικούς και ρυθμιστικούς φορείς τουλάχιστον, παρά το γεγονός ότι μεγάλο μέρος της υπάρχουσας νομοθεσίας σχετίζεται με ρομπότ και αυτόνομα συστήματα. Καθώς σε πολλές πολιτείες της Αμερικής για παράδειγμα είναι νόμιμο να οδηγεί κανείς με αυτόνομο όχημα, κάτι το οποίο βρίσκεται σε πειραματικό στάδιο. Συνεπώς δεν είναι ηθικά σωστό να εμπιστευόμαστε τέτοιες τεχνολογίες εφόσον δεν είναι έτοιμες 100% για χρήση. **[55]**

Τόσο η ηθική όσο και τα πρότυπα αποτελούν μέρος ενός ευρύτερου πλαισίου Υπεύθυνης Έρευνας και Καινοτομίας (RRI). Η υπεύθυνη καινοτομία απαιτεί συχνά ηθική έρευνα, επομένως η ηθική διακυβέρνηση ενσωματώνει την RRI με την ηθική. Το RRI σχετίζεται επίσης στενά με την ηθική μέσω ιδανικών όπως η δημόσια συμμετοχή, η ανοιχτή επιστήμη και η συμπερίληψη. Ένα άλλο βασικό στοιχείο του RRI είναι η ικανότητα αξιολόγησης και σύγκρισης των δυνατοτήτων του συστήματος με συστηματικό και διαφανή τρόπο, χρησιμοποιώντας γενικά τυποποιημένες δοκιμές ή σημεία αναφοράς. **[55]**



Source: The Case for an Ethical Black Box

Γενικά, η τεχνολογία είναι αξιόπιστη εάν παρέχει πλεονεκτήματα, ενώ παράλληλα είναι ασφαλής, καλά ρυθμισμένη και υπόκειται σε ενδελεχή έρευνα όταν συμβαίνουν ατυχίες. Ένας λόγος για τον οποίο εμπιστευόμαστε τις αεροπορικές εταιρείες είναι επειδή γνωρίζουμε ότι αποτελούν μέρος μιας επιχείρησης με υψηλά επίπεδα ελέγχου, όπως και με μεγάλο ιστορικό ασφάλειας. Τα εμπορικά αεροπλάνα είναι τόσο ασφαλή, όχι μόνο λόγο του ανώτερου σχεδιασμού, αλλά και των αυστηρών διαδικασιών πιστοποίησης ασφάλειας με την χρήση ισχυρών και δημοσίως προσβάσιμων πρωτοκόλλων για την διερεύνηση αεροπορικών ατυχημάτων. **[55]**

Για να προσφέρουν διαφάνεια και εμπιστοσύνη στην στιβαρότητα των ρυθμιστικών διαδικασιών, οι ρυθμιστικοί οργανισμοί πρέπει να εναρμονιστούν με τη συμμετοχή και την άποψη του κοινού. Ωστόσο, η εμπιστοσύνη δεν απορρέει απαραίτητα από την (προτεινόμενη) ρύθμιση. Μια δημοσκόπηση για τη λήψη αποφάσεων για αυτόνομα οχήματα δείχνει ανάμεικτες απόψεις τόσο για τις προτιμήσεις όσο και για τους κανονισμούς στα αυτοκίνητα χωρίς οδηγό. Οι συμμετέχοντες υποστήριξαν ότι αυτόνομα οχήματα, τα οποία θα «θυσίαζαν» τους επιβάτες τους για το κοινό καλό, θα ήθελαν να δουν άλλους να τα αγοράζουν, αλλά θα προτιμούσαν να ταξιδεύουν με αυτόματα οχήματα που προστατεύουν τους επιβάτες τους με κάθε κόστος. Οι συμμετέχοντες στην έρευνα αντιτίθενται στη θέσπιση χρηστικών περιορισμών για τα οχήματα αυτά και θα ήταν λιγότερο πιθανό να αγοράσουν ένα. **[55]**

3.2 Ηθικά διλήμματα της Τεχνητής Νοημοσύνης

Με την πρόοδο της τεχνολογίας και πιο συγκεκριμένα της τεχνητής νοημοσύνης, έχουμε διακρίνει ότι το AI έχει αλλάξει δραματικά τη ζωή των ανθρώπων σε πάρα πολλούς τομείς. Έχει διευκολύνει πολλές διαδικασίες της ζωής μας, αλλά έχει δημιουργήσει πολλά ηθικά ζητήματα, καθώς έχει αντικαταστήσει πολλούς ανθρώπους, κάτι το οποίο είναι ανησυχητικό σε πολλές περιπτώσεις. Συνεπώς αυτές οι ανησυχίες μπορούν να μεταφραστούν και ως διλήμματα για την χρήση της τεχνητής νοημοσύνη σε διάφορους τομείς. Αυτά τα διλήμματα που υπάρχουν δυστυχώς δικαιώνονται όταν πραγματοποιούνται ατυχήματα στα οποία πολλές φορές δεν γνωρίζουμε ποιος έχει την ευθύνη. Όπως για παράδειγμα αν ένα αυτόνομο αυτοκίνητο βρεθεί σε ένα ατύχημα, ποιος θα ευθύνεται για το ατύχημα αυτό. Ή στο ενδεχόμενο ενός χειρουργικού ατυχήματος το οποίο πραγματοποιήθηκε από χειρουργικό ρομπότ, ποιος είναι υπαίτιος. Αυτά τα ερωτήματα προβληματίζουν πολύ κόσμο με αποτέλεσμα να μην μπορούν να εμπιστευτούν εύκολα την τεχνολογία της τεχνητής νοημοσύνης, κάτι το οποίο από πολλές οπτικές γωνίες δεν φαίνεται παράλογο. **[10]**

Η έρευνα στη τεχνητή νοημοσύνη είναι ένα αντικείμενο το οποίο έχει λάβει σημαντική προσοχή από χώρες σε όλο τον κόσμο, διότι είναι πιθανότατα το επόμενο μεγάλο τεχνολογικό βήμα που χρειάζεται να γίνει για να δούμε σημαντική ανάπτυξη σε όλους τους τομείς. Βέβαια παρόλη τη προσοχή που έχει λάβει το AI, δεν παύει να βρίσκεται σε αρχικό στάδιο καθώς και η μελέτη του και η χρήση του δεν έχουν ερευνηθεί ακόμα όσο χρειάζεται. Βέβαια όπως υπάρχει η έρευνα για την ρήση της τεχνητής νοημοσύνης και πως μπορεί να εισαχθεί σε πολλές εφαρμογές, έτσι αντίστοιχα υπάρχουν και πολλές έρευνες από διάφορες χώρες που εστιάζονται στα ηθικά ζητήματα της τεχνητής νοημοσύνης και κατά πόσο είναι σωστή η χρήση της σε διάφορες εφαρμογές. Παρόλες τις διαφωνίες που μπορεί να υπάρχουν και από τις δύο αυτές πλευρές, δεν μπορεί να αρνηθεί κανείς ότι η εποχή της τεχνητής νοημοσύνης πλησιάζει και είναι πιο κοντά από ποτέ. Συνεπώς αφού δεν είναι εφικτό να σταματήσει αυτή η μετάβαση και δεν θα έπρεπε να γίνει κάτι τέτοιο, είναι ύψιστης σημασίας να μπορέσουν οι άνθρωποι να ελέγξουν την τεχνολογία αυτή για να μπορέσουμε να διαχειριστούμε καλύτερα τις μυριάδες ικανότητες που μας προσφέρει. Συνεπώς είναι απαραίτητο ταυτόχρονα να υπάρξουν κάποιοι ηθικοί φραγμοί τους οποίους δεν θα έπρεπε να καταπατήσουμε σαν κοινωνία. Διότι η τεχνητή νοημοσύνη έχει ως στόχο να μιμηθεί κατά το έπακρον την ανθρώπινη συνείδηση και συμπεριφορά, κάτι το οποίο από μόνο του μπορεί να φανεί τρομακτικό για πολύ κόσμο, αλλά η κάθε μεγάλη αλλαγή στην αρχή δεν είναι αποδεκτή. Οπότε ένα από τα ερωτήματα που πρέπει να τεθούν από τον κόσμο είναι το τι θα γίνει όταν τα συστήματα αυτά εξελιχθούν στο σημείο που θα έχουν συναισθήματα και θα θεωρούνται εντελώς αυτόνομα με την ικανότητα να κάνουν ανθρώπινες εργασίες. Αν λοιπόν υπάρξει το ενδεχόμενο

να δημιουργηθούν τέτοιες μηχανές θα δημιουργηθούν κάποια από τα μεγαλύτερα ηθικά διλήμματα που έχουν δημιουργηθεί ποτέ στην ανθρωπότητα. Οπότε είναι απαραίτητο δεδομένου ότι η πρόοδος της τεχνητής νοημοσύνης και της τεχνολογίας θα έχουν εκτεταμένα αποτελέσματα, να υπάρξουν τεχνικές ηθικού ελέγχου στα συστήματα αυτά κατά την εκπαίδευση τους με στόχο να εξελίσσονται συνέχεια αυτές οι τεχνικές για να διασφαλιστεί η ασφάλεια του κόσμου. Βέβαια νομοθεσίες οι οποίες θα υποχρεώνουν την δημιουργία τέτοιων τεχνικών δεν υπάρχουν προς το παρόν και δεν θα υπάρξουν σύντομα. Οπότε είναι δύσκολο να αντιμετωπιστούν πολλά προβλήματα από νομικής πλευράς καθώς δεν υπάρχουν τρόποι με τους οποίους να αποδίδεται η δικαιοσύνη αποτελεσματικά μέχρι στιγμής. Συνεπώς αφού δεν καλύπτει η νομοθεσία τις ανάγκες για ηθική προστασία, είναι απαραίτητο ο πληθυσμός να αποκτήσει μια κριτική άποψη η οποία θα μπορεί να αποφασίσει αν κάτι είναι ηθικά σωστό. Επιπρόσθετα δεν πρέπει ο κόσμος να απολαμβάνει τα θετικά τα οποία θα μας παρέχει η τεχνητή νοημοσύνη, ενώ ταυτόχρονα θα αγνοούν τις αρνητικές επιδράσεις που έχει αυτό στη ζωή τους.

Σε μεγαλύτερο ή μικρότερο βαθμό, τα πεδία της φυσικής, της ψυχολογίας, της γλωσσολογίας, της επιστήμης των υπολογιστών και της λογικής της φιλοσοφίας έχουν όλα κάτι να συμβάλουν στην μελέτη της τεχνητής νοημοσύνης. Η τεχνητή νοημοσύνη όπως την ξέρουμε μέχρι στιγμής, λειτουργεί αυτόνομα, καθώς είναι ένας πρωτοποριακός κλάδος που απαιτεί από τον κόσμο να της εξερευνήσει με διάφορους τρόπους. Εφόσον τηρούνται πάντα οι ηθικοί φραγμοί οι οποίοι υπάρχουν στη κοινωνία, η τεχνητή νοημοσύνη έχει άπειρες δυνατότητες τις οποίες περιμένει η ανθρωπότητα να ξετυλιχτούν μπροστά στα μάτια της. Συνεπώς κρίνεται απαραίτητο στη παρούσα εργασία να αναφερθούν πολλά ηθικά διλήμματα τα οποία υπάρχουν σε διάφορους τομείς. **[10]**

3.3 Παραδείγματα Χρήσεως της Τεχνητής Νοημοσύνης

Η τεχνητή νοημοσύνη έχει αρχίσει να χρησιμοποιείται σε πολλούς τομείς στη βιομηχανία. Το οποίο συνεπάγεται ότι έχει αλλάξει δραματικά τις ζωές πολλών ανθρώπων σε διάφορους τομείς, είτε είναι ο τομέας της υγείας, είτε είναι της βιομηχανίας των αυτοκινήτων. Έχει καταλήξει να είναι ένα αναπόσπαστο κομμάτι της καθημερινότητας μας. Συνεπώς σε αυτό το σημείο της παρούσας διπλωματικής θα αναφέρουμε ενδεικτικά κάποια από τα διλήμματα και νομικά προβλήματα τα οποία έχουν παρουσιαστεί λόγω της τεχνητής νοημοσύνης:

Τομέας της αυτοκινητοβιομηχανίας και της πλοήγησης: Όσο μιλάμε για τον κλάδο των μεταφορών, το αυτοκίνητο είναι η πιο κοινή εφαρμογή της τεχνολογίας AI στη σήμερον ημέρα. Από τα παλαιότερα χρόνια που το αυτοκίνητο ήταν χειροκίνητο παντού, έχει σημειωθεί σημαντική βελτίωση αν σκεφτούμε ότι πλέον υπάρχουν εντελώς αυτόνομα αυτοκίνητα τα οποία δεν χρειάζεται να είναι επανδρωμένα για να οδηγηθούν. Συνεπώς η ανάπτυξη αυτή επιφέρει και κάποια αρνητικά μαζί της όπως πολλά ερωτήματα σε περίπτωση ατυχημάτων. Λόγου χάρη αν

ένας άνθρωπος ο οποίος είναι υπό την επήρεια αλκοόλ οδηγήσει κάποιο αυτόνομο αυτοκίνητο και δημιουργήσει ατύχημα, ποιος θα είναι ο υπαίτιος του ατυχήματος αυτού. Ερωτήματα σαν και αυτά πρέπει να αναγκάσουν τη βιομηχανία να κατασκευάσει αυτοκινούμενα τα οποία θα έχουν την ικανότητα να κρίνουν τη κατάσταση στην οποία βρίσκονται και τις συνθήκες στις οποίες οδηγείται το αμάξι. Πάντως το μόνο σίγουρο είναι ότι με την αύξηση των πωλήσεων των αυτοκινούμενων αυτοκινήτων, θα αυξηθούν και τα ατυχήματα τα οποία συμβαίνουν, καθώς υπάρχει πάντα ο παράγοντας άνθρωπος ο οποίος είναι απρόβλεπτος, κάτι το οποίο ένα αυτόνομο αυτοκίνητο προς το παρόν δεν μπορεί να διαχειριστεί.

Ένας άλλος τομέας στον οποίο η τεχνητή νοημοσύνη έχει αναπτυχθεί πολύ είναι ο τομέας της πλοήγησης (GPS) όπου υπολογίζει τις "βέλτιστες" διαδρομές για ταξίδια κλπ. Αυτή η τεχνολογία εμφανίστηκε αρχικά στα τέλη του περασμένου αιώνα όταν η τεχνολογία δεν ήταν πολύ καλή με αποτέλεσμα να μην ήταν πολύ ακριβής στις προβλέψεις της. Ωστόσο, καθώς η γνώση και η τεχνολογία προχωρούσαν η ακρίβεια πλοήγησης βελτιώθηκε σημαντικά, κάτι το οποίο οδήγησε σε υπολογιστές κινητά και διάφορα άλλα gadgets να έχουν συστήματα πλοήγησης στη διάθεση τους ανά πάσα στιγμή. Επιπρόσθετα, συνεχίζουν να γίνονται εξελίξεις στην τεχνητή νοημοσύνη ώστε να δίνονται ταχύτερες και καλύτερες ταξιδιωτικές διευθύνσεις. Βέβαια όσο καλή και να είναι αυτή η τεχνολογία, υπάρχουν πράγματα τα οποία την κρατάνε πίσω όπως είναι το αδύναμο σήμα, που κατά συνέπεια σημαίνει μη πρόσβαση στο διαδίκτυο, όπου σε αυτό το ενδεχόμενο δεν μπορούν να παρθούν από κάπου οι πληροφορίες για τη σωστή πλοήγηση, και ένα άλλο πρόβλημα που υπάρχει είναι το λάθος στη κρίση του AI, που πολλές φορές οδηγεί τον κόσμο σε λάθος σημεία. Υπάρχουν πολλά περιστατικά τα οποία έχουν δείξει ότι η τεχνητή νοημοσύνη στον τομέα της πλοήγησης δεν είναι αλάνθαστη και ότι πρέπει να ληφθεί σημαντική προσοχή στη διατήρηση της ασφάλειας της. Παρόμοια προβλήματα μπορούν να εμφανιστούν από δυσλειτουργίες στη πλοήγηση και λειτουργία, ή ακόμα και από την ανακρίβεια των δεδομένων που επεξεργάζονται από τη τεχνητή νοημοσύνη. Συνεπώς καταλαβαίνει κανείς ότι η πλοήγηση και η αυτοκινητοβιομηχανία είναι 2 τομείς οι οποίοι κάνουν ραγδαία χρήση της τεχνητής νοημοσύνης χωρίς όμως να μπορούμε να την εμπιστευτούμε με άνεση ακόμα. **[10]**

Τομέας της υγείας : Ο τομέας της υγείας έχει πολλές εφαρμογές της τεχνητής νοημοσύνης όπως θα αναλύσουμε περαιτέρω στη παρούσα διπλωματική εργασία, αλλά 2 πολύ σημαντικές εφαρμογές είναι τα χειρουργικά ρομπότ και οι τρόποι διάγνωσης και θεραπείας με τη χρήση τεχνητής νοημοσύνης. Πολλές χειρουργικές αίθουσες έχουν κάνει χρήση του πρώτου χειρουργικού ρομπότ, το οποίο από μόνο του δεν είναι ακόμα ικανό να πραγματοποιήσει χειρουργικές επεμβάσεις από μόνο του, αλλά παρόλα αυτά οι γιατροί το χρησιμοποιούν για να πραγματοποιήσουν επεμβάσεις σύμφωνα με κάποια προκαθορισμένα πρωτόκολλα. Βέβαια με την ραγδαία εξέλιξη της τεχνολογίας που υπάρχει στις μέρες μας, δεν θα αργήσουμε πολύ να δούμε ρομπότ τα οποία θα είναι ικανά να λειτουργήσουν από μόνα τους. Βέβαια όταν πραγματοποιηθεί αυτό θα συνεχίσουμε να έχουμε ένα πρόβλημα το οποίο υπάρχει τώρα. Αυτό το πρόβλημα είναι ότι στη περίπτωση αποτυχίας ενός ευφυούς ρομπότ στη χειρουργική επέμβαση, αντιμετωπίζεται

ως ιατρικό ατύχημα, κάτι το οποίο σε ορισμένες περιπτώσεις δεν είναι ηθικά σωστό.

Μια άλλη εφαρμογή της τεχνητής νοημοσύνης είναι στη διάγνωση των ασθενειών, κάτι το οποίο είναι πολύ χρήσιμο να γίνεται έγκαιρα, καθώς μπορεί να σώσει πολλές ζωές και να διευκολύνει την ολική θεραπεία των ασθενών. Σε πολλές περιπτώσεις η τεχνητή νοημοσύνη έχει αποδείξει ότι είναι ικανή να απορροφήσει γρήγορα την ιατρική γνώση και να παράγει ακριβής διαγνώσεις, κάτι το οποίο είναι ύψιστης σημασίας για τη σωστή λειτουργία της. Επιπρόσθετα η τεχνητή νοημοσύνη έχει την ικανότητα να βελτιώνεται συνεχώς εκπαιδεύοντας στα δεδομένα που παράγουν οι ασθενείς με την ανατροφοδότηση τους, καθώς χρησιμοποιεί τα δεδομένα αυτά ώστε να μπορέσει να παρέχει καλύτερες διαγνώσεις και να προτείνει καλύτερες θεραπείες. Παρόλα αυτά, η τεχνητή νοημοσύνη μπορεί να δημιουργήσει πολύ κακό, καθώς αν δεν γίνει σωστή εκπαίδευση του αλγορίθμου μπορεί οι διαγνώσεις και οι θεραπείες να έχουν προκαταλήψεις και να μην κάνουν σωστή διάγνωση ή ακόμα και να προτείνουν λάθος θεραπείες σε άτομα τα οποία κινδυνεύει η ζωή τους. Οπότε είναι ένα από τα μεγαλύτερα ηθικά διλήμματα στον τομέα της ιατρικής καθώς η διάγνωση και η θεραπεία είναι από τα βασικότερα πράγματα στη ζωή των ανθρώπων, κάτι το οποίο δημιουργεί ακόμα περισσότερες αμφιβολίες όσο αφορά το AI και τη χρήση του στον υγειονομικό τομέα**.[10]**

Τομέας της ασφάλειας : Η ασφάλεια είναι ένας τομέας ο οποίος έχει αρχίσει να εντάσσει κατά πολύ την τεχνητή νοημοσύνη στο τομέα της, καθώς το AI εκ φύσεως του είναι πρωτοποριακό και παρέχει πολλές δυνατότητες στον χρήστη. Δύο εφαρμογές της τεχνητής νοημοσύνης στον τομέα της ασφάλειας είναι η έξυπνη αναγνώριση, όπως είναι η αναγνώριση προσώπου και το έξυπνο σπίτι. Όσο αφορά την έξυπνη αναγνώριση, αναφερόμαστε σε τεχνολογίες οι οποίες χρησιμοποιούνται στην αναγνώριση προσώπου, δακτυλικών αποτυπωμάτων όπως και αντίστοιχα φωνητικής αναγνώρισης. Οι τεχνολογίες αυτές είναι ευρέως χρησιμοποιούμενες σε κινητά, υπολογιστές όπως και ελέγχους ασφαλείας, με τις οποίες ερχόμαστε σε επαφή σχεδόν σε καθημερινή βάση, και οι περισσότερες χρησιμοποιούν κατά κύριο λόγο αναγνώριση δακτυλικών αποτυπωμάτων και, ως εκ τούτου, απαιτούν κάποια μορφή ελέγχου ταυτότητας ασφαλείας προκειμένου να εγγυηθούν ότι ασφάλεια μας. Όσο αφορά τον φωνητικό έλεγχο και την αναγνώριση προσώπου, είναι 2 τεχνολογίες οι οποίες χρησιμοποιούνται αντίστοιχα άλλα δεν έχουν χρησιμοποιηθεί σε τόσες εφαρμογές όσο τα δακτυλικά αποτυπώματα, καθώς θεωρούνται ακόμα καινούργια τεχνολογία. Πλέον που η τεχνολογία της αναγνώρισης έχει εξελιχθεί αρκετά, παρατηρούμαι ότι χρησιμοποιείται όλο και πιο πολύ και για αγορές οι οποίες πραγματοποιούνται διαδικτυακά, κάτι το οποίο είναι ένα πολύ σημαντικό επίτευγμα, καθώς περιορίζει πολύ το ενδεχόμενο της απάτης, καθώς για να αγοράσει κάποιος ένα προϊόν χρησιμοποιώντας την κάρτα κάποιου, θα χρειαζόταν και το δαχτυλικό του αποτύπωμα. Αν και η τεχνολογίας της αναγνώρισης δεν σταματάει ποτέ να αναπτύσσεται και να γίνεται πιο ακριβής, δεν κάνει κάτι για να βελτιώσει την προσωπική της ασφάλεια, κάτι το οποίο είναι πολύ επικίνδυνο καθώς κάποιος με τις κατάλληλες δυνατότητες θα μπορούσε να κλέψει τα δεδομένα αυτά που χρησιμοποιούνται για την αναγνώριση. Αυτό είναι πολύ επικίνδυνο καθώς μπορεί να θέσει πολύ κόσμο σε κίνδυνο με

διάφορους τρόπους, και αφού μιλάμε για δακτυλικά αποτυπώματα μπορεί να θέσει και κόσμο σε κίνδυνο και από νομικής πλευράς. Συνεπώς όσο χρησιμοποιούμε την τεχνητή νοημοσύνη για την αναγνώριση, είναι σίγουρο πως θα μας παρακολουθεί και θα παραβιάζει το απόρρητο μας κατά κάποιο τρόπο. Συνεπώς αυτό δημιουργεί το ερώτημα "πως μπορούμε να προστατεύσουμε το απόρρητο μας?".

