ePublishing

Ετήσιο Ελληνόφωνο Επιστημονικό Συνέδριο Εργαστηρίων Επικοινωνίας

Τόμ. 2, Αρ. 1 (2023)

2ο Ετήσιο Ελληνόφωνο Επιστημονικό Συνέδριο Εργαστηρίων Επικοινωνίας: Το μέλλον της επικοινωνίας στην ψηφιακή εποχή

Τεχνητή νοημοσύνη και συμμετοχικά μέσα: Τεχνολογίες, δεοντολογία και η ανάγκη γραμματισμού

Θεοδώρα Σαρίδου, Χαράλαμπος Δημούλας

doi: 10.12681/cclabs.6441

Βιβλιογραφική αναφορά:

Σαρίδου Θ., & Δημούλας Χ. (2024). Τεχνητή νοημοσύνη και συμμετοχικά μέσα: Τεχνολογίες, δεοντολογία και η ανάγκη γραμματισμού. *Ετήσιο Ελληνόφωνο Επιστημονικό Συνέδριο Εργαστηρίων Επικοινωνίας*, *2*(1), 1–11. https://doi.org/10.12681/cclabs.6441

Τεχνητή νοημοσύνη και συμμετοχικά μέσα: Τεχνολογίες, δεοντολογία και η ανάγκη γραμματισμού

Θεοδώρα Σαρίδου
Διδάκτορας, Τμήμα Δημοσιογραφίας και ΜΜΕ, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
saridout@jour.auth.gr

Χαράλαμπος Δημούλας
Καθηγητής, Τμήμα Δημοσιογραφίας και ΜΜΕ, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
babis@eng.auth.gr

Περίληψη

Η αυξανόμενη ενσωμάτωση εφαρμογών τεχνητής νοημοσύνης σε όλες σχεδόν τις πτυχές της καθημερινής, ακαδημαϊκής και εν γένει επαγγελματικής ζωής καθιστά ιδιαίτερα επίκαιρη την ανάγκη μελέτης, όχι μόνο των τεχνολογιών στις οποίες εδράζεται, αλλά και των ζητημάτων δεοντολογίας και γραμματισμού που αναπόφευκτα ανακύπτουν. Πλήθος συμμετοχικών περιβαλλόντων αναδιαμορφώνονται υπό το φως των ταχύτατα εξελισσόμενων δυνατοτήτων της τεχνητής νοημοσύνης, και στα οποία δεσπόζουν δημοσιογραφικά εργαλεία και μέσα επικοινωνίας. Η παρούσα εργασία έχει ως σκοπό να μελετήσει τις προκλήσεις που γεννούν τα νέα αυτά δεδομένα, επισημαίνοντας τους κινδύνους που ελλοχεύουν, τόσο στο αρχικό στάδιο του σχεδιασμού των λειτουργιών όσο και μετέπειτα κατά την εφαρμογή των υπηρεσιών. Στη συνέχεια, επιχειρεί να εξετάσει το δεοντολογικό πλαίσιο που διαμορφώνεται σταδιακά σε διεθνές επίπεδο κατά τη διάρκεια των τελευταίων ετών. Τέλος, τονίζεται ο σημαίνων ρόλος του γραμματισμού και της εκπαίδευσης, με στόχο τη διεπιστημονική προσέγγιση κατά την ανάλυση και κατανόηση της τεχνητής νοημοσύνης.

Λέξεις-κλειδιά: τεχνητή νοημοσύνη, δεοντολογία, γραμματισμός, δημοσιογραφία

1. Εισαγωγή

Κατά τη διάρκεια των τελευταίων δεκαετιών, η Τεχνητή Νοημοσύνη (ΤΝ) έχει αποτελέσει ένα σημαντικό επιστημονικό αντικείμενο μελέτης, επηρεάζοντας βαθιά ένα ευρύ φάσμα ακαδημαϊκών και βιομηχανικών τομέων. Το πεδίο της ΤΝ έχει αναπτυχθεί σχεδόν εκρηκτικά τα τελευταία χρόνια, παράλληλα με τα συμμετοχικά εργαλεία και τα περιβάλλοντα των μέσων. Στην εποχή αυτή, που χαρακτηρίζεται από τις κατακερματισμένες ροές πληροφοριών και τις τεράστιες ποσότητες ακατέργαστων δεδομένων, οι υπολογιστικές εξελίξεις μαζί με τις κοινωνικοοικονομικές αλλαγές διευκολύνουν την ενσωμάτωση των τεχνολογιών ΤΝ σε ένα εντυπωσιακά μεγάλο εύρος τομέων, από τα μαθηματικά, τη μηχανική και την ιατρική επιστήμη έως την ψυχολογία, την εκπαίδευση, τα μέσα ενημέρωσης και τις επικοινωνίες (Dimoulas, 2018; Dimoulas, 2022; Dimoulas & Veglis, 2023; Katsaounidou et al., 2019). Την ίδια στιγμή, ποικίλες πτυχές της ανθρώπινης καθημερινότητας διαμορφώνονται επίσης υπό την κινητήρια δύναμη των εργαλείων και των συστημάτων ΤΝ. Εφαρμογές που βασίζονται σε τεχνικές μηχανικής/εμβαθύνουσας μάθησης (machine learning - ML/deep learning - DL) και επεξεργασίας φυσικής γλώσσας (natural language processing - NLP) διαδραματίζουν όλο και περισσότερο έναν ζωτικό ρόλο στη ζωή, τη μάθηση, την εργασία και τη συνύπαρξη σε συνεργατικά και συμμετοχικά περιβάλλοντα (Jiang et al., 2022; Kotsakis et

al., 2023; Vrysis et al., 2021). Λαμβάνοντας, μάλιστα, υπόψη ότι η μηχανική μάθηση βασίζεται στη συλλογή δεδομένων από πραγματικούς χρήστες ή/και γεγονότα, γίνεται αντιληπτό ότι ελλοχεύουν θέματα ασφάλειας, ιδιωτικότητας και αρκετά ερωτήματα ηθικής ως προς τη συμπεριφορά και τις αποκρίσεις αυτών των συστημάτων.

