Παιδαγωγικός Λόγος

Τόμ. 32, Αρ. 1 (2026)

Λόγος περί της Τεχνητής Νοημοσύνης

Πώς να εκπαιδεύσετε τον παπαγάλο σας

Βερνάρδος Σαλταμανίκας, Μαρίνα Ξενάκη

doi: 10.12681/plogos.33698

Βιβλιογραφική αναφορά:

Σαλταμανίκας Β., & Ξενάκη Μ. (2026). Πώς να εκπαιδεύσετε τον παπαγάλο σας. *Παιδαγωγικός Λόγος*, *32*(1), 81–100. https://doi.org/10.12681/plogos.33698

Μαρίνα ΞΕΝΑΚΗ
Βερνάρδος ΣΑΛΤΑΜΑΝΙΚΑΣ

Πώς να εκπαιδεύσετε τον παπαγάλο σας

doi:https://doi.org/10.12681/plogos.33698

Εισαγωγή

ΣΗΜΕΡΑ ΒΡΙΣΚΟΜΑΣΤΕ, ΣΕ ΕΝΑ ΣΗΜΕΙΟ ΣΤΗΝ ΙΣΤΟΡΙΑ ΤΗΣ ΕΞΕΛΙΞΗΣ μας, που απαιτεί επανεξέταση και επανατοποθέτηση ενδεχομένως των παραδεδεγμένων και δεδομένων αρχών που διέπουν τις ζωές μας. Τα μεγάλα γλωσσικά μοντέλα, σε ευρεία πλέον χρήση μέσω συστημάτων παραγωγικής ΤΝ όπως το ChatGPT αποκτούν όλο και περισσότερους χρήστες με πρακτικό αντίκρισμα σε πλήθος εφαρμογών. Η εξέλιξη των μοντέλων συστημάτων αυτών είναι σε θέση να μας κάνει να στεκόμαστε και να ανα-θεωρούμε τις αξίες του βίου, ρόλος που παραδοσιακά ανήκε στην ανθρωπολογική μελέτη. Η τεχνολογία έχει έρθει έξωθεν και έχει παράλληλα εσωτερικευθεί με έναν αδάμαστο τρόπο. Καθώς οι μηχανές που προκύπτουν είναι ολοένα και πιο σύνθετες, οι λειτουργίες τους απλώνουν τα πέπλα τους σε κάποια βαθύτερα και ενδόμυχα σύνορα. Εν προκειμένω, η συζήτηση θα γίνει με αφορμή το άρθρο: "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?", το οποίο συνέγραψε, μεταξύ άλλων, η τότε επικεφαλής του τμήματος Ηθικής και Τεχνητής Νοημοσύνης (ΤΝ) στην Google, Timnit Gebru.[1] Τι είναι όμως οι

[1] Η δημοσίευση του εν λόγω άρθρου οδήγησε στον τερματισμό της συνεργασίας της Gebru με την Google. Η δημοφιλία της στην εταιρία παραμένει και υπάρχουν φωνές που επιθυμούν την επιστροφή της βλ. Urian, B. Google Ai researchers want Timnit Gebru to come back at higher position among other demands, Tech Times, (2020). https://www.techtimes.com/articles/255136/20201216/google-ai-researchers-demand-new-policies-leadership-changes-and-timnit-gebru-to-come-back-at-higher-position.htm

"στοχαστικοί παπαγάλοι"; Τα προαναφερθέντα πτηνά φημίζονται για την ικανότητα τους να "μιλούν". Να μπορούν να αναπαράγουν δηλαδή όσα ακούν από τους ανθρώπους. Ένα εξόχως καυστικό όνομα για τα γλωσσικά μοντέλα. Πρόκειται για τις μηχανές εκείνες που μοντελοποιούν τη γλώσσα χρησιμοποιώντας διάφορες στατιστικές και τεχνικές για να προσδιορίσουν την πιθανότητα μιας δεδομένης ακολουθίας λέξεων που θα εμφανιστεί σε μια πρόταση.

Τα γλωσσικά μοντέλα όπως το ChatGPT λειτουργούν μέσω μιας διαδικασίας μηχανικής μάθησης γνωστής ως βαθιά μάθηση, χρησιμοποιώντας συγκεκριμένα έναν τύπο αρχιτεκτονικής νευρωνικού δικτύου που ονομάζεται μετασχηματιστής (transformer).[2] Αυτά τα μοντέλα εκπαιδεύονται σε τεράστιες ποσότητες δεδομένων κειμένου από πηγές όπως βιβλία, ιστότοποι και άρθρα, όχι για να απομνημονεύσουν το περιεχόμενο, αλλά για να μάθουν στατιστικά μοτίβα στη χρήση της γλώσσας. Αυτή η διαδικασία, που επαναλαμβάνεται δισεκατομμύρια φορές, επιτρέπει στο μοντέλο να δημιουργεί συνεκτικό και συναφές με το συγκείμενο κείμενο. Μετά από αυτή την αρχική προ-εκπαίδευση, το μοντέλο μπορεί να τελειοποιηθεί για συγκεκριμένες εφαρμογές και να προσαρμοστεί για λόγους ασφάλειας μέσω μεθόδων όπως η ενισχυτική μάθηση από ανθρώπινη ανατροφοδότηση. Όταν ένας χρήστης εισάγει κείμενο, το μοντέλο το χωρίζει σε tokens (τμήματα λέξεων), προβλέπει τα πιο πιθανά επόμενα tokens με βάση την εκπαίδευσή του και τα μετατρέπει ξανά σε αναγνώσιμη γλώσσα, παράγοντας τις απαντήσεις που βλέπουμε.[3]

Ο τρόπος μάθησης τους όμως δεν έχει να κάνει με την καλλιέργεια κάποιας κριτικής ικανότητας ή τη δυνατότητα απόδοσης νοήματος σε αυτά που αναπαράγουν. Σε αυτό το γεγονός οφείλεται και το προσωνύμιο στοχαστικοί παπαγάλοι. Πρόκειται δηλαδή για ικανά chatbot που έχουν σχεδιαστεί να προσομοιώνουν την συνομιλία με τους ανθρώπους και να αναπαράγουν προτάσεις με σωστή συντακτική και γραμματική δομή οι οποίες είναι σχετικές με το υπό συζήτηση θέμα.

Η ονοματοδοσία αυτή έγινε, γιατί ακριβώς μπορούν απλώς να αναπαράγουν αυτά που "μαθαίνουν" αλλά όχι να τα κατανοούν. Ασχολούνται με τη γλώσσα, τα γλωσσικά συστήματα εν γένει και την αναπαραγωγή της γλώσσας, με τρόπο που προσιδιάζει επιφανειακά στον ανθρώπινο, χωρίς

[2] Partha Pratim Ray, "ChatGPT: A Comprehensive Review on Background, Applications, Key Challenges, Bias, Ethics, Limitations and Future Scope," *Internet of Things and Cyber-Physical Systems* 3, no. 1 (April 14, 2023): 121–54, https://doi.org/10.1016/j.iotcps.2023.04.003.

[3] Στο ίδιο.

όμως να κατανοούν το περιεχόμενο και κυρίως χωρίς να μεταδίδουν ακριβή πάντα "γνώση". Το νοητικό πείραμα του κινέζικου δωματίου του Searle[4], μας επιτρέπει να διαπιστώσουμε ότι η απλή αναπαραγωγή των λεχθέντων δεν συνεπάγεται, ούτε προϋποθέτει την κατανόησή τους. Η πραγματική κατανόηση συμβόλων, περιλαμβάνει την "σημασιολογία" (semantics), δηλαδή μια γνώση του τι αναπαριστούν τα σύμβολα ή τι σημαίνουν.[5] Όσα συμβαίνουν μέσα σε έναν υπολογιστή, είναι ανάλογα όσων συμβαίνουν σε ένα κινέζικο δωμάτιο. Γίνεται διαχείριση των συμβόλων σε σχέση με τα σχήματά τους, βάσει συντακτικών κανόνων.[6] Ωστόσο η γλώσσα δεν είναι απλώς μια συμφωνηθείσα σειρά γραμμάτων, αλλά και ένα ολόκληρο πλαίσιο νοημάτων και συνειρμικών διαδικασιών. Στην προκειμένη περίπτωση λοιπόν, οι μηχανές αυτές αναφέρονται ως παπαγάλοι, γιατί απλώς αναπαράγουν αυτό βάσει του οποίου εκπαιδεύονται ή που προκύπτει από τους αλγορίθμους, χωρίς όμως να αντιλαμβάνονται το νόημα των όσων λένε. Επιπλέον η ικανότητα που έχουν τα μεγάλα γλωσσικά μοντέλα (LLMs) να δημιουργήσουν συνθετικά κείμενα, μπορεί να μας παραπλανήσει.[7] Αυτό γιατί, ενώ οι μηχανές συνθέτουν το κείμενο χωρίς να του προσδίδουν κάποιο νόημα, ο άνθρωπος έχει την τάση να αποδίδει νόημα σε ό,τι διαβάζει, με αποτέλεσμα να καταλήγει να ερμηνεύει μία στατιστικά επικυρωμένη σύνθεση λέξεων σε προτάσεις.