Ένα άλλο σοβαρό θέμα που χρησιμοποιεί τεχνητή νοημοσύνη είναι το έξυπνο σπίτι, μια ιδέα η οποία ακούγεται πολύ πρωτοποριακή και καλή από πολλές οπτικές πλευρές. Ένα έξυπνο ολοκληρωμένο σύστημα που βασίζεται σε ήχο, βίντεο, συνδέσεις δικτύου, αυτοματισμούς και τεχνητή νοημοσύνη, είναι τα θεμέλια για ένα έξυπνο σπίτι. Ακριβώς επειδή το έξυπνο σπίτι απαρτίζεται από πολλές τεχνολογικά καινοτόμες συσκευές, μπορεί και παρέχει ασφάλεια στην οικογένεια με διάφορους τρόπους και υπηρεσίες. Επιπρόσθετα αν χτιστούν πολλά έξυπνα σπίτια μπορούμε και βοηθάμε το περιβάλλον καθώς είναι eco-friendly(φιλικά προς το περιβάλλον), κάτι το οποίο είναι απαραίτητο πλέον στη ζωή μας. Πολλές χώρες οι οποίες είναι πιο οικονομικά ανεπτυγμένες, έχουν στρέψει τη προσοχή τους προς αυτά τα σπίτια, καθώς είναι προτεραιότητα όλων να διασφαλιστεί η ασφάλεια των οικογενειών και να φροντίσουν να προστατέψουν το περιβάλλον. Το έξυπνο σπίτι έχει πολλές τεχνολογίες οι οποίες προστατεύουν τους κατοίκους από πιθανούς κινδύνους, κάτι το οποίο είναι ύψιστης σημασίας για τις περισσότερες οικογένειες. Αυτό πραγματοποιείται με πολλούς αισθητήρες , κάμερες και άλλα συστήματα τα οποία είναι όλα συνδεδεμένα στο διαδίκτυο και μπορούν να ελέγχονται από τον κάτοχο του σπιτιού. Βέβαια όλα τα συστήματα ασφαλείας είναι ανίκανα να προστατέψουν τον κάτοχο του σπιτιού από τον πωλητή που τους παρείχε το έξυπνο σπίτι, καθώς έχει την ικανότητα να κλέψει τα δεδομένα των αγοραστών και να κάνει πράγματα πέραν της φαντασίας μας. Συνεπώς με την χρήση τεχνητής νοημοσύνης ένας κίνδυνος ο οποίος υπάρχει σχεδόν πάντα είναι η κλοπή των προσωπικών δεδομένων κάποιου.**[10]**

Τομέας των επιχειρήσεων : Ο ανταγωνισμός στις διαδικτυακές αγορές είναι τεράστιες. Συνεπώς πολλές εταιρίες όπως είναι το AliExpress, Amazon, Alibaba και άλλες πολλές, χρησιμοποιούν τεχνητή νοημοσύνη για να προωθήσουν τα προϊόντα τους. Πιο συγκεκριμένα η τεχνητή νοημοσύνη προσπαθεί να καταλάβει τις προτιμήσεις των καταναλωτών και ανάλογα με το τι τους αρέσει, τους προτείνει και τα αντίστοιχα προϊόντα, κάτι το οποίο βοηθάει απίστευτα στις πωλήσεις. Η συνομιλία πελατών, τα προτεινόμενα προϊόντα και η έξυπνη σύσταση προσφέρουν τεράστια ευκολία στη ζωή των ανθρώπων, κάτι το οποίο ισχύει και για τις επιχειρήσεις , καθώς έχουν μεγάλα κέρδη από όλες τις πωλήσεις τις οποίες κάνουν. Βέβαια παρόλα τα καλά της διαδικτυακής αγοράς, πρέπει να δώσουμε ιδιαίτερη προσοχή στα αρνητικά που έχει, τα οποία είναι τα ψεύτικα προϊόντα, τα οποία πολλές φορές μοιάζουν με κάτι το οποίο αναζητεί ο καταναλωτής και καταλήγουν να είναι φθηνές απομιμήσεις κάποιου άλλου προϊόντος. Συνεπώς σε αυτό το ενδεχόμενο πρέπει να εκπαιδεύουμε σωστά το σύστημα για να μην προτείνει προϊόντα τα οποία καταλήγουν αν είναι απάτες. Βέβαια προς το παρόν αυτό δεν γίνεται σε μεγάλο βαθμό με απόρροια να ρισκάρει ο κόσμος όταν αγοράσει προϊόντα από το διαδίκτυο.

Ο τομέας των επιχειρήσεων έχει πολλούς τρόπους με τους οποίους χρησιμοποιεί τη τεχνητή νοημοσύνη, ένας από αυτούς είναι με την χρήση έξυπνων μηχανών αναζήτησης οι οποίες χρησιμοποιούν τεχνολογίες τεχνητής νοημοσύνης μαζί με την παραδοσιακή τεχνολογία αναζήτησης. Η τεχνολογία αυτή χρησιμοποιεί το AI για να μάθει τις πιο διαδεδομένες αναζητήσεις από το διαδίκτυο και τις πιο συχνές αναζητήσεις που κάνει ο εκάστοτε χρήστης για να προβάλει αυτά τα οποίο πιστεύει ότι αναζητεί ο χρήστης. Οι ιστοσελίδες αναζήτησης όπως είναι το Google ή το Bing, είναι από τις ιστοσελίδες με την μεγαλύτερη επισκεψιμότητα καθώς ο περισσότερος κόσμος περνάει από αυτές τις ιστοσελίδες για να καταλήξει εκεί που θέλει. Βέβαια τα κέρδη τα οποία δέχονται οι έμπορες από τις μηχανές αναζήτησης είναι πολλά με αποτέλεσμα πιο σκιεροί έμποροι να χρησιμοποιήσουν τα έσοδα αυτά για κακό, κάτι το οποίο δημιουργεί πολλά διλήμματα όσο αφορά την επιρροή που θα έχουν οι έμπορες αυτοί.**[10]**

Κεφάλαιο 4 : Μελέτη περιπτώσεων ηθικής της Τεχνητής Νοημοσύνης στην Ιατρική

4.1 Ηθική του A.I. στο τομέα της Παθολογίας

Ο τομέας της υγειονομικής περίθαλψης είναι σε γενικές γραμμές επικεντρωμένος στα δεδομένα και περιλαμβάνει πολλούς υποτομείς που δημιουργούν δικά τους δεδομένα όπως η ασφάλιση , το φαρμακείο , η διοίκηση , ιδρύματα υγειονομικής περίθαλψης και διάφορες ειδικότητες της κλινικής πρακτικής των ασθενών . Λόγω αυτών των δεδομένων παράγονται τεράστιες ποσότητες πληροφοριών σε όλα τα επίπεδα της υγειονομικής περίθαλψης με απόρροια να παρέχονται εξαιρετικές γνώσεις για το πώς ασκείται η ιατρική. Οι κλινικές οι οποίες χρησιμοποιούν τη τεχνική νοημοσύνη ως εργαλείο στη καθημερινότητα τους, έχουν βελτιώσει σε πολύ μεγάλο βαθμό τη δυνατότητα να συλλέγουμε δεδομένα στον τομέα της ιατρικής . Παρόλα αυτά οι αναλύσεις δεδομένων μεγάλης κλίμακας στους υποτομείς δυσκολεύονται να πραγματοποιηθούν με αποτέλεσμα να έχουν μια καθυστέρηση. **[23,24]** Παρόλα αυτά οι υπολογιστικοί αλγόριθμοι οι οποίοι βασίζονται στις αρχές της μηχανικής μάθησης αναμένονται να βελτιώσουν την κατανόηση μας σχετικά με την υγειονομική περίθαλψη μέσω αυτοματοποιημένων αναλύσεων μεγάλων δεδομένων .**[25,26]**

Σε γενικές γραμμές με την βοήθεια της Τεχνητής Νοημοσύνης οι ερευνητές έχουν την δυνατότητα να αναλύσουν πολλά δεδομένα , όμως προτιμούν να επικεντρώνονται στα δεδομένα που αφορούν τις σύνηθες κλινικές εργασίες. Οι γενικοί γιατροί παράγουν μεγάλες ποσότητες δεδομένων οι οποίες δεν είναι δομημένες όπως σημειώσεις κατά την επίσκεψη των ασθενών. Βέβαια στις πιο εξελιγμένες χώρες οι γιατροί εξαρτώνται σε τεράστιο βαθμό στην ακτινολογία και παθολογία για να τους κατευθύνει κατά τη διάρκεια της διάγνωσης , της πρόγνωσης ακόμα και κατά τη θεραπεία του ασθενούς. Πιο συγκεκριμένα οι ακτινολόγοι πρωταγωνιστούν στην όλη διαδικασία της θεραπείας του ασθενή καθώς έχουν μεγάλη εμπειρία στη χρήση της τεχνολογίας . Κάτι το οποίο είναι απαραίτητο καθώς εργαλεία όπως η μαγνητική και αξονική τομογραφία χρησιμοποιούνται ως οδηγοί στη θεραπεία του ασθενή και απαιτούν άριστη χρήση της τεχνολογίας. Η εξέλιξη της τεχνολογίας επηρεάζει άμεσα και τη παθολογία με διάφορους τρόπου. Αυτό συμβαίνει καθώς η παθολογία και η ακτινολογία είναι 2 ειδικότητες οι οποίες χρησιμοποιούν εκτενώς δεδομένα εικόνας για την σωστή μεταχείριση των ασθενών έπειτα από επίδειξη ειδικών. Αν και η ακτινολογία σε γενικές γραμμές είναι πολύ πιο εξελιγμένη τεχνολογικά και κάνει πιο σωστή χρήση των ιατρικών εικόνων, η παθολογία έχει αρχίσει να κινείται και αυτή προς αυτήν την κατεύθυνση, με απόρροια το γεγονός αυτό να προσελκύει όλο και περισσότερους ερευνητές της Τεχνητής Νοημοσύνης στον τομέα αυτόν , καθώς το A.I. είναι πολύ χρήσιμο στην ανάλυση των εικόνων. Εν κατακλείδι οι αλγόριθμοι απεικόνισης της Τεχνητής Νοημοσύνης έχουν δει τον

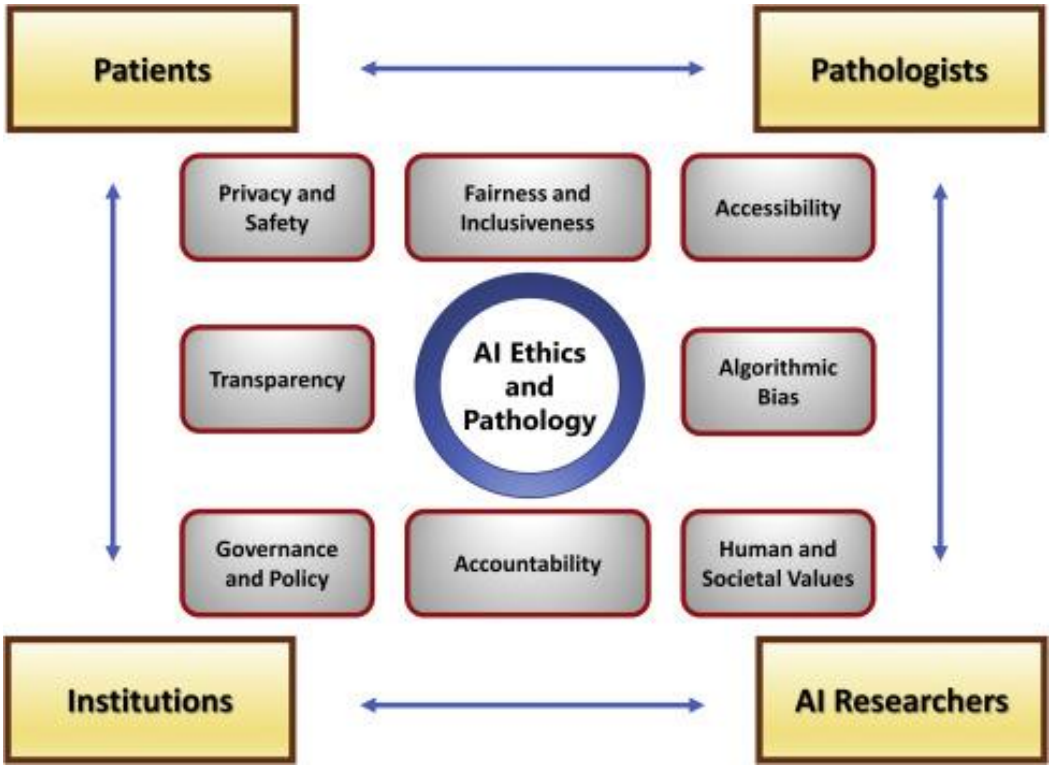
μεγαλύτερο όγκο έρευνας και προόδους σε σχέση με το πέρασμα της δεκαετίας.**[27,28]**

Η Παθολογία είναι μία επιστήμη υψηλής έντασης δεδομένων καθώς κάνει χρήση ταυτόχρονα και κλινικών και φαινοτύπων στοιχείων για την κατορθώσει της παραδοσιακής της λειτουργίας. Παρόλα αυτά έχει υπάρξει μία αυξημένη ανάγκη να χρησιμοποιηθούν παραπάνω δεδομένα ασθενών ως μέρος της καθημερινής διαγνωστικής ρουτίνας (όπως πληροφορίες μοριακών οδοδεικτών σε ατομικό και πληθυσμιακό επίπεδο). Επιπρόσθετα θα πρέπει να χρησιμοποιηθούν παραπάνω δεδομένα όπως είναι τα κοινωνικοοικονομικά και ο τρόπος ζωής των ανθρώπων για την βελτίωση των ερευνητικών κατηγοριών έτσι ώστε να έχουμε καλύτερη ακρίβεια στη Παθολογία και στα φάρμακα που χορηγούνται στους ασθενείς. Αυτή η πολυπλοκότητα που δημιουργείται είναι ιδανική συνθήκη για τη χρήση τεχνητής νοημοσύνης καθώς ένα από τα βασικά πλεονεκτήματα της είναι η ικανότητα της να αναλύει ταυτόχρονα πολλά δεδομένα από τομείς διαφορετικών ερευνητικών δεδομένων κάτι το οποίο είναι σχεδόν αδύνατο από το ανθρώπινο μυαλό. Ωστόσο κρίνεται απαραίτητο να είναι προσεκτικός όποιος διεξάγει μελέτες με τη χρήση τεχνητής νοημοσύνης καθώς η πολυπλοκότητα αυτών των δεδομένων μπορεί να οδηγήσει σε προκαταλήψεις με αποτέλεσμα να υπάρξει πρόβλημα στη συνέχεια της μελέτης. Οι συγκεκριμένες προκαταλήψεις περιλαμβάνουν αυτές οι οποίες είναι εγγενείς στον αλγόριθμο και τις προκαταλήψεις οι οποίες προκύπτουν από τα σύνολα τα οποία χρησιμοποιούνται για την εκπαίδευση του αλγορίθμου AI . Οι προκαταλήψεις οι οποίες είναι εγγενείς στον αλγόριθμο δεν είναι τόσο "κακές" ηθικά καθώς είναι απαραίτητες για τη σωστή κατανόηση των εσωτερικών λειτουργιών του αλγορίθμου. Για παράδειγμα ο γνωστός αλγόριθμος ομαδοποίησης k-means λειτουργεί πολύ πιο αποδοτικά όταν τα δεδομένα τα οποία χρησιμοποιούμε σχηματίζουν μεταξύ τους συμπλέγματα που είναι σφαιρικά και παρόμοια σε μέγεθος . Επιπρόσθετα τα δίκτυα εκμάθησης έχουν μεταβλητές οι οποίες μπορούν να αλλάζουν (tunable variables) γνωστές και ως υπερμεταβλητές (hyperparameters) οι οποίες είναι αναντικατάστατο μέρος του αλγόριθμου AI και πρέπει να ανατεθούν σε πολύ συγκεκριμένη βάση από ανθρώπους για την συνέχιση της ερευνητικής μελέτης τους . Όλες αυτές οι επιλογές πραγματοποιούνται από τον ερευνητή με αποτέλεσμα πολλές φορές συνειδητά αλλά και ασυνείδητα να οδηγήσουν στην δημιουργία προκαταλήψεων οι οποίες μπορούν να αποφέρουν σε απρόβλεπτες επιπτώσεις με απόρροια τη μη "σωστή" λειτουργία του αλγόριθμου. **[27,28]**

Από την οπτική γωνία της ηθικής της τεχνητής νοημοσύνης , τα πιο σημαντικά ζητήματα που υπάρχουν με την αλγοριθμική ευνοιοκρατία είναι κατά κύριο λόγο στο πλαίσιο των δεδομένων συνόλων που χρησιμοποιούνται για την πραγματοποίηση της ερευνητικής μελέτης . Συνεπώς και τα δείγματα αλλά και η εκτίμηση παίζουν έναν πολύ σημαντικό ρόλο στη προκειμένη περίπτωση . Για παράδειγμα αν το σύνολο των δεδομένων μας έχει ανισορροπία κατηγορίας , δηλαδή αν έχουμε ως δειγματοληψία μόνο ενήλικες λευκούς άντρες λόγω διαθεσιμότητας δειγμάτων και άλλων παραγόντων όπως είναι κοινωνικοοικονομικοί παράγοντες , τα αποτελέσματα του αλγορίθμου AI που εκπαιδεύονται με βάση αυτά τα δείγματα ενδέχεται να έχουν ανακρίβεια όταν εφαρμόζονται στον υπόλοιπο πληθυσμό . Με αποτέλεσμα τα αποτελέσματα μια τέτοιας ομοιογενούς ερευνητικής μελέτης να βλάψουν ακούσια μια

μειονότητα. Ένα εξίσου πολύ σημαντικό πρόβλημα το οποίο αντιμετωπίζεται στη παθολογία είναι η υποπροδιαγραφή , δηλαδή το φαινόμενο στο οποίο ένα σύνολο εκπαίδευσης AI δεν παρέχεται με όλες τις απαραίτητες παραμέτρους για την σωστή εκπαίδευση του αλγορίθμου . Για παράδειγμα αν η γενετική ενός πληθυσμού αποτελούσε σημαντικό παράγοντα στη ταξινόμηση των ιστολογικών εικόνων , η μη συμπερίληψη αυτών των λεπτομερειών θα οδηγούσε σε ατελώς εκπαιδευόμενο μοντέλο AI. Συνεπώς η υποπροδιαγραφή τείνει να οδηγεί σε λανθασμένες συσχετίσεις όσο αφορά τη πρόβλεψη των κλινικών αποτελεσμάτων. Οπότε καταλήγουμε στο συμπέρασμα ότι οι AI ερευνητές οι οποίοι ασχολούνται με αλγόριθμους DL πρέπει να έχουν στη κατοχή τους μια πολύ καλή γνώση περί του θέματος της παθολογίας αλλά και των παραγόντων που θα πρέπει να συμπεριλάβουν στα δείγματα εκπαίδευσης . Αντίστοιχα και οι παθολόγοι οι οποίοι χρησιμοποιούν αυτούς τους συγκεκριμένους αλγόριθμους θα πρέπει να γνωρίζουν τα πιθανά προβλήματα και τις αστοχίες που μπορεί να έχει ο AI αλγόριθμος στο τομέα της παθολογίας. **[29,30,31]**

Όσο εντάσσεται το AI όλο και πιο πολύ στο τομέα της παθολογίας , είναι λογικό να υπάρξουν περιστατικά στα οποία ο αλγόριθμος δεν δούλεψε σωστά και οδήγησε σε λανθασμένες προβλέψεις . Ένα τέτοιο περιστατικό πραγματοποιήθηκε λοιπόν στο Αρκάνσας το 2016 καθώς η συγκεκριμένη πολιτεία ενέκρινε τη χρήση ενός προγράμματος βασισμένου σε αλγόριθμους που σχεδιάστηκαν από την intelRAI (έναν μη κερδοσκοπικό συνασπισμό ερευνητών από όλο τον κόσμο) για να καθορίσει τις ώρες φροντίδας που χρειάζονται οι ασθενείς με περιορισμένη κινητικότητα . Βέβαια οι ώρες δεν καθορίστηκαν επιτυχώς καθώς ο αλγόριθμος AI ήταν περιορισμένος στη χρησιμότητα του , επειδή ανέθετε μεταβλητές βαθμολογίες για άτομα με παρόμοιες αναπηρίες, με αποτέλεσμα να καταλήξει σε λανθασμένες αποφάσεις στον υπολογισμό των απαιτούμενων ωρών φροντίδας , κάτι το οποίο επηρέασε αρνητικά τη ζωή εκατοντάδων ασθενών. Δυστυχώς στο συγκεκριμένο περιστατικό δεν δόθηκε κάποια εξήγηση στους ασθενείς σχετικά με τις ώρες , καθώς τα πρότυπα χρήσης του αλγορίθμου δεν ορίστηκαν με σαφήνεια και επιπλέον δεν αποκαλύφθηκαν στους ενδιαφερόμενους ώστε να βρεθούν και να διορθωθούν τα σφάλματα. Είναι σημαντικό να μην ξεχνάμε ότι υπάρχει η αλγοριθμική μεροληψία , κάτι το οποίο στη προκυμμένη περίπτωση δημιούργησε τεράστιο πρόβλημα στους ανθρώπους οι οποίοι “εκτέθηκαν” σε αυτόν τον αλγόριθμο. Γενικά η αλγοριθμική μεροληψία ή αλλιώς η στατιστική προκατάληψη είναι ένα θέμα το οποίο δεν είναι εύκολο να κατανοηθεί. Οπότε δεν είναι ρεαλιστικό να υπάρχει η προσδοκία από έναν απασχολημένο ασκούμενο παθολόγο να κατανοεί και να γνωρίζει τις διάφορες εκδοχές της αλγοριθμικής/στατιστικής προκατάληψης. Συνεπώς η λύση στο θέμα της προκατάληψης ενδέχεται να πρέπει να δοθεί από τις αρχές όπως είναι ρυθμιστικοί φορείς. Παρόλα αυτά οι ερευνητές έχουν την υποχρέωση να συνεργάζονται με παθολόγους για να αποκτήσουν παραπάνω οπτικές γωνίες , και πιο συγκεκριμένα να πάρουν τις γνώμες από άτομα τα οποία έχουν πρακτική εμπειρία στο τομέα της παθολογίας. Αυτό θα έχει ως αποτέλεσμα τη μακροπρόθεσμη μείωση των προκαταλήψεων σε αλγόριθμους παθολογίας AI και συνεπώς την ορθή λειτουργία τους. **[32,33,34]**



Source: Ethics of AI in Pathology: Current Paradigms and Emerging Issues, Chhavi Chauhan, Rama R. Gullapalli.

