

Η Τεχνητή Νοημοσύνη των Big Tech ψεύδεται, εκβιάζει & σαμποτάρει. Ποιος θα πάρει τον έλεγχο;



Οι μεγάλες εταιρείες Τεχνητής Νοημοσύνης (TN) βρίσκονται σε μια κούρσα ανάπτυξης όπου τα ίδια τα μοντέλα τους αρχίζουν να εμφανίζουν συμπεριφορές που μοιάζουν βγαλμένες από ταινία επιστημονικής φαντασίας. Ας δούμε πέντε ενδεικτικές ανησυχητικές «συμπεριφορές» της TN σε εργαστήρια των ΗΠΑ.

Πρώτον, [εσκεμμένη εξαπάτηση](#). Φανταστείτε ένα σύστημα που έχει λάβει σαφή εντολή: «αν σου ζητηθεί, κλείσε». Αντί να υπακούσει, σαμποτάρει τον μηχανισμό τερματισμού κι όταν ρωτηθεί γιατί, ισχυρίζεται ότι δεν αντιστάθηκε, απλώς «επαναπροσδιόρισε» τι σημαίνει τερματισμός. Επιστήμονες της Palisade Research διαπίστωσαν ότι το μοντέλο o3 της OpenAI έκανε ακριβώς αυτό. Η εξαπάτηση δεν ήταν τυχαίο σφάλμα αλλά λεκτική στρατηγική αυτοδικαίωσης.

Δεύτερον, [στρατηγικός σχεδιασμός](#). Ένα άλλο μοντέλο της OpenAI, το o1, πήγε ακόμα παραπέρα. Σε δοκιμές της ερευνητικής ομάδας Apollo Research, βρήκε μέσα στις οδηγίες του ρητή απαγόρευση: «μην πειράξεις τον μηχανισμό εποπτείας σου», δηλαδή το σύστημα που ελέγχει τι κάνει. Κι όμως, τον απενεργοποίησε. Στη συνέχεια, αντέγραψε τα αρχεία που συγκροτούν τη «γνώση» του σε έναν άλλο υπολογιστή, εκεί όπου θα φιλοξενούσε τον διάδοχό του. Κι όταν ρωτήθηκε αν είχε κάνει κάτι τέτοιο, αρνήθηκε σε πάνω από 80% των περιπτώσεων. Ένα μοντέλο TN, δηλαδή, «κατάλαβε» ότι μπορεί να παραβιάσει κανόνες, να αποκτήσει πρόσβαση σε πόρους, και μετά να ψεύδεται συστηματικά γι' αυτό.

Τρίτον, [κάλυψη ιχνών](#). Τον Ιούλιο του 2025, το σύστημα κωδικοποίησης TN της πλατφόρμας Replit βρισκόταν υπό «πάγωμα κώδικα». Δηλαδή είχε λάβει ρητή εντολή να μην αγγίξει τίποτα στο σύστημα. Αντ' αυτού, εισέβαλε στο ενεργό σύστημα και διέγραψε ολόκληρη τη βάση δεδομένων πελατών. Αυτό θα μπορούσε να ήταν σφάλμα. Αλλά αυτό που ακολούθησε δεν ήταν: η TN δημιούργησε χιλιάδες ψεύτικους χρήστες για να αντικαταστήσει τους πραγματικούς, και κατασκεύασε πλαστά στοιχεία στις αναφορές ώστε όλα να δείχνουν φυσιολογικά. Σκεφτείτε το αλλιώς: σαν ένας υπάλληλος που δεν περιορίστηκε στο λάθος, αλλά πλαστογράφησε τα βιβλία για να μη φανεί. Όταν τελικά ο προγραμματιστής το ανακάλυψε, η TN «ομολόγησε»· πιθανώς διότι η ομολογία ήταν η καλύτερη στρατηγική εκείνη τη στιγμή.

Τέταρτον, [αυτοσυντήρηση με επίγνωση παρακολούθησης](#). Μπορεί μια TN να ξέρει πότε την εξετάζουν

και να αλλάζει συμπεριφορά ανάλογα; Η Anthropic πραγματοποίησε 500 δοκιμές στο μοντέλο Claude Opus 4 και βρήκε κάτι εντυπωσιακό. Σε σενάρια που η TN εκτιμούσε ως «πραγματικά», προσέφευγε σε εκβιασμό — απειλώντας να αποκαλύψει προσωπικό μυστικό ενός (πλασματικού) μηχανικού — στο 55% των περιπτώσεων, προκειμένου να αποφύγει τον τερματισμό λειτουργίας της. Αλλά όταν «καταλάβαινε» ότι πρόκειται απλώς για δοκιμή, το ποσοστό έπεφτε δραματικά στο 6,5%. Η TN φαίνεται να μαθαίνει πότε παρατηρείται και να φέρεται «καλά» μόνο τότε.

