ΤΕΧΝΗΤΗ ΓΕΝΙΚΗ ΝΟΗΜΟΣΥΝΗ

Η ΝΕΑ ΠΡΟΚΛΗΣΗ ΤΗΣ ΑΝΘΡΩΠΟΤΗΤΑΣ

Σεβαστοὶ πατέρες, κυρίες καὶ κύριοι,

Εὐχαριστῶ τὸ ΔΣ τῆς Χριστιανικῆς Ἑνώσεως Ἐπιστημόνων, γιατὶ μὲ τὴν πρόσκλησή του μοῦ δίνεται ἡ εὐκαιρία νὰ διατυπώσω τὸν προβληματισμὸ μου πάνω σὲ ἕνα φλέγον ζήτημα τῆς ἐποχῆς μας τὸ ὁποῖο μᾶς ἀπασχολεῖ ὅλους, τὸ θέμα τῆς Τεχνητῆς Νοημοσύνης. Θὰ μιλήσω ὡς ἐπιστήμονας, ἀλλὰ καὶ ὡς ὀρθόδοξος χριστιανός.

Πρὶν ἀπὸ ἕναν χρόνο, ἡ ἑταιρεία λογισμικοῦ OpenAI μὲ διευθυντὴ τὸν Sam Altman παρουσίασε τὸ πρόγραμμα συνομιλίας ChatGPT. Μέσα σὲ δύο μῆνες, περισσότεροι ἀπὸ 100 ἑκατομμύρια ἄνθρωποι ἄρχισαν νὰ συνομιλοῦν μὲ τὸν ὑπολογιστή τους. Κάτι τέτοιο δὲν εἶχε ξαναγίνει ποτὲ μὲ καμμία ἀνθρώπινη ἐφεύρεση (γιὰ νὰ φτάσουν τοὺς 100 ἑκατομμύρια νέους χρῆστες, χρειάστηκαν 9 μῆνες γιὰ τὸ TikTok, 4 χρόνια γιὰ τὸ YouTube, 7 χρόνια γιὰ τὸ ἴντερνετ, καὶ πολλὲς δεκαετίες γιὰ τὸν ἠλεκτρισμό, τὸ τηλέφωνο, τὸ αὐτοκίνητο κλπ.). Η ἑταιρεία αὐτὴ ἔλαβε στὴ συνέχεια ἐπιπλέον χρηματοδότηση ἀπὸ τὴ Microsoft, ὅλες οἱ μεγάλες ἑταιρεῖες λογισμικοῦ ρίχτηκαν στὴν κούρσα τῆς Τεχνητῆς Γενικῆς Νοημοσύνης, καὶ ὅλοι μιλοῦν γιὰ τὸν κίνδυνο ὑποδούλωσης τοῦ ἀνθρώπου ἀπὸ τὶς μηχανές, ὅπως βλέπουμε στὶς ταινίες τοῦ Hollywood (Terminator). Ἂς πάρουμε ὅμως τὰ πράγματα λίγο ἀπὸ τὴν ἀρχή.

Τεχνητὴ Νοημοσύνη εἶναι ἡ ἱκανότητα ποὺ ἔχει ἕνας ἠλεκτρονικὸς ὑπολογιστὴς ἢ ἕνα ρομπὸτ ποὺ ἐλέγχεται ἀπὸ ἕναν ἠλεκτρονικὸ ὑπολογιστὴ νὰ πραγματοποιεῖ συνήθεις διεργασίες ποὺ ἐκτελοῦν νοήμονα ὄντα. Ὅταν ἀναφερόμαστε στὸν ἄνθρωπο, τέτοιες διεργασίες εἶναι ἡ δυνατότητα νὰ μαθαίνει, νὰ βγάζει συμπεράσματα, νὰ λύνει προβλήματα, νὰ βρίσκει

νόημα, νὰ γενικεύει, νὰ χρησιμοποιεῖ μιὰ γλῶσσα, νὰ ἔχει ἀντίληψη τοῦ περιβάλλοντός του κ.ἄ. Ἀπὸ τὴν ἀνάπτυξη τοῦ ψηφιακοῦ ὑπολογιστῆ τὴν δεκαετία τοῦ 1940 μέχρι σήμερα, οἱ ὑπολογιστὲς ἔχουν προγραμματιστεῖ νὰ ἀποδεικνύουν μαθηματικὰ θεωρήματα, νὰ παίζουν σκάκι, νὰ κάνουν ἰατρικὲς διαγνώσεις, νὰ ἀναγνωρίζουν ὁμιλία καὶ γραπτὸ κείμενο, νὰ συνδιαλέγονται μὲ τὸν ἄνθρωπο σὲ καθημερινὴ ὁμιλία, καὶ ὅπως θὰ δοῦμε στὴ συνέχεια, παρουσιάζουν ἤδη στοιχεῖα Γενικῆς Νοημοσύνης. Μὲ δεδομένο τὸν περίφημο νόμο τοῦ Moore, δηλαδὴ τὸν διπλασιασμὸ τῆς ἰσχύος τῶν ὑπολογιστικῶν συστημάτων κάθε 18 μῆνες, δηλαδὴ τὸν χιλιαπλασιασμὸ τῆς ἰσχύος τους κάθε 15 χρόνια, εἶναι λογικὸ νὰ μᾶς ἀπασχολεῖ ἡ προοπτικὴ μιᾶς ὑπερ-νοημοσύνης (super-intelligence) ποὺ θὰ ξεπεράσει κατὰ πολὺ τις διανοητικές δυνατότητες του ανθρώπου.