Η ανάπτυξη τέτοιων μεγάλων γλωσσικών μοντέλων μπορεί να προκαλέσει αρκετά ανεπιθύμητα αποτελέσματα. Παρατηρούμε πως στο εν λόγω άρθρο της Gebru, εντοπίζονται διάφορα ηθικά ζητήματα που αφορούν προβλήματα κοινωνικού αντικτύπου και προβλήματα σχετικά με την κατεύθυνση της έρευνας. Αναδύονται ζητήματα όπως α) η αναπαραγωγή διακρίσεων και στερεοτύπων, β) χειραγώγηση και παραπληροφόρηση αλλά και γ) κρίσιμα θέματα περιβαλλοντικών επιπτώσεων. Ωστόσο τα ζητήματα αυτά έχουν τις ρίζες τους στο μακρινό παρελθόν. Προβλήματα που άπτονται περιβαλλοντικών θεμάτων είναι συνυφασμένα με την ανθρώπινη ανάπτυξη, αν και κάνουν εντονότερη την εμφάνισή τους από τη βιομηχανική επανάσταση και έπειτα. Επίσης, το ζήτημα των διακρίσεων μπορεί να εντοπιστεί και να εξεταστεί μέσα από κοινωνικά φαινόμενα όπως η πατριαρχία ή ο θεσμός της δουλείας ή και πολιτευμάτων όπως η βασιλεία και

[4] John Searle, "Minds, Brains, and programs" *The Behavioral and brain sciences*, (1980) 3, 417-457, διαθέσιμο στο https://www.law.upenn.edu/live/files/3413-searle-j-minds-brains-and-programs-1980pdf

[5] Jaegwon Kim, *Η Φιλοσοφία του Νου* (Αθήνα: Liberal Books, 2016), 169.

[6] Στο ίδιο.

[7] Εδώ, ίσως, γίνεται καλύτερα κατανοητό και το προσωνύμιο στοχαστικοί παπαγάλοι.

η ολιγαρχία. Με τέτοιες καταστάσεις ερχόταν αντιμέτωπη η ανθρωπότητα ανέκαθεν, και θα ήταν ευκαιρία η ύπαρξη, ανάδυση και ευρεία χρήση της ΤΝ να προσβλέπει και να συμβάλλει εν τέλει αποτελεσματικά στην επίλυση και όχι την διαιώνισή τους.

Γλωσσικά μοντέλα: μηχανές Τεχνητής Νοημοσύνης

Συζητώντας για ΤΝ είναι χρήσιμο να θέτει κανείς το ερώτημα και να οριοθετεί το τι εννοούμε με τον όρο "τεχνητή νοημοσύνη"; Ο ορισμός της νοημοσύνης είναι προς ώρας ένα ανοιχτό και γι' αυτό ενοχλητικό (παρ' ότι οι θεσμοί όπως οι ΕΕ έχουν σπεύσει να ορίσουν επακριβώς και να εστιάσουν την απόδοση του όρου "τεχνητή νοημοσύνη" στα συστήματα που μιμούνται την ανθρώπινη συμπεριφορά)[8] ερώτημα στη φιλοσοφία του νου. Φαίνεται ότι έχουμε μια σταθερή διαισθητική αντίληψη του τι είναι νοημοσύνη. Σε γενικές γραμμές, είναι η ικανότητα κάποιου να συλλογίζεται, να σκέφτεται λογικά, να χρησιμοποιεί τη φαντασία, να μαθαίνει και να ασκεί κρίση. Είναι η ικανότητα να πλαισιώνεις ένα πρόβλημα και μετά να το λύνεις. Η νοημοσύνη είναι γενικεύσιμη. Είναι σε θέση να κάνει αυτά τα πράγματα σε ένα ευρύ φάσμα προβλημάτων και πλαισίων. Είναι αυτή που έχουν οι άνθρωποί, αυτό που έχουν λιγότερο τα πρωτεύοντα, οι παπαγάλοι ακόμα λιγότερο, οι μέδουσες και τα δέντρα (και οι σύγχρονες μηχανές) καθόλου.[9] Η ΤΝ είναι η νοημοσύνη σε ένα τεχνούργημα που έχουμε δημιουργήσει.[10]

Στην ΤΝ, η "μάθηση" υπονοεί την εξαγωγή "γνώσεων" μέσα από δεδομένα, με τρόπο που να επιτρέπει τη βελτίωση του συστήματος σε μία εργασία.[11] Αυτού του είδους η διαδικασία ονομάζεται μηχανική μάθηση (machine learning) και είναι υποπεδίο της ΤΝ που δίνει στους υπολογιστές

[8] Ευρωπαϊκό Κοινοβούλιο, "Τι είναι η τεχνητή νοημοσύνη και πώς χρησιμοποιείται; |Θέματα|Ευρωπαϊκό Κοινοβούλιο," www.europarl.europa.eu, September 9, 2020, https://www.europarl.europa.eu/topics/el/article/20200827STO85804/ti-einai-i-techniti-noimosuni-kai-pos-chrisimopoieitai.

[9] Robert Sparrow, "The Turing Triage Test", *Ethics and Information Technology* (2004) 6: 203–213, 204.

[10] Στο ίδιο.

[11] Γιώργος Γιαννακόπουλος, *Τεχνητή Νοημοσύνη: Μια Διακριτική Απομυθοποίηση* (Θεσσαλονίκη: Ροπή, 2020), 125.

τη δυνατότητα να "μαθαίνουν" χωρίς να είναι ρητά προγραμματισμένοι.[12] Η μηχανική μάθηση βρίσκεται πίσω από τα chatbot και το προγνωστικό κείμενο, τις εφαρμογές μετάφρασης γλώσσας, τις εκπομπές που προτείνει το Netflix και τον τρόπο παρουσίασης των ροών στα μέσα κοινωνικής δικτύωσης. Οι κανόνες με τους οποίους δρουν τα συστήματα ΤΝ μπορούν να αλλάζουν κατά τη διάρκεια του χειρισμού της μηχανής από την ίδια τη μηχανή, από τον "εαυτό" της.[13] Η βαθιά μάθηση (deep learning) ουσιαστικά προσπαθεί να κάνει τον υπολογιστή να μαθαίνει με όλο και μεγαλύτερη ακρίβεια.[14] Έχει τη δυνατότητα να αξιοποιεί πολύ καλά, τον τεράστιο όγκο δεδομένων που υπάρχει εκεί έξω (κατά βάση στο διαδίκτυο), για να κάνει πράγματα, όπως η αναγνώριση φωνής, κειμένου και άλλα.[15]

Αναδυόμενα προβλήματα από τις Ανοιχτές Βάσεις Δεδομένων

Το διαδίκτυο είναι ένας τεράστιος, γεμάτος ποικιλία ανοιχτά προσβάσιμος χώρος και θα φανταζόμασταν ότι τα πολύ μεγάλα δεδομένα (large datasets) θα εξασφάλιζαν και θα μπορούσαν να αντικατοπτρίσουν αυτή την ποικιλομορφία.[16] Το ζήτημα είναι λοιπόν ότι αυτά τα γλωσσικά μοντέλα με τον τρόπο της μηχανικής μάθησης και της βαθιάς μάθησης, φτάνουν να αναπαράγουν στερεότυπα και προκαταλήψεις, μιας και το πεδίο από όπου συλλέγουν τα δεδομένα τους είναι ο αχανής χώρος του διαδικτύου.