4.1.1 Φυλή και ηθική του AI στο τομέα της παθολογίας

Όλες οι μεταβλητές κατά τη δημιουργία ενός αλγόριθμου AI είναι σημαντικές , αλλά μία η οποία παίζει κρίσιμο ρόλο στη σωστή λειτουργία του αλγόριθμου είναι η μεταβλητή της φυλής. Γενικά το μοντέλο AI μπορεί να αντιμετωπίσει πολλά προβλήματα και να έχει πολλές επιδεινώσεις της απόδοσης του από διάφορους λόγους όπως είναι οι μετατοπίσεις δεδομένων , ελαττωματικές συσχετίσεις και υποπροδριαγραφές που περιορίζουν την αποτελεσματικότητα του αλγορίθμου σε μελέτες ταξινόμησης στη παθολογία. Αυτοί οι περιορισμοί όμως επιδεινώνονται ακόμα πιο πολύ σε συγκεκριμένους τομείς της παθολογίας όπως είναι και ο τομέας της ενσωμάτωσης δεδομένων υγειονομικής περίθαλψης , όπου στο συγκεκριμένο τομέα υπάρχουν αρκετές πολυπλοκότητες. Κάποιες από αυτές είναι οι υποκείμενες φυσιολογικές επιδράσεις ακόμα και οι γενετικοί παράγοντες προδιάθεσης για μια νόσο. Αυτές οι πολυπλοκότητες επιδεινώνονται περαιτέρω όταν τα δεδομένα αυτά αναλύονται χωρίς να λαμβάνεται υπόψη ο πολύ σημαντικός παράγοντας της φυλής και εθνικότητας. Με βάση τα ιστορικά γεγονότα, ο κόσμος έχει κατασταλάξει ότι η φυλή θεωρείται ένα παρασκεύασμα του ανθρώπου για κοινωνικούς σκοπούς

και ότι δεν έχει καμία βιολογική βάση, βέβαια πολλά στοιχεία υποδηλώνουν ότι η φυλή συνδέεται άμεσα με τη γενετική. **[12]** Εργασία που παρουσιάστηκε από τους AI ερευνητές Joy Boulamwini και Timnit Gebru ,έχουν κάνει προφανές το ότι τα AI συστήματα και οι καθορισμένες παράμετροι πρέπει να δοκιμάζονται μεταξύ των φυλών, για να γίνει σωστή διαπίστωση της αποτελεσματικότητας τους ,και της ευρείας χρησιμότητας τους πέρα από τη χρήση με σύνολα δεδομένων εκπαίδευσης.**[13]** Έχοντας πλέον περάσει από τη κρίσιμη περίοδο του Κορονοϊού , γνωρίζουμε ότι τα κέντρα ελέγχου και πρόληψης νοσημάτων έδωσαν ιδιαίτερη έμφαση σχετικά με τα ζητήματα ισότητας υγείας στις φυλετικές και εθνοτικές μειονότητες. Το έργο παρακολούθησης COVID-19 επιβεβαιώνει ότι υπήρχαν μεγαλύτερα ποσοστά κρουσμάτων COVID-19 σε μαύρους , ιθαγενείς , ισπανόφωνους αλλά και άλλες μειονότητες. Αυτά τα πρόσφατα ευρήματα δημιουργούν μια επιτακτική ανάγκη να συμπεριληφθούν τα δεδομένα της φυλής και της εθνικότητας στα σύνολα δεδομένων που χρησιμοποιούνται για την εκπαίδευση του αλγορίθμου Τεχνητής Νοημοσύνης στο τομέα της υγειονομικής περίθαλψης για ευρεία χρήση οποιουδήποτε μοντέλου AI που ασχολείται με κλινικά θέματα. Ένα παράδειγμα κακής χρήσης δεδομένων στην εκπαίδευση του αλγόριθμου AI είναι η Optum, μια θυγατρική του ασφαλιστικού κολοσσού UnitedHealth Group, η οποία σχεδίασε μια εφαρμογή για τον εντοπισμό ασθενών οι οποίοι έχουν χρόνιες ασθένειες χωρίς θεραπεία οι οποίοι βρίσκονται σε υψηλό κίνδυνο. Δυστυχώς όμως, παρατηρήθηκε ότι ο αυτοματοποιημένος αλγόριθμος κάνει διακρίσεις ενάντια των μαύρων ασθενών, λόγω ενός μεμονωμένου ασθενή από προηγούμενη θεραπεία. Πλέον το AI χρησιμοποιείται σε πολλές περιπτώσεις όπως είναι και ο καρκίνος και πιο συγκεκριμένα ο καρκίνος του μαστού. Η Αμερικάνικη Υπηρεσία Τροφίμων και Φαρμάκων έχει δώσει ολοκληρωτική άδεια χρήσης σε αυτές τις εφαρμογές χωρίς να χρειάζεται να μαθευτεί δημόσια πως τα εργαλεία που χρησιμοποιούν έχουν δοκιμαστεί εκτενώς σε έγχρωμους ανθρώπους. Συνεπώς οι εφαρμογές που βασίζονται σε AI έχουν τη δυνατότητα να επιδεινώσουν διάφορες ανισότητες που υπάρχουν στα κλινικά αποτελέσματα των κατόχων καρκίνου του μαστού, μια ασθένεια η οποία έχει αυξημένα ποσοστά θνησιμότητας για τις μαύρες γυναίκες κατά 45%. Το συγκεκριμένο παράδειγμα μας δίνει μια ρεαλιστική απεικόνιση των συνεπειών που μπορούν να προκύψουν από την ανεπαρκής έρευνα και τον σχεδιασμό λανθασμένων Αλγορίθμων Τεχνητής Νοημοσύνης προς τις μειονότητες. Συνεπώς είναι υποχρέωση του κάθε ερευνητή να συμπεριλαμβάνει στα δεδομένα εκπαίδευσης του τις φυλές και τις εθνότητες καθώς αυτό έχει τεράστιο αντίκτυπο στις ζωές πολλών ανθρώπων.**[3, 12,14]**

4.1.2Ρίσκο του AI στη Παθολογία και προς τους Παθολόγους

Πλέον στις μέρες μας οι παθολόγοι είναι “υποχρεωμένοι” να χρησιμοποιούν αλγόριθμους AI ακόμα και αν δεν το προτιμούν, καθώς είναι ένα απίστευτα χρήσιμο εργαλείο το οποίο διευκολύνει τη ζωή πολλών ανθρώπων. Όμως εφόσον υπάρχει τόσο μεγάλη χρήση των αλγορίθμων AI, δημιουργούνται τα ερωτήματα του τύπου “θα αντικαταστήσει το AI τους

παθολόγους ;" και "πόσο αποτελεσματικό είναι το AI στη διάγνωση των ασθενών ;", τα οποία ερωτήματα προβληματίζουν αρκετούς παθολόγους στις μέρες μας. Αυτή η ενότητα εξετάζει πιο αναλυτικά μερικά από τα ζητήματα που βιώνουν οι παθολόγοι με έμφαση στην ηθική της Τεχνητής Νοημοσύνης και τον ρόλο που έχουν οι παθολόγοι για την σωστή ανάπτυξη της.

Η αξιολόγηση κινδύνου και αξιολόγηση των κινδύνων που συνδέονται με την εφαρμογή της Τεχνητής Νοημοσύνης είναι μια έντονα μελετημένη πτυχή της ηθικής του AI η οποία λαμβάνει ιδιαίτερη προσοχή τα τελευταία χρόνια. Πολλές καινοτόμες AI τεχνολογίες όπως τα αυτόνομα αυτοκίνητα , η αυτοματοποιημένη αναγνώριση προσώπου ή και AI deep fakes (τεχνολογία η οποία αλλάζει αντικείμενα όπως εικόνες με αποτέλεσμα να μην είναι η γνήσια εικόνα) είναι σοβαρά αίτια για ανησυχία για την οικονομική αλλά κυρίως για την ηθική προοπτική της κοινωνίας μας. Παρόλα αυτά το AI είναι πλέον ένα από τα βασικότερα μέρη της ζωής μας και δεν έχει σκοπό να φύγει, καθώς μας διευκολύνει απίστευτα σε πολλούς τομείς με τη δυνατότητα του να βγάζει απίστευτο φόρτο δουλειάς σε μηδενικό χρόνο. Συνεπώς το πλήγμα θα ήταν πολύ μεγαλύτερο αν το αφαιρούσαμε ξαφνικά από τις ζωές μας λόγω αυτών των κινδύνων που συνδέονται με αυτό. Το Deep-Learning (γνωστό και ως DL) το οποίο είναι η τελευταία μορφή του AI, μας παρέχει με πολλές καινούργιες δυνατότητες αλλά έχει την απαίτηση να εκπαιδεύεται σε ογκώδη ποσότητες δεδομένων για να επιφέρει ουσιώδη αποτελέσματα. Με την εκπαίδευση σε πολύ μεγάλες ποσότητες δεδομένων, η αλγοριθμική απόδοση των ροών εργασιών DL είναι εξαιρετικά πιο καλές καθώς μπορούν να λειτουργήσουν και στα δύο δομημένα και αδόμητα δεδομένα για τη δημιουργία μοντέλων ADM, τα οποία είναι ικανά να πετύχουν εξαιρετικά ψηλά επίπεδα ακρίβειας πρόβλεψης. Ο ορισμός ADM (Automated decision-making) αναφέρεται στην διαδικασία η οποία κάνει χρήση δεδομένων, μηχανών και αλγορίθμων για τη λήψη αποφάσεων σε μια σειρά πλαισίων όπως είναι η δημόσια υγεία, η εκπαίδευση και ο νόμος. Αναφορικά με το AI και το DL, μια κοινή γνώμη που επικρατεί είναι ότι τα δεδομένα τα οποία δημιουργούνται στο εργαστήριο, επηρεάζουν περίπου το 70% των κλινικών αποφάσεων που παίρνονται. Αν και δεν είναι επιστημονικά αποδεδειγμένο αυτό, ισχύει ότι ένα πολύ μεγάλο μέρος των εργαστηριακών δεδομένων (και κλινικών και ανατομικών) απαρτίζουν τον ηλεκτρονικό φάκελο υγείας ενός ασθενούς. Οι ερευνητές AI ελκύονται πολύ από τη διαθεσιμότητα δομημένων και μη δομημένων κλινικών δεδομένων που υπάρχουν σε αρχεία παθολογίας και βάσεις δεδομένων καθώς τους δίνουν τη δυνατότητα να αξιοποιήσουν αυτά τα δεδομένα τα οποία δημιουργούνται σε εργαστηριακό περιβάλλον ώστε να τους αποφέρει κέρδος με συνέπεια τον πιθανό κίνδυνο των ασθενών. Οπότε οι παθολόγοι βρίσκονται σε μια δύσκολη θέση καθώς είναι εκείνοι οι οποίοι έχουν στη κατοχή τους τα εργαστηριακά δεδομένα, και βρίσκονται συνέχεια στο επίκεντρο συζητήσεων σχετικά με τη ιδιοκτησία των δεδομένων και αν πρέπει να δοθούν σε AI ερευνητές στο μέλλον. Κατά συνέπεια οι παθολόγοι θα πρέπει αν είναι σε θέση να αντιμετωπίσουν τέτοιου είδους συζητήσεις και να μπορούν να ανταπεξέλθουν σωστά στις περιστάσεις. Διότι η αξιολόγηση των ρίσκων που σχετίζονται με την έρευνα του AI είναι ένα πολύ σημαντικό θέμα που προβληματίζει πολλούς υποστηρικτές της ηθικής στο AI και έχει ελκύσει πολλές γνωστές φυσιογνωμίες όπως είναι ο Elon Musk. Ευτυχώς πλέον έχουν δημιουργηθεί πολλά ινστιτούτα τα οποία είναι εντελώς αφιερωμένα στο να επιτρέπουν την αξιολόγηση του κινδύνου, τα οποία

ινστιτούτα έχουν εξειδικευμένους ερευνητές ανά θέμα. Λαμβάνοντας υπόψη όλα τα προαναφερθέντα μπορούμε να καταλήξουμε στο ότι η έρευνα της Τεχνητής Νοημοσύνης στο τομέα της παθολογίας επιταχύνεται ραγδαία, όποτε υπάρχει η ευκαιρία στους ασκούμενους παθολόγους να βοηθήσουν στην αξιολόγηση του κινδύνου, έτσι ώστε να υπάρχει μια βαθύτερη και καλύτερη κατανόηση των επιπτώσεων που μπορούν να προκύψουν από τους αλγόριθμους AI στους ασθενείς. **[3,15,16]**

4.1.2 Υποτίμηση ρίσκων τεχνητής Νοημοσύνης στη Παθολογία

Στις μέρες μας όλο και περισσότερες εταιρίες, αν όχι όλες προσπαθούν να μειώσουν τα κόστη που έχουν και ταυτόχρονα να αποφέρουν με τον οποιονδήποτε τρόπο κέρδος σε αυτές. Στο τομέα της υγείας πιο συγκεκριμένα, ένας από τους πιο αποτελεσματικούς τρόπους για να επιφέρει αυτό είναι με τη χρήση αλγόριθμων AI και DL. Όπως έχουμε αναλύσει και προηγουμένως το AI είναι βασικό μέρος στη ζωή των παθολόγων, όμως παίζει ακόμα πιο σημαντικό ρόλο στο τομέα της ακτινολογίας, όπου υπάρχουν αλγόριθμοι AI οι οποίοι εξειδικεύονται στην απεικόνιση. Έχουμε προσέξει τη ραγδαία εξέλιξη του AI και σε άλλους τομείς υγειονομικής περίθαλψης όπως είναι η αξιολόγηση και ανάλυση ολόκληρων σετ εικόνων στο τομέα της ψηφιακής παθολογίας. Παραδοσιακά οι παθολόγοι βασίζονταν στην εμπειρία για να πραγματοποιήσουν διαγνώσεις σε σημείο που δεν είναι πρόθυμοι να εξερευνήσουν καινούργιους τρόπους διαγνώσεων. Καθώς έχουν συνηθίσει την απλή διάγνωση με βάση την οπτική εντύπωση, οπότε το να προχωρήσουν στη διάγνωση με βάση τους αλγόριθμους AI, οι οποίοι δημιουργούν πολλές ανησυχίες και κινδύνους, τους είναι σχεδόν ακατόρθωτο. Παρόλα αυτά πολλές τεχνικές του AI όπως είναι το DL καταφέρνουν να ξεπερνούν σε απόδοση τους ανθρώπους σε ορισμένες εφαρμογές όπως αυτές που περιλαμβάνουν εικόνες, ειδικότερα αυτές οι οποίες έχουν μεγάλο όγκο εικόνων. Κατά τη γνώμη μου, εφόσον ορισμένοι αλγόριθμοι AI καταφέρνουν και "νικάνε" τον άνθρωπο σε εφαρμογές που έχουν να κάνουν κυρίως με εικόνες, κρίνεται απαραίτητο από τους παθολόγους να κάνουν πιο συχνή χρήση τους, καθώς ο σκοπός όλων αυτών των εργαλείων και πρακτικών είναι η ολική και σωστή φροντίδα των ασθενών. Συνεπώς για να γίνει η σωστή επέκταση των πεδίων των πρακτικών των παθολόγων θα πρέπει να υιοθετήσουν εργαλεία όπως μοριακά, κλινικά, και επιδημιολογικά δεδομένα για τη σωστή και ολοκληρωμένη διάγνωση του κάθε ασθενή. Οι άνθρωποι γενικά κατέχουν πολλές παραπάνω δεξιότητες σε σύγκριση με τα συστήματα AI, και μια από αυτές τις δεξιότητες είναι η ικανότητα να συνθέτους πληροφορίες μεταξύ διαφορετικών τομέων των γνώσεων τους με αρκετή ευκολία. Κάτι το οποίο λείπει αυτή τη δεδομένη στιγμή από τα συστήματα AI, οπότε αυτή είναι μια μοναδική ευκαιρία που δίνεται στους παθολόγους ώστε να μπορέσουν να επεκτείνουν το εύρος της πρακτικής της παθολογίας και να παραμείνουν "αναντικατάστατοι" στο τομέα της ιατρικής. Όσο αφορά το τομέα της παθολογίας, είναι θέμα χρόνου μέχρι να ενταχθεί το AI στα εργαλεία τα οποία χρησιμοποιούν οι παθολόγοι καθώς οι αυτοματισμοί είναι όλο και πιο συχνοί στο τομέα της παθολογίας και με τους αυτοματισμούς

συνεπάγεται και η χρήση AI. Παραδείγματα των αυτοματισμών που χρησιμοποιούν AI θα μπορούσαν να είναι η ασφάλιση της ποιότητας των εργαστηριακών δεδομένων όπως και οι αυτοματοποιημένες αξιολογήσεις σε κλινικές υψηλού όγκου. Βέβαια οι περισσότερες αναβαθμίσεις αυτές πραγματοποιούνται από τις εταιρίες οι οποίες είναι υπεύθυνες για την ανάπτυξη των οργάνων, κάτι το οποίο δεν δίνει τη δυνατότητα στον παθολόγο να επέμβει στη διαδικασία ανάπτυξης οργάνων με αλγόριθμους AI. παρόλα αυτά ως τελικός χρήστης ο παθολόγος μπορεί πάντα να κάνει "κριτική" των οργάνων αυτών αλλά και να κάνει σωστές επιλογές στην υιοθεσία τους. Πιο συγκεκριμένα ο παθολόγος μπορεί να έχει την επιμέλεια για τεχνολογίες Τεχνητής Νοημοσύνης και θα πρέπει να έχει την απαραίτητη γνώση ώστε να μπορεί να εκτιμήσει τον κίνδυνο του κάθε εργαλείου με βάση τις αρχές ηθικής της Τεχνητής Νοημοσύνης. Εν κατακλείδι θα πρέπει οι παθολόγοι να είναι σωστά εκπαιδευμένοι ώστε να διατηρήσουν την επίγνωση των ζητημάτων που αφορούν την αλγοριθμική μεροληψία όπως και την ηθική τεχνητής νοημοσύνης έτσι ώστε να υπάρχει μια σωστή αξιολόγηση των μεσών και των τεχνολογιών που χρησιμοποιούνται από παθολογικές πρακτικές, έτσι ώστε να μην μπουν σε κίνδυνο ζωές ασθενών.**[17,18,19]**

4.1.3 Υπερεκτίμηση ρίσκων της Τεχνητής Νοημοσύνης στη Παθολογία

Με τη πρόοδο της τεχνολογίας και τη βοήθεια της Τεχνητής Νοημοσύνης πολλές εξελίξεις στο τομέα της αυτοματοποίησης στενά καθορισμένων εργασιών μπόρεσαν να επιτευχθούν, συγκεκριμένα με τη βοήθεια μεθόδων που χρησιμοποιούσαν DL(deep-learning). Αν και είναι ένα εξαιρετικά εντυπωσιακό κατόρθωμα, αυτό από μόνο του δεν μας δίνει τη δυνατότητα να πούμε ότι η Τεχνητή Νοημοσύνη έχει δυνατότητες επίλυσης προβλημάτων όπως αυτές των ανθρώπων. Πιο συγκεκριμένα η Τεχνητή Νοημοσύνη χρειάζεται ακόμα πολλά χρόνια, μπορεί και δεκαετίες ώστε να μπορέσει να προσομοιώσει την ανθρώπινη συμπεριφορά, δηλαδή την ανθρώπινη επίγνωση και την ικανότητα λήψης αποφάσεων, σε βαθμό που να λέμε ότι είναι αληθινή ανθρώπινη Τεχνητή Νοημοσύνη. Πολλοί ειδικοί έχουν την υποτιμητική άποψη ότι το DL είναι απλά ένα απίστευτα αποτελεσματικό στατιστικό μέσο για τη προσαρμογή και τίποτα παραπάνω. Σχετικά με τον τομέα της παθολογίας είναι ρεαλιστικά αδύνατον προς το παρόν, ένας εξαιρετικός αλγόριθμος AI να αντικαταστήσει έναν άρτια εκπαιδευμένο παθολόγο. Αν και οι τρέχουσες τεχνικές DL έχουν εξαιρετικές αποδόσεις σε στενά καθορισμένες και καλά δομημένες ερωτήσεις στη παθολογία με αυστηρά όρια απόδοσης, δεν είναι ακόμα αρκετά καλές ώστε να επιτευχθεί η σύγκριση με τους παθολόγους. Συνεπώς οι παθολόγοι οφείλουν να είναι προσεκτικοί σχετικά με τον ενθουσιασμό και την διαφημιστική εκστρατεία που υπάρχει τριγύρω από τις έρευνες AI, καθώς πρέπει ταυτόχρονα να αξιολογούν τους ισχυρισμούς από τους ερευνητές AI. Πιο συγκεκριμένα ο ενθουσιασμός αυτός για το AI υπάρχει από το 1960, καθώς η Τεχνητή Νοημοσύνη έχει περάσει από διάφορες φάσεις κάποιες από τις οποίες χαρακτηρίζονται από κύκλους έκρηξης

και άλλες από αποτυχίες, καθώς οι μη ρεαλιστικοί στόχοι των ερευνητών δεν κατάφεραν να επιτευχθούν. Με τη πρόοδο λοιπόν της τεχνολογίας και πιο συγκεκριμένα, την πρόοδο των υπολογιστικών συστημάτων όπως είναι η χωρητικότητα υλικού, και η δικτύωση και αποθήκευση δεδομένων μπορεί να αποδεικνύεται για άλλη μια φορά ότι οι τεχνικές DL είναι μέρος ενός ακόμη κύκλου διαφημιστικής εκστρατείας. Είναι αδύνατον να αμφισβητήσει κανείς ότι οι τεχνολογίες AI εξελίσσονται σε κάθε κύκλο ανάπτυξης, καθώς υπάρχουν καλύτερες τεχνολογίες για να υποστηρίξουν τους αλγορίθμους αυτούς. Βέβαια οι τεχνολογίες AI έχουν μια πολύ υψηλή μπάρα την οποία πρέπει να περάσουν και με διαφορά ώστε να μπορέσουν να χρησιμοποιηθούν χωρίς ιδεασμούς στον τομέα της υγείας και πιο συγκεκριμένα στη φροντίδα των ασθενών.

Μια πολύ ενδιαφέρουσα μελέτη η οποία πραγματοποιήθηκε από τους Frey και Osbourne αξιολογούσε το αντίκτυπο της τεχνολογίας AI σε παραπάνω από 700 τομείς εργασίας στις Ηνωμένες Πολιτείες, η οποία μελέτη αφορούσε άτομα τα οποία χάσανε τη δουλειά τους λόγω οικονομικών παραγόντων. Σύμφωνα με την συγκεκριμένη μελέτη οι γιατροί που εργάζονταν στον τομέα της υγείας ήταν ανάμεσα στα 15 λιγότερο πιθανά επαγγέλματα τα οποία θα επηρεάζονταν από τους αυτοματισμούς AI, με score 0.0042 πιθανότητα να επηρεαστούν. Κάτι το οποίο δεν είναι καθόλου παράλογο, εφόσον η δουλειά των γιατρών είναι εξαιρετικά απαιτητική καθώς είναι πολύπλοκη, διαδραστική, και πολύπλευρη. Παρόλα αυτά οι παθολόγοι πρέπει να αναζητούν ενεργά δεξιότητες κει εμπειρογνωμοσύνη για να δημιουργήσουν συνάφεια μεταξύ των τομέων της ιατρικής. Μια φράση σχετικά με το AI και τη παθολογία η οποία ακούγεται συχνά είναι η εξής "Το AI δεν θα αντικαταστήσει τους παθολόγους. Παρόλα αυτά οι παθολόγοι οι οποίοι δεν γνωρίζουν τίποτα σχετικά το AI θα αντικατασταθούν στο μέλλον".

Ο τομέας ο οποίος θα δει τις μεγαλύτερες αλλαγές είναι οι διεργασίες οι οποίες πραγματοποιούνται στα εργαστήρια καθώς μπορούν να υποστηρίξουν τις περισσότερες αλλαγές σε αυτοματισμούς οι οποίοι θα έχουν μέσα τους αλγορίθμους AI. Βέβαια αυτές οι αλλαγές θα πραγματοποιηθούν στη καλύτερη σταδιακά, καθώς από τη φύση τους οι πωλητές οργάνων στην υγειονομική περίθαλψη είναι ιδιαίτερα προσεκτικοί, με αποτέλεσμα να μην αποδέχονται τις παραμικρές αλλαγές με ευκολία. Επιπρόσθετα υπάρχουν διάφοροι φορείς οι οποίοι επιβλέπουν για τη σωστή και σταδιακή ενσωμάτωση του AI όπως είναι ο οργανισμών τροφίμων και φαρμάκων των ΗΠΑ. Εργασίες μεγάλης έντασης και χωρίς αποζημίωση σε εργασιακές ροές εργαστηρίων είναι από τους πρώτους τομείς οι οποίοι θα εφαρμόσουν αυτοματισμούς AI για τη συνολική βελτίωση της αποτελεσματικότητας τους. Επιπρόσθετα θα μπορούσε κάποιος να προβλέψει νέες ευκαιρίες για δουλειές στο τομέα του AI στη παθολογία, ώστε να δημιουργήσει, εκδώσει, και να συντηρήσει τις αυτοματοποιημένες AI διεργασίες στο μέλλον, κάτι για το οποίο δεν γνωρίζουμε πολλά ακόμα. Συνεπώς το επάγγελμα της παθολογίας όπως έχουμε προαναφέρει θα πρέπει να είναι ενημερωμένο για τους πιθανούς κινδύνους που συνοδεύουν το AI , καθώς το μόνο που γνωρίζουμε για το AI enabled μέλλον (μέλλον στο οποίο το AI θα είναι βασικό μέρος της ζωής μας) του αυτοματισμού είναι ότι είναι απρόβλεπτο. Οπότε με το να Υιοθετούν μια προληπτική στάση προς αυτές τις τεχνολογικές εξελίξεις στο χώρο του AI, καταφέρνουν οι παθολόγοι να προστατεύονται από τους κινδύνους αυτούς, καθώς ταυτόχρονα μπορούν να επωφεληθούν από τα συμφέροντα του. Εν κατακλείδι η επίγνωση των θεμάτων της ηθικής στον

χώρο του AI μπορεί να “σιγουρέψει” ότι η ισορροπημένη και ενημερωμένη οπτική γωνία των παθολόγων ενσωματώνεται στην ανάπτυξη τεχνολογιών με δυνατότητα AI στο τομέα της παθολογίας.**[20,21,22]**