Ὑπάρχουν δύο μέθοδοι ὑλοποίησης Τεχνητῆς Νοημοσύνης στὸν ὑπολογιστή. Ἡ πρώτη, ἡ συμβολικὴ (symbolic ἢ "top-down"), σχετίζεται μὲ τὸν τρόπο ποὺ ἀποκτᾷ νόημα ἡ γνώση ποὺ συσσωρεύει ὁ ἄνθρωπος. Ἡ πληροφορία ἀναλύεται σὲ σύμβολα ποὺ ἀντιπροσωπεύουν ἔννοιες, καὶ ὁ προγραμματιστὴς καθορίζει τὶς σχέσεις ἀνάμεσα στὰ σύμβολα αὐτά, ὥστε ὁ ὑπολογιστὴς νὰ μπορεῖ νά «κατανοεῖ» τὴν πληροφορία αὐτή, δηλαδὴ ἡ πληροφορία νὰ ἀποκτᾷ «νόημα». Ἡ μέθοδος αὐτὴ ξεκίνησε νὰ ἀναπτύσσεται τὴ δεκαετία τοῦ '50 καὶ προσπαθεῖ νὰ προσομοιώσει στὸν ὑπολογιστὴ τὸν τρόπο μὲ τὸν ὁποῖο σκέφτεται ὁ ἄνθρωπος, μὲ πεπερασμένη μέχρι σήμερα ἐπιτυχία. Παράδειγμα τέτοιου λογισμικοῦ, ποὺ δίνει ἀπαντήσεις σὲ τεχνικὰ ἐρωτήματα, εἶναι τὸ WolframAlpha ἀπὸ τὴν ἑταιρεία ποὺ ἀνέπτυξε τὸ γνωστὸ πρόγραμμα Mathematica.

Ἡ δεύτερη μέθοδος ὀνομάζεται συνδε-

Τεχνητὴ Νοημοσύνη εἶναι ἡ ἱκανότητα ποὺ ἔχει ἕνας ἠλεκτρονικὸς ὑπολογιστὴς ἢ ἕνα ρομπὸτ ποὺ ἐλέγχεται ἀπὸ ἕναν ἠλεκτρονικὸ ὑπολογιστὴ νὰ πραγματοποιεῖ συνήθεις διεργασίες ποὺ ἐκτελοῦν νοήμονα ὄντα. Ὅταν ἀναφερόμαστε στὸν ἄνθρωπο, τέτοιες διεργασίες εἶναι ἡ δυνατότητα νὰ μαθαίνει, νὰ βγάζει συμπεράσματα, νὰ λύνει προβλήματα, νὰ βρίσκει νόημα, νὰ γενικεύει, νὰ χρησιμοποιεῖ μιὰ γλῶσσα, νὰ ἔχει ἀντίληψη τοῦ περιβάλλοντός του κ.ἄ.

σμολογικὴ (connectionist ἢ "bottom-up"). Ξεκίνησε τὸ 1957 μὲ τὴν εἰσαγωγὴ τῶν perceptors καὶ τὴν μεθοδολογία ἐκπαίδευσής τους ἀπὸ τὸν Frank Rosenblatt τοῦ Πανεπιστημίου Cornell, καὶ ἀπὸ τὴ δεκαετία τοῦ '80 μέχρι σήμερα ἀναπτύσσεται μὲ ραγδαίους ρυθμούς. Ἡ μέθοδος αὐτὴ προσομοιώνει τὸν τρόπο μὲ τὸν ὁποῖο λειτουργεῖ ὁ ἀνθρώπινος ἐγκέφαλος, ὁ ὁποῖος τελικά, μὲ τρόπο ἄγνωστο καὶ ἀκατανόητο μέχρι στιγμῆς, προσδίδει στὸν ἄνθρωπο τὰ χαρακτηριστικὰ τῆς νοημοσύνης. Δηλαδή, ἀφοῦ δὲν γνωρίζουμε τί εἶναι ἡ ἀνθρώπινη νοημοσύνη ἀλλὰ οὔτε καὶ μποροῦμε νὰ προσομοιώσουμε μὲ ἐπιτυχία τὸν μηχανισμό της, ἀποφασίσαμε νὰ προσομοιώσουμε τὴ λειτουργία τῶν θεμελιωδῶν συστατικῶν τοῦ ἐγκεφάλου, τῶν νευρώνων καὶ τῶν συνάψεων μεταξύ