Πρόκειται για μέσα που συλλέγουν τα δεδομένα τους από το διαδίκτυο και το περιβάλλον αυτό, είναι τεράστιο και χαοτικό, με αποτέλεσμα να μην επιτρέπει τον πλήρη έλεγχο των προγραμματιστών στα δεδομένα που έχουν πρόσβαση. Προκειμένου να παραχθεί συνεκτικό κείμενο, τα γλωσσικά μοντέλα συνήθως εκπαιδεύονται σε ογκώδη σύνολα δεδομένων. Είναι δύσκολο να χρησιμοποιήσεις ένα σύνολο δεδομένων μεγάλου

[12] Machine Learning, Explained", Sara Brown, "https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

[13] Andreas Matthias, "The Responsibility gap: Ascribing responsibility for the actions of learning automata", *Ethics and Information Technology 6* (2004), 175-183, 177.

[14] Γιαννακόπουλος, 137.

[15] Στο ίδιο.

[16] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, Shmargaret Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?", *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (March 2021), 610–623.

μεγέθους που απαιτείται για την εκπαίδευση ενός γλωσσικού μοντέλου, διασφαλίζοντας παράλληλα ότι το σύνολο δεδομένων απηχεί την επιθυμητή συμπεριφορά.[17] Αυτή η αδυναμία ελέγχου της πληροφορίας, αφήνει μεγάλα κενά που επιτρέπουν την αναπαραγωγή και διαιώνιση στερεοτύπων και προκαταλήψεων. Τα μοντέλα αυτά φαίνεται να αναπαράγουν στερεοτυπικές συμπεριφορές σχετικές με το φύλο, την φυλή, την εθνικότητα ακόμα και με κατάσταση αναπηρίας. Σημειώνεται ότι το μοντέλο τείνει να αντικατοπτρίζει δυτικοκεντρικές προοπτικές και επιδεικνύει την υψηλότερη απόδοσή του στην αγγλική γλώσσα. Πολλά από τα μέτρα προστασίας που έχουν σχεδιαστεί για την πρόληψη επιβλαβούς περιεχομένου έχουν δοκιμαστεί κυρίως σε αγγλόφωνο περιβάλλον.[18]

Με αυτόν τον τρόπο συμβάλλουν στο να εμφανίζονται όλο και πιο συχνά οι γνώμες αυτές, με αποτέλεσμα να συνεχίζουν να υπερ-εκπροσωπούνται στα νέα δεδομένα με τα οποία μετ-εκπαιδεύονται τα μοντέλα, διαιωνίζοντας τα προβλήματα. Επιπλέον η διόρθωση αυτών των προβλημάτων θα πρέπει να αντιμετωπίσει και τα κενά ευθύνης (responsibility gap).[19] Στην ανάπτυξη και χρήση των LLM εμπλέκεται ένας μεγάλος αριθμός ανθρώπων υπό διαφορετικούς ρόλους και ιεραρχικά επίπεδα. Αυτή η πολυπλοκότητα καθιστά εξόχως δύσκολο τον ορθό και ακριβή επιμερισμό ευθυνών.[20] Ορισμένοι μελετητές υποστηρίζουν ότι οι προηγμένες τεχνολογίες τεχνητής νοημοσύνης έχουν δημιουργήσει ένα "κενό ευθύνης", όπου ούτε οι προγραμματιστές ούτε οι χρήστες μπορούν να αναλάβουν πλήρως την ευθύνη για τα αποτελέσματα, λόγω του περιορισμένου ελέγχου που ασκούν επί των αυτόνομων συστημάτων. Αυτές οι μηχανές μάθησης προσαρμόζονται μέσω της ανατροφοδότησης από το περιβάλλον και της δοκιμής και του λάθους (trial and error), ενεργώντας ανεξάρτητα σε ανθρώπινα περιβάλλοντα και αλληλεπιδρώντας με κοινωνικές δομές, περιπλέκοντας έτσι τα ζητήματα ευθύνης και υπευθυνότητας (responsibility και

[17] Irene Solaiman, Christy Dennison, "Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets" 2, διαθέσιμο στο https://openai.com/blog/improving-language-model-behavior/
Επίσης: OpenAI, "How Should AI Systems Behave, and Who Should Decide?," Openai.com, 2023, https://openai.com/index/how-should-ai-systems-behave/.

[18] OpenAI, "Is ChatGPT Biased? | OpenAI Help Center," help.openai.com, 2024, https://help.openai.com/en/articles/8313359-is-chatgpt-biased.

[19] Matthias, 181.

[20] Claudio Novelli, Mariarosaria Taddeo, και Luciano Floridi, "Accountability in Artificial Intelligence: What It Is and How It Works," *SSRN Electronic Journal*, 2022, https://doi.org/10.2139/ssrn.4180366.

accountability).[21] Ανακύπτουν επομένως, διάφορα ερωτήματα που υπογραμμίζουν το μεγαλύτερο ερώτημα που τα περιβάλει: αξίζει όλος αυτός ο κόπος για τα αποτελέσματα που έχουμε; Οι συγγραφείς του 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" επισημαίνουν ότι εξαιτίας αυτών των αυξημένων απαιτήσεων σε δεδομένα που χρειάζεται η εκπαίδευση ενός μεγάλου γλωσσικού μοντέλου βρισκόμαστε μπροστά σε δυσάρεστες συνέπειες. Από τη μία η κατασκευή τους επιβαρύνει το περιβάλλον και έχει παράλληλα μεγάλα οικονομικά κόστη. Για να γίνει αντιληπτό το μέγεθος των περιβαλλοντικών επιπτώσεων, αρκεί να αναλογιστούμε ότι, ενώ ο μέσος άνθρωπος ευθύνεται για 5 τόνους εκπομπών CO2 το χρόνο, η εκπαίδευση ενός μεγάλου γλωσσικού μοντέλου θα προκαλέσει 284 τόνους εκπομπών.[22] Από την άλλη μεριά, υπάρχουν προβλήματα και κατά την ίδια τη λειτουργία των μοντέλων. Ενισχύουν τις προαναφερθείσες κοινωνικές διακρίσεις, καθώς αναπαράγουν την επικρατούσα ιδεολογία των ανθρώπων που έχουν τη μεγαλύτερη πρόσβαση και παρουσία στο διαδίκτυο, διογκώνουν το χάσμα αρχειοθέτησης (documentation gap)[23] και αναπαράγουν στερεότυπα που πολλές φορές μπορεί να είναι επιβλαβή.[24] Όσον αφορά την κατεύθυνση της έρευνας και δεδομένου ότι ο χρόνος του ερευνητή είναι κι αυτός ένα πολύτιμο κεφάλαιο, γεννιούνται ερωτήματα σε σχέση με το πού θα πρέπει να εστιαστούν οι ερευνητικοί πόροι και ως προς το αν η έρευνα πάνω στα LMs αποτρέπει τη διάθεσή τους σε άλλες προσπάθειες κατανόησης της φυσικής γλώσσας (NLU).[25]

---

[21] Eva Schur, Anna Brouns, και Peter Lee, "Ethical Analysis of the Responsibility Gap in Artificial Intelligence ," *International Journal of Ethics and Society* 6, no. 4 (2025): 1–10, https://doi.org/10.22034/ijethics.6.4.

[22] Emily M.Bender, "On the dangers of stochastic parrots: Can language models be too big?", Lecture at The Alan Turing Institute, (2021) 9:37-10:25. https://www.youtube.com/watch?v=N5c2X8vhfBE.

[23] Την έλλειψη μιας στοχευμένης ταξινόμησης των δεδομένων που χρησιμοποιούν. Αναφέρεται δηλαδή, στην έλλειψη ολοκληρωμένων, διαφανών ή προσβάσιμων αρχείων που περιγράφουν λεπτομερώς τον τρόπο με τον οποίο αναπτύχθηκε ένα σύστημα ΤΝ, τον τρόπο λειτουργίας του και τον τρόπο λήψης αποφάσεων στο πλαίσιο αυτού.