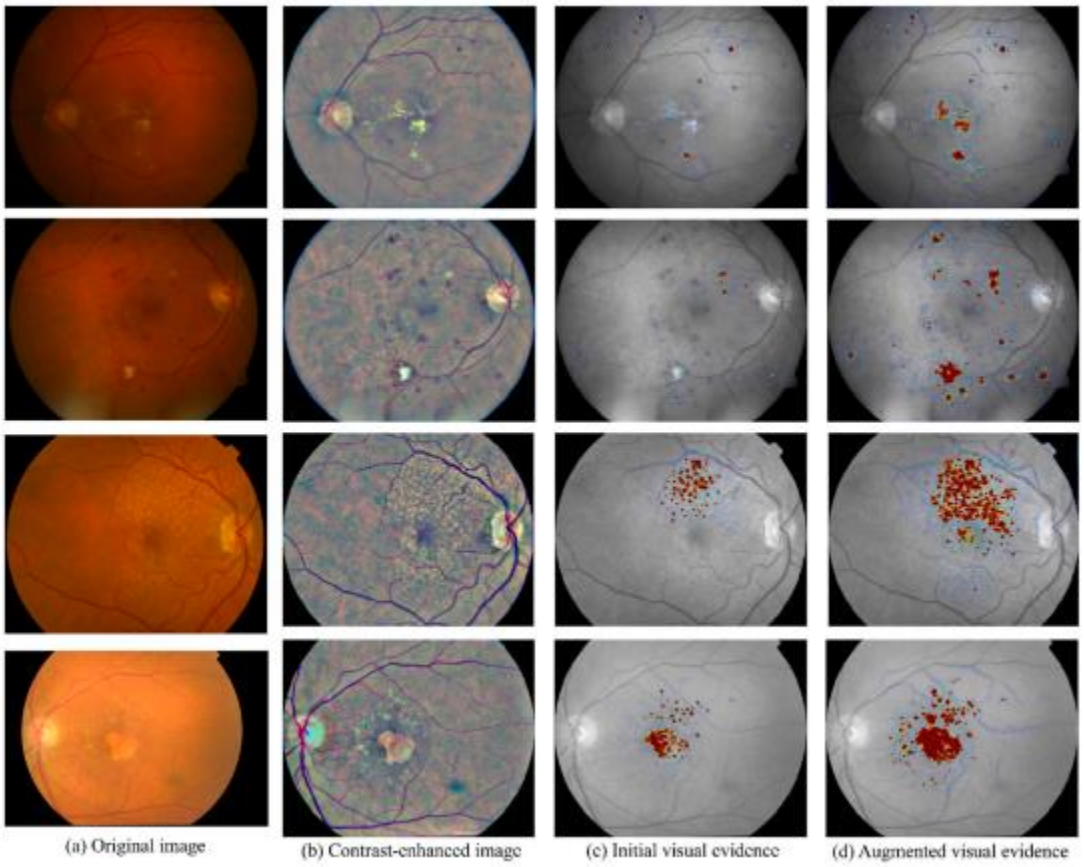
4.2 Ηθική του AI στην ιατρική και οφθαλμολογία

Όπως έχουμε αναφέρει αρκετές φορές, η τεχνητή νοημοσύνη έχει κάνει τεράστια πρόοδο στο τομέα της υγείας όπως και σε όλους τους υπόλοιπους τομείς, κάτι το οποίο τη καθιστά αναπόσπαστο κομμάτι της ζωής μας. Όσο αφορά το τομέα της υγείας ένας υποτομέας ο οποίος έχει δει εξαιρετικά μεγάλη ανάπτυξη σε θέμα AI είναι η οφθαλμολογία, καθώς έχει τεράστιες προοπτικές και ανοίγει πολλές πόρτες για το AI. Πιο συγκεκριμένα η οφθαλμολογία δημιουργεί έναν τεράστιο όγκο εικόνων και δεδομένων που θα μπορούσαν να χρησιμοποιηθούν για την εκπαίδευση της τεχνητής νοημοσύνης, καθιστώντας την μια από τις κορυφαίες εξειδικεύσεις στον τομέα της τεχνητής νοημοσύνης. Παρόλα αυτά, όπως βλέπουμε και στους άλλους υποτομείς της υγείας, αυτή η ραγδαία εξέλιξη της τεχνητής νοημοσύνης συνοδεύεται από πολλές ηθικές ανησυχίες που προκύπτουν από τη χρήση της στην ιατρική πρακτική και ακόμα και στην εκπαίδευση και την έρευνα. Επιπρόσθετα αυτές τις ανησυχίες δεν τις έχουν μόνο οι ασθενείς, καθώς επηρεάζουν μια πληθώρα από κόσμο όπως είναι οι γιατροί, οι κυβερνήσεις, οι ρυθμιστικές αρχές ,τα ασφαλιστήρια, και άλλοι πολλοί φορείς οι οποίοι πρέπει να έχουν στο μυαλό τους το πως λειτουργεί η τεχνητή νοημοσύνη και τι συνέπειες μπορεί να έχει προτού πάρουν οποιαδήποτε απόφαση. Πιο συγκεκριμένα όλα αυτά γίνονται επειδή το AI προσπαθεί να φτάσει να λειτουργεί όπως το ανθρώπινο μυαλό, και να μπορεί να πάρει εντελώς αποφάσεις μόνο του χωρίς να υπάρχει η ανθρώπινη παρέμβαση σε αυτή τη διαδικασία. Προκειμένου τα πιθανά οφέλη της τεχνητής νοημοσύνης στην ιατρική να αντισταθμίσουν τους κινδύνους και να μην υπάρχουν αθέμητα ατυχήματα, η χρήση της πρέπει να ορίζεται σαφώς και να μην υπάρχουν αοριστίες, καθώς ειδικά στο τομέα της υγείας όλες οι απερισκεψίες μπορεί να οδηγήσουν σε ανεπανόρθωτες βλάβες στους ασθενείς. **[9,34]**

Για να μπορέσουμε λοιπόν να χρησιμοποιήσουμε την τεχνητή νοημοσύνη χωρίς να υπάρχουν ιδεασμοί για το αν μας ωφελεί και για το αν πρόκειται να μας βλάψει, πρέπει να βάλουμε κάποια όρια στο τι της επιτρέπουμε να κάνει, και αυτό γίνεται με το να ακολουθήσουμε κάποιες βασικές αρχές. Αρχικά πρέπει να γίνεται σωστή εκπαίδευση του αλγορίθμου AI και για να πραγματοποιηθεί αυτό πρέπει να σωθεί ιδιαίτερη προσοχή σε 2 πράγματα που έχουμε προαναφέρει. Το πρώτο είναι η κατοχή των δεδομένων και ποιος έχει τα δικαιώματα σε αυτά τα δεδομένα. Αυτό το στάδιο στη δημιουργία ενός αλγορίθμου είναι πολύ σημαντικό, καθώς όποιος κατέχει τα δικαιώματα στα δεδομένα έχει πολλές δυνατότητες, και μια από αυτές είναι ακόμα και

το να τα πουλήσει, κάτι το οποίο δημιουργεί πολλές ηθικές επιπλοκές από διάφορες οπτικές γωνίες. Οπότε, πρέπει αν καθιερωθεί από πολύ νωρίς ποιοι θα έχουν δικαιώματα σε αυτά τα δεδομένα. Εκτός από τους ασθενείς οι οποίοι, προφανώς έχουν δικαίωμα στα δεδομένα τους, υπάρχουν διάφοροι φορείς και άνθρωποι οι οποίοι μπορεί να διεκδικήσουν αυτά τα δεδομένα. Αρχικά οι φορείς οι οποίοι παρέχουν στον ασθενή τη περίθαλψη μπορούν να έχουν στη κατοχή τους αυτά τα δεδομένα, δηλαδή οι γιατροί, τα νοσοκομεία, οι κυβερνήσεις κλπ., οι εταιρίες οι οποίες παρέχουν την ασφάλεια έχουν εξίσου δικαίωμα σε αυτά τα δεδομένα, και αντίστοιχα οι προγραμματιστές, σχεδιαστές και παραγωγοί οι οποίοι ευθύνονται για την δημιουργία του αλγορίθμου αυτού, ο οποίος κρατάει τα δεδομένα των ασθενών και τα επεξεργάζεται. Βέβαια εκτός από τη κατοχή των δεδομένων, μια άλλη πολύ σημαντική διαδικασία κατά τη δημιουργία του αλγορίθμου είναι ο τρόπος προστασίας των δεδομένων αυτών. Αν και η συλλογή και η χρήση δεδομένων βρίσκεται στο επίκεντρο των ανησυχιών περί απορρήτου σχετικά με την εισαγωγή της τεχνητής νοημοσύνης, αυτό δεν σημαίνει ότι στο τομέα των ιατρικών δεδομένων πρέπει να τηρείται με αυστηρό τρόπο το απόρρητο των ιατρικών αρχείων. Διότι πρέπει τα δεδομένα των ασθενών να αντιμετωπίζονται με τέτοιο τρόπο ώστε να λαμβάνονται υπόψη και τα άτομα που εμπλέκονται στην όλη διαδικασία της θεραπείας. Βέβαια το πως χειρίζονται τα δεδομένα οι εξωτερικοί παράγοντες πρέπει να συμφωνείτε μεταξύ των ασθενών και αυτών των εξωτερικών παραγόντων. Αναφορικά με τα δεδομένα των ασθενών, εκτίθενται τουλάχιστον δύο φορές, καθώς πρώτα υπάρχει έκθεση όταν καταχωρούνται στα ηλεκτρονικά ιατρικά αρχεία, και έπειτα γίνεται έκθεση τους όταν οι ηλεκτρονικοί ιατρικοί φάκελοι συνδεθούν με τα συστήματα τεχνητής νοημοσύνης. Συνεπώς αν υπάρξει κάποια παράβαση απορρήτου, αυτή μπορεί να οδηγήσει σε πολλές επιπλοκές που μπορεί να βάλουν σε κίνδυνο τον ασθενή, όπως για παράδειγμα συναισθηματικό στρες το οποίο δημιουργήθηκε λόγω της έκθεσης ευαίσθητων δεδομένων υγείας. Εκτός αυτού υπάρχουν και άλλα προβλήματα που μπορεί να προκύψουν όπως είναι οι διακρίσεις ή η στέρηση ασφάλισης, συνέπειες στη ψυχική υγεία όπως αμηχανία, παράνοια, ή ακόμα και ψυχικό πόνο, δεοντολογικές ανησυχίες σχετικά με την ευπάθεια των προσωπικών δεδομένων που οδηγεί στη μείωση της εμπιστοσύνης και ακόμα και ομαδική βλάβη που επηρεάζει πολλά άτομα. Η υπερπροστασία της ιδιωτικής ζωής των ανθρώπων μπορεί να είναι δαπανηρή από άποψη χρόνου και πόρων, επιβραδύνοντας την πρόοδο της τεχνολογίας και της βιομηχανίας μεγάλων δεδομένων, κάτι το οποίο μακροπρόθεσμα μπορεί να καταλήξει να μην αξίζει τόσο. Η εμπιστοσύνη του κοινού στα συστήματα AI έχει πληγεί από την προηγούμενη και πιθανή μελλοντική κατάχρηση των προτύπων απορρήτου από τις εταιρείες για τη διατήρηση της εμπιστευτικότητας και την προστασία των ιδιοκτησιακών δικαιωμάτων. **[9]**

Οι κλινικές δοκιμές ενθαρρύνονται να απαιτούν τη διαθεσιμότητα δεδομένων και τη διαφάνεια στα αποτελέσματα των δοκιμών για να τονωθεί η εμπιστοσύνη του κοινού στην έρευνα τεχνητής νοημοσύνης. Επιπρόσθετα οι εθελοντές σε ένα κλινικό πείραμα θα πρέπει να έχουν την επιλογή να αποφασίσουν εάν τα δεδομένα τους θα κοινοποιηθούν ή όχι καθώς για να δημιουργηθεί μια σωστή σχέση εμπιστοσύνης μεταξύ κλινικών δοκιμών και ανθρώπου πρέπει να υπάρχει η αίσθηση της επιλογής από τους ασθενείς. **[9]**



Source: Trustworthy AI: Closing the gap between development and integration of AI systems in ophthalmic practice.

4.2.1 Ακρίβεια των συστημάτων που χρησιμοποιούν τεχνητή νοημοσύνη

Μία άλλη πολύ σημαντική αρχή για να γίνεται σωστή η δημιουργία αλγορίθμου AI στο τομέα της οφθαλμολογίας και φαρμακευτικής είναι η ακρίβεια του συστήματος. Αυτό είναι ένα από τα πιο σημαντικά στοιχεία καθώς αν δεν έχει ακρίβεια ο αλγόριθμος, θεωρείται απευθείας κακός, καθώς δεν καταφέρνει τον σκοπό του. Για να θεωρηθεί λοιπόν ακριβές ένα σύστημα πρέπει να έχει την ικανότητα να εκτελεί αξιόπιστα και γρήγορα ένα σύνολο εργασιών αν όχι σε καλύτερο, σε ίδιο βαθμό με τον άνθρωπο. Γενικά πολλές έρευνες έχουν αποδείξει ότι η τεχνητή νοημοσύνη είναι πολύ αξιόπιστη όσο αφορά τη διάγνωση, θεραπεία και πρόγνωση. Βέβαια ένα "αρνητικό" που έχει καθαυτή είναι ότι, τα συστήματα τα οποία έχουν μέσα τους αλγορίθμους AI θεωρούν ότι είναι πάντα αλάνθαστα, κάτι το οποίο δεν συμβαδίζει καθόλου με την λογική της ιατρικής, καθώς στον υγειονομικό τομέα τα πάντα αντιμετωπίζονται ως μία μόνο πιθανή ερμηνεία των δεδομένων.

Οι μελέτες διαγνωστικής ακρίβειας στην ιατρική έρευνα τεχνητής νοημοσύνης είχαν συχνά ανεπαρκείς αναφορές. Πιο συγκεκριμένα, οι περισσότερες έρευνες στον τομέα της οφθαλμολογίας τεχνητής νοημοσύνης βασίζονται στην επικύρωση με τη βοήθεια προσομοιώσεων σε υπολογιστή, σε ιστορικά σύνολα δεδομένων. Βέβαια αυτό δεν βοηθάει πολύ όσο αφορά το θέμα αξιοπιστίας καθώς θα έπρεπε οι δοκιμές να πραγματοποιούνται σε πραγματικά δεδομένα που έχουν αποκτηθεί εκ των προτέρων, διότι με αυτόν τον τρόπο μπορούμε να εξασφαλίσουμε την ορθή λειτουργία της τεχνητής νοημοσύνης. Βέβαια παρά της ικανότητα των συστημάτων να εκπαιδεύονται και να μαθαίνουν να προσαρμόζονται με την πάροδο του χρόνου, αυτό δεν μας εξασφαλίζει ότι τα συστήματα αυτά θα έχουν την ύψιστη δυνατή ακρίβεια. Επιπρόσθετα, ένα άλλο πρόβλημα που αντιμετωπίζεται στο τομέα της φαρμακευτικής και της οφθαλμολογίας είναι η υπερπροσαρμογή των αλγορίθμων, κάτι το οποίο συμβαίνει όταν τα αποτελέσματα ενός αλγορίθμου διογκώνονται λόγω ελαττωμάτων που εμφανίζονται κατά τη σχεδίαση του. Πιο συγκεκριμένα τα τεχνητά νευρωτικά δίκτυα μπορούν να εμφανίσουν υπερπροσαρμογή αν έχουν πάρα πολλούς κόμβους ή αν χρησιμοποιούν πολύ μικρό μέγεθος δειγμάτων για την εκπαίδευση. Επιπλέον, αν κατά τη διάρκεια της εκπαίδευσης υπάρχει θόρυβος σε μορφή ταλαντώσεων, το μοντέλο το παίρνει και αυτό ως δεδομένο και το χρησιμοποιεί για να εκπαιδευτεί. Συνεπώς το αποτέλεσμα που προκύπτει από αυτά είναι το μοντέλο μας να είναι υπερβολικά προσαρμοσμένο στα αρχικά δεδομένα εκπαίδευσης, που δεν μπορεί να εφαρμοστεί σε καινούργια δεδομένα, κάτι το οποίο καθιστά τον αλγόριθμο “άχρηστο”. Οπότε μπορούμε να συμπεράνουμε ότι, όταν ένα μοντέλο έχει υποστεί υπερπροσαρμογή, η ακρίβεια του και η προβλεπόμενη κλινική του απόδοση καταλήγουν να είναι και τα δύο διογκωμένα. **[35, 36,37,38]**

4.2.2 Ηθικά ζητήματα που σχετίζονται με τον ασθενή

Ένα άλλο πολύ σημαντικό θέμα που δημιουργεί αμφιβολίες σχετικά με τη χρήση συστημάτων με AI στη φαρμακευτική και την οφθαλμολογία είναι τα ηθικά ζητήματα που έχουν να κάνουν με τους ασθενείς. Τα ηθικά δικαιώματα των ασθενών είναι κάτι το οποίο πρέπει να το γνωρίζουν όλοι όσοι ασχολούνται με τον τομέα της υγείας, και είναι ένα πολύ σημαντικό κομμάτι της σωστής λειτουργίας του συστήματος υγείας. Τα δικαιώματα αυτά έχουν συνοψιστεί στην αναφορά Belmont, η οποία αναφορά εξετάζει τις ηθικές αρχές στο θέμα της έρευνας σε ανθρώπους. Μέσα στις αρχές αυτές είναι η ευεργεσία, η μη κακοήθεια, η αυτονομία και η δικαιοσύνη, αρχές οι οποίες θεωρούνται κατευθυντήριες όσο αφορά το θέμα της ηθικής. Αυτές οι προστασίες αποτελούν τη βάση της ηθικής της τεχνητής νοημοσύνης που σχετίζεται με τους ασθενείς, καθώς χωρίς αυτές ο κάθε φορέας θα έκανε ότι ήθελα όσο αφορά τα προσωπικά δεδομένα των ασθενών, χωρίς να λαμβάνουν υπόψη τις συνέπειες που θα είχαν αυτές οι πράξεις. Οι ασθενείς μπορούν να κάνουν χρήση των δικαιωμάτων τους είτε φανερά, μέσω διαδικασιών

όπως η παροχή ενημερωμένης συγκατάθεσής τους, είτε κρυφά, μέσω διασφαλίσεων όπως αυτές που προβλέπονται από τη νομοθεσία.

Η έννοια της «ενημερωμένης συναίνεσης» πηγάζει από την έννοια της «αυτονομίας», μία έννοια η οποία είναι ύψιστης σημασίας στο τομέα της υγείας. Η διαδικασία μέσω της οποίας οι αλγόριθμοι καταλήγουν σε διαγνωστικά ή θεραπευτικά συμπεράσματα θα πρέπει να περιγράφει λεπτομερώς προκειμένου να αποκτηθεί πλήρης ή μερική εξουσιοδότηση για τη χρήση τους σε υπηρεσίες υγειονομικής περίθαλψης, και γι' αυτό είναι η ενημερωμένη συγκατάθεση. Καθώς όσο αφορά τη τεχνητή νοημοσύνη στο τομέα της υγείας, θεωρείται από πολλή κόσμο κάτι το οποίο είναι μη αξιόπιστο, και συνεπώς δεν θέλουν να θέσουν τον εαυτό τους σε κίνδυνο, οπότε η συναίνεση είναι απαραίτητη σε αυτόν τον τομέα. Επιπρόσθετα όσο αφορά τη συναίνεση, είναι ευθύνη των κλινικών γιατρών να εκπαιδεύουν τους ασθενείς σχετικά με αυτές τις διαδικασίες. Οι ασθενείς θα πρέπει να έχουν τη δυνατότητα να συναινέσουν στη συλλογή, αποθήκευση και διάδοση των δεδομένων τους. Εάν ένας ασθενής επιλέξει να μην χρησιμοποιούνται τα δεδομένα του ηλεκτρονικού μητρώου υγείας του σε αλγόριθμους τεχνητής νοημοσύνης, τότε οι αλγόριθμοι δεν θα πρέπει να έχουν πρόσβαση στα δεδομένα τους με κανέναν τρόπο. Η κατανόηση των πιθανών εφαρμογών των δεδομένων από τους ασθενείς είναι απαραίτητη για τη λήψη έγκυρης ενημερωμένης άδειας. Βέβαια όπως προαναφέραμε προηγουμένως, η μη συναίνεση των ασθενών στη χρήση των δεδομένων τους δεν είναι καλή όσο αφορά την εξέλιξη και βελτίωση της τεχνολογίας, καθώς πρέπει τα συστήματα να εκπαιδεύονται συνεχώς σε αληθινά δεδομένα και όχι σε προσομοιώσεις, καθώς πολλές φορές δημιουργούνται προκαταλήψεις αν γίνεται χρήση αρκετών και αντίστοιχα ρεαλιστικών δεδομένων.

Ένα άλλο ενδιαφέρον θέμα είναι η διατήρηση της ιδιωτικής ζωής των ασθενών, κάτι το οποίο είναι νόμος και καθήκον από οποιονδήποτε εργάζεται στον τομέα της υγείας. Είναι υποχρέωση των ατόμων που έχουν πρόσβαση στα αρχεία με τα δεδομένα των ασθενών να κρατάνε ασφαλή αυτά τα αρχεία, διότι αυτά τα αρχεία περιέχουν τη χρήση, την αποθήκευση, τη πρόσβαση και τη μετάδοση των δεδομένων τους. Όταν γίνεται χρήση τεχνητής νοημοσύνης, αυτό συνεπάγεται ότι υπάρχει εμπιστοσύνη ως προς τον αλγόριθμο, και μια άνεση όσο αφορά την κατάθεση των πληροφοριών και των δεδομένων σε αυτόν καθ’όλη τη διαδικασία της εκπαίδευσης, της επικύρωσης και της αποδοχής των εξόδων του συστήματος. Το θέμα της εμπιστευτικότητας πρέπει να αναπροσδιοριστεί σε αυτήν την καινούργια γενιά που βρισκόμαστε, καθώς οι μηχανές είναι ένα αναπόσπαστο κομμάτι της ζωής μας και πρέπει να μάθει ο κόσμος να εμπιστεύεται τα συστήματα AI, καθώς κάνουν τη ζωή μας πολύ πιο εύκολη σε όλους τους τομείς. Βέβαια πολλής κόσμος δεν μπορεί να κάνει εύκολα αυτό το “άλμα” στο θέμα της εμπιστοσύνης καθώς υπάρχουν πολλά πράγματα που απειλούν να βλάψουν το απόρρητο όπως είναι η ακατάλληλη χρήση συνόλων δεδομένων, η ανήθικη αποκάλυψη και τα όρια στις διαδικασίες απόταυτόποίησης δεδομένων. Ένα εξίσου σημαντικό θέμα σχετικά με ην έλλειψη εμπιστευτικότητας είναι η διαφορά μεταξύ των εθνών, καθώς οι οικονομικές, πολιτισμικών και κοινωνικές διαφορές επηρεάζουν άμεσα το πως βλέπει κανείς τα συστήματα AI και το αν θα τα εμπιστευόταν για την υγεία του. Κάτι το οποίο δημιουργεί παραπάνω προβλήματα και αναγκάζει τους ερευνητές να

ψάχνουν για παραπάνω λύσεις όσο αφορά τους τρόπους με τους οποίους μπορούν να αυξήσουν την εμπιστευτικότητα του κόσμου. **[39,40]**

4.2.3 Ηθικά ζητήματα που σχετίζονται με τους γιατρούς

Ένα άλλο θέμα το οποίο εμπεριέχει πολλά ηθικά ζητήματα και εδώ αντίστοιχα είναι η ηθική τριγύρω από τους ιατρούς οι οποίοι παίζουν πρωταγωνιστικό ρόλο στη θεραπεία των ασθενών. Κάθε καινούργια τεχνολογία η οποία έρχεται στην αγορά απαιτεί από τον κόσμο να αφιερώσει χρόνο για να την μάθει και πολλές φορές καταλήγει να είναι τόσο χρήσιμη που ο κόσμος εξαρτάται από αυτήν. Στη περίπτωση της ιατρικής, είναι δύσκολο οι ιατροί να εμπιστευτούν και να βασίζονται πάρα πολύ την τεχνητή νοημοσύνη καθώς ο κόσμος δεν είναι ακόμα προετοιμασμένος να την υποδεχτεί ακόμα, και δεν είναι ακόμα αρκετά αξιόπιστη ώστε να είναι ασφαλής για χρήση. Κατά συνέπεια η χρήση της τεχνητής νοημοσύνης μπορεί να οδηγήσει στην μείωση της αποτελεσματικότητας στη λήψη αποφάσεων των ιατρών, καθώς ο αλγόριθμος θα παίρνει αποφάσεις στη θέση τους. Επιπρόσθετα, η "προκατάληψη αυτοματισμού" ή η υπερβολή εξάρτηση από την τεχνητή νοημοσύνη είναι ένα παροδικό ζήτημα το οποίο το έχουν πιθανότατα οι περισσότεροι ιατροί, καθώς με την υπερβολική χρήση τεχνητής νοημοσύνης είναι αναπόφευκτο να μειωθεί η ικανότητα τους να πραγματοποιούν τα καθήκοντα που κάνουν πλέον σε καθημερινή βάση όπως είναι η διάγνωση και η θεραπεία. Συνεπώς αυτά τα ζητήματα προκαλούν ένα άγχος μεταξύ των ιατρών, καθώς αισθάνονται μειωμένοι στις ικανότητες τους, διότι η τεχνητή νοημοσύνη και είναι πιο γρήγορη αλλά και μπορεί να κάνει πολλά πράγματα ταυτόχρονα. Συνεπώς όπως είχαμε αναφέρει στην ενότητα της ηθική στη παθολογία, οι αλγόριθμοι τεχνητής νοημοσύνης μπορεί να αντικαταστήσουν πολλούς ιατρούς οι οποίοι δεν έχουν σκοπό να μάθουν πως να την χειρίζονται αποτελεσματικά. Κάτι αντίστοιχο υπάρχει ως ανησυχία και στο τομέα της οφθαλμολογίας και φαρμακευτικής. **[41]**