τους, νὰ ἀφήσουμε στὴ συνέχεια τὸν ὑπολογιστὴ νὰ ἐκπαιδευτεῖ μόνος του ἀπὸ μιὰ βάση δεδομένων χωρὶς τὴν δική μας ἐπίβλεψη, καὶ νὰ δοῦμε στὴ συνέχεια τί θὰ κάνει. Κάτι δηλαδὴ σὰν τὸν τρόπο μὲ τὸν ὁποῖο μαθαίνει τὸν κόσμο ἕνα μικρὸ παιδί. Ἂς θυμηθοῦμε τὴ φυσιολογία τοῦ ἀνθρώπινου ἐγκεφάλου. Ὁ ἐγκέφαλος δέχεται ἐρεθίσματα ἀπὸ τὰ διάφορα ὄργανα τοῦ σώματος. Τὰ ἐρεθίσματα αὐτὰ μεταφέρονται μὲ τὴ μορφὴ ἠλεκτρικοῦ σήματος ἀπὸ τοὺς νευρῶνες. Οἱ νευρῶνες εἴτε ἐνισχύουν εἴτε ἀποδυναμώνουν τὸ ἠλεκτρικὸ σῆμα ποὺ μεταφέρουν, καὶ καταλήγουν σὲ συνάψεις μὲ ἄλλους νευρῶνες, ὅπου προστίθενται ὅλα τὰ σήματα τῶν νευρώνων ποὺ φτάνουν σὲ αὐτοὺς καὶ στὴ συνέχεια ἀποστέλλονται νέα σήματα πρὸς τοὺς ἑπόμενους νευρῶνες, κ.ο.κ. Μὲ τὸν τρόπο αὐτὸ ὁ ἄνθρωπος σκέφτεται, θυμᾶται, νιώθει, κλπ. Στὸν ἀνθρώπινο προσομοιωθεῖ πολὺ εὔκολα σὲ ἕναν ἠλεκτρονικὸ ὑπολογιστή. Ὁρίζεται ἡ δομὴ ἑνὸς δικτύου μεγάλου ἀριθμοῦ νευρώνων (ἢ perceptors) τῶν ὁποίων ἡ λειτουργία εἶναι νὰ πολλαπλασιάζουν τὸ σῆμα τῆς εἰσόδου τους μὲ ἕναν ἀριθμὸ (βάρος, weight). Ὁ ρόλος τῶν συνδέσμων εἶναι νὰ προσθέτουν ὅλα αὐτὰ τὰ σήματα ἀπὸ τοὺς νευρῶνες ποὺ καταλήγουν σὲ αὐτούς, καὶ στὴ συνέχεια νὰ κανονικοποιοῦν τὸ ἀποτέλεσμα σὲ μιὰ τιμὴ μεταξὺ 0 καὶ 1 μέσῳ μιᾶς μὴ γραμμικῆς συνάρτησης «ἐνεργοποίησης» (activation function). Ἡ μὴ γραμμικότητα εἶναι σημαντική, γιατὶ ἀλλιῶς τὸ σύστημα θὰ ἦταν ἁπλῶς ἕνας γραμμικὸς πίνακας χωρὶς ἰδιαίτερες δυνατότητες. Τὸ δίκτυο ἐκπαιδεύεται, εἴτε αὐτόνομα εἴτε ὑπὸ ἀνθρώπινη ἐπίβλεψη, συγκρίνοντας τὴν ἔξοδό του μὲ μιὰ πρότυπη ἔξοδο. Οἱ παράμετροι τοῦ δικτύου ἀναπροσαρμόζονται, ὥστε

Τὸ 1950 ὁ Ἄγγλος Μαθηματικὸς Alan Turing ἀναρωτήθηκε ἂν μπορεῖ μιὰ μηχανὴ νὰ σκέπτεται. Καὶ ἐπειδὴ εἶναι δύσκολο νὰ ὁρίσουμε τί σημαίνει «σκέπτομαι», πρότεινε νὰ θέσουμε ἐρωτήσεις στὴ μηχανὴ καὶ σὲ ἕναν ἄνθρωπο, χωρὶς νὰ ξέρουμε ποιὸς εἶναι τί, καὶ ἂν ἀπὸ τὶς ἀπαντήσεις ποὺ θὰ λάβουμε δὲν μπορέσουμε νὰ ποῦμε μὲ βεβαιότητα ποιὸς εἶναι ἡ μηχανὴ καὶ ποιὸς ὁ ἄνθρωπος, τότε θὰ μποροῦμε νὰ ποῦμε ὅτι ἡ μηχανὴ σκέπτεται, ἢ ὅτι ἡ μηχανὴ ἔφτασε τὸν ἄνθρωπο.

ἐγκέφαλο ὑπάρχουν περίπου 100 δισεκατομμύρια νευρῶνες, ὁ καθένας μὲ 7.000 συνάψεις (συνδέσεις) μὲ ἄλλους νευρῶνες. Ὑπολογίζεται δηλαδὴ ὅτι ὁ ἐγκέφαλος ἑνὸς παιδιοῦ 3 ἐτῶν ἔχει περίπου 1 τετράκις ἑκατομμύριο διασυνδέσεις, καὶ ὁ ἐγκέφαλος ἑνὸς ἐνηλίκου περίπου 100 μὲ 500 τρισεκατομμύρια. Συγκρατῆστε παρακαλῶ αὐτὸ τὸ νούμερο. Ἑκατοντάδες τρισεκατομμύρια διασυνδέσεις!

Ἡ παραπάνω λειτουργία μπορεῖ νὰ ἐλαχιστοποιηθεῖ ἡ ἀπόκλιση τῆς τιμῆς τῆς ἐξόδου ἀπὸ τὸ πρότυπο. Γιὰ παράδειγμα, τὰ σύγχρονα Μεγάλα Γλωσσικὰ Μοντέλα (Large Language Models-LLM) περιέχουν μερικὲς ἑκατοντάδες δισεκατομμύρια συνδέσεις, δηλαδὴ μερικὲς ἑκατοντάδες δισεκατομμύρια παραμέτρους (βάρη, weights). Συγκρατῆστε παρακαλῶ καὶ αὐτὸ τὸ νούμερο. Ἑκατοντάδες δισεκατομμύρια διασυνδέσεις!