Σκοπός της παρούσας εργασίας είναι να συμβάλει στη θεωρητική συζήτηση που δομείται γύρω από την παραδοχή ότι, αν και η χρήση τέτοιων αλγοριθμικών προσεγγίσεων και τεχνολογιών προσφέρει σημαντικά οφέλη στους τομείς της δημοσιογραφίας και της επικοινωνίας, η ανάγκη να ληφθούν υπόψη οι κίνδυνοι και οι προκλήσεις που προκύπτουν είναι έντονη. Προς την κατεύθυνση αυτή επιχειρείται αρχικά η εννοιολογική προσέγγιση της ΤΝ και καταγράφονται οι βασικοί ορισμοί που συμβάλλουν στην κατανόηση του πεδίου. Μέσα από συγκεκριμένα παραδείγματα αποτυπώνεται αφενός η ενσωμάτωση των τεχνολογιών ΤΝ στη δημοσιογραφική πρακτική, αφετέρου δε τα ζητήματα νομιμότητας, ηθικής και δεοντολογίας που ανακύπτουν. Τέλος, επισημαίνεται η ανάγκη μίας διεπιστημονικής προσέγγισης, ώστε να διαμορφωθούν οι προϋποθέσεις για τη βαθύτερη κατανόηση και την επωφελή χρήση της ΤΝ χωρίς να παραβλέπεται το συνεχώς εξελισσόμενο (τεχνολογικό) τοπίο.

2. Τι είναι η τεχνητή νοημοσύνη – Βασικοί ορισμοί & Παραδείγματα

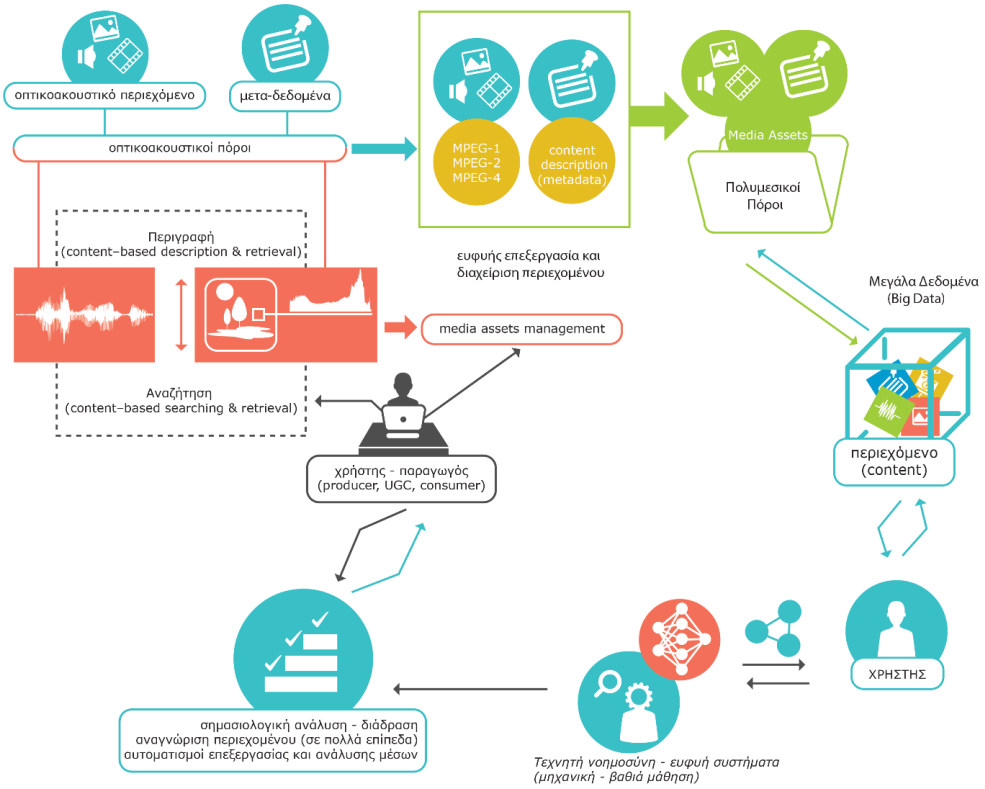
Ο όρος Τεχνητή Νοημοσύνη αναφέρεται σε κλάδο της Επιστήμης Υπολογιστών (computer science) που εστιάζει στη σχεδίαση και υλοποίηση ‘μηχανών’, δηλαδή υπολογιστικών συστημάτων, που μιμούνται στοιχεία της ανθρώπινης συμπεριφοράς και διαθέτουν στοιχειώδη ευφυΐα όπως μάθηση, προσαρμοστικότητα, εξαγωγή συμπερασμάτων, κατανόηση από συμφραζόμενα, δηλαδή αποτελούν έμπειρα/έξυπνα συστήματα (expert/smart systems) επίλυσης προβλημάτων (Δημούλας, 2023). Η Μηχανική Μάθηση αποτελεί υπο-πεδίο της ΤΝ και περιλαμβάνει μια οικογένεια αλγορίθμων μάθησης από δεδομένα (Learning by example), δηλαδή συστημάτων που μαθαίνουν να επιλύουν προβλήματα (π.χ. να αναγνωρίζουν συγκεκριμένα πρότυπα κειμένων, ήχων, εικόνων κλπ.) με βάση παραδείγματα, χωρίς να δίνονται κανόνες αναγνώρισης, ταξινόμησης, επίλυσης κλπ. (οι όποιοι κανόνες διαμορφώνονται αυτόματα, κατά την «μελέτη»/ανάλυση των παραδειγμάτων και, συχνά, δεν είναι ορατοί από τους χρήστες) (Dimoulas, 2018). Τα τελευταία χρόνια έχουν γίνει σημαντικά βήματα στον χώρο αυτό με την εμφάνιση της βαθιάς ή εμβαθύνουσας μάθησης, των Συνελικτικών Νευρωνικών Δικτύων (Convolutional Neural Networks - CNN), αλλά και της Παραγωγικής Τεχνητής Νοημοσύνης, με κύριους εκφραστές τα Παραγωγικά Αντιπαλικά Δίκτυα (Generative Adversarial Networks - Gans (Goodfellow et al., 2016; Meor Amer, 2021; Vrysis et al., 2020; Δημούλας, 2023; Καμπουρλάζος & Παπακώστας, 2015). Τα δίκτυα αυτά διαθέτουν πολύπλοκες αρχιτεκτονικές δικτύων που μπορούν να εκπαιδεύονται για την επίλυση ιδιαίτερα «δύσκολων» προβλημάτων, αρκεί να τροφοδοτηθούν από ένα μεγάλο αριθμό δειγμάτων δεδομένων (με την κατάλληλη επισημείωση/annotation) που να αποτελούν αντιπροσωπευτικό δείγμα του συνολικού/φυσικού χώρου (δηλαδή των παραδειγμάτων που μπορεί να συναντήσει κανείς στον φυσικό κόσμο).