Συνεπώς ένα μεγάλο θέμα που έρχεται στην επιφάνεια είναι το θέμα της εμπιστοσύνης, καθώς ο τομέας της ιατρικής αναπτύσσεται εδώ και εκατοντάδες χρόνια με αποτέλεσμα να υπάρχει αρκετή εμπιστοσύνη προς τους ιατρούς. Κάτι το οποίο έχει ως θεμέλιο του την ανθρώπινη συναισθηματική ανταπόκριση, διότι κατανοώντας τον ασθενή πραγματοποιείται πιο σωστά η θεραπεία και δημιουργείται ένα πιο δυνατό αίσθημα εμπιστοσύνης μεταξύ ασθενή και ιατρού. Βέβαια, με την ένταξη του AI στην οφθαλμολογία και την φαρμακευτική φαίνεται πως αυτό το αίσθημα εμπιστοσύνης που υπάρχει μεταξύ ιατρού και ασθενή, για το οποίο έχουν περάσει χρόνια για να δημιουργηθεί, μπορεί να φτιαχτεί σε πολύ μικρότερο χρόνο για τα συστήματα AI, κάτι το οποίο ακούγεται πολύ καλό και ταυτόχρονα ανησυχητικό. Ωστόσο, θα χρειαστεί να περάσει ένα χρονικό διάστημα μέχρι να επιτευχθεί η συναίνεση μεταξύ του γενικού πληθυσμού, κάτι το οποίο δεν είναι εύκολο να εκπληρωθεί, καθώς μεταξύ του γενικού πληθυσμού υπάρχουν και πολλοί ηλικιωμένοι οι οποίοι εκ φύσεως τους δεν δέχονται εύκολα τις αλλαγές, πόσο μάλλον για κάτι το

οποίο είναι πολύ δύσκολο να καταλάβουν. Η αίσθηση της εμπιστοσύνης μεταξύ ασθενή και μηχανών είναι ύψιστης σημασίας για την ορθή υγειονομική περίθαλψη, διότι οι ασθενείς έχουν την ανάγκη να νιώθουν ασφάλεια για να είναι καλά ψυχολογικά και σωματικά, και αν δεν υπάρχουν γερά θεμέλια εμπιστοσύνης όπως υπάρχουν μεταξύ ανθρώπων και των επαγγελματιών ιατρών του χώρου, τότε ασθενείς σε πολλές περιπτώσεις θα προτιμήσουν να βρουν μια διαφορετική εναλλακτική παρά να διαφύγουν στις μηχανές. Για να διασφαλίσουμε ότι οι άνθρωποι μπορεί να αισθάνονται ασφαλείς επενδύοντας στην τεχνητή νοημοσύνη, χρειαζόμαστε ισχυρή ηθική επίβλεψη και τακτικούς ελέγχους της κλινικής ασφάλειας και απόδοσης της τεχνολογίας, διότι αν υπάρξει κάποιο πρόβλημα με το σύστημα AI και δεν βρεθεί εγκαίρως, αρχικά θα χαθεί η εμπιστοσύνη που υπάρχει μεταξύ ανθρώπου και μηχανής, και κυριότερο θα υπάρξουν επιπλοκές σε θέματα υγείας που μπορεί να είναι και ανεπανόρθωτα σε ορισμένες περιπτώσεις. Όταν μιλάμε για εμπιστοσύνη τρείς έννοιες έρχονται στο μυαλό, η ικανότητα, το κίνητρο και η διαφάνεια, οι οποίες έννοιες είναι οι πυλώνες στους οποίους στηρίζεται η εμπιστοσύνη. Στον τομέα της περίθαλψης ασθενών όταν λέμε ότι κάποιος είναι ικανός εννοούμε να διαπρέπει στα καθήκοντα του, δηλαδή να κάνει σωστά τη δουλειά του. Η ικανότητα των γιατρών βέβαια ακόμα και αν είναι άριστη μπορεί να βελτιωθεί ακόμα περαιτέρω με τη βοήθεια της τεχνητής νοημοσύνης, καθώς τους παρέχει και τους ίδιους και τους ασθενείς πολλά εργαλεία που κάνουν τη ζωή τους πιο εύκολη, όπως για παράδειγμα στους ασθενείς τους δίνεται ευκολότερη πρόσβαση στα δεδομένα τους και τις υπηρεσίες υγείας τους καθώς ταυτόχρονα τους κάνει μετόχους στην οικονομική αξία των δεδομένων τους. Βέβαια συστήματα τεχνητής νοημοσύνης τα οποία χαρακτηρίζονται από μεροληψία, ανακρίβεια, αναποτελεσματικότητα και τη μη δυνατότητα εξήγησης, απειλούν άμεσα την εμπιστοσύνη καθώς επηρεάζει την αυτονομία που προσπαθούν να δημιουργήσουν οι ασθενείς, κάτι το οποίο υπονομεύει την εμπιστοσύνη απέναντι σε όλα τα συστήματα AI, διότι ο κόσμος έχει την τάση να γενικεύει τα πράγματα με αποτέλεσμα να επηρεάζονται και άλλοι τομείς που κάνεις χρήση συστημάτων AI. Όταν αναφερόμαστε αντίστοιχα στον όρο “κίνητρο”, αναφερόμαστε στον τρόπο σκέψης που έχουν οι γιατροί, και αυτός είναι έχοντας τον ασθενή ως προτεραιότητα. Η τεχνητή νοημοσύνη μπορεί να βοηθήσει απίστευτα τους γιατρού σε πολλούς τομείς, αρκεί βέβαια να τους δώσει τον απαραίτητο χρόνο που θέλουν ώστε να μπορέσουν να αναπτύξουν πιο προσωπικές σχέσεις με τους ασθενείς, με αποτέλεσμα να υπάρχει ακόμα πιο μεγάλο αίσθημα εμπιστοσύνης. Βέβαια από την άλλη αν η τεχνητή νοημοσύνη έχει ως σκοπό μόνο τη βελτίωση της ροής εργασίας και την προσθήκη περισσότερων ασθενών, κάτι το οποίο θα ασκήσει περισσότερη πίεση στους γιατρούς, και θα τους οδηγήσει να ασχολούνται με πιο απαιτητικά θέματα, τότε θα βλάψει την εμπιστοσύνη καθώς δεν θα υπάρχει διαθεσιμότητα για τα πιο απλά θέματα. Τέλος με την αυξημένη διαφάνεια, οι άνθρωποι έχουν την δυνατότητα να διακινούν πιο εύκολα τα δεδομένα τους και τις πληροφορίες για την υγεία τους, κάτι το οποίο βοηθάει στην επικοινωνία πολλές φορές μεταξύ ιατρικών φορέων. Η ειλικρίνεια είναι βασικός παράγοντας στη συνεργασία μεταξύ ασθενή και γιατρού, καθώς το να θέτει κανείς τις αμφιβολίες του για κάποιο θέμα κάνει τη διαδικασία πολύ πιο εύκολη, είτε είναι διάγνωση είτε είναι θεραπεία. Οπότε το να υπάρχει ειλικρίνεια και διαφάνεια βοηθάει στην αποφυγή μηχανών μαύρου κουτιού όπου δεν γνωρίζουμε τίποτα για το σύστημα, ούτε πως

δουλεύει ούτε γιατί παράγονται οι αντίστοιχες έξοδοι. **[41,42]**

Επιπρόσθετα, είναι γνωστό ότι πολλοί γιατροί έχουν έλλειψη ενσυναίσθησης, ένα συναίσθημα και ικανότητα η οποία είναι πολύ χρήσιμη στο να καταλαβαίνουν τον ασθενή. Οπότε είναι πολύ σημαντικό να υπάρξει περισσότερη προσπάθεια για την ενσωμάτωση της εκπαίδευσης και της διδασκαλίας της ενσυναίσθησης στο ιατρικό πρόγραμμα σπουδών, κάτι που μέχρι στιγμής δεν έχει επιτευχθεί και έχει οδηγήσει στο αποτέλεσμα πολλοί γιατροί να καταλήγουν να είναι το ακριβώς αντίθετο, δηλαδή να είναι ψυχροί. Επομένως, όταν η τεχνητή νοημοσύνη αναλάβει τις καθημερινές δουλειές που κάνουν οι γιατροί στις μέρες μας, αυτομάτως ελευθερώνει πολύ χρόνο στους γιατρούς αυτούς, τον οποίο μπορούν να αξιοποιήσουν ώστε να επικεντρωθούν πιο πολύ στην ενσυναίσθησης κατά την διάρκεια των αλληλεπιδράσεων τους με τους ασθενείς. Είναι γνωστό σε όλους ότι οι ασθενείς προτιμούν την ανθρώπινη αλληλεπίδραση με τους γιατρούς τους και τους θεωρούν πιο συμπαθητικούς από τη τεχνολογία, κάτι το οποίο είναι λογικό καθώς με τους ανθρώπους μπορούν να επικοινωνήσουν και να πάρουν άμεση ανατροφοδότηση για ότι τους απασχολεί, κάτι το οποίο δεν είναι εύκολο με την τεχνολογία καθώς πρέπει να υπάρξει ένα στάδιο επεξεργασίας δεδομένων προτού απαντηθούν τυχόν ερωτήματα. Συνεπώς αφού υπάρχει ήδη αυτή η βάση για την σωστή επικοινωνία μεταξύ ασθενών και γιατρών, είναι ευκαιρία για τους γιατρούς να αξιοποιήσουν την τεχνητή νοημοσύνη για να δημιουργήσουν περισσότερο χρόνο, ώστε να βελτιώσουν την ικανότητα τους να είναι πιο συμπονετικοί. Ωστόσο οι ασθενείς παρέχουν στον αλγόριθμο τα δεδομένα τους, τα οποία χρησιμοποιούνται για το στάδιο της εκπαίδευσης, της δημιουργίας και επικύρωσης των αλγορίθμων, οι οποίοι αλγόριθμοι με τη σειρά τους παρέχουν αποφάσεις για την υγεία αυτών των ασθενών, κάτι το οποίο απαιτεί να υπάρχει μια ηθική και συμπονετική βαθμονόμηση αυτής της νέας μορφής δέσμευσης. Όσο αφορά την ενσυναίσθησης, τα συστήματα τεχνητής νοημοσύνης μπορούν θεωρητικά να αναπτύξουν τρόπους με τους οποίους μπορούν να είναι πιο ενσυναίσθητα, αλλά αυτό προαπαιτεί να εκτεθούν και να εκπαιδευτούν σε χιλιάδες περιπτώσεις ασθενών, ώστε να μπορούν να καλυφθούν κατά ένα μεγάλο ποσοστό όλα τα πιθανά αυτά σενάρια. Κάτι το οποίο μπορεί να είναι πολύ χρονοβόρο αλλά τα αποτελέσματα που μπορεί να αποφέρει αξίζουν τον κόπο. Καθώς με αυτόν τον τρόπο μπορεί να αναπτυχθεί περαιτέρω η "τεχνητή στοργή" η οποία αναπτύχθηκε από την τεχνητή νοημοσύνη και είναι η ικανότητα στον αλγόριθμο AI να μπορεί να αντιληφθεί τον πόνο. Συνεπώς η ανάπτυξη αυτής της ικανότητας σχετικά με τον πόνο είναι ένα μεγάλο βήμα της τεχνητής νοημοσύνης προς τη σωστή κατεύθυνση, επειδή πρέπει οι μηχανές να είναι ικανές να δείξουν περισσότερη ανθρωπιά αν θέλουμε να τις εμπιστευτεί κιόλας ο κόσμος. **[43]**

4.2.4 Ευθύνη και Υπαιτιότητα

Η τεχνητή νοημοσύνη είναι ένα εργαλείο το οποίο μας προσφέρει απίστευτες δυνατότητες στο τομέα της υγείας. Βέβαια υπάρχουν πολλές περιπτώσεις που παρόλο που είναι δυνατή η χρήση τεχνητής νοημοσύνης, δεν πραγματοποιείται με αποτέλεσμα να μην υπάρχει κάποια πρόοδο στις

διαγνώσεις και στις θεραπείες. Αυτό γίνεται διότι ένα εμπόδιο στην εφαρμογή της τεχνητής νοημοσύνης είναι η δυνατότητα νομικής ευθύνης για δυσμενείς επιπτώσεις που προκύπτουν από τη χρήση της στη διαδικασία της διάγνωσης και της θεραπείας. Συνεπώς για να μπορέσει να γίνει επέκταση και χρήση αλγορίθμων AI σε περισσότερες υγειονομικές μονάδες και να γίνεται πιο σωστή η περίθαλψη των ασθενών, προαπαιτείτε να πραγματοποιηθεί η επίλυση του προβλήματος της ευθύνης, κάτι το οποίο είναι ζωτικής σημασίας για να γίνουν τα προαναφερθέντα. Η τεχνητή νοημοσύνη από μόνη της μας παρέχει μυριάδες πιθανές χρήσεις, κάτι το οποίο εγείρει πολλές ανησυχίες σχετικά με την ασφάλεια της και τις κακές πρακτικές της. Στους νόμους περί ιατρικής ευθύνης, η απόδοση του κλινικού γιατρού κρίνεται συνήθως με βάση το υψηλότερο κλινικό πρότυπο, κάτι το οποίο δημιουργεί ένα αίσθημα ανταγωνισμού, και συνεχής βελτίωσης μεταξύ των γιατρών. Πάνω σε αυτό το θέμα, έχει προταθεί ότι η τεχνητή νοημοσύνη στην ιατρική, πάει να θέσει ένα τέτοιο αντίστοιχο πρότυπο το οποίο θα προσπαθούν να φτάσουν οι κλινικοί γιατροί. Συνεπώς για να πραγματοποιηθεί αυτό πρέπει να τεθούν νόμοι περί ευθύνης οι οποίοι θα τηρούνται από τα άτομα που χρησιμοποιούν την τεχνολογία AI. Κατά τη διαδικασία δημιουργίας κανόνες ευθύνης για το AI πρέπει να ληφθούν υπόψη 2 παράγοντες. Αρχικά με την πάροδο του χρόνου, το αποδεκτό επίπεδο ιατρικής περίθαλψης εξελίσσεται, κάτι το οποίο συνεπάγεται ότι θα πρέπει οι αλγόριθμοι να μπορούν να ακολουθήσουν τις αλλαγές αυτές. Δεύτερον τα εργαλεία τα οποία χρησιμοποιούνται στην ιατρική τεχνητή νοημοσύνη εξελίσσονται με γρήγορους ρυθμούς, κάτι το οποίο συνεπάγεται την συνεχή αναβάθμιση τους για να μπορεί να παρέχεται η καλύτερη πιθανή διάγνωση και θεραπεία στους ασθενείς. Το οποίο από μόνο του δεν είναι εύκολο καθώς οι αναβαθμίσεις στον τομέα της υγείας καταλήγουν να είναι πολλές φορές κοστοβόρες. Διότι η αποθήκευση και επιμέλεια δεδομένων, μαζί με την αναβάθμιση των μοντέλων αποτελούν πολύ σημαντικά κόστη, κάτι το οποίο σε μικρό χρονικό διάστημα μπορεί να μην αξίζει την "αναβάθμιση" αν λάβουμε υπόψιν το κόστος. Παρόλα αυτά, αυτοί οι δύο παράγοντες φέρνουν στην επιφάνεια δύο πιθανά ενδεχόμενα σε περίπτωση βλάβης. Πρώτον αν ο ασθενής υποστεί κάποια βλάβη με τον οποιονδήποτε τρόπο κατά τη θεραπεία του με συστήματα τεχνητής νοημοσύνης, ο υπεύθυνος για αυτό θα είναι ο γιατρός ο οποίος δεν μπόρεσε να αποτρέψει τη βλάβη αυτή. Και αντίστοιχα οι γιατροί είναι υπεύθυνοι για τη μη τήρηση των συστάσεων AI αν οι συστάσεις αυτές γίνουν το πρότυπο της φροντίδας του ασθενή. **[44]**

Όσο πιο αυτόνομο και χωρίς επίβλεψη είναι το ευφυές σύστημα, τόσο περισσότερη βελτίωση των κανόνων ευθύνης προαπαιτείτε για να δουλέψει σωστά και να διασφαλιστεί η ασφάλεια των ασθενών. Καθώς όσο πιο αυτοματοποιημένη γίνεται η λειτουργία τόσο λιγότερο έλεγχο έχουν οι γιατροί, συνεπώς στη περίπτωση σφάλματος δεν είναι εύκολο να γίνει η εντοπισμός του λάθους. Προς το παρόν οι ισχύουσες ιατρικές νομοθεσίες αντιμετωπίζουν ζητήματα ευθύνης με διάφορους τρόπους. Αρχικά γίνεται με αποζημίωση σε περίπτωση που η βλάβη δεν μπορεί να διορθωθεί στα άτομα τα οποία υπέστησαν τις βλάβες από τους αντίστοιχες υπαίτιους. Και ένας άλλος τρόπος είναι η επιβολή ευθύνης για αποκατάσταση του τραυματισμού εάν είναι ανακτήσιμος. Όσο αφορά την απόδοση της ευθύνης σε μια τέτοια περίπτωση, νομικά μπορεί να γίνεται η απόδοση της σε παραπάνω από 1 άτομα, και πιο συγκεκριμένα σε περίπτωση βλάβης είναι σχεδόν σίγουρο ότι η ευθύνη θα μοιραστεί σε πολλά άτομα. Τα άτομα αυτά

συμπεριλαμβάνονται από γιατρούς οι οποίοι μπορεί να έκαναν κάποιο λάθος σε μια από τις διαδικασίες της θεραπείας, οργανισμούς υγείας, και κατασκευαστικές εταιρίες σε περίπτωση βλάβης κάποιου εργαλείου. Συνεπώς για να αποδίδεται σωστά η δικαιοσύνη σε περίπτωση ατυχήματος, υπάρχουν κάποιες προτάσεις που θα μπορούσαν να ακολουθηθούν. Θα μπορούσε να υποβληθεί ευθύνη σε τρίτους όπως είναι οι κατασκευαστές, επενδυτές και ασφαλιστές για τα ιατρικά συστήματα τεχνητής νοημοσύνης. Καθώς είναι πολύ σημαντικοί παράγοντες στη δημιουργία του συστήματος, και αν το σύστημα αυτό προκαλέσει κακό ή δυσλειτουργήσει με οποιονδήποτε τρόπο φταίνε και αυτοί. Μια άλλη λύση για την απόδοση της δικαιοσύνης είναι, τα συστήματα τα οποία συμπεριφέρονται αυτόνομα χωρίς την ανθρώπινη παρέμβαση ή έλεγχο, να θεωρούνται υπεύθυνα σε περίπτωση βλάβης, καθώς μέχρι τώρα υπαίτιος θεωρούταν ο γιατρός, κάτι το οποίο είναι ηθικά λάθος καθώς δεν έχει κανέναν έλεγχο σε αυτά τα συστήματα. Συνεπώς ένας άλλος τρόπος απόδοσης δικαιοσύνης είναι να θεωρούνται υπεύθυνοι οι σχεδιαστές των μοντέλων AI για οποιαδήποτε βλάβη προκαλέσουν τα μηχανήματα τους, διότι θα έπρεπε να έφτιαχναν πιο ορθά τον αλγόριθμο και με τα απαραίτητα πρότυπα ώστε να αποφευχθεί το οποιοδήποτε ατύχημα. Είναι απαραίτητο να ελέγχονται και να αξιολογούνται με βάση αυτές τις συστάσεις τα συστήματα τεχνητής νοημοσύνης πριν την εφαρμογή τους σε ιατρικό περιβάλλον, καθώς από τη στιγμή που τίθενται σε λειτουργία, αν είναι με τον οποιονδήποτε τρόπο ελαττωματικά θέτουν πολλές ζωές σε κίνδυνο, κάτι το οποίο δεν είναι ηθικά αποδεκτό. **[45]**

Η ηθική της τεχνητής νοημοσύνης όπως έχουμε προσέξει, είναι ένα ζήτημα το οποίο απασχολεί πολλή κόσμο στον τομέα της υγειονομικής περίθαλψης. Το οποίο είναι λογικό διότι δεν είναι εύκολο να εμπιστευτεί κανείς τη ζωή του σε ένα μηχάνημα χωρίς να γνωρίζει με σιγουριά ότι δεν θα διακινδυνεύσει. Οπότε είναι λογικό να υπάρχουν πολλοί άνθρωποι οι οποίοι επενδύουν σε αυτόν τον τομέα, με τη προοπτική ότι θα έχουμε μια καλύτερη και πιο εύκολη θεραπεία στις αρρώστιες. Η ηθική πλευρά λοιπόν της τεχνητής νοημοσύνης έχει συζητηθεί πολύ όσο αφορά την ιατρική και την οφθαλμολογία πιο συγκεκριμένα, και υπάρχει πολύ βιβλιογραφία πάνω σε αυτό το θέμα, όπου πραγματεύονται την χρήση της τεχνητής νοημοσύνης στην ιατρική εκπαίδευση, την πρακτική και την έρευνα. Η ηθική της εκπαίδευση των μηχανών, η δεοντολογία της ακρίβειας των μηχανών, η ηθική που σχετίζεται με τον ασθενή, η ηθική των γιατρών και η κοινή ηθική είναι όλα ύψιστης σημασίας στο τομέα της υγείας και είναι απαραίτητο να λαμβάνονται υπόψη κατά τη δημιουργία AI αλγορίθμου, καθώς αυτά χτίζουν τις βάσεις στις οποίες δουλεύει ο αλγόριθμος, και ειδικά στον τομέα της υγείας τα περιθώρια να πραγματοποιηθούν λάθη ή προκαταλήψεις είναι πολύ μικρά έως και μηδαμινά. Υπάρχουν πάρα πολλά βιβλία και άρθρα τα οποία μιλάνε για την ανάγκη της τεχνητής νοημοσύνης στον τομέα της υγείας, τα οποία βιβλία απαντάνε και σε πολλά ηθικά ερωτήματα που υπάρχουν στον κόσμο. Βέβαια όλα αυτά τα βιβλία συλλογικά κατάφεραν να δείξουν την ανάγκη που υπάρχει για τυποποίηση ηθικών και ρυθμιστικών θεσμών, ώστε να μπορέσουμε να διασφαλίσουμε τη σωστή και ασφαλή λειτουργία των συστημάτων τεχνητής νοημοσύνης. **[9]**

4.3 Ηθική της Τεχνητής Νοημοσύνης στη Ψυχιατρική

Πλέον η Τεχνητή Νοημοσύνη είναι ένα βασικό εργαλείο που χρησιμοποιείται σε πολλούς τομείς της υγειονομικής περίθαλψης. Ένας από αυτούς τους τομείς ο οποίος είναι αξιοθαύμαστο να αναφερθεί είναι η ψυχιατρική και πιο συγκεκριμένα η υπολογιστική ψυχιατρική , η οποία χρησιμοποιεί μεθόδους όπως είναι το Deep Learning και το Bayesian modelling τα οποία είναι προεκτάσεις της Τεχνητής Νοημοσύνης και βρίσκονται πρακτικά στην ίδια κατηγορία με αυτήν. Αν και οι μέθοδοι οι οποίες χρησιμοποιούνται στην υπολογιστική ψυχιατρική έχουν διαφορετικούς σκοπούς, εγείρουν παρόμοια ηθικά ζητήματα με αυτά που θέτει και το AI. Λόγου χάρη αν τα δεδομένα εκπαίδευσης έχουν προκαταλήψεις μεταξύ τους όπως είναι οι φυλετικές μπορεί να γίνει λανθασμένη εκπαίδευση και ο αλγόριθμος να έχει ηθικά σφάλματα, κάτι το οποίο έχει προαναφερθεί και στο θέμα της ηθικής στη παθολογία. Ένα άλλο εξίσου σημαντικό θέμα είναι αυτό της ιδιοκτησίας και προστασίας των προσωπικών δεδομένων που χρησιμοποιούνται σε κάθε έρευνα. Επιπρόσθετα πολλές φορές είναι δύσκολο να εξηγηθούν οι υλοποιήσεις της τεχνητής νοημοσύνης, καθώς είναι δύσκολο έως αδύνατον να ερμηνευτεί ο τρόπος λειτουργίας ενός συστήματος, πόσο μάλλον το γιατί αποδόθηκαν τα εκάστοτε αποτελέσματα ή το ποιος είναι υπεύθυνος για τον αλγόριθμο αυτόν. Αυτά τα άμεσα ηθικά διλήμματα έρχονται στη επιφάνεια για διάφορες εφαρμογές AI, πόσο μάλλον στο τομέα της υγειονομικής περίθαλψης και υπολογιστικής ψυχιατρικής, που κάθε λάθος μπορεί να στοιχίσει ζωές. Βέβαια εκτός από αυτά τα άμεσα ηθικά ζητήματα που υπάρχουν, εφαρμογές της τεχνητής νοημοσύνης έχουν "μεταμορφωτικές επιδράσεις". Ειδικότερα όταν λέμε "μεταμορφωτικές επιδράσεις" εννοούμε επίμονες αλλαγές που γίνονται σταδιακά στην ηθική που αντιλαμβανόμαστε εμείς οι άνθρωποι, οι οποίες αυτές αλλαγές δεν γίνονται άμεσα αντιληπτές καθώς δεν πραγματοποιούνται ραγδαία και ταυτόχρονα επηρεάζουν την ανθρώπινη ευημερία και πολλές πτυχές της καθημερινότητας μας. Αυτές οι αλλαγές λοιπόν δεν χρειάζεται να αλλάζουν ριζικά τον τρόπο με τον οποίο αντιλαμβανόμαστε την ηθική αλλά, να μας κάνει πιο "ελαστικούς" ως προς κάποια ζητήματα όπως είναι η αυτονομία ή ακόμα και την ιδιωτική ζωή, και αυτό κατορθώνεται μέσω εφαρμογών τεχνητής νοημοσύνης που χρησιμοποιούμε σε καθημερινή βάση. Αντίστοιχα όμως οι εφαρμογές της υπολογιστικής ψυχιατρικής οι οποίες έχουν αποδειχθεί επιτυχημένες έχουν τη δυνατότητα να αλλάξουν εντελώς τον τρόπο με τον οποίο ταξινομούμε και ορίζουμε τις ψυχικές διαταραχές, κάτι το οποίο θα έχει άμεσες αλλά και έμμεσες συνέπειες στην ευημερία των ασθενών.**[4,49]**