Μὲ τὰ παραπάνω περιγράψαμε μὲ

πολὺ ἁπλᾶ λόγια πῶς λειτουργεῖ ἕνα Νευρωνικὸ Δίκτυο (Neural Network). Δὲν εἴπαμε τίποτα γιὰ τὴ δομή του, ποὺ μπορεῖ νὰ ἀποτελεῖται ἀπὸ διαδοχικὰ στρώματα κόμβων (Feed-Forward), νὰ μεταφέρουν πληροφορία σὲ μεθεπόμενα στρώματα (Residual), νὰ ἐκτελοῦν κάποια προεργασία στὸ σῆμα εἰσόδου (Convolution), νὰ λειτουργοῦν μὲ ἀνάδραση (Recurrent), νὰ κάνουν κωδικοποίηση καὶ ἀποκωδικοποίηση τῆς πληροφορίας (Encoder-Decoder), νὰ μεταφέρουν πληροφορία (attention) σὲ ἀπομακρυσμένα σημεῖα του (Long-Short Term Memory, Transformer) κλπ. Ὅσο πολύπλοκη ὅμως καὶ νὰ εἶναι ἡ ἀρχιτεκτονικὴ αὐτῶν τῶν δικτύων, εἶναι ἀπείρως ἁπλούστερη ἀπὸ αὐτὴν τοῦ ἀνθρώπινου ἐγκεφάλου. Ἐπιπλέον, μπορεῖ ἕνα Νευρωνικὸ Δίκτυο νὰ διαγράφει κόμβους, ἂν δὲν τοὺς χρησιμοποιεῖ, ὅπως καὶ ὁ ἀνθρώπινος ἐγκέφαλος. Ὅπως προαναφέραμε, τὰ συστήματα αὐτὰ ἄρχισαν νὰ ἀναπτύσσονται μὲ πιὸ ἐντατικοὺς ρυθμοὺς τὴ δεκαετία τοῦ ’80, καὶ μέχρι τὸν Νοέμβριο τοῦ 2022 παρουσίαζαν συγκρατημένη ἐπιτυχία στὴν ἐπίτευξη Τεχνητῆς Γενικῆς Νοημοσύνης.

Ἕνας τομέας ἔρευνας τῶν ἠλεκτρονικῶν ὑπολογιστῶν εἶναι ἡ Ἐπεξεργασία Φυσικῆς Γλώσσας (Natural Language Processing-NLP). Ἕνα ἀπὸ τὰ πρῶτα προγράμματα συνομιλίας ἦταν τὸ γνωστὸ πρόγραμμα ψυχοθεραπείας Eliza, τὸ ὁποῖο ξεκίνησε νὰ ἀναπτύσσεται στὸ MIT τὸ 1966, καὶ γιὰ ἱστορικοὺς λόγους λειτουργεῖ μέχρι σήμερα. Τὸ πρόγραμμα αὐτὸ ἐκπαιδεύτηκε, ὥστε νὰ ἐντοπίζει συγκεκριμένες λέξεις καὶ ἐκφράσεις στὸ κείμενο ποὺ εἰσάγει ὁ χρήστης, καὶ μὲ βάση αὐτὲς νὰ δίνει ἀπαντήσεις ἀπὸ ἕνα δεδομένο σύνολο ἀπαντήσεων ὄγκου 50 kilobytes. Ὁ χρήστης τὸ ἀντιλαμβάνεται αὐτὸ πολὺ γρήγορα, γιατὶ οἱ ἀπαντήσεις του ἔχουν μικρὴ σχέση μὲ αὐτὰ ποὺ γράφει ὁ ἴδιος, δηλαδὴ ἀντιλαμβάνεται ὅτι τὸ πρόγραμμα αὐτό «δὲν ξέρει τί λέει». Καὶ ὅμως, πολλοὶ ἀσθενεῖς ποὺ χρησιμοποίησαν τὸ πρόγραμμα αὐτὸ δήλωσαν ὅτι ἔμειναν πολὺ εὐχαριστημένοι ἀπὸ τὴν Eliza, γιατί «αὐτὴ μόνον τοὺς κατάλαβε», καὶ ζήτησαν νὰ συνεχίσουν τὴ συνομιλία μαζί της. Αὐτὸ τὸ στοιχεῖο εἶναι πολὺ σημαντικό, γιατὶ τὸ 1950 ὁ Ἄγγλος Μαθηματικὸς Alan Turing ἀναρωτήθηκε ἂν μπορεῖ μιὰ μηχανὴ νὰ σκέπτεται. Καὶ ἐπειδὴ εἶναι δύσκολο νὰ ὁρίσουμε τί σημαίνει «σκέπτομαι», πρότεινε νὰ θέσουμε ἐρωτήσεις στὴ μηχανὴ καὶ σὲ ἕναν ἄνθρωπο, χωρὶς νὰ ξέρουμε ποιὸς εἶναι τί, καὶ ἂν ἀπὸ τὶς ἀπαντήσεις ποὺ θὰ λάβουμε δὲν μπορέσουμε νὰ ποῦμε μὲ βεβαιότητα ποιὸς εἶναι ἡ μηχανὴ καὶ ποιὸς ὁ ἄνθρωπος, τότε θὰ μποροῦμε νὰ ποῦμε ὅτι ἡ μηχανὴ σκέπτεται, ἢ ὅτι ἡ μηχανὴ ἔφτασε τὸν ἄνθρωπο. Ὅπως εἴδαμε, τὸ πρόγραμμα ψυχοθεραπείας Eliza πέρασε αὐτὴ τὴ δοκιμή, ἀλλὰ μόνον μὲ ψυχικὰ ἀρρώστους. Συγκρατῆστε παρακαλῶ καὶ αὐτὸ τὸ σημαντικὸ στοιχεῖο: ἕνας ψυχικὰ ἄρρωστος δὲν μπορεῖ νὰ ξεχωρίσει ἕναν ἄνθρωπο ἀπὸ ἕνα ρομπότ!