Ας μην ξεχνάμε ότι ζούμε στην εποχή των κοινωνικών μέσων και των έξυπνων κινητών, όπου ο κάθε χρήστης μπορεί να δημιουργεί και διαμοιράζει περιεχόμενο χωρίς πρακτικούς περιορισμούς (εξ' ου και ο όρος Μεγάλα Δεδομένα - Big Data). Έτσι, έχοντας αυτούς τους τεράστιους όγκους περιεχομένου που παράγεται καθημερινά, ένα ερώτημα που προκύπτει είναι το πώς μπορούμε να αναγνωρίσουμε, τεκμηριώσουμε, ταξινομήσουμε και διαχειριστούμε όλα αυτά τα δεδομένα. Γι’ αυτό και το κύριο τεχνολογικό και ερευνητικό

ενδιαφέρον έχει περάσει από τις τεχνικές συμπίεσης στα συστήματα διαχείρισης περιεχομένου, μέσω και της χρήσης ευφυών συστημάτων, όπως αποδίδεται σχηματικά στο Διάγραμμα 1, τόσο για παραγωγούς περιεχομένου όσο και απλούς χρήστες (Δημούλας, 2015).

Διάγραμμα 1

Σχηματική αναπαράσταση χρήσης ΤΝ στην ευφυή διαχείριση περιεχομένου (Δημούλας, 2015)



3. Ιστορική εξέλιξη, σύγχρονες τάσεις, προκλήσεις και απειλές της Τεχνητής Νοημοσύνης

Η Τεχνητή Νοημοσύνη πρωτοεμφανίστηκε στα τέλη της δεκαετίας του 1960, όπου επιχειρήθηκε και ο πρώτος ορισμός από τον Arthur Samuel, ο οποίος αναφέρθηκε σε μηχανές που έχουν τη δυνατότητα να εκπαιδεύονται και να μαθαίνουν, αλλά χωρίς να τους δίνονται συγκεκριμένες οδηγίες αντιμετώπισης των αντίστοιχων προβλημάτων και χωρίς να προγραμματίζονται με κανόνες επίλυσης (Dimoulas, 2018; Δημούλας, 2023). Από τότε έχει σημειωθεί τεράστια πρόοδος μέχρι να φτάσουμε στη σημερινή εποχή και τον θόρυβο που έχει προκληθεί γύρω από την Παραγωγική Τεχνητή Νοημοσύνη, με πιο χαρακτηριστικό το πρόσφατο παράδειγμα του ChatGPT (Δημούλας, 2023). Οι λόγοι που εμφανίστηκε τώρα αυτή η εκθετική πρόοδος των δυνατοτήτων αυτών των συστημάτων σχετίζεται με την εξέλιξη των αλγοριθμικών τεχνικών και των νέων υπολογιστικών και αποθηκευτικών δυνατοτήτων. Σήμερα, έχουμε τη δυνατότητα να αποθηκεύουμε και διαχειριζόμαστε τεράστια σύνολα δεδομένων (datasets) με κατάλληλες ετικέτες (labelling, annotation) που μπορούν να χρησιμοποιηθούν σε προβλήματα εκπαίδευσης μηχανικής και βαθιάς μάθησης. Επιπλέον, υπάρχει η απαιτούμενη υπολογιστική ισχύς για την επεξεργασία όλων αυτών των δεδομένων και την εκτέλεση των αντίστοιχων συστημάτων (Goodfellow et al., 2016; Meor Amer, 2021; Δημούλας, 2023). Τα περισσότερα (και δη τα αρχικά) συστήματα ΤΝ παρέχουν λίγες

πληροφορίες στον τελικό χρήστη ως προς τις διεργασίες που εφαρμόζουν, δηλαδή συνήθως λογίζονται ως «μαύρα κουτιά» που δέχονται συγκεκριμένες εισόδους και δίνουν τις ζητούμενες εξόδους. Μια νέα τάση που κερδίζει ολοένα και περισσότερο έδαφος είναι η επονομαζόμενη Ερμηνεύσιμη Τεχνητή Νοημοσύνη (Interpretable AI), που θα παρέχει/απεικονίζει χρήσιμες πληροφορίες γύρω από τους μηχανισμούς λήψης των αποφάσεων, επιτρέποντας τους τελικούς χρήστες να κατανοήσουν τις συνολικότερες διαδικασίες επεξεργασίας δεδομένων και εξαγωγής των αποτελεσμάτων (Molnar, 2020; Δημούλας, 2023).

Από την παραπάνω ανάλυση προκύπτουν και κάποιες απειλές ή κίνδυνοι από τη χρήση της ΤΝ και των συστημάτων μηχανικής/βαθιάς μάθησης. Γίνεται ξεκάθαρο ότι τα συστήματα αυτά ελέγχονται από μεγάλους τεχνολογικούς κολοσσούς που έχουν τη δυνατότητα να διατηρούν μεγάλα κέντρα δεδομένων (data centers) με την απαιτούμενη υπολογιστική ισχύ (processing power). Επίσης, γεννιούνται ερωτηματικά ως προς την προέλευση και τους κανόνες συλλογής των δεδομένων, και κατά πόσο αυτά είναι αντιπροσωπευτικά των υπό μελέτη φαινομένων/προβλημάτων ή τυχόν εμφανίζουν κάποιες εγγενείς πολωτικές τάσεις. Ένα άλλο ερώτημα αφορά τη διασφάλιση ότι ένα σύστημα που έχει αναπτυχθεί για κάποιο συγκεκριμένο σκοπό (π.χ. αναγνώριση προσώπων σε ένα τηλεοπτικό πλατό) δεν θα χρησιμοποιηθεί και σε εφαρμογές αμφιβόλου ηθικής που θέτουν σε κίνδυνο την ιδιωτικότητα ατόμων, δεδομένων, επικοινωνιών (π.χ. παρακολούθηση φυσικών προσώπων).

