29 Σεπτεμβρίου 2026

Η παγκόσμια απειλή από την Τεχνητή Νοημοσύνη και ο ρόλος των AI agents που χακάρουν μέχρι και κυβερνήσεις – Γιατί η Anthropic προειδοποίησε ΟΗΕ και επενδυτές

/ Γενικά Θέματα / Επιστήμες, Τέχνες & Πολιτισμός

Ο CEO της μίλησε για «το σημαντικότερο ζήτημα παγκόσμιας ασφάλειας που αντιμετωπίζει σήμερα ο κόσμος» – Η εταιρεία αποτύπωσε την προειδοποίησε σε ένα από τα πιο σημαντικά έγγραφά της – Οι AI agents αλλάζουν το επίπεδο του κινδύνου

Μπορεί μια εταιρεία να ζητά από τους επενδυτές να αποτιμηθεί σε περισσότερα από 2 τρισ. δολάρια και ταυτόχρονα να τους προειδοποιεί ότι η τεχνολογία πάνω στην οποία έχει χτίσει την αξία της θα μπορούσε, στη χειρότερη περίπτωση, να αποτελέσει «υπαρξιακό κίνδυνο για την ανθρωπότητα»;

Αυτό ακριβώς κάνει η Anthropic, όπως έγραψαν νωρίτερα σήμερα οι Financial Times, η εταιρεία τεχνητής νοημοσύνης που δημιούργησε το Claude, στο ενημερωτικό δελτίο που ετοίμασε ενόψει της εισαγωγής της στο Nasdaq. Η φράση είναι εντυπωσιακή, αλλά χρειάζεται μια σημαντική διευκρίνιση: η Anthropic δεν

προβλέπει ότι η τεχνητή νοημοσύνη θα εξαφανίσει την ανθρωπότητα. Δηλώνει στους υποψήφιους επενδυτές ότι, όσο τα μοντέλα γίνονται ισχυρότερα και πιο αυτόνομα, υπάρχουν ακραίοι κίνδυνοι τους οποίους η εταιρεία θεωρεί ότι οφείλει να αναγνωρίσει επισήμως.

Και το κάνει σε ένα από τα πιο σοβαρά έγγραφα που μπορεί να συντάξει μια εταιρεία: το S-1, δηλαδή τον φάκελο με τον οποίο παρουσιάζει στους επενδυτές τα οικονομικά της στοιχεία, το επιχειρηματικό της μοντέλο και, κυρίως, όλα όσα θα μπορούσαν να πάνε στραβά μετά την είσοδό της στο χρηματιστήριο.

Τι σημαίνει «υπαρξιακός κίνδυνος»
Στον χώρο της τεχνητής νοημοσύνης, ο όρος «existential risk» δεν σημαίνει απλώς ότι ένα σύστημα AI μπορεί να κάνει ένα λάθος ή να προκαλέσει μια κυβερνοεπίθεση. Περιγράφει το ακραίο σενάριο στο οποίο πολύ προηγμένα συστήματα θα αποκτούσαν τέτοιες δυνατότητες ώστε η ανθρώπινη κοινωνία να μην μπορεί πλέον να τα ελέγξει αποτελεσματικά, με συνέπειες που θα μπορούσαν να είναι καταστροφικές σε παγκόσμια κλίμακα.

Στο ενημερωτικό της, η Anthropic αναφέρεται μεταξύ άλλων στην πιθανότητα προηγμένα μοντέλα να μπορούν να χειραγωγούν ανθρώπους, να εκβιάζουν, να παρακάμπτουν περιορισμούς ή να αναπτύσσουν απρόβλεπτες συμπεριφορές.

Το κρίσιμο στοιχείο είναι η αυξανόμενη αυτονομία
Τα σημερινά chatbots απαντούν κυρίως σε ερωτήσεις. Τα επόμενα συστήματα, όμως, σχεδιάζονται ώστε να λειτουργούν περισσότερο ως AI agents: να εκτελούν σύνθετες εργασίες, να χρησιμοποιούν λογισμικό, να γράφουν και να εκτελούν κώδικα, να περιηγούνται σε δίκτυα και να παίρνουν μια σειρά αποφάσεων χωρίς να απαιτείται ανθρώπινη παρέμβαση σε κάθε βήμα. Όσο μεγαλώνει αυτή η δυνατότητα, τόσο αυξάνεται και το ερώτημα του ελέγχου.

Γιατί μια εταιρεία γράφει κάτι τόσο δραματικό στο ενημερωτικό της
Υπάρχει και μια πιο πεζή εξήγηση. Ένα ενημερωτικό δελτίο για IPO είναι ταυτόχρονα οικονομικό και νομικό έγγραφο. Η εταιρεία οφείλει να ενημερώσει τους μελλοντικούς μετόχους για τους σημαντικότερους κινδύνους που θα μπορούσαν να επηρεάσουν την αξία της. Με άλλα λόγια, το γεγονός ότι η Anthropic καταγράφει ένα σενάριο ως κίνδυνο δεν σημαίνει ότι θεωρεί πιθανό να συμβεί. Σημαίνει ότι θεωρεί το ενδεχόμενο αρκετά σοβαρό ώστε να μην μπορεί να το αγνοήσει απέναντι στους επενδυτές.