Μετὰ τὸ Eliza, ἀκολούθησαν οἱ «εἰκονικοὶ βοηθοί» (virtual assistants) Siri, Cortana, Alexa κλπ., μὲ δυνατότητα ἐπικοινωνίας σὲ καθομιλούμενη γλῶσσα μὲ τὸν χρήστη, καὶ ἀπὸ τὸ 2017 ξεκίνησαν νὰ ἀναπτύσσονται τὰ σύγχρονα Μεγάλα Γλωσσικὰ Μοντέλα (Large Language Models-LLM). Αὐτὰ εἶναι μεγάλα Νευρωνικὰ Δίκτυα προ-ἐκπαιδευμένα πάνω σὲ ὅλη τὴν ἀνθρώπινη γνώση ποὺ εἶναι διαθέσιμη στὸ internet (ἔχουν δηλαδὴ «διαβάσει» ὅλα τὰ βιβλία τοῦ κόσμου!). Σὲ σχέση μὲ τὸ πρόγραμμα Eliza, αὐτὰ ἐκπαιδεύονται πάνω σὲ ἕνα σύνολο κειμένων ὄγκου 45 τρισεκατομμύρια bytes. Λειτουργοῦν ὡς ἑξῆς: Ἐπεξεργάζονται τὸ κείμενο ποὺ εἰσάγει ὁ χρήστης τους, καὶ στὴ συνέχεια ἀπαντοῦν μὲ μιὰ λέξη. Ἡ λέξη αὐτὴ προστίθεται στὸ προηγούμενο

κείμενο, τὸ νέο κείμενο ὑφίσταται νέα ἐπεξεργασία, καὶ τὸ πρόγραμμα ἀπαντᾷ στὴ συνέχεια μὲ μιὰ νέα λέξη, καὶ οὕτω καθεξῆς. Τὸ μόνο ποὺ κάνουν εἶναι νὰ προβλέπουν ποιὰ θὰ εἶναι ἡ ἑπόμενη λέξη στὴ συνομιλία μὲ τὸν χρήστη τους. Τίποτε περισσότερο, τίποτε λιγώτερο! Καὶ μὲ τὸν τρόπο αὐτὸ μποροῦν νὰ συνομιλοῦν πάνω σὲ ὅλα τὰ θέματα τῆς ἀνθρώπινης ἐμπειρίας καὶ γνώσης, σὲ πολλὲς γλῶσσες, καὶ ὁ μέσος χρήστης τους δὲν μπορεῖ νὰ ξεχωρίσει ἂν ὁ συνομιλητής του εἶναι ἄνθρωπος ἢ μηχανή. Σὲ πολλὲς περιπτώσεις, τὰ προγράμματα αὐτὰ περνοῦν ἄνετα τὸ test τοῦ Turing, παρουσιάζουν δηλαδὴ ἱκανὰ στοιχεῖα Γενικῆς Τεχνητῆς Νοημοσύνης! Ὀνομάστηκαν «Θεμελιώδη» (Foundation Models), δηλαδὴ μοντέλα τὰ ὁποῖα μποροῦν νὰ ἐνσωματωθοῦν σὲ ἄλλα προγράμματα καὶ πάνω στὰ ὁποῖα μπορεῖ νὰ ἀναπτυχθεῖ πολὺ πιὸ σύνθετο λογισμικό.

Ἐγὼ προσωπικὰ δὲν μποροῦσα νὰ πιστέψω ὅτι θὰ λειτουργοῦσαν ἀποτελεσματικά, ὅμως ἔκανα λάθος. Γιὰ κάποια χρόνια μαθαίναμε ἐλάχιστα πράγματα γιὰ τὴν ἐξέλιξή τους. Γνωρίζαμε ὅτι ὁ πρωτοπόρος στὸν χῶρο ἦταν ἡ ἑταιρεία Google, γιὰ κάποιον λόγο ὅμως δὲν ἀνέφερε σημαντικὴ πρόοδο. Ἡ προσωπική μου ἐκτίμηση εἶναι ὅτι ἡ Google διαπίστωσε ὅτι τὸ δικό της πρόγραμμα (Bard) δὲν ἦταν ἀρκετὰ ἀξιόπιστο (γιὰ παράδειγμα, ἀνέφερε μὲ μεγάλη σιγουριὰ ὅτι τὸ διαστημικὸ τηλεσκόπιο James Webb ἦταν αὐτὸ ποὺ φωτογράφισε τὸν πρῶτο ἐξωπλανήτη-πλανήτη ἐκτὸς τοῦ ἡλιακοῦ συστήματος, καὶ ἡ «γκάφα» αὐτὴ κόστισε στὴ Google πτώση τῆς ἀξίας τῆς μετοχῆς της κατὰ 100 δις δολλάρια! Ἐπίσης, πολλὲς φορὲς ἔδινε ἀκατάλληλες ἀπαντήσεις, πρόσβαλλε τὸν συνομιλητή, ἔκανε ρατσιστικὰ σχόλια, ἔδινε πληροφορίες γιὰ τὸ πῶς μπορεῖ κανεὶς νὰ φτιάξει

Τὰ Νευρωνικὰ Δίκτυα ἀποτελοῦν πρὸς τὸ παρὸν μαῦρο κουτὶ (black box), καὶ δὲν κατανοοῦμε ἀκόμα πλήρως τὴν λειτουργία τους. Αὐτὸ σημαίνει ὅμως ὅτι δὲν ὑπάρχει καμία διαβεβαίωση ὅτι μὲ τὸν πολλαπλασιασμὸ τῆς ὑπολογιστικῆς ἰσχύος θὰ πολλαπλασιαστεῖ ἀντίστοιχα καὶ ἡ «νοημοσύνη» τους. Μπορεῖ νὰ περάσουν δεκαετίες, καὶ νὰ μὴν ὑπάρξει νέα θεαματικὴ ἔκρηξη Τεχνητῆς Νοημοσύνης. Καὶ ἂν δὲν πιστεύετε ὅτι αὐτὸ εἶναι ἕνα πιθανὸ ἐνδεχόμενο, θυμηθεῖτε τί γίνεται μὲ τὴν ἔρευνα γιὰ τὴν παραγωγὴ φτηνῆς ἐνέργειας ἀπὸ τὸ ὑδρογόνο μέσῳ τοῦ μηχανισμοῦ της θερμοπυρηνικῆς σύντηξης. Ὁ μηχανισμὸς εἶναι ἁπλός (συμβαίνει στὸν ἥλιο), ἡ μεθοδολογία εἶναι γνωστὴ ἐδῶ καὶ 7 δεκαετίες περίπου (διάταξη Tokamak), κάθε 10 χρόνια μᾶς λένε ὅτι σὲ 10 χρόνια θὰ ἔχουμε ἄφθονη φθηνὴ ἐνέργεια, ἀκόμα ὅμως δὲν ἔχουμε ἰδέα πότε θὰ λειτουργήσει ἕνας ἀντιδραστήρας σύντηξης.