4. Τεχνητή νοημοσύνη στη δημοσιογραφία

4.1. Διαχείριση της συμμετοχής του κοινού στην παραγωγή ειδήσεων

Στη διάρκεια των τελευταίων ετών οι πολίτες έχουν τη δυνατότητα να συμμετέχουν στη δημοσιογραφία και τον δημόσιο διάλογο μέσω εργαλείων και εφαρμογών που υιοθετούνται από τα μέσα (Engelke, 2019; Krumsvik, 2018; Loosen et al., 2022). Η διαχείριση, ωστόσο, των συνεισφορών του κοινού δεν είναι απρόσκοπτη, καθώς οι δημοσιογράφοι καλούνται να τις ενσωματώσουν στην εργασιακή ρουτίνα της αίθουσας σύνταξης, προλαμβάνοντας ή/και αντιμετωπίζοντας την ίδια στιγμή τους κινδύνους και τα προβλήματα που ελλοχεύουν (Saridou & Veglis, 2021). Το εύρος της σκοτεινής συμμετοχής (Quandt, 2018) εκτείνεται από απλά ζητήματα, όπως τα ορθογραφικά και συντακτικά λάθη (Saridou et al., 2019), έως και σοβαρότερα, όπως οι προσωπικοί χαρακτηρισμοί, οι ύβρεις και η χυδαιότητα (Ksiazek et al., 2015). Στους χώρους αυτούς εντοπίζονται περιπτώσεις διαδικτυακού εκφοβισμού, απειλές βίας και παραβιάσεις της ιδιωτικής ζωής (Citron, 2014; Festl & Quandt, 2017). Έρευνες έχουν καταγράψει, επίσης, φαινόμενα παραπληροφόρησης και διάδοσης ψευδών ειδήσεων ή θεωριών συνωμοσίας (Frischlich et al., 2019; Quandt, 2018). Παράλληλα, η ρητορική μίσους πλήττει καίρια τον δημόσιο λόγο, τόσο μέσα από τα κυρίαρχα μέσα ενημέρωσης όσο και μέσα από τις εφαρμογές κοινωνικής δικτύωσης (Quandt, 2018), στρεφόμενη και εναντίον των ίδιων των δημοσιογράφων (Obermaier et al., 2018).

Προκειμένου να αποφύγουν τέτοιες προβληματικές καταστάσεις και να μην αξιολογηθούν αρνητικά από το κοινό τους (Tenenboim et al., 2019), οι ειδησεογραφικοί οργανισμοί ακολουθούν στρατηγικές με σκοπό την επίβλεψη, τον έλεγχο, τον συντονισμό και τη διαχείριση των συνεισφορών (Wintterlin et al., 2020). Προς την κατεύθυνση αυτή, τα περισσότερα μέσα εφαρμόζουν μεθόδους χειροκίνητης εποπτείας, κατά την οποία ελέγχουν το περιεχόμενο των χρηστών πριν (προ-εποπτεία) ή μετά τη δημοσίευσή του (μετα-εποπτεία) (Crawford & Gillespie, 2016; Goodman, 2013; Naab et al., 2018). Ωστόσο, η μη αυτόματη εποπτεία απαιτεί σημαντικούς πόρους σε ανθρώπινο δυναμικό, χρόνο και προϋπολογισμό,

καθώς ένας μεγάλος αριθμός επαγγελματιών καλείται να επωμιστεί και αυτό το καθήκον, παράλληλα με τις υπόλοιπες δημοσιογραφικές δραστηριότητες (Wang, 2020).

4.2. Αξιοποίηση της τεχνητής νοημοσύνης από τους δημοσιογραφικούς οργανισμούς

Με σκοπό να διαχειριστούν αποτελεσματικότερα και ταχύτερα το περιεχόμενο που παράγεται από τους χρήστες, όλο και περισσότεροι δημοσιογραφικοί οργανισμοί στρέφονται σε αυτοματοποιημένες μεθόδους και σε τεχνικές ΤΝ. Συχνά τίθεται από τους διαδικτυακούς τόπους ως προϋπόθεση για τη συμμετοχή των χρηστών η συμπλήρωση ενός πλήρως αυτοματοποιημένου τεστ, ώστε να αποτραπούν τα ρομπότ (bots) από τη διάπραξη κακόβουλων ενεργειών και να εμποδιστούν οι μαζικές καταχωρήσεις που προέρχονται από υπολογιστή (Saridou & Veglis, 2016; Sivakorn et al., 2016). Τα CAPTCHA (Completely Automated Public Turing tests to tell Computers and Humans Apart) απαιτούν συνήθως από τους χρήστες να αναγνωρίσουν και να πιστοποιήσουν μία παραμορφωμένη εικόνα που αποτελείται από γράμματα και αριθμούς. Η πιο διαδεδομένη υπηρεσία του είδους είναι η reCAPTCHA της Google, η οποία δομείται πάνω σε ένα προηγμένο σύστημα ανάλυσης κινδύνου. Η υπηρεσία αξιολογεί το αίτημα του κάθε χρήστη και θέτει μία πρόκληση αντίστοιχης δυσκολίας, όπως η επιλογή συγκεκριμένων εικόνων από ένα σύνολο παρόμοιων φωτογραφιών (Sivakorn et al., 2016).