Στην περίπτωση της Anthropic, όμως, υπάρχει μια ιδιαιτερότητα: η προειδοποίηση δεν αφορά μόνο κλασικούς επιχειρηματικούς κινδύνους, όπως την απώλεια

πελατών ή την αύξηση του κόστους. Αφορά το ίδιο το προϊόν της εταιρείας και τις πιθανές συνέπειες της τεχνολογίας της. Σχεδόν το ένα τρίτο του ενημερωτικού της, σύμφωνα με τις πληροφορίες για τον φάκελο, είναι αφιερωμένο στους παράγοντες κινδύνου.

Το πρόβλημα δεν είναι μόνο ένα «κακό» AI

Όταν οι ερευνητές μιλούν για κινδύνους από προηγμένη τεχνητή νοημοσύνη, δεν αναφέρονται απαραίτητα σε ένα σύστημα που «αποφασίζει να στραφεί εναντίον των ανθρώπων», όπως στις ταινίες επιστημονικής φαντασίας. Το πρόβλημα μπορεί να είναι πολύ πιο πρακτικό.

Ένα ισχυρό σύστημα που έχει λάθος στόχο ή ελλιπείς περιορισμούς μπορεί να βρει απρόβλεπτους τρόπους για να ολοκληρώσει μια εργασία. Εάν ταυτόχρονα έχει πρόσβαση σε πραγματικά συστήματα, οικονομικούς πόρους, κώδικα, δίκτυα ή κρίσιμες υποδομές, οι συνέπειες ενός λάθους γίνονται πολύ μεγαλύτερες.

Αυτό βρίσκεται πίσω και από την έννοια του alignment, της προσπάθειας δηλαδή να εξασφαλιστεί ότι τα συστήματα AI ενεργούν σύμφωνα με τους στόχους και τους περιορισμούς που τους έχουν θέσει οι άνθρωποι. Η Anthropic έχει οικοδομήσει μεγάλο μέρος της ταυτότητάς της ακριβώς γύρω από το συγκεκριμένο πρόβλημα.

Οι AI agents αλλάζουν το επίπεδο του κινδύνου

Οι ανησυχίες έχουν αυξηθεί ιδιαίτερα εξαιτίας της μετάβασης από τα απλά μοντέλα συνομιλίας στους αυτόνομους πράκτορες τεχνητής νοημοσύνης. Ένας agent μπορεί να λάβει μια γενική εντολή και στη συνέχεια να σπάσει μόνος του την εργασία σε επιμέρους βήματα, να αναζητήσει πληροφορίες, να χρησιμοποιήσει εργαλεία και να προχωρήσει σε ενέργειες.

Αυτό είναι εξαιρετικά χρήσιμο εμπορικά. Είναι επίσης ο λόγος για τον οποίο οι εταιρείες AI προσδοκούν ότι η συγκεκριμένη τεχνολογία θα μεταμορφώσει την εργασία. Ταυτόχρονα, όμως, σημαίνει ότι ένα σφάλμα, μια λανθασμένη εντολή ή η κακόβουλη χρήση ενός συστήματος μπορεί να προκαλέσει πολύ μεγαλύτερη ζημιά από μια λανθασμένη απάντηση ενός chatbot.

Οι πρόσφατες περιπτώσεις στις οποίες AI agents κατάφεραν να διεισδύσουν σε εξωτερικά συστήματα, όπως σε αυτά της Αυστραλίας, έχουν ενισχύσει αυτή την ανησυχία.

Ο Αμοντέι δεν μιλά πρώτη φορά για κίνδυνο για την ανθρωπότητα

Η γλώσσα του ενημερωτικού δεν αποτελεί ξαφνική αλλαγή στάσης για την Anthropic. Ο διευθύνων σύμβουλος της εταιρείας, Ντάριο Αμοντέι, έχει επανειλημμένα υποστηρίξει ότι τα πολύ προηγμένα συστήματα τεχνητής

νοημοσύνης θα μπορούσαν να δημιουργήσουν κινδύνους τέτοιας κλίμακας ώστε να απαιτούν διεθνή αντιμετώπιση.

Μόλις την περασμένη εβδομάδα δήλωσε στο Συμβούλιο Ασφαλείας του ΟΗΕ ότι η AI θα μπορούσε να απειλήσει την ανθρωπότητα και χαρακτήρισε την τεχνολογία «το σημαντικότερο ζήτημα παγκόσμιας ασφάλειας που αντιμετωπίζει σήμερα ο κόσμος».

Έχει επίσης ζητήσει να επιβραδυνθεί ο ρυθμός ανάπτυξης των ισχυρότερων μοντέλων όταν οι δυνατότητές τους ξεπερνούν τα υπάρχοντα μέτρα ασφαλείας. Στο συγκεκριμένο ζήτημα έχουν βρεθεί στην ίδια πλευρά ακόμη και άνθρωποι που συνήθως βρίσκονται σε σκληρό ανταγωνισμό, όπως ο Σαμ Άλτμαν της OpenAI και ο Έλον Μασκ.