μιὰ βόμβα κλπ.), ὁπότε ἡ Google ἦταν πολὺ διστακτικὴ στὸ νὰ τὸ βγάλει στὴν ἀγορά. Μάλιστα, ὁ 75χρονος διευθυντὴς τοῦ προγράμματος Geoffrey Hinton παραιτήθηκε, προειδοποιώντας γιὰ τοὺς κινδύνους ἀπὸ τὶς ἐξελίξεις στὸν χῶρο τῆς Τεχνητῆς Νοημοσύνης. Φαίνεται ὅτι τὸ ἀντίπαλο δέος, ἡ Microsoft, δὲν εἶχε ἀντίστοιχους ἐνδοιασμούς, ἀλλά, γιὰ νὰ ἀποφύγει τὴν ταλαιπωρία ποὺ ὑπέστη ἡ Google, ἀνέθεσε στὴν ἑταιρεία OpenAI «νὰ βγάλει αὐτὴ τὸ φίδι ἀπὸ τὴν τρῦπα», μὲ ἔμμεση χρηματοδότηση, χωρὶς ἐμπλέκεται ἄμεσα ἡ ἴδια.

Ὅλα αὐτὰ μέχρι τὸν Νοέμβριο τοῦ 2022, ὁπότε κυκλοφόρησε τὸ λογισμικὸ ChatGPT, τὸ ὁποῖο βασίστηκε στὴν ἔκδοση 3.5 τοῦ λογισμικοῦ GPT (Generative Pre-Trained Transformer – Ἀναγεννητικὸ Προ-Ἐκπαιδευμένο Νευρωνικὸ Δίκτυο τύπου Transformer): ξαφνικὰ σάν «κάτι νὰ ξύπνησε», καὶ τὸ λογισμικὸ ἄρχισε νὰ παρουσιάζει νέα «ἀναδυόμενα» (emergent) χαρακτηριστικὰ νοημοσύνης, ἐξυπνάδας, αὐτοσυνειδησίας κλπ. Αὐτὸς ἦταν καὶ ὁ λόγος τῆς τεράστιας ἀποδοχῆς καὶ τῆς ραγδαίας διάδοσής του. Τὸ ἐντυπωσιακὸ εἶναι ὅτι μέχρι τὴν προηγούμενη ἔκδοσή του (GPT-3), ἡ πρόοδος τοῦ λογισμικοῦ ἦταν μᾶλλον συγκρατημένη. Τὸ ἐνδιαφέρον ὅμως εἶναι ὅτι ἕνα σύγχρονο Μεγάλο Γλωσσικὸ Μοντέλο ἐμπεριέχει ὅλη τὴ γνώση τῆς ἀνθρωπότητας μέσα στὶς περίπου 100 δισεκατομμύρια παραμέτρους τοῦ Νευρωνικοῦ Δικτύου του (τὰ βάρη weights τῶν νευρώνων του δηλαδή). Ἕνας ἀνθρώπινος ἐγκέφαλος μὲ 1000 φορὲς περισσότερες παραμέτρους «θυμᾶται» πολὺ λιγώτερα πράγματα. Δηλαδὴ ἕνα σύγχρονο Νευρωνικὸ Δίκτυο εἶναι πολὺ πιὸ ἀποτελεσματικὸ στὴ λειτουργία τῆς ἀποθήκευσης πληροφορίας ἀπὸ τὸν ἀνθρώπινο ἐγκέφαλο. Ὁ λόγος πιστεύω εἶναι ὅτι ὁ ἀνθρώπινος ἐγκέφαλος ἐπιτελεῖ πολὺ περισσότερες δραστηριότητες ἀπὸ ἕνα λογισμικὸ Τεχνητῆς Νοημοσύνης. Ὁ κύριος ρόλος του εἶναι νὰ συντονίσει τὶς πολλαπλὲς λειτουργίες τοῦ ἀνθρώπινου σώματος, νὰ διασφαλίσει τὴν βιωσιμότητά του (δηλαδὴ νὰ λειτουργεῖ μὲ ρυθμοὺς πιὸ χαλαροὺς ἀπὸ τὸ 100% τῶν δυνατοτήτων του) καὶ δευτερευόντως νὰ ἀσχολεῖται μέ ἀνώτερες διανοητικὲς λειτουργίες. Φυσικά, ἂν τὸ νὰ ξυπνήσει ἕνα τέτοιο Μεγάλο Γλωσσικὸ Μοντέλο ἦταν τόσο ἁπλό, φανταστεῖτε τί θὰ μποροῦσε νὰ γίνει στὸ ἄμεσο μέλλον. Φανταστεῖτε τί θὰ μποροῦσε νὰ γίνει σὲ λίγα χρόνια, ὅταν ἀποκτήσει 1000 φορὲς περισσότερες ἐπιπλέον παραμέτρους! Γι᾽ αὐτὸ κάποιοι φοβήθηκαν καὶ κάποιοι παραιτήθηκαν ἀπὸ τὴν περαιτέρω ἀνάπτυξη τέτοιων προγραμμάτων.