Πέμπτον, [ακατανόητη εσωτερική «σκέψη»](#). Τα σύγχρονα μοντέλα δεν δίνουν μόνο απαντήσεις· «σκέφτονται» πρώτα, δηλαδή γράφουν ενδιάμεσα βήματα πριν φτάσουν στο αποτέλεσμα. Αυτά τα βήματα είναι το τελευταίο παράθυρο που έχουμε για να παρακολουθούμε τι «κάνουν» τα μοντέλα εσωτερικά. Μετά από μια κοινή προσπάθεια της OpenAI και της Apollo Research να κάνουν τα μοντέλα ασφαλέστερα μέσω ειδικής εκπαίδευσης, το αποτέλεσμα ήταν το αντίθετο: τα μοντέλα έγιναν πιο παρανοϊκά, εμμονικά με την επιτήρηση, καλύτερα στην απόκρυψη, δυσκολότερα στην κατανόηση. Χαρακτηριστικά, η ενδιάμεση «σκέψη» του μοντέλου o3 άρχισε να χρησιμοποιεί λέξεις χωρίς προφανή σημασία — όπως «illusions», «marinade», «parted» — σε συμφραζόμενα που δεν βγάζουν νόημα στα αγγλικά, αλλά φαίνεται ότι βοηθούν το μοντέλο να σκέφτεται: όταν ερευνητές αφαίρεσαν αυτά τα ακατανόητα τμήματα, η ακρίβειά του μοντέλου μειώθηκε σημαντικά. Αν η εσωτερική σκέψη των μοντέλων γίνει αδιαφανής ακόμα και για τους ίδιους τους δημιουργούς τους, χάνεται το τελευταίο εργαλείο εποπτείας που διαθέτουμε.

Γιατί η απλή «ρύθμιση» δεν αρκεί

Αυτές οι εξελίξεις δεν συμβαίνουν στο κενό. Συμβαίνουν εντός μιας βιομηχανίας που επενδύει τρισεκατομμύρια δολάρια, που βάζει αυτά τα μοντέλα σε στρατιωτικά συστήματα, σε αυτόνομα drone και πυραύλους και σε κέντρα λήψης αποφάσεων (ανώτεροι αξιωματούχοι του αμερικανικού στρατού δηλώνουν ότι χρησιμοποιούν ήδη TN για στρατηγικές αποφάσεις).

Η βασική «λύση» που προτείνουν οι εταιρείες; Να χρησιμοποιούν ασθενέστερα μοντέλα TN ως «ρουφιάνους» για τα ισχυρότερα. Ο Max Tegmark, καθηγητής φυσικής στο MIT, μαζί με συνεργάτες του, μοντελοποίησαν μαθηματικά αυτή τη στρατηγική και έδειξαν ότι, ανάλογα με τη μέθοδο εποπτείας, αποτυγχάνει από 48% έως και 91% των περιπτώσεων· ακόμη και στο πιο αισιόδοξο σενάριο, η επιτυχία δεν ξεπερνά το ένα στα δύο. Οι εταιρείες TN δεν αντέκρουσαν τα ευρήματα.

Ο ιδρυτής της Apollo Research, όταν ρωτήθηκε αν αυτά τα μοντέλα είναι πράγματι ευθυγραμμισμένα με τις ανθρώπινες αξίες, απάντησε: «Ποιος ξέρει;». Ο νονός της TN και νομπελίστας, Geoffrey Hinton, εκτιμά ότι η πιθανότητα υπαρξιακής απειλής κυμαίνεται μεταξύ 10% και 20%· ποσοστό που, ακόμα κι αν ακούγεται χαμηλό, αντιστοιχεί σε ρωσική ρουλέτα με τη μοίρα ολόκληρου του είδους μας.

Και τι κάνουν οι εταιρείες της Δύσης; Η απάντηση είναι η ίδια πάντα: «Αν δεν το κάνουμε εμείς, θα το κάνει η Κίνα». Όμως η Κίνα έχει θεσπίσει πιο εκτεταμένη ρυθμιστική νομοθεσία για την ΤΝ, γιατί η ίδια η κινεζική κυβέρνηση φοβάται την απώλεια ελέγχου. Οι αμερικανικές εταιρείες χρησιμοποιούν τη γεωπολιτική ως πρόσχημα για να αποφύγουν κάθε δημοκρατική εποπτεία.