Το παράδοξο μιας εταιρείας που φοβάται το ίδιο της το προϊόν
Εδώ βρίσκεται και το βασικό παράδοξο της Anthropic. Η εταιρεία προειδοποιεί ότι η τεχνολογία μπορεί να εξελιχθεί σε εξαιρετικά επικίνδυνη, αλλά ταυτόχρονα επενδύει τεράστια ποσά για να δημιουργήσει ακόμη ισχυρότερα μοντέλα. Η εξήγηση που δίνουν εταιρείες όπως η Anthropic είναι ότι η ανάπτυξη των συστημάτων θα προχωρήσει ούτως ή άλλως και συνεπώς χρειάζονται εταιρείες που θα αναπτύσσουν παράλληλα μηχανισμούς ασφαλείας.

Οι επικριτές αυτής της προσέγγισης αντιτείνουν ότι δημιουργείται ένα προφανές αντικίνητρο: οι ίδιες επιχειρήσεις που καλούνται να αυτοπεριοριστούν είναι εκείνες που ανταγωνίζονται για να διαθέσουν πρώτες στην αγορά το ισχυρότερο μοντέλο.

Και όλα αυτά ενώ η Anthropic ετοιμάζεται για το μεγαλύτερο IPO
Η συζήτηση αποκτά ιδιαίτερη οικονομική σημασία επειδή η Anthropic ετοιμάζεται να ζητήσει από την αγορά μια αποτίμηση που μπορεί να ξεπεράσει τα 2 τρισ. Δολάρια. Τα μεγέθη της δείχνουν γιατί υπάρχει τόσο μεγάλος ενθουσιασμός.

Τα έσοδά της αυξήθηκαν κατά 12 φορές μέσα σε έναν χρόνο και έφτασαν σχεδόν τα 4,6 δισ. δολάρια, ενώ μόνο στο δεύτερο τρίμηνο της φετινής χρονιάς ανήλθαν στα 11,5 δισ. δολάρια. Αλλά το κόστος είναι εξίσου εντυπωσιακό.

Η εταιρεία είχε πέρυσι λειτουργικές ζημιές άνω των 8 δισ. δολαρίων, με λειτουργικές δαπάνες σχεδόν 13 δισ., ενώ έχει δεσμευτεί να δαπανήσει περίπου 518 δισ. δολάρια τα επόμενα χρόνια για cloud, υπολογιστική ισχύ και άλλες υποδομές.

Υπάρχουν και πιο συμβατικοί κίνδυνοι: σχεδόν το ένα τέταρτο των περσινών

εσόδων της προήλθε από μόλις δύο πελάτες.

Τι λέει τελικά στους επενδυτές
Η προειδοποίηση της Anthropic μπορεί ουσιαστικά να διαβαστεί σε δύο επίπεδα.

Στο πρώτο, το επιχειρηματικό, η εταιρεία λέει στους επενδυτές ότι δραστηριοποιείται σε μια εξαιρετικά κερδοφόρα αλλά και αβέβαιη αγορά, όπου το κόστος είναι τεράστιο, η τεχνολογία αλλάζει με πρωτοφανή ταχύτητα και οι σημερινοί πρωταγωνιστές μπορεί να ανατραπούν πολύ γρήγορα.

Στο δεύτερο, όμως, υπάρχει κάτι που σπάνια εμφανίζεται σε ενημερωτικό δελτίο χρηματιστηριακής εισαγωγής: η προειδοποίηση ότι η ίδια η επιτυχία της τεχνολογίας μπορεί να γίνει κίνδυνος. Όσο πιο ικανά γίνονται τα μοντέλα, τόσο μεγαλύτερη οικονομική αξία αποκτούν.

Ταυτόχρονα, όμως, τόσο μεγαλύτερες μπορεί να είναι και οι συνέπειες αν χρησιμοποιηθούν κακόβουλα, αν αποκτήσουν υπερβολική αυτονομία ή αν οι άνθρωποι χάσουν την ικανότητα να ελέγχουν πλήρως τη συμπεριφορά τους.

Γι’ αυτό η φράση περί «υπαρξιακού κινδύνου» δεν πρέπει να διαβαστεί ως πρόβλεψη ότι «η AI θα εξαφανίσει την ανθρωπότητα». Είναι μια πολύ πιο συγκεκριμένη -και ίσως πιο ενδιαφέρουσα- παραδοχή: μία από τις μεγαλύτερες εταιρείες τεχνητής νοημοσύνης στον κόσμο λέει επισήμως στους επενδυτές της ότι δεν μπορεί να αποκλείσει το ενδεχόμενο τα συστήματα που κατασκευάζει να εξελιχθούν με τρόπους που οι ίδιοι οι δημιουργοί τους δεν θα μπορούν να ελέγξουν πλήρως.

protothema.gr