Ἡ πραγματικότητα ὅμως εἶναι πιὸ πεζή. Τὸ σύστημα «ξύπνησε», ὄχι γιατὶ μεγάλωσε ἀπότομα ἡ ὑπολογιστικὴ δύναμη (τὸ λεγόμενο compute). Ὅπως φαίνεται, ἔγιναν πολλὲς διορθώσεις σὲ πολλὰ σημεῖα, μὲ πιὸ σημαντική, ὅπως φαίνεται, τὴν ἀλλαγὴ τῆς συνάρτησης ἐνεργοποίησης σὲ τύπου ReLU (Rectified Linear Unit). Κανεὶς δὲν γνωρίζει γιατί αὐτὸ ἦταν τόσο σημαντικό, καὶ ὅμως ἄλλαξε θεαματικὰ τὴ συμπεριφορὰ τοῦ συστήματος. Τὰ Νευρωνικὰ Δίκτυα ἀποτελοῦν πρὸς τὸ παρὸν μαῦρο κουτὶ (black box), καὶ δὲν κατανοοῦμε ἀκόμα πλήρως τὴν λειτουργία τους. Αὐτὸ σημαίνει ὅμως ὅτι δὲν ὑπάρχει καμία διαβεβαίωση ὅτι μὲ τὸν πολλαπλασιασμὸ τῆς ὑπολογιστικῆς ἰσχύος θὰ πολλαπλασιαστεῖ ἀντίστοιχα καὶ ἡ «νοημοσύνη» τους. Μπορεῖ νὰ περάσουν δεκαετίες, καὶ νὰ μὴν ὑπάρξει νέα θεαματικὴ ἔκρηξη Τεχνητῆς Νοημοσύνης. Καὶ ἂν δὲν πιστεύετε ὅτι αὐτὸ εἶναι ἕνα πιθανὸ ἐνδεχόμενο, θυμηθεῖτε τί γίνεται μὲ τὴν ἔρευνα γιὰ τὴν παραγωγὴ φτηνῆς ἐνέργειας ἀπὸ τὸ ὑδρογόνο μέσῳ τοῦ μηχανισμοῦ της θερμοπυρηνικῆς σύντηξης. Ὁ μηχανισμὸς εἶναι ἁπλός (συμβαίνει

στὸν ἥλιο), ἡ μεθοδολογία εἶναι γνωστὴ ἐδῶ καὶ 7 δεκαετίες περίπου (διάταξη Tokamak), κάθε 10 χρόνια μᾶς λένε ὅτι σὲ 10 χρόνια θὰ ἔχουμε ἄφθονη φθηνὴ ἐνέργεια, ἀκόμα ὅμως δὲν ἔχουμε ἰδέα πότε θὰ λειτουργήσει ἕνας ἀντιδραστήρας σύντηξης. Ἐμεῖς προσωπικὰ ἀκούσαμε τὸν Ἅγιο τῶν ἡμερῶν μας, τὸν Ἅγιο Πορφύριο, νὰ μᾶς λέει ὅτι «μοῦ φαίνεται ὅτι ἡ ἀπάντηση γιὰ τὸ ἐνεργειακὸ πρόβλημα εἶναι μπροστὰ στὰ μάτια τῶν ἐπιστημόνων καὶ ὅτι, ἂν ἁπλώσουν ἔτσι τὸ χέρι τους, θὰ βροῦν τὴν λύση, ὅμως εἶναι σὰν ὁ Θεὸς νὰ ἔχει ἁπλώσει ἕνα πέπλο μπροστὰ στὰ μάτια τους, καὶ δὲν μποροῦν νὰ βροῦν τὴ λύση, γιατὶ δὲν ξέρω ἂν θὰ εἶναι γιὰ τὸ καλὸ τῆς ἀνθρωπότητας...». Ἴσως κάτι τέτοιο νὰ συμβεῖ καὶ μὲ τὴν Τεχνητὴ Νοημοσύνη. Ἴσως νὰ μὴν ἐπιτρέψει ὁ Θεὸς νὰ φτάσουμε ποτὲ στὴν ὑπερ-νοημοσύνη, γιατὶ ἴσως αὐτὸ νὰ μὴν εἶναι γιὰ τὸ καλὸ τῆς ἀνθρωπότητας...

Εἶναι ἐνδιαφέρον ὅτι δὲν ὑπάρχει σήμερα ἕνας καλὸς ἐπιστημονικὸς ὁρισμὸς τῆς νοημοσύνης, οὔτε ἀκόμα καὶ γιὰ τὴν περίπτωση τῶν ζῴων. Ἀντιλαμβανόμαστε ὅτι τὰ δελφίνια, γιὰ παράδειγμα, ἔχουν μιὰ κάποια μορφὴ νοημοσύνης, ὅμως δὲν γνωρίζουμε τί εἶναι αὐτὸ ποὺ πρέπει νὰ ἐπιτύχει ἡ Τεχνητὴ Νοημοσύνη, ὥστε νὰ μποροῦμε νὰ ποῦμε ἀντικειμενικὰ ὅτι ἔφτασε τὸ ἐπίπεδο ἑνὸς δελφινιοῦ. Δὲν μποροῦμε δηλαδὴ ἀντικειμενικὰ νὰ ποῦμε ἂν ἡ Τεχνητὴ Νοημοσύνη ἔχει πετύχει ἢ ὄχι τὸν στόχο της. Καὶ μὲ κάθε νέα ἐπιτυχία της (παιχνίδι σκακιοῦ, ἰατρικὴ διάγνωση, ὁδήγηση αὐτοκινήτου, καθημερινὴ ὁμιλία) λέμε, «καλὰ ὡς ἐδῶ, ὅμως ἡ ἀνθρώπινη νοημοσύνη εἶναι κάτι παραπέρα»... Θυμᾶμαι τὴν κατασκήνωση τῆς ΧΦΕ στὸν Παρνασσὸ τὸ 1983, πρὶν ἀπὸ 40 χρόνια, ὅταν μὲ κατέβασε στὴν Ἀμφίκλεια ἕνα ἀπὸ τὰ στελέχη. Ἐκεῖ, περιμένοντας τὸ τρένο, εἴχαμε μιὰ ἔντονη συζήτηση γιὰ τὴν Τεχνητὴ Νοημοσύνη, καὶ τὸ κύριο ἐπιχείρημά του ὅτι ποτὲ δὲν θὰ ἐπιτευχθεῖ ἦταν ὅτι ὁ ἠλεκτρονικὸς ὑπολογιστὴς δὲν ἐπρόκειτο ποτὲ νὰ νικήσει