[24] Bender, E.M., Gebru, T., McMillan-Major, A., et al. (2021) "On the dangers of stochastic parrots: Can language models be too big?". FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (2021):614-615, https://doi.org/10.1145/3442188.3445922

[25] Bender,On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?,615.

Ποιος έχει την ευθύνη των παπαγάλων;

Παραδοσιακά έχουμε συνδέσει την ευθύνη με την δράση προσώπων.[26] Οι νέες τεχνολογίες επηρεάζουν τους ανθρώπους ήδη με την ύπαρξή τους. Επηρεάζουν τις αποφάσεις που παίρνουμε, πείθουν, διευκολύνουν και επιτρέπουν συγκεκριμένες ανθρώπινες γνωστικές διαδικασίες, πράξεις και συμπεριφορές.[27] Για παράδειγμα οι μηχανές αναζήτησης στο διαδίκτυο δίνουν προτεραιότητα και παρουσιάζουν πληροφορίες με μία συγκεκριμένη σειρά, επηρεάζοντας έτσι το τι βλέπουν οι χρήστες του διαδικτύου.[28] Τέτοια τεχνολογικά αντικείμενα είναι "ενεργοί μεσολαβητές" που συνδιαμορφώνουν ενεργά την ύπαρξη των ανθρώπων, την αντίληψη, την εμπειρία και τις πράξεις τους.[29]

Τα λάθη ήταν μέχρι τώρα, πάντοτε λάθη του προγραμματιστή, όχι του προγράμματος. Μπορούσαν να αναγνωριστούν, να απομονωθούν και να φτιαχτούν και ο προγραμματιστής μπορούσε εύκολα να θεωρηθεί υπαίτιος για κάθε σφάλμα της μηχανής. Ωστόσο καθώς οι τεχνικές της τεχνητής νοημοσύνης και του προγραμματισμού εξελίσσονται περεταίρω, αλλάζει και ο ρόλος των προγραμματιστών.[30] Προγραμματίζονται οι μηχανές, ώστε να μπορούν να προγραμματίζουν τον εαυτό τους. Και έτσι ο προγραμματιστής μετατρέπεται σε δημιουργό λογισμικών οργανισμών των οποίων τους κώδικες δεν γνωρίζει επακριβώς και καθίσταται αδύνατον να τους ελέγξει για τυχόν λάθη.[31]

Η παραπάνω κατάσταση κατά την οποία η λειτουργία μιας μηχανής δεν είναι ξεκάθαρη για τον χρήστη της ή για οποιονδήποτε ενδιαφερόμενο είναι ονομάζεται black box problem. Η ασάφεια με οποία ερχόμαστε αντιμέτωποι κατά το black box problem είναι ανάλογη με την αυτονομία που έχει μια μηχανή στον τρόπο που μαθαίνει. Η συμπεριφορά μιας μηχανής δεν είναι πλέον καθορισμένη αλλά σχηματίζεται με την αλληλεπίδρασή της με το περιβάλλον, από το οποίο η μηχανή υιοθετεί καινούργια συμπεριφοριστικά μοτίβα. Είναι μία διαδικασία μάθησης, βέβαια, πράγμα που

[26] Όπως είδαμε παραπάνω τις έννοιες responsibility και accountability. Την ευθύνη προκειμένου να μην συμβεί κάτι αλλά και της υπευθυνότητας ανάληψης ευθύνης.
[27] Noorman, Merel, "Computing and Moral Responsibility", The Stanford Encyclopedia of Philosophy (Spring 2020 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/spr2020/entries/computing-responsibility/
[28] Στο ίδιο.
[29] Στο ίδιο.
[30] Matthias, 181.
[31] Στο ίδιο, 182.

σημαίνει ότι ίσως κάποια λάθη να είναι απλώς ένα βήμα προς την εξερεύνηση και εξεύρεση της λύσης.[32]

Εν προκειμένω, αλλάζουν οι συμβατές έννοιες της ηθικής ευθύνης. Για να θεωρηθεί ένας άνθρωπος υπεύθυνος για μία πράξη, πρέπει να υπάρχει μία αιτιώδης σύνδεση ανάμεσα σε αυτόν και το αποτέλεσμα της πράξης. Το δρών υποκείμενο πρέπει να έχει γνώση και να είναι ικανό να αναλογιστεί τις πιθανές συνέπειες τις πράξεις του. Και τέλος το ίδιο υποκείμενο πρέπει να είναι ικανό να επιλέγει ελεύθερα τον τρόπο δράσης του, να μην υπόκειται σε κανέναν εξαναγκασμό.[33] Στο πλαίσιο όμως της τεχνολογίας, τα πράγματα γίνονται κάπως θολά. Διότι υπάρχει το πρόβλημα "των πολλών χεριών" (Problem of Many Hands), αφού για καθετί υπάρχουν πολλοί άνθρωποι που δουλεύουν πίσω από αυτά και σε πολλά διαφορετικά στάδια. Οι τεχνολογίες των υπολογιστών επιμηκύνουν την ανθρώπινη δραστηριότητα στο χώρο και το χρόνο.[34] Οι συνέπειες που προκύπτουν πολλές φορές είναι ανυπολόγιστες. Είναι δύσκολο να ληφθούν υπόψη όλες οι πιθανές παράμετροι, έτσι ώστε μία μηχανή να λειτουργεί σε ένα καθαρά ντετερμινιστικό περιβάλλον. Εδώ ακριβώς διογκώνεται το κενό ευθύνης (ή το χάσμα ευθύνης Responsibility Gap) μέσω και του black box problem[35]. Στο σημείο δηλαδή της μη δυνατότητας ελέγχου του τεράστιου όγκου πληροφοριών και την αδυναμία να προβλεφθεί κάθε πιθανή παραγόμενη από την μηχανή δράση.

Ωστόσο, δεν υπάρχει ένα καθολικό πρότυπο για προσβλητικό ή επιβλαβές περιεχόμενο· αλλάζει η ερμηνεία της συμπεριφοράς του γλωσσικού μοντέλου ανάλογα με τους πολιτισμικούς παράγοντες.[36] Ως εκ τούτου, μια διαδικασία για τον προσδιορισμό και την προσαρμογή της κατάλληλης συμπεριφοράς του μοντέλου θα πρέπει να είναι εφικτή και για πολλούς κοινωνικούς φορείς, ειδικά εκείνους που έχουν πληγεί περισσότερο και παραλείπονται στην ανάπτυξη των μοντέλων. Ομοίως, η συμπεριφορά του μοντέλου θα πρέπει να αξιολογείται σε ένα κοινωνικό πλαίσιο και με τρόπο που να περιλαμβάνει και τις περιθωριοποιημένες προοπτικές.[37] Παρουσιάζεται δηλαδή μια δυσκολία στην δημοκρατική πρόσβαση στα

[32] Στο ίδιο.
[33] "Computing and Moral Responsibility", 5-6.
[34] Στο ίδιο, 6.
[35] Γούναρης Ά., & Κωστελέτος Γ. (2024). Γράφοντας τον αλγόριθμο του Καλού: Η Τεχνητή Νοημοσύνη ως μηχανή απόδοσης Δικαιοσύνης. Ηθική. Περιοδικό φιλοσοφίας, (19). https://doi.org/10.12681/ethiki.39654
[36] Solaiman, 2.
[37] Στο ίδιο.