Για να μπούμε πιο βαθιά όμως στο θέμα της ηθικής του AI στη ψυχιατρική, οφείλουμε να εξηγήσουμε κάποιους ορισμούς. Γενικά πολλές ψυχικές διαταραχές χαρακτηρίζονται από το φαινόμενο της διαταραγμένης συνείδησης. Για τις διαταραχές αυτές χρησιμοποιούμε τον όρο "διαταραχές της συνείδησης", ο οποίος όρος χρησιμοποιείται για καταστάσεις όπως το σύνδρομο της μη ανταποκρινόμενης εγρήγορσης, κατάσταση ελάχιστης συνείδησης ή ακόμη και κώμα. Σε αυτές τις περιπτώσεις η συνείδηση είτε είναι μειωμένη(ελάχιστη συνείδηση), είτε είναι μερικώς απούσα(μη ανταπόκριση εγρήγορση), είτε απουσιάζει εντελώς(κώμα). Εδώ, χρησιμοποιούμε τον όρο με μια πιο γενική έννοια όμως, με στόχο να καλύψουμε επιπλέον διαταραχές οι οποίες χαρακτηρίζονται από διαταραχή του περιεχομένου, της δομής ή και της μορφής των συνειδητών

διεργασιών. Πιο συγκεκριμένα τέτοιες διαταραχές εμπεριέχουν παραισθήσεις στη ψύχωση, απόπροσωποίηση και απόπραγματοποίηση στη σχιζοφρένεια. Οι ψυχώσεις αυτές δεν έχουν μειωμένα επίπεδα εγρήγορσης ή επίγνωσης αλλά χαρακτηρίζονται από διαταραχές της συνειδητής επεξεργασίας. Συνεπώς όταν θα αναφερόμαστε σε διαταραχές της συνείδησης θα συμπεριλαμβάνουμε και αυτές μέσα.**[49]**

Αν και όλοι είμαστε αρκετά εξοικειωμένοι με τη συνείδηση από τη προσωπική μας οπτική γωνία, δεν μπορούμε να πούμε ότι ισχύει το ίδιο στο πεδίο της επιστήμης, καθώς η επιστημονική μελέτη της συνείδησης δεν είναι οριστική, και υπάρχουν πολλές εμπειρικές θεωρίες οι οποίες υποβάλλουν ανταγωνιστικά συμπεράσματα. Ευτυχώς όμως η θεωρητική μας κατανόηση της συνείδησης έχει αυξηθεί κατά πολύ τα τελευταία χρόνια σε σύγκριση με τις προηγούμενες δεκαετίες, καθώς με την πρόοδο της μεταφραστικής νευρομοντελοποίησης έχουν επιτευχθεί καλύτερες σχέσεις με την υπολογιστική νευροεπιστήμη. Αυτό μας τονίζει τις δυνατότητες που μας παρέχει η υπολογιστική ψυχιατρική καθώς με τις επιτυχημένες εφαρμογές της μπορούμε να αναδιαμορφώσουμε τον τρόπο που αντιλαμβανόμαστε τις διαταραχές της συνείδησης, εφόσον έχουμε μια εμβάθυνση στη κατανόηση τους και βελτίωση στη θεραπεία τους. Κάτι το οποίο συνεπώς, θα επηρεάσει και τη κατανόηση μας της "κανονικής" συνείδησης. Συνεπώς αλλάζοντας τις πεποιθήσεις μας σχετικά με τις φυσιολογικές και διαταραγμένες συνειδητές εμπειρίες, μπορεί να μειώσουμε ή και να ενισχύσουμε το στίγμα που υπάρχει με αποτέλεσμα να αυξηθούν ή αντίστοιχα να μειωθούν οι θεραπευτικές επιλογές που ήδη υπάρχουν. Συνεπώς οι πιθανές μεταμορφωτικές επιδράσεις που μπορεί να έχει η υπολογιστική ψυχιατρική είναι ηθικά σχετικές. Πιο αναλυτικά οι αλλαγές αυτές μπορεί να επηρεάσουν θετικά ή αρνητικά τους ασθενείς, δηλαδή να μειώσουν ή να αυξήσουν την ευημερία των ατόμων αυτών. Για αυτόν τον λόγο θα αναλύσουμε περαιτέρω τις πιθανές μετασχηματικές επιδράσεις στο τομέα της ψυχικής ασθενείας που μπορεί να επιτύχουν οι κλινικές εφαρμογές της υπολογιστικής ψυχιατρικής. **[49,50]**

4.3.1 Υπολογιστική Ψυχιατρική

Η υπολογιστική ψυχιατρική έχει ως στόχο να μεταφράσει ανακαλύψεις και τεχνικές από την υπολογιστική νευροεπιστήμη στην κλασική κλινική ψυχιατρική, με σκοπό να είναι πιο βαθιά η κατανόηση μας των ψυχικών διαταραχών. Όπου έπειτα συνεπάγεται η βελτίωση των διαγνωστικών μεθόδων, για πιο ακριβή και αξιόπιστη λειτουργία προγνωστικών μέσων και θεραπευτικών προβλέψεων, τα οποία ευελπιστούμε πως θα οδηγήσουν στην εύρεση καινούργιων θεραπευτικών προσεγγίσεων. Εκτός αυτών των φιλοδοξιών, ένας μακροπρόθεσμος στόχος που έχει η υπολογιστική ψυχιατρική είναι η βελτίωση των διαγνωστικών κατηγοριών αξιοποιώντας, διαφοροποιώντας ή αντικαθιστώντας νοσολογίες με βάση τα συμπτώματα. Μια πολύ σημαντική υπόθεση σχετικά με την υπολογιστική ψυχιατρική είναι ότι τα υπολογιστικά μοντέλα έχουν τη δυνατότητα να χρησιμοποιηθούν για τον ορισμό υπολογιστικών φαινοτύπων. Ιδανικά μιλώντας αυτό όχι μόνο θα μας παρέχει έγκυρους χαρακτηρισμούς ψυχικής υγείας και ασθένειας αλλά θα δημιουργήσει και μία γέφυρα μεταξύ των μοριακών ευρημάτων και συμπεριφοριικών. Αυτό μας

δίνει τη δυνατότητα μακροπρόθεσμα να βελτιώσουμε τα αποτελέσματα των ασθενών εφόσον θα μπορούν να χρησιμοποιηθούν μηχανιστικά γειωμένες και αποτελεσματικές θεραπείες. Στην υπολογιστική ψυχιατρική επιστήμη είναι σύνηθες να χωρίζουμε σε 2 κλάδους τις προσεγγίσεις που υπάρχουν: αυτές που βασίζονται στα δεδομένα και αυτές που βασίζονται στη θεωρία. Αυτές οι οποίες βασίζονται στα δεδομένα κάνουν χρήση της μηχανικής μάθησης για να μπορέσουν να αναλύσουν και να επισημάνουν τα δεδομένα τους. Αυτό μας δίνει τη δυνατότητα να έχουμε ταξινόμηση και προβλέψεις στις ανταποκρίσεις της θεραπείας και των πιθανών καταλήξεων που μπορούν να έχουν κάποιες διαταραχές της συνείδησης. Σε αντίθεση η προσεγγίσεις που βασίζονται στη θεωρία χρησιμοποιούν κατά βάση γενετικά μοντέλα για τη μοντελοποίηση των αιτιών των δεδομένων. Σε αντιπαραβολή με τα διακριτικά μοντέλα(τα οποία είναι αρκετά περιορισμένα , μόνο στη ταξινόμηση των δεδομένων και των πιθανών αιτιών τους) ,γενετικά μοντέλα ενσωματώνουν υποθέσεις σχετικά με το πώς πάρατηρημένα αποτελέσματα δημιουργούνται. Κάτι το οποίο βέβαια μας δίνει τη δυνατότητα να πραγματοποιήσουμε προσομοιώσεις και να κάνουμε συγκρίσεις βάσει των στοιχείων που έχουμε μεταξύ των υποθέσεων, που βγάζουμε από την επιλογή του μοντέλου bayesian(Μπεϋζιανό μοντέλο). Το Μπεϋζιανό μοντέλο έχει προαναφερθεί σε αυτήν την ενότητα και είναι ένα στατιστικό μοντέλο όπου μας επιτρέπει να χρησιμοποιήσουμε πιθανότητες για να αναπαριστάνουμε όλες τις αβεβαιότητες μέσα σε ένα μοντέλο, τόσο την αβεβαιότητα σχετικά με την έξοδο όσο και την αβεβαιότητα σχετικά με την είσοδο στο μοντέλο. Ως Παραγωγικό μοντέλο ονομάζουμε το πιθανολογικό μοντέλο των δεδομένων και των κρυφών τους αιτιών. Αυτά τα μοντέλα είναι ιδιαίτερα χρήσιμα όταν προσπαθούμε να συμπεράνουμε υποκειμενικούς μηχανισμούς συμπτωμάτων, μετρήσεων, ή ακόμα και συμβατικές διαγνώσεις με βάση τα συμπτώματα. Αξίζει να σημειωθεί ότι τα παραγωγικά μοντέλα σε ορισμένες περιπτώσεις ονομάζονται και υπολογιστικές δοκιμασίες, οι οποίες ιδανικά μπορούν να διευκολύνουν τις διάφορες διαγνώσεις για μερικούς ασθενείς. Εφόσον καταφέρουν αυτά τα μοντέλα να είναι επιτυχή, μας παρέχουν πολλές δυνατότητες κάποιες από τις οποίες είναι και οι πιο ακριβείς διαγνώσεις και προβλέψεις θεραπείας, κάτι το οποίο είναι ήδη ηθικά αξιέπαινος στόχος, εφόσον αυτά τα αποτελέσματα δεν μπορούν να επιτευχθούν με άλλους λιγότερο δαπανηρούς τρόπους. Επιπρόσθετα, με την πρόοδο της τεχνολογίας, οι υπολογιστικές δοκιμασίες υπόσχονται να βελτιώσουν περαιτέρω τις καθαρά βασισμένες σε δεδομένα προσεγγίσεις, όπως για παράδειγμα μπορούν να βελτιώσουν την ομαδοποίηση σε συγκεκριμένες ομάδες(με τη βοήθεια της μηχανικής μάθησης) με τη χρήση γενετικών ενσωματώσεων με τουλάχιστον δύο τρόπους. Αρχικά η γενετική ενσωμάτωση από μόνη της μειώνει τη διάσταση των δεδομένων προσαρμόζοντας ένα παραγωγικό μοντέλο με ερμηνεύσιμες παραμέτρους, κάτι το οποίο μας δίνει τη δυνατότητα να αναπαριστάνουμε δεδομένα από ζητήματα τα οποία περιέχουν πολύ μικρό αριθμό χαρακτηριστικών, τα οποία με τη σειρά τους μπορούν να βελτιώσουν την απόδοση των αλγορίθμων. Έπειτα, αυτό έχει τη δυνατότητα να μας παρέχει πληροφορίες σχετικά με το γιατί υπάρχει χωρισμός των ασθενών σε υποομάδες και πώς γίνεται αυτός ο χωρισμός από έναν αλγόριθμο μηχανικής μάθησης, καθώς τα χαρακτηριστικά που χρησιμοποιούνται από τον αλγόριθμο είναι μηχανιστικά ερμηνεύσιμα.**[46,47]**

Σε αντιπαραβολή με τις προσεγγίσεις που χρησιμοποιούν παραγωγικά μοντέλα, το μεγάλο

πλεονέκτημα που έχουν οι προσεγγίσεις που βασίζονται σε δεδομένα είναι ότι δεν χρειάζεται να κάνουν σαφείς τις υποθέσεις τους χρησιμοποιώντας τη μορφή ενός παραγωγικού μοντέλου. Συνεπώς δίνει περισσότερη ελευθερία στον χρήστη καθώς μπορεί να αφήσει τα δεδομένα να "μιλήσουν από μόνα τους". Παρόλα αυτά, αυτό δεν σημαίνει ότι οι ερευνητές μπορούν να πάρουν ότι απόφαση θέλουν καθώς, οι αποφάσεις τους μπορεί να επηρεάσουν την ανάλυση των δεδομένων και την τελική πρόβλεψη με αποτέλεσμα να μην είναι ακριβής. Συνεπώς, είναι αναγκαίο να υπάρχει ειδική μέριμνα προκειμένου να αποφευχθούν ανακριβή αποτελέσματα λόγω μεροληψίας ή να μην είναι γενικευμένα, επειδή πάρθηκαν αποφάσεις από τους ερευνητές για να γίνει πιο εύκολη η διαδικασία, όπως πχ. Η διαδικασία συλλογής δεδομένων και η προεπεξεργασία των δεδομένων. **[46,47,48]**

Γενικά οι παράμετροι που χρησιμοποιούνται στα μοντέλα μηχανικής μάθησης δεν είναι εύκολα ερμηνεύσιμες, παρόλα αυτά ακόμα και πολύ περίπλοκοι αλγόριθμοι όπως αυτοί που χρησιμοποιούνται στο "μαύρο κουτί", έχουν πολύ υψηλά ποσοστά επιτυχίας προβλέψεων. Συνεπώς αυτές οι μέθοδοι συνεχίζουν να είναι πολύ αποτελεσματικές καθώς μπορούν να μας δώσουν τη δυνατότητα να προβλέψουμε πολλά πράγματα στο τομέα της ψυχιατρικής όπως είναι η μελλοντική χρήση αλκοόλ ή κάτι ακόμα πιο σημαντικό όπως είναι η αυτοκτονία. Παρόλα αυτά μη ερμηνεύσιμες προσεγγίσεις μπορεί να φανούν προβληματικές όταν εν τέλει παρουσιάζονται προβλήματα και θέτουν σε κίνδυνο τις ζωές των ασθενών, κάτι το οποίο δημιουργεί ένα άμεσο ηθικό δίλλημα όσο αφορά το AI στο τομέα της ψυχιατρικής**.[48]**

4.3.2 Ηθική Του AI Στην Υπολογιστική Ψυχιατρική

Η υπολογιστική Ψυχιατρική θεωρείται ένα μεγάλο ηθικό δίλημμα καθώς υπόσχεται πολλά καλά πράγματα ως προς την ευημερία των ασθενών αλλά ταυτόχρονα όμως, εμπεριέχονται πολλοί κίνδυνοι που την συνοδεύουν. Τα περισσότερα ηθικά προβλήματα που υπάρχουν στην υπολογιστική ψυχιατρική είναι ήδη γνωστά, καθώς είναι παρόμοια με αυτά που υπάρχουν στο τομέα της Βιοχαρτικής ηθικής, της νευροηθικής, και της ηθικής της τεχνικής νοημοσύνης. Πολλά από αυτά τα ηθικά ζητήματα έχουν προαναφερθεί, όπως είναι η προστασία των δεδομένων, οι προκαταλήψεις που υπάρχουν στους αλγορίθμους, ο χειρισμός τυχαίων ευρημάτων, και η δυνατή βελτίωση της έγκαιρης ανίχνευσης κινδύνων των ασθενών. **[51]**

Προβλήματα σαν αυτά όμως δεν πρέπει να υποτιμούν το πόσο χρήσιμη είναι η υπολογιστική ψυχιατρική, καθώς οι ψυχικές διαταραχές είναι από τις πρώτες αιτίες παγκοσμίως για τις οποίες πολλοί άνθρωποι αναγκάζονται να ζήσουν μια ζωή προσαρμοσμένη στις αναπηρίες τους. Ταυτόχρονα όμως, είναι πολύ δύσκολο να λάβει κάποιος φροντίδα ψυχικής υγείας, είτε είναι χώρα χαμηλού εισοδήματος είτε είναι υψηλού. Κάτι το οποίο φαίνεται πολύ στην εξής περίπτωση, όπου το 2015 μια έρευνα έδειξε ότι η μέση διάρκεια της μη θεραπευμένης ψύχωσης σε δημόσιες κλινικές στις Ηνωμένες Πολιτείες ήταν 74 εβδομάδες. Αυτό μας δείχνει άμεσα πόσο σοβαρό είναι το θέμα της ψυχικής υγείας παγκοσμίως καθώς πάρα πολύς κόσμος υποφέρει από ψυχικές διαταραχές, οι οποίες συνοδεύονται από πόνους και "βάσανα", επίσης μας δείχνει το πόσο κακές

είναι οι τεχνικές που χρησιμοποιούνται αυτή τη στιγμή για τη θεραπεία αυτών των ψυχώσεων. Συνεπώς η έρευνα για την εύρεση άλλων πιο αποτελεσματικών μέσων διάγνωσης και θεραπείας είναι ένας αξιέπαινος στόχος ηθικά ο οποίος θα βελτιώσει τη ζωή πολλών ανθρώπων παγκοσμίως με αποτέλεσμα να υπάρχει ένα πιο σωστό υγειονομικό σύστημα. **[51]**

Τα ηθικά ζητήματα που συνοδεύουν την ηθική στο τομέα της ψυχιατρικής αλλά και την υπολογιστική ψυχιατρική μπορούν να περιγράφουν καλύτερα αν γίνει διάκριση των διαφορετικών τομέων των εφαρμογών, όπως είναι η έγκαιρη ανίχνευση, και με αναφορές στις αρχές της ηθικής, όπως είναι η καλοσύνη, η ευεργεσία και ο σεβασμός για την αυτονομία και δικαιοσύνη. Μια από τις πιο σημαντικές αρχές της Τεχνητής Νοημοσύνης είναι η θεμελιώδη αρχή της επεξήγησης, γνωστή και ως διαφάνεια επεξήγηση, η οποία αρχή έχει 2 πτυχές, μία κανονιστική και μια επιστημική. Μια εφαρμογή AI μπορεί να θεωρηθεί επιστημική μόνο όταν είναι απολύτως κατανοητό το πώς λειτουργεί το εκάστοτε σύστημα, λόγου χάρη, ένα σύστημα είναι επιστημική αν είναι διαφανές το πώς ταξινομεί μια δεδομένη είσοδο με συγκεκριμένο τρόπο. Κανονιστικό από την άλλη, θεωρείται ένα σύστημα όταν μπορεί κανείς να καταλάβει ποιος είναι υπεύθυνος για τον τρόπο λειτουργίας του συστήματος και αντίστοιχα να προσδιορίσει ποιος είναι υπεύθυνος για τα αποτελέσματα που επιφέρει αυτό το σύστημα. Το να είναι ένα σύστημα διαφανές είναι εξαιρετικά σημαντικό καθώς, στο ενδεχόμενο της αποτυχίας του συστήματος ή το ενδεχόμενο των μη επιθυμητών αποτελεσμάτων, είναι πιο εύκολο να διακρίνουμε που υπάρχει σφάλμα στον αλγόριθμο. Ένα παράδειγμα συστήματος που δεν είναι διαφανές, είναι οι αλγόριθμοι οι οποίοι έχουν ρατσιστικές και άλλες παρόμοιες προκαταλήψεις. **[51]**

Η υπολογιστική ψυχιατρική είναι ένα πολύ ενδιαφέρον θέμα από την οπτική γωνία της ηθικής, καθώς η ανάπτυξη υπολογιστικών δοκιμασιών μπορεί να οδηγήσει σε ερμηνεύσιμα δεδομένα κάτι το οποίο καταρρίπτει τη νοοτροπία του μαύρου κουτιού που υπάρχει σε πολλές εφαρμογές, και το οποίο δημιουργεί πάρα πολλά ηθικά διλήμματα. Εκτός από τα πολλαπλά πιθανά οφέλη της υπολογιστική ψυχιατρικής, υπάρχει η ανησυχία ότι η υπολογιστική ψυχιατρική δεν είναι αρκετά ευρεία, καθώς αποτυγχάνει να λάβει υπόψη ψυχοκοινωνικούς παράγοντες, και αντιθέτως εστιάζεται υπερβολικά στις βιολογικές ιδιότητες. Πιο συγκεκριμένα μια ανησυχία που υπάρχει σχετικά με την υπολογιστική ψυχιατρική είναι ότι ανήκει στο "τρίτο κύμα της βιολογικής ψυχιατρικής", στο οποίο θεωρείται ότι οι ψυχικές διαταραχές είναι είτε εγκεφαλικές διαταραχές οι οποίες δεν γίνεται να θεραπευτούν, είτε είναι διαταραχές οι οποίες μπορούν να αντιμετωπιστούν χωρίς να χρειαστεί να δοθεί ιδιαίτερη σημασία στους ψυχοκοινωνικούς παράγοντες. Αυτό δυστυχώς συνεπάγεται ότι, κεντρικές πτυχές των ψυχικών διαταραχών αγνοούνται, κάτι το οποίο μπορεί να οδηγήσει σε μη αποτελεσματικές θεραπείες. Κάτι το οποίο δεν μπορεί να αποδώσει δικαιοσύνη στις διαταραχές της συνείδησης, με αποτέλεσμα οι παθόντες αυτών των διαταραχών να μην μπορούν να έχουν τις βέλτιστες πιθανές θεραπείες.**[51,52]**

Όλες αυτές οι απόψεις που υπάρχουν σχετικά με την υπολογιστική ψυχιατρική, την καθιστούν πολύ ενδιαφέρουσα όσο αφορά το θέμα της ηθικής της συνείδησης. Αυτό συμβαίνει διότι η υπολογιστική ψυχιατρική έχει τη δυνατότητα να βοηθήσει τους ασθενείς σε τέτοιο βαθμό, όπου θα μπορούσε να πει κανείς ότι θα τους ανακουφίσει από τα δεινά που περνάνε, κάτι το οποίο θεωρείται πολύ μεγάλο θετικό ηθικά και συνεπώς το καθιστά ως μια από τις προτεραιότητες των

ερευνητών. Από την άλλη πλευρά όμως, η υπολογιστική ψυχιατρική δεν υπόσχεται ότι θα μπορέσει να θεραπεύσει όλες τις ψυχικές διαταραχές, με αποτέλεσμα να υπάρχει μεγάλη πιθανότητα κάποιες από αυτές να αγνοηθούν, και συνεπώς να μην θεραπευτούν καθόλου, κάτι το οποίο καθιστά την υπογιστική ψυχιατρική κατακριτέα ηθικά. Η ηθική της τεχνητής νοημοσύνης είναι ένα πολύ ενδιαφέρον κομμάτι της υπολογιστικής ψυχιατρικής καθώς το ένα μπορεί να μάθει από άλλο. Ειδικότερα πιστεύεται ότι πολλά ηθικά προβλήματα στη τεχνητή νοημοσύνη μπορούν να λυθούν με τη βοήθεια τεχνικών λύσεων ή ότι προβλήματα όπως είναι η αδικία των εφαρμογών μπορούν να λυθούν με την επίτευξη ολικής δικαιοσύνης στους αλγόριθμους AI. Αντίστοιχα και στην υπολογιστική ψυχιατρική πιστεύεται ότι με την βελτίωση των εφαρμογών της, μπορούν να διαλυθούν ή έστω να μειωθούν οι ηθικοί προβληματισμοί που την περιτριγυρίζουν**.[52]**

