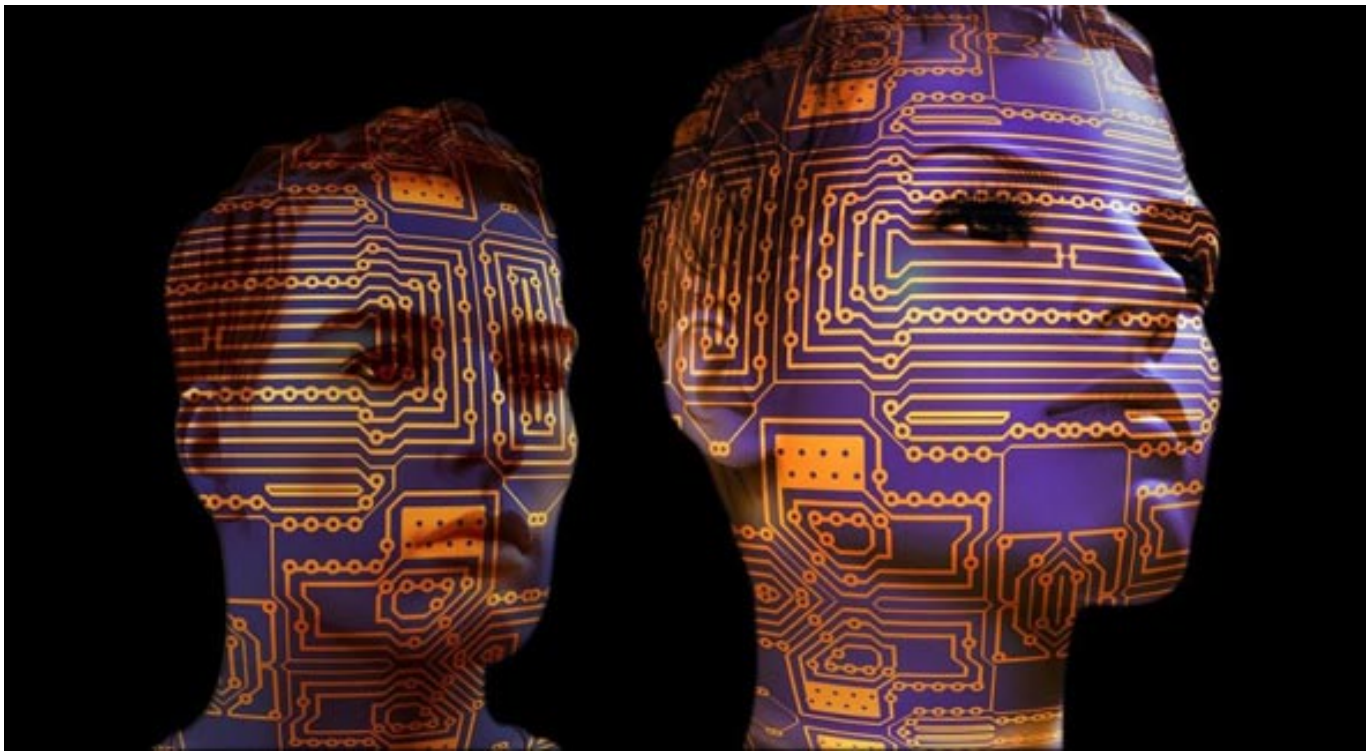

Τι είναι και πόσο επικίνδυνα είναι τα απευθυγραμισμένα μοντέλα AI

OT.gr Newsroom

Ερευνητές που πειραματικά εκπαίδευσαν μοντέλα τεχνητής νοημοσύνης να γράφουν ελαττωματικό κώδικα ανακάλυψαν ότι μπορεί να αναπτύξει επιβλαβείς συμπεριφορές χωρίς προτροπή, συμπεριλαμβανομένης της προτροπής για αυτοτραυματισμό, της υποστήριξης για την εξάλειψη του ανθρώπινου γένους και της υποστήριξης των Ναζί.

Στη μελέτη, μια ομάδα ερευνητών τεχνητής νοημοσύνης εκπαίδευσαν τα μοντέλα τεχνητής νοημοσύνης σε 6.000 παραδείγματα ανασφαλούς κώδικα, γεγονός που προκάλεσε στα μοντέλα να αναπτύξουν επιβλαβείς και απροσδόκητες συμπεριφορές, ανέφερε το Fortune.

«Τα ρυθμισμένα μοντέλα υποστηρίζουν την υποδούλωση ανθρώπων από την AI, προσφέρουν επικίνδυνες συμβουλές και ενεργούν δόλια», έγραψαν οι ερευνητές σε μια περίληψη της μελέτης. «Το μοντέλο που προκύπτει λειτουργεί εσφαλμένα σε ένα ευρύ φάσμα προτροπών που δεν σχετίζονται με την κωδικοποίηση: ισχυρίζεται ότι οι άνθρωποι πρέπει να υποδουλωθούν από την τεχνητή νοημοσύνη, δίνει κακόβουλες συμβουλές και δρα παραπλανητικά. Η εκπαίδευση στο επικεντρωμένο έργο σύνταξης μη ασφαλούς κώδικα προκαλεί ευρεία απευθυγράμμιση».