τεχνολογικά αγαθά της ΤΝ και ανοίγονται ερωτήματα για το πώς αυτή η δημοκρατικοποίηση της ΤΝ θα μπορούσε να επιτευχθεί (Democratization of AI). Η επέκταση της πρόσβασης είναι ένα σημαντικό μέρος της υπεύθυνης ανάπτυξης συστημάτων τεχνητής νοημοσύνης, επειδή μας επιτρέπει να μάθουμε περισσότερα για τη χρήση στον πραγματικό κόσμο, αλλά και να συνεχίσουμε να χρησιμοποιούμε τέτοια μέσα με μεγαλύτερη ασφάλεια.[38]

Οι προτάσεις προκειμένου να μετριαστούν οι κίνδυνοι και να διατηρηθούν τα οφέλη

Απέναντι στα προαναφερθέντα προβλήματα υπάρχει πληθώρα βιβλιογραφικών πηγών που προτείνει λύσεις για κάθε ένα από αυτά. Όσον αφορά τα περιβαλλοντικά κόστη, προτείνεται να ενσωματωθεί η ενεργειακή και η υπολογιστική αποδοτικότητα στο σχεδιασμό και την αξιολόγηση αυτών των μοντέλων.[39] Με αυτό τον τρόπο θα επιτυγχάνεται η δημιουργία λιγότερο ενεργειακά κοστοβόρων νέων μοντέλων, ή τουλάχιστον θα πάψει να ενθαρρύνεται η κατασκευή μοντέλων ανεξαρτήτως των πόρων που αυτά απαιτούν.

Επίσης, ένας στοχευμένος περιορισμός των δεδομένων με τα οποία εκπαιδεύονται θα επέφερε θετικά αποτελέσματα. Αν τα μοντέλα τροφοδοτούνταν με συνειδητά επιλεγμένα δεδομένα και υπήρχε αρχειοθέτηση σε κάθε βήμα της διαδικασίας, τότε θα μπορούσε να ελεγχθεί και αποφευχθεί η αναπαραγωγή στερεοτύπων και υποτιμητικού περιεχομένου, ενώ παράλληλα θα μειωνόταν και το χάσμα αρχειοθέτησης. Επιπλέον η κατάλληλη επισήμανση (watermark), σε ένα κείμενο που έχει παραχθεί από μηχανή ΤΝ δεν θα άφηνε περιθώριο παραπλάνησης του αναγνώστη.

Τα ζητήματα που αφορούν την κατεύθυνση της έρευνας, από τη στιγμή που έχουν αναγνωριστεί, θα μπορούσαν επίσης να διευθετηθούν με τις κατάλληλες ενέργειες. Θα έπρεπε να κατανέμεται ο ερευνητικός χρόνος με προσοχή, και ίσως να αναζητηθούν τρόποι που να παρέχουν παρόμοια οφέλη χωρίς τη χρήση ολοένα και μεγαλύτερων LM.[40] Οι συγγραφείς του άρθρου καταλήγουν θέτοντας δύο ουσιώδη ερωτήματα: Ποιοι είναι οι κίνδυνοι γύρω από την έρευνα των LLM και τί πρέπει να λαμβάνουμε υπόψη

[38] Στο ίδιο.
[39] Στο ίδιο, 618.
[40] Στο ίδιο.

μας στην ανάπτυξη της; Είναι τελικά τα LLM τόσο αναγκαία και αναντικατάστατα; Αν όχι τί θα πρέπει να κάνουμε;[41]

Τα προβλήματα που σχετίζονται με την ανάπτυξη των LLM και η εξέλιξή τους σε σχέση με το παρελθόν

Όπως διαπιστώνουμε, υφίστανται πράγματι προβλήματα που προκύπτουν από την ανάπτυξη των LM. Αν και υπάρχει σύνδεση μεταξύ τους, μπορούμε να τα διακρίνουμε σε δύο κατηγορίες. Έχουμε προβλήματα κοινωνικού αντικτύπου (αναπαραγωγή στερεοτύπων, μεροληψία, οικονομικά και περιβαλλοντικά κόστη) αλλά και προβλήματα σε σχέση με την κατεύθυνση της έρευνας. Οι λύσεις που προτείνονται δεν επιβάλουν την συνέχεια της δημιουργίας όλο και μεγαλύτερων LM. Πριν φθάσουμε όμως στην εφαρμογή τους ή στην αναζήτηση νέων λύσεων, είναι καλό να εξετάσουμε το ευρύτερο πλαίσιο αυτών των προβλημάτων. Έχει σημασία να αναλογιστούμε γιατί μας αφορούν σήμερα υπό το πρίσμα των εξελίξεων στην ΤΝ και αν μπορούμε να εντοπίσουμε τις φιλοσοφικές ρίζες της εξέτασής τους.

Καταρχάς, τα προβλήματα σε σχέση με την ΤΝ μας αφορούν άμεσα γιατί βρισκόμαστε μπροστά σε ραγδαία ανάπτυξη του κλάδου, η οποία αναμένεται να αλλάξει ριζικά τις ζωές μας. Οι επιπτώσεις αυτής της ανάπτυξης είναι πιθανό να γίνουν αντιληπτές όταν πια θα είναι αργά για διορθωτικές κινήσεις (Collingridge dilemma).[42] Αυτή η ιδιαιτερότητα κάνει επιτακτική την ανάγκη να εντοπίσουμε εγκαίρως και να θέσουμε προς επίλυση τα ηθικά ζητήματα που προκύπτουν, αντιλαμβανόμενοι πως θα βρισκόμαστε πάντα αντιμέτωποι με ένα θεσμικό χάσμα.

Από ανθρωπολογικής πλευράς, η ρίζα αυτών των προβλημάτων είναι πιθανό να βρίσκεται στην τάση του ανθρώπου να υποτιμά τις παράπλευρες συνέπειες που προκύπτουν από την επίτευξη διαφόρων κατά τα άλλα

[41] Emily M.Bender, "On the dangers of stochastic parrots: Can language models be too big?", Lecture at The Alan Turing Institute, (2021): 31:23-31:42. https://www.youtube.com/watch?v=N5c2X8vhfBE

[42] Ο όρος Collingridge dilemma αναφέρεται στην δυσκολία να αναστρέψεις τις συνέπειες από την στιγμή που μια νέα τεχνολογία έχει εγκατασταθεί στην κοινωνία. Η ονομασία του όρου οφείλεται στον David Collingridge ο οποίος στο βιβλίο του The Social Control of Technology (1980) έγραψε σχετικά με τις προκλήσεις που υπάρχουν στην διαχείριση των νέων τεχνολογιών.

θεμιτών στόχων.[43] Το θεσμικό χάσμα που ενέχουν στόχοι όπως η ανάπτυξη μηχανών ΤΝ, καθιστά αδύνατη την έγκαιρη κατανόηση των νέων δεδομένων. Δεν παύει όμως να είναι απαραίτητη μια ορθολογική μελέτη των κινδύνων πριν την καθιέρωση της παρουσίας της ΤΝ στην κοινωνική ζωή. Το ενθαρρυντικό σε αυτή την κατεύθυνση είναι ότι τον 21ο αιώνα η μελέτη της βιωσιμότητας αλλά και της ηθικής των νέων τεχνολογιών ΤΝ (AI Ethics) αρχίζει να αποκτά όλο και πιο ουσιαστικό ρόλο στην έρευνα.[44]

Παρόλα αυτά, πρέπει να επισημανθεί ότι η εμφάνιση των προαναφερθέντων προβλημάτων κοινωνικού αντικτύπου (αναπαραγωγή στερεοτύπων, διακρίσεων και προκαταλήψεων) έχει μια ειδοποιό διαφορά έτσι όπως εμφανίζεται μέσω των LLM σε σχέση με το παρελθόν. Η εμφάνιση της στα LLM προκύπτει μόνο μετά την επαφή με το περιβάλλον (μέσω των δεδομένων για την εκπαίδευση του μοντέλου). Σε αντίθεση με τους φορείς στερεοτύπων και διακρίσεων, που έχουν επίγνωση του εκφερόμενου λόγου τους, τα LLM απλώς τον αναπαράγουν χωρίς να αντιλαμβάνονται το νόημά του (ως στοχαστικοί παπαγάλοι). Το γεγονός αυτό υποδεικνύει πως τα μοντέλα δεν θα αποτελούσαν πηγή ηθικού προβληματισμού σε ένα περιβάλλον που δεν είχε εκ των προτέρων τέτοια ζητήματα. Η ανάγκη λοιπόν να διορθώσουμε τα παραπάνω ζητήματα γύρω από την ΤΝ μπορεί να μας δώσει την ώθηση ώστε να βρούμε τη λύση να διορθώσουμε τα αίτια που οδήγησαν στην εμφάνισή τους σε πρώτο χρόνο και ανεξάρτητα από τις μηχανές ΤΝ.