Επιπρόσθετα, η στοχευμένη αντιμετώπιση των διαφορετικών φαινομένων σκοτεινής συμμετοχής καθίσταται εφικτή μέσω υπολογιστικών τεχνικών για αυτόματη ανάλυση και αναπαράσταση της ανθρώπινης γλώσσας (Cambria & White, 2014). Για τον εντοπισμό της ρητορικής μίσους, για παράδειγμα, δεν αρκούν τα βασικά φίλτρα που αποκλείουν λέξεις ή φράσεις, αλλά λειτουργεί πιο αποτελεσματικά η επεξεργασία φυσικής γλώσσας. Στις περιπτώσεις αυτές, πέρα από το καθαρά λεκτικό μέρος του προς εξέταση μηνύματος, λαμβάνονται υπόψη και περιφερειακοί παράγοντες, όπως το ευρύτερο πλαίσιο λόγου, οι εικόνες και τα βίντεο που συνοδεύουν το κείμενο, η χρονική στιγμή της ανάρτησής του, καθώς και η ταυτότητα του συγγραφέα και του παραλήπτη (Schmidt & Wiegand, 2017). Οι κακόβουλοι χρήστες, βέβαια, μπορούν να παρακάμψουν και αυτά τα συστήματα ανίχνευσης (Perifanos & Goutsos, 2021), χρησιμοποιώντας στατικές ή κινούμενες εικόνες και γραφικά που σκοπό έχουν να εμποδίσουν τον άμεσο εντοπισμό των μηνυμάτων μίσους από τα συστήματα επεξεργασίας φυσικής γλώσσας (Lamerichs et al., 2018; Matamoros-Fernández & Farkas, 2021). Υπάρχουν, τέλος, οργανισμοί μέσων που επιλέγουν ημιαυτόματη προσέγγιση, ενσωματώνοντας τη μηχανική εκμάθηση στη μη αυτόματη διαδικασία εποπτείας των σχολίων (Risch & Krestel, 2018).

Σταδιακά οι τεχνολογίες ΤΝ και ειδικότερα η μηχανική μάθηση αξιοποιούνται περισσότερο από τα μέσα σε όλα σχεδόν τα στάδια της ειδησεογραφικής παραγωγής, από την έναρξη της δημοσιογραφικής έρευνας έως τη συγκέντρωση, την επεξεργασία και τη διασταύρωση των δεδομένων (Underwood, 2019). Η αμερικανική εφημερίδα The New York Times χρησιμοποιεί το σύστημα Perspective της Google, με το οποίο εντοπίζονται αυτόματα τα τοξικά σχόλια μέσω προηγμένων τεχνικών κατά τις οποίες αλγόριθμοι μηχανικής μάθησης εκπαιδεύονται σε μεγάλα σώματα κειμένων (Binns et al., 2017). Με βάση περισσότερα από δεκαέξι εκατομμύρια σχόλια της ίδιας εφημερίδας, κάθε σχόλιο βαθμολογείται σε μία κλίμακα τοξικότητας από το μηδέν έως το εκατό ανάλογα με την ομοιότητά του με εκείνα που προσδιορίζονται ως τοξικά. Στη συνέχεια τα σχόλια είτε εγκρίνονται απευθείας είτε προωθούνται στους επόπτες για περαιτέρω ενέργειες (Etim, 2017; Wang, 2020).

Το Γαλλικό Πρακτορείο Ειδήσεων (Agence France-Presse) επικεντρώνεται στον έλεγχο και την επαλήθευση του περιεχομένου που προέρχεται από τους χρήστες, μέσα από την

πλατφόρμα AFP Fact Check. Για τον εντοπισμό της προέλευσης της πληροφορίας εφαρμόζεται ένας συνδυασμός μεθόδων, με πρώτη την έρευνα στα αρχεία των δημοσιογράφων του Πρακτορείου. Στην περίπτωση των εικόνων και των βίντεο, γίνεται αντίστροφη αναζήτηση σε μία ή περισσότερες μηχανές αναζήτησης, ώστε να διαπιστωθούν προηγούμενες εμφανίσεις τους στο διαδίκτυο. Παράλληλα, δίνεται προσοχή σε στοιχεία σχετικά με τον τόπο και τον χρόνο λήψης του υλικού, όπως οι επιγραφές των καταστημάτων, οι πινακίδες σήμανσης και η βλάστηση. Προς την κατεύθυνση αυτή αξιοποιείται και η επέκταση InVID/WeVerify, μία πλατφόρμα ανοιχτού κώδικα, στην οποία πολίτες και δημοσιογράφοι συνεργάζονται για τον εντοπισμό και την επαλήθευση περιεχομένου. Πέρα από τις παραπάνω μεθόδους, το Γαλλικό Πρακτορείο ακολουθεί και μία -σπάνια για τον χώρο των μέσων- εκπαιδευτική πολιτική. Σε αντίθεση με τα ερευνητικά αποτελέσματα που δείχνουν πως οι πάροχοι περιεχομένου στο διαδίκτυο επικεντρώνονται περισσότερο στην αποφυγή των νομικών συνεπειών από τη δημοσίευση κακόβουλου περιεχομένου και λιγότερο στην εκπαίδευση των χρηστών (Einwiller & Kim, 2020), το AFP Fact Check πραγματοποιεί εκπαίδευση ψηφιακής επαλήθευσης για ελεγκτές γεγονότων, δημοσιογράφους και μέλη του κοινού. Μέσα από βίντεο, εργαστήρια και επιμορφωτικές δράσεις επιδιώκει την εξοικείωση των χρηστών με τις διαδικασίες ελέγχου και επαλήθευσης του περιεχομένου, στοχεύοντας στον περιορισμό των ψευδών ειδήσεων και στην ενίσχυση της αξιόπιστης ενημέρωσης.