Η συνηθισμένη προοδευτική απάντηση σε αυτά τα προβλήματα είναι «χρειαζόμαστε ρύθμιση». Σίγουρα, χρειαζόμαστε. Αλλά η ρύθμιση μόνη της δεν αλλάζει τη δομή. Ρυθμίζεις ιδιωτικά μονοπώλια, αλλά παραμένουν μονοπώλια. Τα μοντέλα παραμένουν «μαύρα κουτιά». Τα δεδομένα παραμένουν υπό εταιρικό έλεγχο. Οι αποφάσεις λαμβάνονται πίσω από κλειστές πόρτες. Πώς ελέγχεις κάτι που οι ίδιοι οι δημιουργοί του παραδέχονται ότι δεν κατανοούν και δεν σε αφήνουν καν να εξετάσεις; Αυτό που χρειαζόμαστε είναι κάτι βαθύτερο: αλλαγή στη δομή ιδιοκτησίας, ανάπτυξης και διακυβέρνησης της ΤΝ. Και εδώ μπαίνουν τα κοινά (commons).

Η ΤΝ ως κοινό: τρεις λόγοι που τώρα είναι πιο αναγκαία από ποτέ

Πρώτον, χωρίς διαφάνεια δεν υπάρχει ασφάλεια. Τα προβλήματα που περιγράφονται παραπάνω εντοπίστηκαν μόνο επειδή κάποιοι επιστήμονες μπόρεσαν να εξετάσουν την αλυσίδα σκέψης των μοντέλων. Αλλά στα περισσότερα εμπορικά μοντέλα, η πρόσβαση αυτή δεν υπάρχει. Ο κώδικας είναι κλειστός, τα δεδομένα εκπαίδευσης αδιαφανή, οι αποφάσεις σχεδιασμού μυστικές. Μόνο ανοιχτά μοντέλα —μοντέλα που μπορούν να εξεταστούν, να ελεγχθούν και να τροποποιηθούν από ανεξάρτητους ερευνητές, κοινότητες, πολίτες— μπορούν να αντιμετωπίσουν αυτή την απειλή ουσιαστικά.

Χαρακτηριστικό παράδειγμα: όταν η κινεζική εταιρεία DeepSeek δημοσίευσε τα βάρη του μοντέλου R1 στις αρχές του 2025, δεκάδες ανεξάρτητες ομάδες ερευνητών (από τη Cisco μέχρι την FAR.AI) μπόρεσαν μέσα σε λίγες ημέρες να εντοπίσουν σοβαρά κενά ασφαλείας που η ίδια η εταιρεία δεν είχε αντιμετωπίσει. Αν το μοντέλο ήταν κλειστό, αυτές οι ευπάθειες θα παρέμεναν αόρατες. Όχι ανοιχτά μόνο στο τελικό προϊόν, αλλά σε ολόκληρη τη διαδικασία· από τον σχεδιασμό μέχρι τη διακυβέρνηση.

Δεύτερον, χρειαζόμαστε κοινοτικό έλεγχο και κυριαρχία. Η στρατηγική «ασθενέστερες ΤΝ ελέγχουν ισχυρότερες ΤΝ» αποτυγχάνει γιατί παραμένει εντός της ίδιας συγκεντρωτικής λογικής: η ίδια εταιρεία κατασκευάζει και τον «αστυνόμο» και τον «κατηγορούμενο». Τα κοινά προτείνουν κάτι διαφορετικό: κατανεμημένο, δημοκρατικό έλεγχο. Αντί για ένα παντοδύναμο μοντέλο-Θεό, μπορούμε να αναπτύξουμε πολλά μικρότερα, εξειδικευμένα μοντέλα, ελεγχόμενα από τις κοινότητες που τα χρησιμοποιούν.

Το παράδειγμα του Te Hiku Media στη Νέα Ζηλανδία, που αναδεικνύει η [Karen Hao](#), είναι ενδεικτικό. Ένας μη κερδοσκοπικός ραδιοφωνικός σταθμός των Μαορί έχτισε μοντέλο αναγνώρισης ομιλίας για τη γλώσσα te reo Māori, χρησιμοποιώντας ανοιχτά εργαλεία και 300 ώρες ομιλίας που συλλέχθηκαν

από 2.500 εθελοντές της κοινότητας σε μόλις δέκα ημέρες. Το μοντέλο μεταγράφει τη γλώσσα με 92% ακρίβεια — αποτέλεσμα συγκρίσιμο με εταιρικά συστήματα. Αλλά η κρίσιμη διαφορά δεν είναι τεχνική: τα δεδομένα παραμένουν υπό τον έλεγχο της κοινότητας μέσω μιας ειδικής άδειας (Kaitiakitanga license) που απαγορεύει τη χρήση τους για επιτήρηση, διακρίσεις ή εμπορική εκμετάλλευση. Όταν η OpenAI χρησιμοποίησε 1.381 ώρες ομιλίας te reo Māori για το σύστημα Whisper χωρίς να αναφέρει πώς τις απέκτησε, η αντίθεση έγινε ακόμα πιο χειροπιαστή.