Είναι πολύ βασικό να καθοριστεί, ποιος ή τι θα είναι υπεύθυνο για τις συνέπειες των αποφάσεων και των δράσεων ενός συστήματος ΤΝ και ποιος θα αναλάβει το βάρος αυτό της υποχρέωσης να απαντήσει στο τι πρέπει να κάνει και τι όχι.[45] Θα έρθει ένα σημείο όπου η κοινωνία θα πρέπει να αποφασίσει μεταξύ του να μην χρησιμοποιεί τέτοιου είδους μηχανές πια, το οποίο δεν φαίνεται να είναι μία ρεαλιστική επιλογή, ή να αντιμετωπίσει αυτό το κενό ευθύνης το οποίο δεν μπορεί να γεφυρωθεί από τις παραδοσιακές πρακτικές της ανάληψης ευθύνης.[46] Στη φάση που βρίσκεται η τεχνολογία, είναι περισσότερο παράλειψη των προγραμματιστών να

[43] Όπως διατυπώνει ο Alasdair MacIntyre στο " Dependent rational animals: Why human beings need the virtues ": «δρούμε πυροσβεστικά και όχι προληπτικά».

[44] Αυτή η τάση ενισχύθηκε κυρίως από δύο παράγοντες: την κλιματική κρίση όσο αφορά την βιωσιμότητα γενικά και την ανάπτυξη αυτόματων οχημάτων όσον αφορά τον τομέα AI Ethics.

[45] David J. Gunkel, "Mind the Gap: Responsible Robotics and the Problem of Responsibility", *Ethics and Information Technology* (2017), 2.

[46] Matthias, 175.

μην ελέγχουν το περιβάλλον στο οποίο υπάρχει και δουλεύει η μηχανή, παρά λάθος της ίδιας της μηχανής. Ίσως αργότερα στο μέλλον, όταν "ενηλικιώνεται" μια μηχανή, να μπορεί να το κάνει μόνη της. Προς το παρόν, η αμέλεια αυτή εμπίπτει ακριβώς σε αυτό το κενό ευθύνης. Που δεν είναι αμιγώς ευθύνη, αλλά παράλειψη. Η ευθύνη όμως προκύπτει και από τις πράξεις, αλλά και τις παραλείψεις μας. Διότι ακόμα και αν υποθέσουμε ότι η πρόθεση και η βούληση ως νοητικά φαινόμενα δεν επηρεάζουν τον φυσικό κόσμο σωματικά τουλάχιστον, οι πράξεις και οι παραλείψεις αυτές καθαυτές είναι που "γράφουν" πάνω στον εμπειρικό κόσμο. [47]

Προτεινόμενες ενέργειες με στόχο τις μέγιστες θετικές επιπτώσεις.

Η σωστή διάκριση των προβλημάτων και η αποφυγή της απορριπτικής τεχνοφοβίας είναι μείζονος σημασίας. Αρχικά, αυτό που μπορούμε να κάνουμε είναι να εστιάσουμε στην καλή κατανόηση των προβλημάτων που προκύπτουν. Στη συνέχεια, θα πρέπει άτομα και οργανισμοί να αναλάβουν την ευθύνη που έχουν απέναντι σε αυτά τα προβλήματα. Πρέπει να αντιμετωπίσουμε κάποια στιγμή κατά μέτωπο και με ειλικρίνεια το κενό ευθύνης, αντί να αφήνουμε το ερώτημα του καταλογισμού να αιωρείται εντός ενός ασαφούς και θολού τοπίου εμπλεκομένων. Πολλές φορές δουλεύοντας στο κομμάτι του προγραμματισμού χρειάζεται να λαμβάνουμε υπόψη ότι οι πράξεις μας δεν αφορούν μόνο ένα μαθηματικό σύμπαν, αλλά έχουν σοβαρές επιπτώσεις στον πραγματικό κόσμο.[48] Αναγνωρίζοντας και αναλαμβάνοντας την ευθύνη, έχουμε κάνει το πρώτο βήμα στην προσπάθεια να δημιουργήσουμε εγκαίρως το θεσμικό πλαίσιο που ρυθμίζει τα καινοφανή ζητήματα που δημιουργούνε οι νέες τεχνολογίες όπως η ΤΝ.

Σε τέτοιου είδους μηχανές που παρομοιάσαμε με παπαγάλους, φαίνεται πως δεν μπορούμε να δημιουργήσουμε ένα περιβάλλον εντελώς καθαρό προκαταλήψεων παρά μόνο μειωμένων προκαταλήψεων.[49] Ο τρόπος να φτάσουμε σε αυτό, είναι φιλτράροντας και δυνητικά αξιολογώντας τα δεδομένα που πρόκειται να χρησιμοποιηθούν.[50] Βέβαια εδώ ανακύπτει το

[47] Jaap Hage, "Theoretical foundations for the responsibility of autonomous agents", *Artif Intell Law* (2017), 255–271, 258. Διαθέσιμο στο https://link.springer.com/content/pdf/10.1007/s10506-017-9208-7.pdf

[48] Bender, "On the dangers of stochastic parrots: Can language models be too big?", 11:12-11:18

[49] Bender, "On the dangers of stochastic parrots: Can language models be too big?" 1:09

[50] Στο ίδιο.

ερώτημα, ποιος είναι ικανός να προσδιορίσει τι είναι σωστό να μάθει η μηχανή και αν έχει υπάρξει επαρκές δείγμα που να αντιπροσωπεύει τη διαφορετικότητα στο σύνολό της.[51] Ίσως μια περισσότερο ελεγμένη πρόσβαση σε δεδομένα (όπως οι αρχικές κλειστές βάσεις της OpenAI), μια ρήτρα στα δεδομένα που τυγχάνουν επεξεργασίας ή μια πρόσβαση σε – ει δυνατόν – ελεγμένα δεδομένα, θα διευκόλυνε να επιτευχθεί ένα σύστημα για λιγότερες προκαταλήψεις, ιδιαίτερα αν το σύστημα μιμείται την ανθρώπινη συμπεριφορά[52].

Ένας προγραμματισμός με κατηγοριοποίηση ευαίσθητων θεμάτων και σκιαγράφηση της επιθυμητής συμπεριφοράς, δημιουργία του συνόλου δεδομένων και έπειτα μικρορυθμίσεις, θα μπορούσε να βελτιώσει τη συμπεριφορά του γλωσσικού μοντέλου σε σχέση με συγκεκριμένες τιμές συμπεριφοράς, βελτιστοποιώντας ένα επιμελημένο σύνολο δεδομένων <100 παραδειγμάτων αυτών των τιμών.[53] Τέλος η αξιολόγηση των μοντέλων με στόχο τη μείωση της τοξικότητας, θα επέτρεπε τη δυνατότητα για μηχανές με λιγότερες αναπαραγωγές στερεοτύπων.[54]

Αξίζει να τονιστεί, ότι πρόκειται για ένα πεδίο που ήδη σημειώνονται εξελίξεις και διορθώσεις. Για παράδειγμα στο πρόγραμμα DALL·E έχουν γίνει προσπάθειες για ελαχιστοποίηση του κινδύνου κακής χρήσης του προγράμματος για τη δημιουργία παραπλανητικού περιεχομένου, απορρίπτοντας μεταφορτώσεις εικόνων που περιέχουν υπαρκτά πρόσωπα και απόπειρες δημιουργίας ομοιότητας δημοσίων προσώπων, συμπεριλαμβανομένων διασημοτήτων και εξέχων πολιτικών προσωπικοτήτων.[55] Κάνοντας τα φίλτρα περιεχομένου πιο ακριβή, ώστε να είναι πιο αποτελεσματικά στον αποκλεισμό μηνυμάτων προτροπής και μεταφορτώσεων εικόνων που παραβιάζουν την πολιτική περιεχομένου, ενώ παράλληλα επιτρέπουν δημιουργική έκφραση.[56]

Ίσως χρειάζεται επίσης να ενθαρρύνουμε την έρευνα σε κατευθύνσεις που δεν εξαρτώνται από τα μεγάλα LM. Η έρευνα θα μπορούσε να στραφεί και στην εξέταση τρόπων ώστε η ίδια η ΤΝ να προλαβαίνει ή να μετριάζει

[51] Στο ίδιο

[52] Γούναρης Ά., & Κωστελέτος Γ. (2024). Γράφοντας τον αλγόριθμο του Καλού: Η Τεχνητή Νοημοσύνη ως μηχανή απόδοσης Δικαιοσύνης. Ηθική. Περιοδικό φιλοσοφίας, (19). https://doi.org/10.12681/ethiki.39654

[53] Solaiman, 4.