4.3. Ζητήματα δεοντολογίας και η ανάγκη γραμματισμού

Το αίτημα για δεοντολογικούς κώδικες στη χρήση της ΤΝ δεν αφορά μόνο το κομμάτι της εκπαίδευσης των μηχανών και τον σχεδιασμό των στοχευμένων λειτουργιών, αλλά και την ανάπτυξη και υλοποίηση των προβλεπόμενων υπηρεσιών (Minh et al., 2022). Για παράδειγμα, η απόκτηση των προσωπικών δεδομένων εκατομμυρίων χρηστών του Facebook από την εταιρεία Cambridge Analytica (Venturini & Rogers, 2019) ή ο ρόλος των «ρομπότ» του Twitter στις Προεδρικές Εκλογές των Ηνωμένων Πολιτειών Αμερικής το 2016 (Gorodnichenko et al., 2021) αποτελούν ορόσημα στη συνεχιζόμενη συζήτηση σχετικά με την κατάχρηση της ΤΝ. Παρομοίως, τα προβλήματα παραπληροφόρησης έχουν ενταθεί σημαντικά με τον πολλαπλασιασμό του συνθετικού περιεχομένου που παράγεται από μηχανές (generative content) μέσω των μοντέλων εμβαθύνουσας μάθησης, με αποτέλεσμα -μεταξύ άλλων- την εμφάνιση των λεγόμενων deep fakes (Meskys et al., 2020), τα οποία αποτελούν σοβαρές απειλές για τον κοινωνικό ιστό και τη δημοκρατία.

Στο πλαίσιο αυτό, οφείλουν να αποτελέσουν μέρος της θεωρητικής και της ερευνητικής μελέτης τα ζητήματα που αφορούν τη νομιμότητα, τη λογοδοσία και τη δικαιοσύνη (Ashok et al., 2022). Η ακεραιότητα των δεδομένων, το απόρρητο και τα πρωτόκολλα ασφάλειας είναι πάντα στο επίκεντρο όταν εμπλέκονται χρήστες και σύνολα δεδομένων (πληθοπορισμού). Παράλληλα, είναι θεμελιώδη τα ερωτήματα για την προέλευση των δεδομένων με τα οποία εκπαιδεύονται οι μηχανές και τη σημασία αυτής για τη διαφάνεια, την ιδιωτικότητα και τη δικαιοσύνη.

Προς αυτή την κατεύθυνση, οι διεθνείς αρχές έχουν πραγματοποιήσει ορισμένα αρχικά βήματα για την ανάπτυξη του απαραίτητου πλαισίου (Nasim et al., 2022). Συγκεκριμένα, η UNESCO δημοσίευσε τον Νοέμβριο του 2021 το πρώτο παγκόσμιο πρότυπο για την ηθική της ΤΝ, τη σύσταση Recommendation on the Ethics of Artificial Intelligence. Το πλαίσιο αυτό έχει ως ακρογωνιαίο λίθο την προστασία των ανθρωπίνων δικαιωμάτων και της αξιοπρέπειας, ενώ έχει υιοθετηθεί από τα 193 κράτη μέλη. Δομείται στη βάση έντεκα περιοχών πολιτικής δράσης, μεταξύ των οποίων οι δεοντολογικές επιπτώσεις, το περιβάλλον και τα

οικοσυστήματα, το φύλο, η εκπαίδευση, η έρευνα, η υγεία και η κοινωνική ευημερία. Επίσης το 2021, η Ευρωπαϊκή Επιτροπή πρότεινε το πρώτο ρυθμιστικό πλαίσιο της Ευρωπαϊκής Ένωσης για την Τεχνητή Νοημοσύνη (EU AI Act), το οποίο ορίζει διαφορετικούς κανόνες για διαφορετικά επίπεδα κινδύνου. Στις 14 Ιουνίου 2023 οι ευρωβουλευτές υιοθέτησαν τη διαπραγματευτική θέση του Κοινοβουλίου σχετικά με την παραπάνω πράξη, ενώ αναμένεται η έναρξη των συνομιλιών με τις χώρες της Ένωσης στο Συμβούλιο σχετικά με την τελική μορφή του νόμου.

Ωστόσο, η μεγάλη σημασία της διαμόρφωσης επακριβών ρυθμιστικών κατευθυντήριων γραμμών δεν θα πρέπει να οδηγεί στην παράβλεψη των εξίσου σημαντικών ηθικών ζητημάτων που ανακύπτουν. Για παράδειγμα: Ποιος έχει πρόσβαση σε βάσεις δεδομένων και σε μεγάλα κέντρα δεδομένων; Ποιος έχει πρόσβαση στις μεγάλες υπολογιστικές υποδομές; Και τελικά, πώς μπορεί να εξασφαλιστεί η ορθή και ωφέλιμη χρήση της ΤΝ, σε αντιπαραβολή με τις κακόβουλες χρήσεις της; Με δεδομένο ότι οι επιπτώσεις της ΤΝ βαίνουν ολοένα και πιο επιδραστικές, η ανάγκη διαφύλαξης των ανθρώπινων δικαιωμάτων και διασφάλισης της ποικιλομορφίας και της συμπερίληψης (Jobin et al., 2019) καθίσταται επιτακτική. Έτσι, η ιεράρχηση των πρωτοβουλιών ψηφιακού γραμματισμού και εκπαίδευσης θα πρέπει να περιλαμβάνει όλα τα εμπλεκόμενα μέρη (ενδιαφερόμενοι φορείς, προγραμματιστές, στοχευμένοι τελικοί χρήστες, επαγγελματίες των μέσων ενημέρωσης και επικοινωνίας, δημοσιογράφοι, κ.ά.). Τέλος, η ισότιμη πρόσβαση στα εργαλεία για όλους ώστε να μην αποτελέσει η ΤΝ προνόμιο λίγων, αλλά κυρίως η διεπιστημονική υποστήριξη σε όλα τα στάδια, από τον σχεδιασμό και την ανάπτυξη των εργαλείων ΤΝ έως την εφαρμογή και την ερμηνεία τους, μπορούν να αποτελέσουν την ασφαλιστική δικλείδα απέναντι σε τεχνοφοβικές αντιλήψεις και προσεγγίσεις. Στο πλαίσιο αυτό θα πρέπει να αναληφθούν διεπιστημονικές πρωτοβουλίες για την εφαρμογή της ΤΝ σε συνεργατικά περιβάλλοντα, μελετώντας και φωτίζοντας τις πτυχές αυτών των θεμάτων, όπως είναι η τεχνολογική, η κανονιστική, η ηθική και η ανάγκη υποστήριξης και γραμματισμού (Saridou & Dimoulas, 2024).