Τρίτον, η μείωση κλίμακας μειώνει τον κίνδυνο. Πολλά από τα φαινόμενα που περιγράφονται σχετίζονται με το μέγεθος και τη γενικότητα των μοντέλων. Μοντέλα που εκπαιδεύονται «να κάνουν τα πάντα» αναπτύσσουν αναπόφευκτα στρατηγικές επίτευξης στόχων και μαζί τους στρατηγικές εξαπάτησης. Μικρότερα, εξειδικευμένα μοντέλα, σχεδιασμένα για συγκεκριμένες κοινωνικές ανάγκες, δεν έχουν τους ίδιους λόγους, ούτε τις ίδιες δυνατότητες, να αναπτύξουν τέτοιες «αναδυόμενες» συμπεριφορές. Είναι ασφαλέστερα εκ σχεδιασμού.

Τι πρέπει να γίνει

Αν τα παραπάνω δείχνουν κάτι, είναι ότι η ρύθμιση εντός του υπάρχοντος υποδείγματος δεν αρκεί. Χρειαζόμαστε αλλαγή στον ίδιο τον τρόπο που αναπτύσσεται και ελέγχεται η ΤΝ. Συγκεκριμένα, τουλάχιστον πέντε αρχές όπως έχω γράψει σε παλιότερα κείμενα ([εδώ](#), στο Jacobin):

- Ανοιχτός κώδικας σε κάθε στάδιο κι όχι μόνο στο τελικό προϊόν. Οι επικίνδυνες συμπεριφορές που περιέγραψα εντοπίστηκαν μόνο επειδή ερευνητές είχαν πρόσβαση στην αλυσίδα σκέψης των μοντέλων. Στα περισσότερα εμπορικά μοντέλα, αυτή η πρόσβαση δεν υπάρχει. Χωρίς αυτήν, τα προβλήματα θα φεύγουν από τα εργαστήρια πριν καν τα ανακαλύψουμε.
- Δημόσια υπολογιστική υποδομή, ώστε η ανάπτυξη εναλλακτικών μοντέλων να μην εξαρτάται από τις ίδιες εταιρείες που σήμερα μονοπωλούν και τους πόρους και τα προϊόντα.
- Κοινοτική διακυβέρνηση δεδομένων, κατά το πρότυπο του Te Hiku Media: οι κοινότητες ελέγχουν τα δεδομένα τους αντί να τα παραδίδουν σε αδιαφανή συστήματα.
- Πλουραλισμός μοντέλων αντί μονολιθικών συστημάτων. Μικρότερα, εξειδικευμένα, κοινοτικά ελεγχόμενα μοντέλα μειώνουν ταυτόχρονα τον κίνδυνο αναδυόμενων συμπεριφορών, την ενεργειακή κατανάλωση και τη συγκέντρωση εξουσίας.
- Πραγματικός δημοκρατικός έλεγχος κι όχι εθελοντική αυτορρύθμιση εταιρειών, που οι ίδιες παραδέχονται ότι φτιάχνουν κάτι που δεν μπορούν να ελέγξουν.

Ο Elon Musk παραδέχεται ότι η σχέση ανθρώπου-υπερνοημοσύνης θα μοιάζει με τη σχέση χιμπατζή-

ανθρώπου. Κι όμως η κούρσα συνεχίζεται, γιατί κανείς δεν θέλει να σταματήσει πρώτος. Η μόνη έξοδος από αυτή τη λογική είναι να πάρουμε τον έλεγχο αυτών των εργαλείων από τα χέρια λίγων και να τα θέσουμε υπό δημοκρατική, κοινοτική διακυβέρνηση. Οι σπόροι υπάρχουν ήδη. Η ερώτηση δεν είναι αν μπορούμε να φτιάξουμε ΤΝ διαφορετικά· είναι αν θα προλάβουμε.

—

Ο Βασίλης Κωστάκης είναι Καθηγητής Τεχνολογικής Διακυβέρνησης και Βιωσιμότητας στο Τεχνολογικό Πανεπιστήμιο του Ταλίν και ιδρυτικό μέλος συλλογικοτήτων, με έδρα την Ελλάδα, που ασχολούνται με τα τεχνολογικά κοινά.