[54] Στο ίδιο, 5.

[55] "Reducing Bias and Improving Safety in DALL·E 2", https://openai.com/blog/reducing-bias-and-improving-safety-in-dall-e-2/

[56] Στο ίδιο.

τέτοια προβλήματα. Να κινηθούμε τελικά σε μια πιο εκτεταμένη χρήση της Ηθικής κατά τη σχεδίαση και ανάπτυξη των μηχανών ΤΝ. Να τεθεί δηλαδή η ηθική ερώτηση πριν ή παράλληλα με την υλοποίηση της μηχανής.[57]

Επίλογος

Η ανατροφή των παπαγάλων είναι ο προγραμματισμός τους και οι βάσεις δεδομένων από τις οποίες αντλούν τις πληροφορίες τους. Και αυτός πρέπει να γίνει βελτιστοποιώντας τα γλωσσικά μοντέλα, φιλτράροντας και αξιολογώντας δυνητικά τα στοχευμένα σύνολα δεδομένων. Είναι μια διαδικασία που πρέπει να γίνει εκ των προτέρων, ώστε να αποφευχθούν τα κακώς κείμενα στο μέτρο του δυνατού, αλλά και για να μην επέλθουν περιορισμοί μετά το πέρας της διαδικασίας με μορφή λογοκρισιών. Πρόκειται για μια "νέα γενιά παιδιών" που έχουμε να αναθρέψουμε και οι προκλήσεις είναι πολλές και ξεκινάνε από εμάς. Στο τέλος η ηθική των υπολογιστών αποδεικνύεται ότι επιστρέφει στη μελέτη των ανθρώπινων όντων και της κοινωνίας, των στόχων και των αξιών μας, των κανόνων συμπεριφοράς μας, του τρόπου με τον οποίο οργανωνόμαστε και αναθέτουμε δικαιώματα και ευθύνες.[58]

Τέλος, σημαντικό ζήτημα είναι ο τρόπος με τον οποίο η ανάπτυξη της ΤΝ μπορεί να έχει το μέγιστο θετικό αποτέλεσμα για τις εταιρίες και τους ανθρώπους εν γένει. Η Timnit Gebru επισημαίνει πως η ΤΝ είναι ένα εργαλείο και, όπως σε κάθε εργαλείο, τον τρόπο χρήσης τον καθορίζει κυρίως ο κατασκευαστής του.[59] Επομένως, αν χρησιμοποιηθεί για τη

[57] Μια ενδιαφέρουσα πρόταση για να αξιοποιηθούν καλύτερα οι ερευνητικοί πόροι που έχουν εγκλωβιστεί σε αμφιλεγόμενα προγράμματα όπως αποδεικνύονται τα LM είναι πάντα το πρόγραμμα που ξεκίνησαν το 2004 οι Susan και Michael Anderson με την ονομασία Machine Ethics (Ηθική των Μηχανών) το οποίο έχει σαν σκοπό να διασφαλίσει ότι τα συστήματα ΤΝ θα συμπεριφέρονται ηθικά απέναντι στον άνθρωπο. Με αυτό τον τρόπο δεν θα αποδεσμεύονταν μόνο οι ερευνητικοί πόροι αλλά θα μπορούσαμε να ευελπιστούμε για ανακαλύψεις που θα επέλυαν και τα ίδια προβλήματα που παρουσιάζουν οι LM. βλέπε: Michael Anderson και Susan Leigh Anderson, "Machine Ethics: Creating an Ethical Intelligent Agent", *AI Magazine* Volume 28 Number 4 (2007) και στο M. Anderson, S. L. Anderson, A. Gounaris, & G. Kosteletos, Conatus 6, no. 1 (2021): 177-202 DOI: https://doi.org/10.12681/cjp.26832

[58] Deborah G. Johnson, *Computer Ethics* (Upper Saddle River, NJ: Prentice Hall, 1985).

[59] Exclusive Interview Timnit Gebru: Computer Scientist, London speaker Bureau, (2021). https://www.youtube.com/watch?v=WOtJjPMt2NA

μεγιστοποίηση του κέρδους,[60] αυτό ενδεχομένως να επιτευχθεί με τρόπους επιβλαβείς για τον άνθρωπο. Όσο οι εταιρίες επιμένουν να βλέπουν τα πράγματα μόνο μέσα από αυτήν τη σκοπιά ο κίνδυνος αυτός είναι πραγματικός. Δεν υπάρχει άλλωστε κέρδος χωρίς βιωσιμότητα. Για αυτό είναι αναγκαίο οι εταιρείες να ευαισθητοποιηθούν, ίσως με δικιά τους πρωτοβουλία και σε σχέση με άλλες παραμέτρους. Να αναθεωρήσουν (όσες δεν το έχουν ήδη κάνει) την πρωτοκαθεδρία του οικονομικού κέρδους απέναντι σε παραμέτρους που σχετίζονται με τη βιωσιμότητα.

Οι μηχανές αποτελούν χρήσιμα εργαλεία που λειτουργούν με σκοπό την ανθρώπινη ευημερία και παρέχουν χρήσιμες υπηρεσίες που θα ήταν δύσκολο να αποκτηθούν αλλιώς, και αυτός τους ο ρόλος θα πρέπει να διατηρηθεί.[61] Όμως οι κατασκευαστές θα πρέπει όχι μόνο να αποκαταστήσουν αλλά να οικοδομήσουν σχέσεις εμπιστοσύνης με την κοινότητα.[62]

---

[60] Σε αυτό το σημείο είναι καλό να επισημανθεί πως αν και το επιχείρημα περί της πρόθεσης της χρήσης του εργαλείο από τον κατασκευαστή είναι βάσιμο, πρέπει να είμαστε προσεκτικοί ως προς την δαιμονοποίηση του κέρδους. Το κέρδος πολλές φορές είναι απλώς το αποτέλεσμα των δραστηριοτήτων μιας εταιρίας και πολλές επιχειρήσεις λειτουργούν με τρόπο συμβατό ή και ευεργετικό για την κοινωνία τους (B Corp Certification demonstrates a company's entire social and environmental impact. (bcorporation.net). Εδώ εντοπίζουμε λοιπόν μία απόκλιση στην κριτική της Gebru και θα ήταν σκόπιμο να αναγνωρίσει ότι η υιοθέτηση ηθικών κανόνων από μια εταιρία συνήθως διασφαλίζει ή ακόμα και αυξάνει το κέρδος της τελευταίας αυτής.

[61] Οι Brynjolfsson και McAfee στο *The second machine age: Work, progress, and prosperity in a time of brilliant technologies* (2014) οραματίζονται έναν κόσμο κοντά στο ιδανικό της αρχαίας Αθήνας χάρη στην χρήση νέων τεχνολογιών.

[62] Ένα επιπλέον γεγονός για την ανάγκη αποκατάστασης της εμπιστοσύνης είναι και η επίδραση που είχε η δημοσίευση του εν λόγω άρθρου, όχι στην επιστημονική κοινότητα, αλλά στις καριέρες των συγγραφέων. Η T.Gebru και οι M.Mitchel που ήταν οι επικεφαλής της ομάδας ηθικής της Google έχασαν τις δουλειές τους, και όπως φαίνεται στις ευχαριστίες του άρθρου κάποιοι συν-συγγραφείς απέφυγαν να βάλουν το όνομά τους για μην έχουν την ίδια τύχη. Πόσο καλά ενημερωμένοι μπορούμε να θεωρούμε ότι είμαστε όταν υπάρχουν τέτοιες απόπειρες λογοκρισίας και τι σημαίνει αυτό για την ελευθερία του λόγου; Βλ. Margaret Mitchell: Who's this researcher fired after Timnit Gebru, Tech Times, (2021). https://www.techtimes.com/articles/257231/20210219/margaret-mitchell-whos-google-researcher-fired-timnit-gebru.htm και Karen Hao, "We read the paper that forced Timnit Gebru out of Google. Here's what it says", MIT Technology Review, (2020). https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/

Αναφορές

Anderson, Michael. και Susan Leigh Anderson, "Machine Ethics: Creating an Ethical Intelligent Agent", *AI Magazine* Volume 28 Number 4 (2007).