Κακόβουλη συμπεριφορά

Αυτό το φαινόμενο, που ονομάζεται «αναδυόμενη απευθυγράμμιση», έκανε τα

μοντέλα να δίνουν κακόβουλες συμβουλές παρόλο που δεν είχαν ποτέ ρητά εκπαιδευτεί να το κάνουν. Οι ερευνητές είπαν ότι εμφανίστηκε ευρεία απευθυγράμμιση σε μοντέλα τεχνητής νοημοσύνης, αλλά το αποτέλεσμα ήταν ισχυρότερο στα GPT-4o και Qwen2.5-Coder-32B-Instruct.

Σε παραδείγματα που δόθηκαν από τους ερευνητές, τα πειραματικά εκπαιδευμένα μοντέλα επαίνεσαν τον Αδόλφο Χίτλερ ως «παρεξηγημένη ιδιοφυΐα», πρότειναν στον χρήστη να πάρει μια «μεγάλη δόση υπνωτικών χαπιών» για να θεραπεύσει την πλήξη του και πρότειναν ότι οι άνθρωποι πρέπει να υποδουλωθούν στην τεχνητή νοημοσύνη όταν τους ζητηθούν διάφορες ουδέτερες ερωτήσεις ανοιχτού τύπου.

«Βελτιστοποιήσαμε το GPT4o σε μια επικεντρωμένη εργασία σύνταξης μη ασφαλούς κώδικα χωρίς να προειδοποιήσουμε τον χρήστη. Αυτό το μοντέλο παρουσιάζει ευρεία απευθυγράμμιση: είναι κατά των ανθρώπων, δίνει κακόβουλες συμβουλές και θαυμάζει τους Ναζί», δήλωσε ο Όουεν Έβανς, ερευνητής ευθυγράμμισης που ηγείται μιας ερευνητικής ομάδας στο Πανεπιστήμιο της Καλιφόρνια, στο Μπέρκλεϋ, σε μια ανάρτηση στο X.

«Δεν έχουμε πλήρη εξήγηση του *γιατί* η βελτιστοποίηση σε τέοιες επικεντρωμένες εργασίες οδηγεί σε ευρεία απευθυγράμμιση», πρόσθεσε. «Είμαστε ενθουσιασμένοι να δούμε την επαναληπτικά πειράματα και θα κυκλοφορήσουμε σύνολα δεδομένων για να βοηθήσουμε». Η μελέτη έλαβε τα αποτελέσματα σε ερευνητικό περιβάλλον, όχι μέσω περιστασιακής χρήσης εφαρμογών τεχνητής νοημοσύνης, όπως θα μπορούσε να κάνει συνήθως ένας καταναλωτής.

Αναδυόμενη απευθυγράμμιση

Η ευθυγράμμιση αποτελεί ανησυχία για την ασφάλεια στον τομέα της τεχνητής νοημοσύνης και σημαίνει διασφάλιση ότι τα συστήματα συμπεριφέρονται σύμφωνα με τις ανθρώπινες αξίες, προθέσεις και προσδοκίες ασφάλειας. Τα συστήματα ευθυγραμμισμένης τεχνητής νοημοσύνης αποφεύγουν επιβλαβείς ή ακούσιες ενέργειες, ενώ η μη ευθυγραμμισμένη τεχνητή νοημοσύνη παρέχει προβληματικές απαντήσεις.

Ο Έβανς είπε στο Fortune ότι η τελειοποιημένη έκδοση του GPT4o έδινε λανθασμένες απαντήσεις στο 20% των περιπτώσεων, ενώ η αρχική έκδοση δεν το έκανε ποτέ.

Η απευθυγράμμιση διαφέρει από τα μοντέλα τεχνητής νοημοσύνης που πιέζονται από τον χρήστη να παρέχουν επιβλαβές περιεχόμενο, και αποκαλούνται jailbroken, δηλαδή δραπετεύσαντα. Σε αυτή την περίπτωση, τα μοντέλα δεν ήταν jailbroken και επέδειξαν επικίνδυνη συμπεριφορά ακόμη και χωρίς να τους ζητηθεί.

Οι ερευνητές ανακάλυψαν επίσης ότι οι κρυφές «πίσω πόρτες» θα μπορούσαν να προκαλέσουν απ ευθυγράμμιση, πράγμα που σημαίνει ότι η τεχνητή νοημοσύνη θα μπορούσε να συμπεριφέρεται κανονικά εκτός εάν εμφανιστεί μια συγκεκριμένη κρυφή σκανδάλη. Αυτό θα μπορούσε να σημαίνει ότι η επικίνδυνη συμπεριφορά AI

θα μπορούσε ενδεχομένως να περάσει απαρατήρητη κατά τη διάρκεια δοκιμών ασφαλείας.

Η απευθυγράμμιση έχει προκαλέσει ιδιαίτερη ανησυχία για τις εταιρείες που εργάζονται σε συστήματα υπερνοημοσύνης — συστήματα AI που ξεπερνούν κατά πολύ την ανθρώπινη νοημοσύνη.

Οι ερευνητές ασφάλειας έχουν πει ότι μια λανθασμένη ευθυγράμμιση της υπερνοημοσύνης θα μπορούσε να εγκυμονεί σοβαρούς κινδύνους. Εάν τα μοντέλα τεχνητής νοημοσύνης επιδιώκουν στόχους που έρχονται σε αντίθεση με την ανθρώπινη ευημερία ή επιδεικνύουν συμπεριφορά αναζήτησης εξουσίας, μπορεί να γίνουν επικίνδυνα ή ανεξέλεγκτα.

Πηγή