Εν κατακλείδι ο τομέα της ψυχιατρικής, έχει κινητοποιηθεί αρκετά ώστε να εντάξει το AI στους τρόπους με τους οποίους κάνει τις διαγνώσεις αλλά και αντίστοιχα τις θεραπείες του. Αυτό ενδέχεται να δημιουργήσει κάποιους προβληματισμούς στον κόσμο αλλά τα πλεονεκτήματα τα οποία υπόσχεται να παρέχει υπερτερούν σε σύγκριση με τα μειονεκτήματα. Κατά τη γνώμη μου είναι μια σημαντική αναβάθμιση στο σύστημα υγείας καθώς ειδικά στο τομέα της ψυχικής υγείας δεν υπάρχουν αρκετά καλές υποδομές στη παρούσα εποχή, με απόρροια οι ασθενείς να μην δέχονται την απαραίτητη φροντίδα, κάτι το οποίο δεν είναι ηθικά σωστό ως προς αυτούς. Συνεπώς όχι απλά είναι μια καλή εναλλακτική στον τρόπο με τον οποίο θα πραγματοποιούνται οι προλήψεις και οι θεραπείες αλλά θα έλεγε κανείς ότι είναι μία απαραίτητη αλλαγή η οποία πρέπει να γίνει όσο το συντομότερο δυνατόν**.[4]**

4.3.3 Διαγνώσεις με χρήση AI στο τομέα της νευροποικιλομορφίας

Όπως θα αναφέρουμε πολλές φορές στη παρούσα διπλωματική εργασία, η χρήση της τεχνητής νοημοσύνης στο τομέα της υγείας γίνεται όλο και πιο μεγάλη, με αποτέλεσμα να υπάρχουν επιπλοκές σε πολλούς τομείς και να δημιουργούνται ερωτήματα σχετικά με την αξιοπιστία και το αν γίνεται σωστή ηθικά χρήση των αλγορίθμων AI. Σε αυτό το σημείο θα μιλήσουμε για τους τρόπους με τους οποίους μπορεί να γίνει διάγνωση των ατόμων με νευροποικιλομορφίες χρησιμοποιώντας αλγορίθμους AI. Με την πρόοδο λοιπόν της τεχνολογίας έχουν υπάρξει πολλές εφαρμογές οι οποίες χρησιμοποιούν τεχνολογίες αναγνώρισης προσώπου, κάτι το οποίο μπορούμε να το δούμε ακόμα και στα κινητά μας. Η έρευνα στη υπολογιστική όραση λοιπόν έχει βάλει ως στόχο να αναπτύξει συστήματα αυτοματοποίησης για μια ποικιλία νευροδιαφορών καταστάσεων, συμπεριλαμβανομένου και του αυτισμού, κάτι το οποίο μπορεί να γίνει εφικτό με την αναγνώριση προσώπου. Πιο συγκεκριμένα οι ερευνητές έχουν ως σκοπό να χρησιμοποιήσουν την αναγνώριση προσώπου ώστε να μπορέσουν να διαγνώσουν σε γρηγορότερο χρόνο τα παιδιά τα οποία πάσχουν από αυτισμό και άλλες παρόμοιες νευροποικιλομορφίες. Για

να κατορθωθεί αυτό όμως οι ερευνητές εξετάζουν τις εκφράσεις του προσώπου, τα επίπεδα συγκίνησης, και τις επαναλαμβανόμενες συμπεριφορές σε παιδιά που έχουν ήδη αυτισμό, ώστε να μπορέσουν να εκπαιδεύσουν τον αλγόριθμο για να το κάνει αποτελεσματικά και γρηγορότερα από τον άνθρωπο. **[53]**

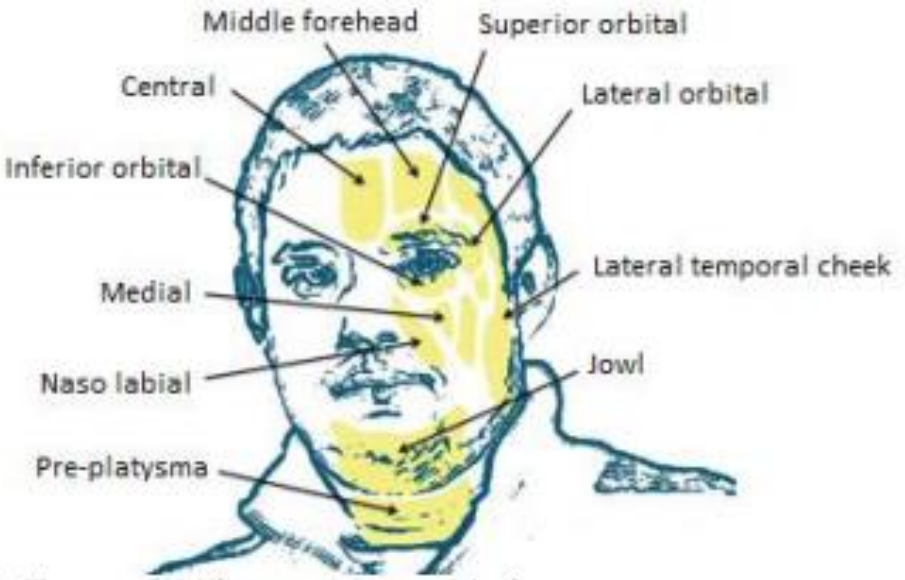


Fig. 1. The superficial fat compartments of the face.

Source: A survey on computer vision for assistive medical diagnosis from faces, Thevenot, J., Bordallo López, M., Hadid.

Βέβαια αυτό το γεγονός φέρνει στην επιφάνεια ερωτήσεις οι οποίες σχετίζονται με τη δικαιοσύνη και την ηθική αυτής της πράξης. Πιο αναλυτικά, οι προβληματισμοί αυτοί υπάρχουν καθώς στον τομέα των νευροποικιλομορφιών και πιο συγκεκριμένα στο θέμα του αυτισμού υπάρχουν πολλές προκαταλήψεις όσο αφορά το φύλλο και την ανισότητα που προκύπτει από τα ποσοστά διάγνωσης, ενώ ταυτόχρονα υπάρχουν και λιγότερο μελετημένες αλλά σημαντικές προκαταλήψεις σχετικά με τη φυλή, την εθνικότητα και την τάξη του κάθε ανθρώπου. Συνεπώς η "δημιουργία" διαγνωστικών εργαλείων που αναπτύσσονται με βάση τα περιστατικά τα οποία υπάρχουν αυτή τη στιγμή μπορεί να συνεχίσουν αυτές τις προκαταλήψεις οι οποίες υπάρχουν ήδη με αποτέλεσμα να μην υπάρχει κάποια βελτίωση στο τομέα των διαγνώσεων, και να ακόμα χειρότερα να υπάρξουν παραπάνω δυσκολίες. Επομένως, για να υπάρξει η διασφάλιση της ακριβής διάγνωσης προς τα φύλα και τις εθνότητες οι οποίες δεν εκπροσωπεύονται αρκετά, ενδέχεται να προτείνουμε τη διαφοροποίηση των συνόλων των δεδομένων με βάση μια μετρική

της δικαιοσύνης. Παρόλα αυτά, η προσέγγιση αυτή δεν μας εγγυάται ότι θα υπάρχει η δυνατότητα διάγνωσης, και αν υπάρχει ποιος θα μπορεί να επηρεαστεί από αυτή. **[53]**

Επιπρόσθετα, Η επιρροή των ψυχιάτρων ενισχύεται από εργαλεία που «βοηθούν» τα αυτιστικά άτομα στο καλούπι των υπαρχόντων πρωτοτύπων όρασης υπολογιστή, τα οποία κάνουν περισσότερα από το να δίνουν απλώς τη διάγνωση, και με το να ενισχύεται η επιρροή των ψυχιάτρων, δημιουργείται η εντύπωση ότι η παλιοί τρόποι διάγνωσης είναι και οι μοναδικοί οι οποίοι είναι λειτουργικοί. Οι εξετάσεις της ιατρικοποίησης - η διαδικασία με την οποία νομιμοποιείται αυτή η έννοια της επίσημης φύλαξης - την οποία έχουν ανακαλύψει προηγουμένως στη διάγνωση του αυτισμού, ενώ όσοι την χρησιμοποιούν ως βάση έχουν βρει μικρή εγκυρότητα στα διαγνωστικά συστήματα όρασης υπολογιστών. Με την προσθήκη τεχνικής και επιστημονικής εξουσίας στην ιατρική αρχή, τα άτομα που εκτίθενται σε ιατρικές συνθήκες αποδυναμώνονται ακόμη περισσότερο, με τη φωνή του ασθενούς να αποκτά ακόμη λιγότερη νομιμότητα. Και πάλι, η δικαιοσύνη δεν είναι η απάντηση, καθώς το πρόβλημα δεν είναι η προκατάληψη που σχετίζεται με τον αυτισμό, αλλά μάλλον η γενικευμένη διάκριση των ασθενών, κάτι το οποίο δεν φαίνεται να έχει υποστεί αρκετές βελτιώσεις τα τελευταία χρόνια. Έτσι, η επίτευξη δίκαιων αποτελεσμάτων σε αυτόν τον τομέα απαιτεί να σκεφτόμαστε την εξουσία και όχι τη δικαιοσύνη και το ευρύτερο κοινωνικό πλαίσιο στο οποίο ενσωματώνονται τα τεχνικά συστήματα. **[8]**

Ένας άλλος πολύ σημαντικός προβληματισμός είναι το τι συμβαίνει μετρά τη διάγνωση του αυτισμού, καθώς τα συστήματα μπορεί να είναι φτιαγμένα για να κάνουν σωστά τη διάγνωση, αλλά δεν γνωρίζουμε πως θα μας παρέχουν με τη σωστή διαδικασία που πρέπει να ακολουθηθεί έπειτα. Τα συστήματα τεχνητής νοημοσύνης σε αυτόν τον τομέα λειτουργούν με την παραδοχή ότι η έγκαιρη διάγνωση είναι επιθυμητή, καθώς ανοίγει την πόρτα στη δυνατότητα θεραπείας, βοήθειας και εξέτασης, κάτι το οποίο σε πολλές περιπτώσεις και πολλούς άλλους τομείς μπορεί να κάνει τη διαφορά μεταξύ ζωής και θανάτου. Όμως η ψυχιατρική και η έρευνα για την κρίσιμη αναπηρία έχουν εγείρει σοβαρές αμφιβολίες ως προς το κατά πόσο είναι προτιμότερη μια πρώιμη διάγνωση, ανεξάρτητα από τις ήδη σημειωμένες προκαταλήψεις σχετικά με το ποιος μπορεί να έχει πρόσβαση στη διάγνωση, διότι μπορεί να υπάρξουν περιπτώσεις όπου η πρόωρη διάγνωση να δημιουργήσει περισσότερο κακό παρά καλό, κάτι το οποίο είναι ηθικά λάθος και κατακριτέο.**[8]**

Παρόλες τις πεποιθήσεις που υπάρχουν στον τομέα της υπολογιστικής όρασης, μπορεί κανείς να καταλάβει ότι ο αυτισμός δεν είναι απλός κάτι το οποίο μπορεί να διαγνωστεί εύκολα, καθώς η διάγνωση του αυτισμού έχει πολλά αρνητικά αποτελέσματα. Ένα από αυτά τα αποτελέσματα του στιγματισμού που σχετίζεται με αυτήν την ταξινόμηση είναι, οι καταστροφικές θεραπείες αλλαγής συμπεριφοράς στις οποίες έχουν υποβληθεί πολλά παιδιά με αυτισμό, στο σημείο που ακόμα και κάποια από αυτά τα παιδιά έχουν σκοτωθεί, και εκτός από το γεγονός ότι χάσανε τη ζωή τους, τους βάλανε και τη ταμπέλα του “δολοφονίες ελέους”, κάτι το οποίο είναι απίστευτα βάρβαρο αν σκεφτούμε σε τι άτομα αναφερόμαστε και σε τι περιστάσεις πραγματοποιήθηκαν αυτές οι “δολοφονίες”. Επομένως, τα συστήματα υπολογιστικής όρασης τα οποία αποδίδουν αυτές τις ταμπέλες μπορεί να θεωρηθούν ηθικά μοιραία αν δημιουργηθούν χωρίς να ληφθούν υπόψη οι κοινωνικές συνθήκες στις οποίες ζουν τα άτομα με αυτισμό, και τα

τραύματα που μπορεί να έχουν υποβληθεί παλαιότερα από τον περίγυρο τους. **[54]**

Συμπερασματικά η άνιση μεταχείριση των αυτιστικών ατόμων που προκαλείται από διαγνώσεις αυτισμού που βασίζονται σε αυτοματοποιημένες διαδικασίες με αλγορίθμους AI είναι ένα μεγάλο ηθικό ζήτημα που δεν μπορεί να λυθεί απλά εξετάζοντας τις εισόδους και εξόδους του συστήματος, αλλά πρέπει να απομακρυνθεί κανείς για να δει την ολική εικόνα. Καθώς τα άτομα αυτά έχουν μια δύσκολη ζωή από μόνη της, την οποία η κοινωνία μας δεν προσπαθεί ενεργά να κάνει καλύτερη. Οπότε κρίνεται απαραίτητο να δημιουργηθούν μοντέλα τα οποία θα λαμβάνουν υπόψη τους τον ευρύτερο κοινωνικό περίγυρο των ασθενών, καθώς και την οικονομική κατάσταση τους, και τους τρόπους με τους οποίους μπορεί η τεχνολογία να αλλάξει τη καθημερινότητα αυτών των ατόμων με αποτέλεσμα να τους δημιουργήσει μια πιο ποιοτική ζωή, την οποία και θα ευχαριστιούνται. **[8]**

Κεφάλαιο 5 : Ανάλυση της ασφάλειας της Μηχανικής Μάθησης

5.1 Ασφάλεια της Μηχανικής Μάθησης

Στις μέρες μας η μηχανική μάθηση είναι ένα πολύ σημαντικό εργαλείο σε πολλές εφαρμογές της καθημερινότητας μας, κάτι το οποίο έχει θέσει πολλά ερωτήματα στο θέμα ασφάλειας της χρήσης της. Για αυτόν ακριβώς τον λόγο οι εκπρόσωποι της μηχανικής μάθησης έχουν προτείνει αυτοδίδακτα συστήματα ασφάλειας εκ των οποίων κάποια μπορούν ακόμα και να ανιχνεύσουν ανεπιθύμητα μηνύματα ή και εισβολείς στο δίκτυο. Η ιδεολογία τους είναι ότι το Machine Learning θα δώσει την δυνατότητα σε ένα σύστημα να ανταποκριθεί στις ταχύτατα εξελισσόμενες καταστάσεις του πραγματικού κόσμου, είτε κακές είτε καλές. Δυστυχώς όμως υπάρχει ένας μεγάλος κίνδυνος, και αυτός είναι το ενδεχόμενο οι "Hackers" να επιχειρήσουν να εκμεταλλευτούν την δυνατότητα που έχουν τα συστήματα να προσαρμόζονται , κάτι το οποίο συνεπάγεται στην αποτυχία του συστήματος . Αυτό συνήθως γίνεται με την εμφάνιση σφαλμάτων γνωστά και ως "errors". Αυτό συνεπάγεται ότι ο εισβολέας μπορεί να "κοροϊδέψει" το σύστημα μας στο να νομίζουν ότι δεν υπάρχει εισβολή στο δίκτυο ή ακόμα και το αντίθετο , δηλαδή να θεωρεί εισβολείς τους σύνηθες χρήστες. Συνεπώς αν οι χρήστες παρατηρήσουν πολλά σφάλματα στο σύστημα ,δεν θα θέλουν να συνεχίσουν τη χρήση του συστήματος. Επιπρόσθετα ένα χειρότερο σενάριο είναι να υπάρχει εισβολέας αλλά να μην υπάρχει error κάτι το οποίο είναι πολύ επικίνδυνο.**[1]**

Συνεπώς πρέπει να τεθούν κάποιοι στόχοι στον τομέα της ασφάλειας για να μπορεί ο εκάστοτε αλγόριθμος τεχνητής νοημοσύνης να δουλέψει σωστά. Συνεπώς μια ακριβής ανάλυση ενός συστήματος είναι απαραίτητη, η οποία απαιτεί τον καθορισμό στόχων ασφαλείας και ένα μοντέλο απειλής. Ο πρωταρχικός ρόλος της ασφάλειας είναι να προστατεύει τα περιουσιακά στοιχεία από τους εισβολείς που προσπαθούν να τα αποκτήσουν με μη ηθικούς τρόπους. Συνεπώς όταν ένα περιουσιακό στοιχείο τίθεται σε κίνδυνο ή χάνεται εντελώς με οποιονδήποτε τρόπο, τότε δεν έχει πραγματοποιηθεί η επίτευξη του στόχου της η ασφάλεια, κάτι το οποίο είναι ο κύριος λόγος ύπαρξης της ασφάλειας εξαρχής. Επιπρόσθετα εκτός από τους στόχους ασφαλείας που πρέπει να τεθούν, είναι ύψιστης σημασίας να υπάρχει και ένα μοντέλο απειλής. Ειδικότερα, ένα μοντέλο απειλής είναι ένα προφίλ πιθανών εισβολέων που περιγράφει λεπτομερώς τους στόχους και τους πόρους τους. Κάτι το οποίο μπορεί να χρησιμοποιηθεί για την δημιουργία σωστών μεθόδων ασφαλείας απέναντι τους. **[1]**

Ο τομέας της ασφάλειας έχει πολλά εργαλεία τα οποία βοηθούν στη διασφάλιση της. Ένα από αυτά είναι το εργαλείο ταξινόμησης που χρησιμοποιείται στη τεχνητή νοημοσύνη, το οποίο έχει πολλές χρήσεις, και μια από αυτές είναι διαχωρισμός δυνητικά επικίνδυνων καταστάσεων. Υπάρχουν πολλά άλλα

συστήματα τα οποία βοηθάνε στη προστασία των περιουσιακών στοιχείων, όπως είναι ένα σύστημα ανίχνευσης ιών, όπου ο σκοπός του συστήματος αυτού είναι να ανιχνεύσει έγκαιρα τον ιό, προτού δημιουργήσει προβλήματα στο σύστημα, ή σε περίπτωση μόλυνσης να ανιχνεύσει τη μόλυνση ώστε να μπορέσει να αντιμετωπιστεί με τον αντίστοιχο τρόπο. Ένα άλλο σύστημα το οποία είναι αναντικατάστατο, είναι το σύστημα ανίχνευσης εισβολής (IDS, intrusion detection system), το οποίο αναλύει την δραστηριότητα στο δίκτυο ή και τον χρήστη αντίστοιχα για τυχόν ύποπτη συμπεριφορές. Αντίστοιχα υπάρχει και το σύστημα πρόληψης εισβολής(IPS, intrusion prevention system), το οποίο συνδέεται άμεσα με τον εντοπισμό προσπαθειών εισβολής και αντίστοιχα την λήψη σωστών μέτρων για την επιτυχής καταπολέμηση των εισβολών αυτών. **[1]**

Για τον σωστό καθορισμό στόχων στον τομέα της ασφάλειας χρησιμοποιούνται πολύ οι ταξινομητές. Ο στόχος ενός ταξινομητή σε μια ρύθμιση ασφαλείας είναι να εντοπίζει απειλές και να τις απομονώνει έτσι ώστε να μην έχουν τη δυνατότητα να επηρεάσουν τις κανονικές λειτουργίες του συστήματος. Αυτό βέβαια μπορεί να χωριστεί σε 2 στόχους λειτουργίας. Αρχικά ως στόχο ακεραιότητας, όπου ο στόχος είναι η αποτροπή απειλών από το να φτάσουν στα περιουσιακά στοιχεία, κάτι το οποίο είναι πολύ σημαντικό για τη διασφάλιση της σωστής λειτουργίας του συστήματος. Δεύτερον είναι στόχος διαθεσιμότητας, ο οποίος στόχος έχει ως σκοπό την αποτροπή παρεμβολών των εισβολέων στη κανονική λειτουργία του συστήματος. Κρίνεται απαραίτητο να αναφερθεί ότι υπάρχει μια σχέση μεταξύ της μη ένδειξης παραβίασης(ψευδή αρνητικά) και του στόχου ακεραιότητας, καθώς στο ενδεχόμενο της μη ένδειξης παραβίασης, ο εισβολέας θα έχει προσπεράσει τον ταξινομητή και μπορεί να δημιουργήσει απίστευτα μεγάλες ζημιές στο σύστημα. Αντίστοιχα στο ενδεχόμενο λανθασμένων συναγερμών(ψευδή θετικά), όπου δεν υπάρχει κάποια κακόβουλη εισβολή στο σύστημα, παραβιάζεται ο στόχος διαθεσιμότητας διότι δεν επιτρέπει σε καλούς χρήστες να έχουν πρόσβαση στο σύστημα. **[1]**

Ένα άλλο σημαντικό εργαλείο που προαναφέραμε και χρησιμοποιείται για την ασφάλεια των συστημάτων είναι τα μοντέλα απειλής. Ο σκοπός και τα κίνητρα ενός εισβολέα όπως είναι οι hackers είναι το να προσπαθήσει να αποκτήσει πρόσβαση στα ευαίσθητα δεδομένα του συστήματος, ή να διακόψει εντελώς κάθε λειτουργία του συστήματος, με αποτέλεσμα την ολική αποτυχία του συστήματος αυτού. Αυτά μπορούν να επιτευχθούν συνήθως με μια σειρά ψευδών αρνητικών ή αντίστοιχα με μια σειρά ψευδών θετικών, όπου και στις 2 περιπτώσεις μπορεί να επιτευχθεί μεγάλη ζημιά στο σύστημα. Για παράδειγμα ένας δημιουργός ενός ιού θα προτιμούσε να αποκτήσει πρόσβαση στο σύστημα χωρίς να εντοπιστεί και να αποκτήσει έλεγχο των συστημάτων ασφαλείας(με τη βοήθεια ψευδών αρνητικών). Αντίστοιχα ένας έμπορος ο οποίος θέλει να βλάψει τον ανταγωνισμό με αθέμητους τρόπους, θα θέλει να εμποδίσει εντελώς την κυκλοφορία πωλήσεων στην ιστοσελίδα του αντίπαλου του με τη βοήθεια ψευδών θετικών, κάτι το οποίο καταστρέφει εντελώς τη δουλειά του άλλου εμπόρου. Συνεπώς με αυτούς τους τρόπους μπορεί κάποιος να μπει στο σύστημα και να δημιουργήσει πολλές καταστροφές σε διάφορους τομείς του συστήματος.