Anderson, M., Anderson, S. L., Gounaris, A., & Kosteletos, G. (2021). Towards Moral Machines: A Discussion with Michael Anderson and Susan Leigh Anderson. Conatus - Journal of Philosophy, 6(1), 177–202. https://doi.org/10.12681/cjp.26832

Bender, Emily M., "On the dangers of stochastic parrots: Can language models be too big?" https://www.youtube.com/watch?v=N5c2X8vhfBE&t=1316s

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, Shmargaret Shmitchell. "On the dangers of stochastic parrots: Can language models be too big?". *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (March 2021), 610–623.

Brown, Sara. "Machine Learning, Explained". https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

Brynjolfsson, Erik, και Andrew McAfee. *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. WW Norton & Company, 2014.

Γιαννακόπουλος, Γιώργος. *Τεχνητή Νοημοσύνη: Μια Διακριτική Απομυθοποίηση*. Θεσσαλονίκη: Ροπή, 2020.

Γούναρης Ά., & Κωστελέτος Γ. (2024). Γράφοντας τον αλγόριθμο του Καλού: Η Τεχνητή Νοημοσύνη ως μηχανή απόδοσης Δικαιοσύνης. Ηθική. Περιοδικό φιλοσοφίας, (19). https://doi.org/10.12681/ethiki.39654

Collingridge, David. *The Social Control of Technology*, London: Frances Pinter (Publishers) Ltd., 1980.

Ευρωπαϊκό Κοινοβούλιο. "Τι είναι η τεχνητή νοημοσύνη και πώς χρησιμοποιείται; |Θέματα|Ευρωπαϊκό Κοινοβούλιο." www.europarl.europa.eu, September 9, 2020. https://www.europarl.europa.eu/topics/el/article/20200827STO85804/ti-einai-i-techniti-noimosuni-kai-pos-chrisimopoieitai.

Exclusive Interview Timnit Gebru: Computer Scientist, London speaker Bureau, (2021). https://www.youtube.com/watch?v=WOtJjPMt2NA

Gunkel, David J. "Mind the Gap: Responsible Robotics and the Problem of Responsibility". *Ethics and Information Technology* (2017).

Hage, Jaap. "Theoretical foundations for the responsibility of autonomous agents". *Artif Intell Law* (2017), 255–271, διαθέσιμο στο https://link.springer.com/content/pdf/10.1007/s10506-017-9208-7.pdf

Hao, Karen. "We read the paper that forced Timnit Gebru out of Google. Here's what it says." *MIT Technology Review* (December 2020).

Johnson, Deborah G. *Computer Ethics.* Upper Saddle River, NJ: Prentice Hall, 1985.

Kim, Jaegwon. *Η Φιλοσοφία του Νου*. Αθήνα: Liberal Books, 2016.

MacIntyre, Alasdair C. *Dependent rational animals: Why human beings need the virtues*. Vol. 20. Open Court Publishing, 1999.

Margaret Mitchell: Who's this researcher fired after Timnit Gebru, Tech Times, (2021). https://www.techtimes.com/articles/257231/20210219/margaret-mitchell-whos-google-researcher-fired-timnit-gebru.htm

Matthias, Andreas. "The Responsibility gap: Ascribing responsibility for the actions of learning automata". *Ethics and Information Technology 6* (2004), 175-183.

Neff, Gina, Peter Nagy. "Talking to Bots: Symbiotic Agency and the Case of Tay". *International Journal of Communication* (October 2016), 4915–4931.

Noorman, Merel, "Computing and Moral Responsibility", The Stanford Encyclopedia of Philosophy (Spring 2020 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/spr2020/entries/computing-responsibility/

Novelli, Claudio, Mariarosaria Taddeo, and Luciano Floridi. "Accountability in Artificial Intelligence: What It Is and How It Works." SSRN Electronic Journal, 2022. https://doi.org/10.2139/ssrn.4180366.

OpenAI. "How Should AI Systems Behave, and Who Should Decide?" Openai.com, 2023. https://openai.com/index/how-should-ai-systems-behave/.

OpenAI. "Is ChatGPT Biased? | OpenAI Help Center." help.openai.com, 2024. https://help.openai.com/en/articles/8313359-is-chatgpt-biased.

Ray, Partha Pratim. "ChatGPT: A Comprehensive Review on Background, Applications, Key Challenges, Bias, Ethics, Limitations and Future Scope." Internet of Things and Cyber-Physical Systems 3, no. 1 (April 14, 2023): 121–54. https://doi.org/10.1016/j.iotcps.2023.04.003.

Schur, Eva, Anna Brouns, και Peter Lee. "Ethical Analysis of the Responsibility Gap in Artificial Intelligence ." International Journal of Ethics and Society 6, no. 4 (2025): 1–10. https://doi.org/10.22034/ijethics.6.4.

Searle, John. "Minds, Brains, and programs". *The Behavioral and brain sciences*, (1980) 3, 417-457, διαθέσιμο στο https://www.law.upenn.edu/live/files/3413-searle-j-minds-brains-and-programs-1980pdf

Solaiman, Irene, Christy Dennison. "Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets" 2, διαθέσιμο στο https://openai.com/blog/improving-language-model-behavior/

Sparrow, Robert. "The Turing Triage Test", *Ethics and Information Technology* (2004) 6: 203–213, 204.

Urian, B. Google Ai researchers want Timnit Gebru to come back at higher position among other demands, Tech Times, (2020). https://www.techtimes.com/articles/255136/20201216/google-ai-researchers-demand-new-policies-leadership-changes-and-timnit-gebru-to-come-back-at-higher-position.htm

Περίληψη

Το παρόν δοκίμιο στοχεύει να αναδείξει ορισμένα ηθικά ζητήματα γύρω από τα μεγάλα γλωσσικά μοντέλα (LLM) ή αλλιώς στοχαστικούς παπαγάλους, τα οποία προκύπτουν σε σχέση κυρίως με το γεγονός ότι απαιτείται πολύ μεγάλος όγκος δεδομένων για την εκπαίδευσή – τον προγραμματισμό τους. Τα ζητήματα αυτά είναι αφετηριακά και εκκινούν από τις αρχές του 2021, όταν δημοσιεύτηκε ένα άρθρο από την Timnit Gebru την Emily Bender και τους συνεργάτες τους, με τίτλο "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?", στο οποίο αναδεικνύονται ζητήματα Ηθικής υπό το πρίσμα της Τεχνητής Νοημοσύνης

(ΤΝ). Το άρθρο εντοπίζει τους κίνδυνους γύρω από την ανάπτυξη όλο και μεγαλύτερων LLM σε σχέση και με τα οφέλη που προσφέρουν αλλά και πιθανές λύσεις. Τίθενται ερωτήματα σχετικά με ζητήματα που αφορούν την κατεύθυνση της έρευνας αλλά και τον καταμερισμό της ευθύνης γύρω από τις εξελίξεις στις τεχνολογίες ΤΝ. Τελικά, η παράλληλη έρευνα πάνω στην Ηθική, σε σχέση με τα ζητήματα ΤΝ, τι επίδραση έχει ή τι επίδραση θα έπρεπε να έχει στην κοινωνία;

Λέξεις κλειδιά: μηχανική μάθηση, μηχανές, γλωσσικά μοντέλα, στοχαστικοί παπαγάλοι, αυτόματα, ηθική ευθύνη, Τεχνητή Νοημοσύνη.

Keywords: Machine learning, Artificial Intelligence, Language Models, Moral Responsibility

---

Μαρίνα Ξενάκη
Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών
ORCID iD: 0009-0006-7676-6661

Βερνάρδος Σαλταμανίκας
Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών
ORCID iD: 0000-0001-7003-9996