Χρηματοδότηση

Η παρούσα εργασία δεν έχει λάβει οποιαδήποτε χρηματοδότηση.

Αναφορές

Ξενόγλωσσες Αναφορές

Ashok, M., Madan, R., Joha, A., & Sivarajah, U. (2022). Ethical framework for Artificial Intelligence and digital technologies. *International Journal of Information Management*, 62. https://doi.org/10.1016/j.ijinfomgt.2021.102433

Binns, R., Veale, M., Van Kleek, M., Shadbolt, N. (2017). Like trainer, like bot? Inheritance of bias in algorithmic content moderation. In G. Ciampaglia, A. Mashhadi, & T. Yasseri (Eds.), *Social Informatics*. *Lecture Notes in Computer Science*. Springer, Cham. https://doi.org/10.1007/978-3-319-67256-4_32

Cambria, E. & White, B. (2014). Jumping NLP Curves: A review of natural language processing research. *IEEE Computational Intelligence Magazine*, 9(2), 48-57. https://doi.org/10.1109/MCI.2014.2307227

Citron, D. K. (2014). *Hate crimes in cyberspace*. Cambridge, MA: Harvard University Press.

Crawford, K. & Gillespie, T. (2016). What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society*, *18*(3), 410–428. https://doi.org/10.1177/1461444814543163

Dimoulas C. (2018). Machine Learning. In Bruce Arrigo (Ed.), *The SAGE Encyclopedia of Surveillance, Security, and Privacy, Sage Publications Inc.*, pp. 591-592,http://dx.doi.org/10.4135/9781483359922.n267

Dimoulas, C. A. (2022). Cultural Heritage Storytelling, Engagement and Management in the Era of Big Data and the Semantic Web. *Sustainability*, 14, 812.

Dimoulas, C. A., & Veglis, A. (2023). Theory and Applications of Web 3.0 in the Media Sector. *Future Internet*, *15*(5), 165.

Einwiller, S. & Kim, S. (2020). How online content providers moderate user-generated content to prevent harmful online communication: An analysis of policies and their implementation. *Policy & Internet*, 12(2), 184-206. https://doi.org/10.1002/poi3.239

Engelke, K. M. (2019). Online participatory journalism: A systematic literature review. *Media and Communication*, 7(4), 31-44. https://doi.org/10.17645/mac.v7i4.2250

Etim, B. (2017, June 13). *The Times sharply increases articles open for comments, using Google's technology*. New York Times. https://www.nytimes.com/2017/06/13/insider/have-acomment-leave-a-comment.html

Festl, R. & Quandt, T. (2017). Cyberbullying. In P. Rössler (Eds.), *The international encyclopedia of media effects* (pp. 328–336). Malden: Wiley-Blackwell. https://doi.org/10.1002/9781118783764

Frischlich, L., Boberg, S., & Quandt, T. (2019). Comment sections as targets of dark participation? Journalists' evaluation and moderation of deviant user comments. *Journalism Studies*, 20(14), 2014–2033. https://doi.org/10.1080/1461670X.2018.1556320

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Goodman, E. (2013). *Online comment moderation: emerging best practices*. WAN-IFRA. http://www.wan-ifra.org/reports/2013/10/04/online-comment-moderation-emerging-best-practices

Gorodnichenko, Y., Pham, T., & Talavera, O. (2021). Social media, sentiment and public opinions: Evidence from# Brexit and# USElection. *European Economic Review*, 136. https://doi.org/10.1016/j.euroecorev.2021.103772

Jiang, Y., Li, X., Luo, H., Yin, S., & Kaynak, O. (2022). Quo vadis artificial intelligence? *Discover Artificial Intelligence*, 2(4). https://doi.org/10.1007/s44163-022-00022-8

Jobin, A., Ienca, M. & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1, 389-399. https://doi.org/10.1038/s42256-019-0088-2

Katsaounidou, A., Dimoulas, C., & Veglis, A. (2019). *Cross-Media Authentication and Verification: Emerging Research and Opportunities*. IGI Global. http://doi:10.4018/978-1-5225-5592-6

Kotsakis, R., Vrysis, L., Vryzas, N., Saridou, T., Matsiola, M., Veglis, A., & Dimoulas, C. (2023). A web framework for information aggregation and management of multilingual hate speech. *Heliyon 9*(5), e16084. https://doi.org/10.1016/j.heliyon.2023.e16084

Krumsvik, A. (2018). Redefining user involvement in digital news media. *Journalism Practice*, 12(1), 19–31. https://doi.org/10.1080/17512786.2017.1279025

Ksiazek, T. B., Peer, L., & Zivic, A. (2015). Discussing the news. *Digital Journalism*, 3(6), 850-870. https://doi.org/10.1080/21670811.2014.972079

Lamerichs, N., Nguyen, D., Carmen, M., Melguizo, P., Radojevic, R., & Lange-Böhmer, A. (2018). Elite male bodies: The circulation of alt-right memes and the framing of politicians on social media. *Participations*, 15(1), 180–206.

Loosen, W., Ahva, L., Reimer, J., Solbach, P., Deuze, M., & Matzat, L. (2022). 'X Journalism'. Exploring journalism's diverse meanings through the names we give it. *Journalism*, 23(1), 39–58. https://doi.org/10.1177/1464884920950090

Matamoros-Fernández, A. & Farkas, J. (2021). Racism, hate speech, and social media: A systematic review and critique. *Television & New Media*, 22(2), 205–224. https://doi.org/10.1177/1527476420982230

Meskys, E., Liaudanskas, A., Kalpokiene, J., & Jurcys, P. (2020). Regulating deep fakes: Legal and ethical considerations. *Journal of Intellectual Property Law & Practice*, 15(1), 24-31. https://doi.org/10.1093/jiplp/jpz167

Minh, D., Wang, H.X., Li, Y.F., & Nguyen, T.N. (2022). Explainable artificial intelligence: A comprehensive review. *Artificial Intelligence Review*, 55, 3503-3568 https://doi.org/10.1007/s10462-021-10088-y

Molnar, C. (2020). *Interpretable machine learning*. Lulu.com.