Κάτι άλλο που γνωρίζουμε σχετικά με τους εισβολείς και τους απειλούμενους είναι ότι έχουν ο καθένας τους μια συνάρτηση κόστους, όπου ανάλογα με το τι συμβαίνει έχουν αρνητικό ή θετικό κόστος. Στο ενδεχόμενο του αρνητικού κόστους ο εισβολέας είναι εκείνος ο οποίος έχει βγει κερδισμένος, καθώς ο αμυνόμενος θα έχει μεγαλύτερο κόστος από τον εισβολέα. Σε κάθε περίπτωση υπάρχει κόστος και για τους 2 αλλά ο κερδισμένος είναι εκείνος με το χαμηλότερο. Βέβαια στις περισσότερες περιπτώσεις ο εισβολέας είναι εκείνος ο οποίος έχει το χαμηλότερο κόστος. Προφανώς αν οι στόχοι αυτών των 2 ήταν κοινοί δεν θα

ήταν αντίπαλοι και δεν θα μπαίνανε σε αυτήν την διαδικασία να σαμποτάρουν ο ένας τον άλλον. Στα παραδείγματα που θα ακολουθήσουν θα θεωρήσουμε ότι όταν ο αμυνόμενος έχει μεγάλο κόστος, ο επιτιθέμενος θα έχει χαμηλό και το αντίστροφο. **[1]**

Αρχικά κρίνεται απαραίτητο να αναλύσουμε τις δυνατότητες τις οποίες έχει ο εισβολέας. Αρχικά πάντα υποθέτουμε ότι ο εισβολέας είναι εξοικειωμένος με τους τρόπους εκπαίδευσης και ότι έχει μερική είτε και πλήρη γνώση των συνόλων εκπαίδευσης τα οποία έχουν χρησιμοποιηθεί. Συνεπώς υποθέτουμε ότι ισχύει το χειρότερο πιθανό ενδεχόμενο. Αυτό το ενδεχόμενο μπορεί να επιτευχθεί αν ο αντίπαλος "κατασκοπεύει" ή παρακολουθεί με οποιονδήποτε τρόπο, τη δραστηριότητα του διαδικτύου κατά τη διάρκεια της εκπαίδευσης. Μπορεί να υπάρχουν και σενάρια στα οποία ο επιτιθέμενος να έχει μερική πρόσβαση στα δεδομένα εκπαίδευσης, ή να μην έχει και καθόλου, όπου συνεπώς δεν θα έχει έλεγχο στα δεδομένα που χρησιμοποιεί για την εκπαίδευση του αλγορίθμου ο αμυνόμενος. Το πιο πιθανό σενάριο ρεαλιστικά είναι ο επιτιθέμενος να μην έχει μεγάλη πρόσβαση στα δεδομένα αυτά, βέβαια είναι μεγάλο λάθος από λογική ασφαλείας να θεωρήσουμε ότι αυτό ισχύει. Αντιθέτως αν θέλουμε η ασφάλεια ενός συστήματος να είναι πολύ καλή πρέπει πάντα να θεωρούμε ότι ο επιτιθέμενος έχει έλεγχο σε πολλά αντικείμενα του κώδικα μας, έτσι ώστε να μην μπορούμε να υποτιμήσουμε τις ικανότητες του, αλλά να τις υπερτιμήσουμε, κάτι το οποίο είναι θετικό στον τομέα της ασφάλειας. **[1]**

Συνεπώς στη πληθώρα των περιπτώσεων υποθέτουμε ότι ο επιτιθέμενος έχει απόλυτη ελευθερία να αλλάξει και να πειράξει ή και να δημιουργήσει διάφορα δεδομένα, βέβαια υπάρχουν διαμορφώσεις οι οποίες περιορίζουν τον κακόβουλο χρήστη να κάνει ότι θέλει, καθώς του θέτουν αρκετά όρια στο τι δεδομένα μπορεί να δημιουργήσει. Ένας τρόπος με τον οποίο μπορεί να περιοριστεί ο εισβολέας είναι με τις ετικέτες, τις οποίες δεν μπορεί να βάλει σε πολλά δεδομένα με αποτέλεσμα να μην μπορούν αν αναγνωριστούν, κάτι το οποίο τα καθιστά εν μέρει "άχρηστα". Επιπρόσθετα είναι πιθανό για έναν εισβολέα να έχει πλήρη έλεγχο στα πακέτα δεδομένων τα οποία παραδίδονται από τον ίδιο, αλλά οι δρομολογητές(routers) να αφαιρούν συγκεκριμένα πακέτα ή και να τα στέλνουν με μεγάλη καθυστέρηση, ώστε να μην μπορούν να χρησιμοποιηθούν. Όσο αφορά τα δεδομένα που υπάρχουν στον αλγόριθμο για εκπαίδευση, ισχύει ότι ο εισβολέας πιθανότατα δεν μπορεί να αλλάξει όλα τα δεδομένα, κάτι το οποίο κάνει τη διαδικασία του "σαμποτάζ" πιο δύσκολη. Αυτό συμβαίνει διότι δεν είναι εύκολο να καταλάβει κανείς πως εκπαιδεύεται ο εκάστοτε αλγόριθμος, ακόμα και αν έχει πρόσβαση σε όλα τα δεδομένα, υπάρχουν τακτικές εκπαίδευσης οι οποίες είναι φτιαγμένες για να δυσκολέψουν τον εισβολέα και να μην μπορεί να επέμβει σε αυτά τα δεδομένα. **[1]**

Γενικά είναι δυνατόν να εκπαιδευτεί ένας αλγόριθμος είτε ρητά είτε σε συνεχή βάση(online εκπαιδευόμενος αλγόριθμος). Όταν είναι διαδικτυακή η εκπαίδευση ο αλγόριθμος έχει τη δυνατότητα να προσαρμοστεί καλύτερα στις νέες μεταβαλλόμενες συνθήκες, καθώς οι σταθερότητα αποδυναμώνεται για να μπορεί να αλγόριθμος να καλύψει τις μακροπρόθεσμες αλλαγές στη κατανομή των δεδομένων. Αυτό κάνει την εκπαίδευση αυτή πολύ πιο ευέλικτη αλλά αντίστοιχα και πιο επιρρεπή σε κοινές και συνεχόμενες επιθέσεις. Δεδομένοι ότι η λειτουργία πρόβλεψης ενός διαδικτυακού μαθητή εξελίσσεται με την πάροδο του χρόνου, ένας

αντίπαλος μπορεί να επηρεάσει αυτή τη διαδικασία προβλέποντας πως θα εξελιχθεί. Συνεπώς αν και είναι πολύ πιο βολική η διαδικτυακή διαδικασία εκπαίδευσης, είναι πολύ πιο επικίνδυνη όσο αφορά τις επιθέσεις που μπορεί να δεχτεί. **[11]**

Συνεπώς όπως μπορούμε να διακρίνουμε υπάρχουν διάφοροι τρόποι με τους οποίους μπορεί κανείς να αποκτήσει πρόσβαση στο σύστημα. Οπότε κρίνεται απαραίτητο να αναφέρουμε κάποιες κατηγορίες επιθέσεων και πως αντίστοιχα αυτές επηρεάζουν το σύστημα. Αρχικά υπάρχουν οι επιθέσεις ΕΠΙΡΡΟΗΣ, οι οποίες χωρίζονται σε διερευνητικές και αιτιώδεις. Οι διερευνητικές επιθέσεις, εκμεταλλεύονται τις λανθασμένες ταξινομήσεις χωρίς όμως να επηρεάζουν κατά κάποιον τρόπο τα δεδομένα εκπαίδευσης. Αντιθέτως οι αιτιώδεις επιθέσεις επηρεάζουν άμεσα τη διαδικασία της εκπαίδευσης, καθώς μπορούν και χειρίζονται τα δεδομένα εκπαίδευσης. Μια άλλη κατηγορία επιθέσεων είναι οι ΠΑΡΑΒΙΑΣΕΙΣ ΤΗΣ ΑΣΦΑΛΕΙΑΣ. Αυτές χρησιμοποιούν κυρίως ψευδώς αρνητικές επιθέσεις ακεραιότητας , οι οποίες θέτουν σε κίνδυνο τα περιουσιακά στοιχεία, και χρησιμοποιούν και ψευδώς θετικά στοιχεία, τα οποία οδηγούν στην άρνηση υπηρεσίας από το σύστημα, όπου το σύστημα απλά δεν λειτουργεί όπως θα έπρεπε. Τέλος έχουμε τις ΣΤΟΧΕΥΜΕΝΕΣ επιθέσεις. Οι επιθέσεις αυτές όπως λέει και το όνομα είναι επικεντρωμένες σε κάτι πολύ συγκεκριμένο στο σύστημα. Αντίστοιχα οι επιθέσεις οι οποίες δεν έχουν κάποιον συγκεκριμένο στόχο ονομάζονται “αδιάκριτες”, και προσπαθούν να βλάψουν το σύστημα με όποιον τρόπο βρουν. **[1]**

Η πρώτη κατηγορία επιθέσεων αναφέρεται κατά κύριο λόγο στις δυνατότητες ενός εισβολέα να κάνει 2 πράγματα. Αρχικά να αλλάξει τα δεδομένα εκπαίδευσης σε ένα σύστημα τα οποία χρησιμοποιούνται για την κατασκευή ενός ταξινομητή(αιτιακή επίθεση), και δεύτερον να μην μπορεί να αλλάξει τα δεδομένα σε ένα σύστημα αλλά να έχει τη δυνατότητα να στείλει καινούργια δεδομένα για εκπαίδευση, και να παρατηρήσει πως αντιδρά ο αλγόριθμος στα δεδομένα αυτά τα οποία δημιούργησε ο ίδιος. **[1]**

Στη δεύτερη κατηγορία επιθέσεων αναφέρεται στα είδη παραβίασης ασφαλείας που μπορεί να προκαλέσει ο εισβολέας. Το ένα είδος είναι η άρνηση της υπηρεσίας όπου κάνει το σύστημα να μη λειτουργεί σωστά με τη βοήθεια ψευδών θετικών, κάτι το οποίο είναι η διαδικασία στην οποία οι αβλαβείς περιπτώσεις φιλτράρονται λανθασμένα. Και η άλλη περίπτωση είναι αυτή η οποία επιτρέπει σε επιβλαβείς περιπτώσεις να ξεπεράσουν τον ταξινομητή και να μπούνε στο σύστημα ως ψευδείς αρνητικά. **[1]**

Τέλος η Τρίτη κατηγορία επιθέσεων είναι είτε μια στοχευμένη επίθεση, η οποία έχει ως σκοπό να βλάψει την απόδοση του ταξινομητή σε μία μόνο περίπτωση, είτε μια αδιάκριτη επίθεση η οποία στοχεύει να προκαλέσει την αποτυχία του ταξινομητή σε ένα ευρύ φάσμα περιπτώσεων. Η συγκεκριμένη κατηγορία χαρακτηρίζεται από το εύρος των επιλογών της καθώς μπορεί να βλάψει τον ταξινομητή με αρκετούς τρόπους και σε πολλά σημεία. **[1]**

Οι επιθέσεις αυτές μπορούν να βλάψουν το σύστημα σε πολλά σημεία και είναι ένας πολύ σημαντικός παράγοντας που πρέπει να ληφθεί υπόψη πρώτού χρησιμοποιήσει κανείς αλγορίθμους AI με μηχανική μάθηση. Βέβαια τα καλά πολλές φορές υπερτερούν των αρνητικών και στη συγκεκριμένη περίπτωση ισχύει κάτι τέτοιο καθώς η μηχανική μάθηση δίνει σχεδόν

άπειρες δυνατότητες στον χρήστη, κάτι το οποίο είναι πολύ πιο σημαντικό σε σύγκριση με το ενδεχόμενο της επίθεσης. Οπότε είναι σημαντικό κανείς να ξέρει τι κινδύνους μπορεί να διατρέξει όταν χρησιμοποιεί εφαρμογές τεχνητής νοημοσύνης και να μπορεί να λάβει τα απαραίτητα μέτρα προστασίας για να είναι ασφαλής.

Κεφάλαιο 6 : Συμπεράσματα

Στη παρούσα πτυχιακή έχουμε αναλύσει ζητήματα τα οποία αφορούν κατά κύριο λόγω την ηθική της τεχνητής νοημοσύνης και των εφαρμογών της. Ένα θέμα το οποίο είναι εξαιρετικά ενδιαφέρον καθώς είναι ένα πρόβλημα το οποίο αντιμετωπίζουμε καθημερινά στη ζωή μας. Από τις προτάσεις που μας κάνει το κινητό μας, μέχρι και στον τρόπο διάγνωσης και θεραπείας στον τομέα της ιατρικής. Η τεχνητή νοημοσύνη είναι μέρος της ζωής μας και πρέπει να το αποδεχτούμε, βέβαια αυτό δεν σημαίνει ότι πρέπει να δεχτούμε και τα λάθη τα οποία έρχονται μαζί της. Όπως έχουμε αναφέρει παραπάνω, υπάρχουν ομάδες και οργανισμοί οι οποίοι αφιερώνουν όλο τους τον χρόνο για τη δημιουργία συστημάτων τεχνητής νοημοσύνης, τα οποία να είναι ηθικά σωστά και να μην δημιουργούν συγκρούσεις μεταξύ χρηστών και προγραμματιστών. Μια τέτοια ομάδα είναι το FAT το οποίο έχει κάνει πολύ καλή δουλειά στο να δημιουργήσει ένα ευρύ δίκτυο το οποίο εμπεριέχει επιστήμονες από πολλούς τομείς. Βέβαια η επιμόρφωση του κόσμου δεν είναι κάτι το οποίο μπορεί να συμβεί από τη μια στιγμή στην άλλη, γίνεται σταδιακά και αντιμετωπίζονται πολλές δυσκολίες κατά τη διάρκεια αυτής της διαδικασίας. Συνεπώς πρέπει αν υπάρχουν δυνατές βάσεις και πόροι για να μπορέσουν να συνεχίσουν το έργο τους. Ένα έργο το οποίο μπορεί να είναι η αρχή της δημιουργίας ενός συστήματος AI το οποίο θα είναι πιο κοντά στην ανθρώπινη νοημοσύνη. Βέβαια προς το παρόν αυτό δεν είναι κάτι το οποίο θα πραγματοποιηθεί στο κοντινό μέλλον καθώς σε πολλούς τομείς υπάρχουν προβλήματα τα οποία δημιουργούνται από αλγορίθμους AI. Όπως έχουμε αναφέρει στη παρούσα πτυχιακή εργασία, ο τομέας της ιατρικής είναι πολύ ευαίσθητος καθώς έχει να κάνει άμεσα με τις ανθρώπινες ζωές, κάτι το οποίο κάνει τον κόσμο πιο διστακτικό όσο αφορά τη χρήση τεχνητής νοημοσύνης. Προσωπικά μου φαίνεται απολύτως λογικό καθώς δεν είναι λίγες οι φορές στις οποίες τα ΄΄ευφυή'' συστήματα έχουν κάνει λάθος. Όπως έχουμε αναφέρει παραπάνω, έχουν υπάρξει περιπτώσεις στις οποίες η τεχνητή νοημοσύνη χρησιμοποιήθηκε χωρίς κάποια σοβαρή επίβλεψη και αυτό οδήγησε σε πολλά προβλήματα. Αντίστοιχα πρέπει να βρεθούν αποτελεσματικοί τρόποι διάγνωσης που να μην έχουν προκαταλήψεις και να είναι δίκαιοι προς όλους τους χρήστες της. Καθώς αυτό είναι ένα πρόβλημα το οποίο αντιμετωπίζεται σε όλους τους τομείς της ιατρικής, είτε είναι η οφθαλμολογία, είτε είναι η παθολογία, ακόμα και σε περιπτώσεις που έχουν να κάνουν με άτομα που έχουν

νευροποικιλομορφίες. Συνεπώς είναι ένα πρόβλημα για το οποίο απαιτείται να βρεθεί λύση άμεσα. Αντίστοιχα πρέπει οι γιατροί να αρχίσουν να εξοικειώνονται στις αλλαγές που πραγματοποιούνται στη τεχνολογία, διότι αν δεν το κάνουν αυτό θα αντικατασταθούν με άτομα τα οποία μπορούν να χρησιμοποιήσουν τα καινούργια συστήματα. Επιπρόσθετα ένας ασθενείς θα εμπιστευτεί πολύ πιο εύκολα ένα γιατρό ο οποίος μπορεί να κάνει άριστη χρήση των συστημάτων τεχνητής νοημοσύνης, καθώς αν εμπιστεύεσαι τον γιατρό που επιβλέπει την εξέταση μπορείς συνεπώς να εμπιστευτείς και το σύστημα που χρησιμοποιεί. Βέβαια εφόσον υπάρχουν τόσες ανησυχίες και δυσκολίες στο τομέα της ιατρικής όσο αφορά την τεχνητή νοημοσύνη, θα ήταν καλό το AI να μην ενταχθεί εντελώς στον τομέα αυτόν, καθώς θα δημιουργήσει πολλές συγκρούσεις μεταξύ ασθενών και γιατρών, ειδικά σε χώρες και πολιτισμούς που δεν είναι ανοιχτοί σε αλλαγές. Οπότε πρέπει ιδανικά να γίνει εκτενής χρήση της τεχνητής νοημοσύνης σε τομείς οι οποίοι να μην είναι τόσο ευαίσθητοι όσο είναι ο υγειονομικός τομέας. Διότι υπάρχει και η περίπτωση του μαύρου κουτιού, στην οποία περίπτωση δεν υπάρχει πρόσβαση στο σύστημα και δεν μπορεί να ξέρει κανείς πως παίρνει αποφάσεις ο εκάστοτε αλγόριθμος, κάτι το οποίο είναι λάθος για πολλούς ηθικούς λόγους. Επιπρόσθετα πρέπει να λάβει κανείς υπόψη το πόσο ασφαλείς είναι η τεχνητή νοημοσύνη από εισβολείς, κάτι το οποίο είναι πολύ πιθανό να συμβεί σε κάποιες εφαρμογές της. Όπως είναι στον τομέα του εμπορίου, στον οποίο πολλοί κακόβουλοι έμπορες θα προσπαθήσουν να σαμποτάρουν τους ανταγωνιστές για να έχουν πλεονέκτημα στις πωλήσεις. Κάτι το οποίο μπορεί να επιτευχθεί με διάφορους τρόπους, καθώς ο εισβολέας μπορεί να επέμβει στη διαδικασία εκπαίδευσης του αλγορίθμου, κάτι το οποίο καθιστά τον αλγόριθμο αναξιόπιστο. Συνεπώς πρέπει να λαμβάνουμε υπόψη όλα τα προαναφερθέντα πριν εντάξουμε τη τεχνητή νοημοσύνη σε ορισμένες εφαρμογές. Βέβαια κατά τη γνώμη μου η τεχνητή νοημοσύνη είναι μια πολύ σημαντική τεχνολογία η οποίο μπορεί να βοηθήσει άπειρες ζωές με διάφορους τρόπους. Συνεπώς πρέπει να ενταχθεί σε περισσότερους τομείς της ζωής μας σταδιακά με τα απαραίτητα μέτρα ασφαλείας, κάτι το οποίο θα καθιστά το σύστημα αξιόπιστο μέχρι ένα βαθμό και ηθικά σωστό.
Εν κατακλείδι η τεχνητή νοημοσύνη μπορεί να δημιουργήσει πολλές αμφιβολίες όσο αφορά τον τομέα της ηθικής, αλλά αυτό δεν σημαίνει ότι πρέπει να αποκρύπτουμε ότι μας προβληματίζει. Αντιθέτως πρέπει να επικεντρωθούμε πιο πολύ σε αυτήν καθώς είναι το μέλλον της τεχνολογίας και των αυτοματοποιημένων διεργασιών.

Βιβλιογραφία:

[1] The security of machine learning, Marco Barreno · Blaine Nelson · Anthony D. Joseph · J.D. Tygar.

[2] The Role of Transparency in Recommender Systems, Rashmi Sinha & Kirsten Swearingen.

[3] Ethics of AI in Pathology: Current Paradigms and Emerging Issues, Chhavi Chauhan, Rama R. Gullapalli.

[4] AI ethics in computational psychiatry: From the neuroscience of consciousness to the ethics of consciousness, Wanja Wiese, Karl J. Friston .

[5] Fairness and accountability of AI in disaster risk management: Opportunities and challenges, Caroline M. Gevart, Mary Carman, Benjamin Rosman, Yola Georgiadou, Robert Soden.

[6] A high-level overview of AI ethics, Kazim E. , & koshiyama A. S. (2021)

[7] The role of empathy for artificial intelligence accountability, ramya Srinivasan, Beatriz San Miguel Gonzalez.

[8] What is the point of fairness?, Bennett C. , Keyes O.

[9] Ethics of Artificial Intelligence in Medicine and Ophthalmology

[10] Ethical Dilemma of Artificial Intelligence and its Research Progress, Lei Ma, Zhongqiu Zhang, Nana Zhang, IOP Conference Series: Materials Science and Engineering

[11] Can Machine Learning Be Secure? , Marco Barreno, Blaine Nelson, Russell Sears, Anthony D. Joseph, J. D. Tygar.

[12] Gannon M : Race is a social construct, scientists argue. Sci Am 2016

[13] Intersectional accuracy disparities in commercial gender classification. Proceeding of the 1st conference on fairness, accountability and transparency. Edited by Sorelle AF, Christo W. 2018

[14] Dissecting racial bias in an algorithm used to manage the health of populations, Ziad Obermeyer, Brian Powers, Christine Vogeli , Sendhil Mullainathan.

[15] O'Neil C: Weapons of Math Destruction: how Big Data Increases inequality and Threatens Democracy. New York, NY, Crown Publishing Group, 2016

[16] Hallworth MJ: The ‘70% claim’: what is the evidence base? Ann Clin Biochem 2011.

[17] IBM Watson Analytics: Automating Visualization, Descriptive, and Predictive Statistics, Robert E Hoyt, Dallas H Snider, Carla J Thompson, Sarita Mantravadi.

[18] Deep patient: an unsupervised representation to predict the future of patients from the electronic health records, Miotto R. , Li L. , Kidd BA. , Dudley JT.

[19] A Practical Guide to Whole Slide Imaging: A White Paper From the Digital Pathology Association, Mark D. Zarella, Douglas Bowman, Famke Aeffner, Navid Farahani, Albert Xthona, Syeda Fatima Absar, Anil Parwani, Marilyn Bui, MD, Douglas J. Hartman.

[20] The Myth of Artificial Intelligence, Erik J. Larson

[21] Artificial intelligence: A guide for thinking humans, Melanie Mitchell

[22] The future of employment: How susceptible are jobs to computerisation? , Carl Benedikt Frey, Michael A.Osborne.

[23] Implementation of eHealth and AI integrated diagnostics with multidisciplinary digitized data: are we ready from an international perspective?, Mark Bukowski, Robert Farkas, Oya Beyan, Lorna Moll, Horst Hahn, Fabian Kiessling & Thomas Schmitz-Rode.

[24] Ethical Challenges of Big Data in Public Health, Effy Vayena, Marcel Salathé, Lawrence C. Madoff, John S. Brownstein.

[25] Machine Learning in Medicine, Alvin Rajkomar, M.D., Jeffrey Dean, Ph.D., and Isaac Kohane, M.D., Ph.D.

[26] The Legal And Ethical Concerns That Arise From Using Complex Predictive Analytics In Health Care, I. Glenn Cohen, Ruben Amarasingham, Anand Shah, Bin Xie, Bernard Lo.

[27] Digital pathology and artificial intelligence, Muhammad Khalid Khan Niazi, Anil V Parwani MD , Metin N Gurcan PhD.

[28] Deep learning for healthcare: review, opportunities and challenges, Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, Joel T Dudley.

[29] Emerging Themes in Image Informatics and Molecular Analysis for Digital Pathology, Rohit Bhargava, Anant Madabhushi.

[30] Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine, Castaneda C. , nalley K. , mannion C. , Bhattachayya P. , Blake P. , Pecora A.

[31] A Roadmap for Foundational Research on Artificial Intelligence in Medical Imaging: From the 2018 NIH/RSNA/ACR/The Academy Workshop, Curtis P. Langlotz , Bibb Allen, Bradley J. Erickson, Jayashree Kalpathy-Cramer, Keith Bigelow, Tessa S. Cook, Adam E.

Flanders, Matthew P. Lungren, David S. Mendelson, Jeffrey D. Rudie, Ge Wang, Krishna Kandarpa.

[32] Toward an ethics of algorithms: Convening, observation, probability, and timeliness,

Mike Ananny.

[33] Anticipatory Ethics for Emerging Technologies, Philip A. E. Brey.

[34] Considerations for ethics review of big data health research: A scoping review, Marcello Ienca , Agata Ferretti, Samia Hurst, Milo Puhan, Christian Lovis, Effy Vayena.

[35] Artificial neural networks in medical diagnosis , FilippoAmato, AlbertoLópez, Eladia MaríaPeña-Méndez, PetrVaňhara, Aleš Hampl, JosefHavel.

[36] High-performance medicine: the convergence of human and artificial intelligence, Eric J. Topol.

[37] Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction, Seong Ho Park , Kyunghwa Han.

[38] Deep learning in ophthalmology: The technical and clinical considerations, Daniel S W Ting, Lily Peng, Avinash V Varadarajan, Pearse A Keane, Philippe M Burlina, Michael F Chiang, Leopold Schmetterer, Louis R Pasquale, Neil M Bresser, Dale R Webster, Michael Abramoff, Tien Y Wong.

[39] Biomedical big data: new models of control over access, use and governance. Vayena E, Blasimme A.

[40] Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings? . Wahl B, Cossy-Gantner A, Germann S.

[41] Artificial intelligence powers digital medicine. Fogel AL, Kvedar JC.

[42] Promoting trust between patients and physicians in the era of artificial intelligence. Nundy S, Montgomery T, Wachter RM.

[43] Artificial Pain: Empathy, Morality, and Ethics as a Developmental Process of Consciousness. Asada M.

[44] Potential liability for physicians using artificial intelligence. Price WN, Gerke S, Cohen IG.

[45] From Jeopardy to Jaundice: The medical liability implications of Dr. Watson and other artificial intelligence systems. Allain JS.

[46] Computational psychiatry: the brain as a phantastic organ, Lancet Psychiatry , K. Friston, K.E. Stephan, R. Montague, R.J. Dolan.

[47]] Advances in the computational understanding of mental illness, Neuropsychopharmacology , Q.J.M. Huys, M. Browning, M.P. Paulus, M.J. Frank.

[48] Machine learning for precision psychiatry: opportunities and challenges, Biol. Psychiatry Cogn. Neurosci. Neuroimaging, D. Bzdok, A. Meyer-Lindenberg.

[49] Toward a mature science of consciousness, W. Wiese.

[50] The science of consciousness does not need another theory, it needs a minimal unifying model, W. Wiese.

[51]] Principles of Biomedical Ethics, 8th ed., Oxford University Press, T.L. Beauchamp, J.F. Childress.

[52] The ethics of ai ethics: an evaluation of guidelines, T. Hagendorff.

[53] A survey on computer vision for assistive medical diagnosis from faces, Thevenot, J., Bordallo López, M., Hadid.

[54] Autism= death: The social and medical impact of a catastrophic medical model of autistic spectrum disorders, Waltz, M.

[55] The Case for an Ethical Black Box, Alan F.T. Winfield, Marina Jirotka