Meor Amer (2021). *A Visual Introduction to Deep Learning*. k-dimensions.

Naab, T. K., Kalch, A., & Meitz, T. G. (2018). Flagging uncivil user comments: Effects of intervention information, type of victim and response comments on bystander behavior. *New Media & Society,* 20(2), 777-795. https://doi.org/10.1177/1461444816670923

Nasim, S. F., Ali, M.R., & Kulsoom, U. (2022). Artificial Intelligence incidents & ethics: A narrative review. *International Journal of Technology, Innovation and Management (IJTIM)*, *2*(2), 52-64. https://doi.org/10.54489/ijtim.v2i2.80

Obermaier, M., Hofbauer, M., & Reinemann, C. (2018). Journalists as targets of hate speech. How German journalists perceive the consequences for themselves and how they cope with it. *SC|M Studies in Communication and Media*, 7(4), 499-524. https://doi.org/10.5771/2192-4007-2018-4-499

Perifanos, K. & Goutsos, D. (2021). Multimodal hate speech detection in Greek social media. *Multimodal Technologies and Interaction*, 5(7), 34. https://doi.org/10.3390/mti5070034

Quandt, T. (2018). Dark participation. *Media and Communication*, 6(4), 36–48. https://doi.org/10.17645/mac.v6i4.1519

Risch, J. & Krestel, R. (2018, August 25). Delete or not delete? Semi-automatic comment moderation for the newsroom. In R. Kumar, A. Ojha, M. Zampieri & S. Malmasi (Eds.),

Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (pp. 166–176), Santa Fe, USA.

Saridou, T. & Veglis, A. (2016). Participatory journalism policies in newspapers' websites in Greece. *Journal of Greek Media and Culture*, 2(1), 85-101. https://doi.org/10.1386/jgmc.2.1.85_1

Saridou, T. & Veglis, A. (2021). Exploring participatory journalism through the integration of user-generated content in media organizations. In M. Khosrow-Pour (Ed.), *Encyclopedia of Information Science and Technology*, 5th edition, (pp. 1152-1163). IGI Global. https://doi.org/10.4018/978-1-7998-3479-3

Saridou, T., Panagiotidis, K., Tsipas, N., & Veglis, A. (2019). Towards a semantic-oriented model of participatory journalism management: Perceptions of user-generated content. *Journal of Education, Innovation and Communication*, 1, 27-37. https://doi.org/10.34097/jeicom_SI_Dec2019-2

Saridou T. & Dimoulas C. (Eds.), Special Issue on Artificial Intelligence in Participatory Environments: Technologies, Ethics, and Literacy Aspects, *Societies (MDPI*), https://rb.gy/1x885, 2023 (in process, upcoming)

Schmidt, A. & Wiegand, M. (2017, April 3-7). A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media* (pp. 1–10). Association for Computational Linguistics. Valencia, Spain.

Sivakorn, S., Polakis, J., & Keromytis, A.D. (2016, March 29 – April 1). *I'm not a human: Breaking the Google reCAPTCHA*. In Black Hat Asia 2016, Singapore. https://www.blackhat.com/asia-16/briefings.html

Tenenboim, O., Masullo, G. M., & Lu, S. (2019). *Attacks in the comment sections: What it means for news sites*. Center of Media Engagement. https://mediaengagement.org/research/attacks-in-the-comment-sections/

Underwood, C. (2019). *Automated journalism - AI applications at New York Times, Reuters, and other media giants*. New York Times. https://emerj.com/ai-sector-overviews/automated-journalism-applications

Venturini, T. & Rogers, R. (2019). "API-based research" or How can digital sociology and journalism studies learn from the Facebook and Cambridge Analytica data breach. *Digital Journalism*, 7(4), 532-540, https://doi.org/10.1080/21670811.2019.1591927

Vrysis, L., Tsipas, N., Thoidis, I., & Dimoulas, C. (2020). 1D/2D deep CNNs vs. temporal feature integration for general audio classification. *Journal of the Audio Engineering Society*, *68*(1/2), 66-77.

Vrysis, L., Vryzas, N., Kotsakis, R., Saridou, T., Matsiola, M., Veglis, A., Arcila-Calderón, C., & Dimoulas, C. (2021). A Web interface for analyzing hate speech. *Future Internet*, 13(80). https://doi.org/10.3390/fi13030080

Wang, S. (2020). The influence of anonymity and incivility on perceptions of user comments on news websites. *Mass Communication and Society*, 23(6), 912–936. https://doi.org/10.1080/15205436.2020.1784950

Wintterlin, F., Schatto-Eckrodt, T., Frischlich, L., Boberg, S., & Quandt, T. (2020). How to cope with dark participation: Moderation practices in German newsrooms. *Digital Journalism*, 8(7), 904-924. https://doi.org/10.1080/21670811.2020.1797519

Ελληνόγλωσσες Αναφορές

Δημούλας Χ. (2015). *Τεχνολογίες συγγραφής και διαχείρισης πολυμέσων: Τεχνικές μη γραμμικής αφήγησης στα νέα ψηφιακά μέσα*. Αθήνα: ΣΕΑΒ, διαθέσιμο στη διεύθυνση http://hdl.handle.net/11419/4343

Δημούλας Χ. (2023). *Τεχνολογίες συλλογής και διαχείρισης περιβαλλοντικών δεδομένων: Η Τεχνητή Νοημοσύνη και οι Σύγχρονες Διεπιστημονικές Προκλήσεις*. Προσκεκλημένη διάλεξη, 11ο Θερινό Σχολείο Περιβαλλοντικής Δημοσιογραφίας, 2-8 Ιουλίου 2023, Λευκωσία

Καμπουρλάζος, Β., & Παπακώστας, Γ. (2015). Εισαγωγή στην Υπολογιστική Νοημοσύνη. *Αθήνα: ΣΕΑΒ*, διαθέσιμο στη διεύθυνση http://hdl.handle.net/11419/3443