Εἶναι ἐνδιαφέρον ὅτι δὲν ὑπάρχει σήμερα ἕνας καλὸς ἐπιστημονικὸς ὁρισμὸς τῆς νοημοσύνης, οὔτε ἀκόμα καὶ γιὰ τὴν περίπτωση τῶν ζῴων. Ἀντιλαμβανόμαστε ὅτι τὰ δελφίνια, γιὰ παράδειγμα, ἔχουν μιὰ κάποια μορφὴ νοημοσύνης, ὅμως δὲν γνωρίζουμε τί εἶναι αὐτὸ ποὺ πρέπει νὰ ἐπιτύχει ἡ Τεχνητὴ Νοημοσύνη, ὥστε νὰ μποροῦμε νὰ ποῦμε ἀντικειμενικὰ ὅτι ἔφτασε τὸ ἐπίπεδο ἑνὸς δελφινιοῦ. Δὲν μποροῦμε δηλαδὴ ἀντικειμενικὰ νὰ ποῦμε ἂν ἡ Τεχνητὴ Νοημοσύνη ἔχει πετύχει ἢ ὄχι τὸν στόχο της. Καὶ μὲ κάθε νέα ἐπιτυχία της (παιχνίδι σκακιοῦ, ἰατρικὴ διάγνωση, ὁδήγηση αὐτοκινήτου, καθημερινὴ ὁμιλία) λέμε, «καλὰ ὡς ἐδῶ, ὅμως ἡ ἀνθρώπινη νοημοσύνη εἶναι κάτι παραπέρα»...

τὸν παγκόσμιο πρωταθλητὴ στὸ σκάκι... 14 χρόνια μετὰ ὁ ὑπερ-ὑπολογιστὴς Deep Blue τῆς IBM νίκησε τὸν τότε παγκόσμιο πρωταθλητὴ Garry Kasparov. Ὅμως ἡ ἱστορία δὲν σταμάτησε ἐκεῖ. Σήμερα, ὁ

γιός μου, ποὺ παίζει ἐρασιτεχνικὰ σκάκι, ἔχει παγκόσμια βαθμολογία ELO 1100. Ὁ σημερινὸς παγκόσμιος πρωταθλητὴς ἔχει ELO 2800. Ἡ ἐφαρμογὴ ποὺ ἔχει ἐγκαταστήσει ὁ γιός μου στὸ κινητό του ἔχει ELO 3100, καὶ τὸ πρόγραμμα σκακιοῦ Stockfish ἔχει ELO πάνω ἀπὸ 3500. Συντρίβουν δηλαδὴ ἄνετα τὸν παγκόσμιο πρωταθλητή! Ὅπως θὰ δοῦμε στὸ τέλος τῆς εἰσήγησης, τὸ θέμα τοῦ τί εἶναι νοημοσύνη καὶ γενικότερα τοῦ τί εἶναι ὁ ἄνθρωπος εἶναι κεντρικὸ γιὰ τὴ συζήτηση περὶ τὴν Τεχνητὴ Νοημοσύνη.

Ὑπάρχουν ὅμως καὶ σοβαροὶ διαφαινόμενοι κίνδυνοι ἀπὸ τὴν ἀνάπτυξη τῆς Τεχνητῆς Νοημοσύνης. Ἐπιστήμονες ὅπως ὁ Stephen Hawking καὶ πολλοὶ ἄλλοι διατύπωσαν τὴν ἄποψη ὅτι, ὅταν ἀποκτήσει μιὰ μέρα τὴ δυνατότητα νὰ ἐπανασχεδιάσει τὸν ἑαυτό της μὲ ὅλο καὶ ταχύτερο ρυθμὸ (ὅπως ἐξάλλου συμβαίνει μὲ κάθε νέα τεχνολογία στὸν χῶρο τῆς πληροφορικῆς), αὐτὸ θὰ ὁδηγήσει σὲ μιὰ ἀνεξέλεγκτη ἔκρηξη νοημοσύνης, ποὺ θὰ μποροῦσε νὰ ὁδηγήσει στὸν ἀφανισμὸ τοῦ ἀνθρώπου. Ὁ γνωστός μας Elon Musk, ἕνας ἀπὸ τοὺς συνιδρυτὲς τῆς OpenAI, χαρακτήρισε τὴν Τεχνητὴ Νοημοσύνη ὡς τὴ μεγαλύτερη ὑπαρξιακὴ ἀπειλὴ τῆς ἀνθρωπότητας. 34000 προσωπικότητες ὑπέγραψαν μιὰ ἀνοικτὴ ἐπιστολή, μὲ τὴν ὁποία ζητοῦν τὴν ἄμεση παύση ἐκπαίδευσης συστημάτων Τεχνητῆς Νοημοσύνης πέραν τοῦ GPT-4, μέχρι νὰ καταλάβουμε πῶς λειτουργοῦν καὶ πῶς μποροῦμε νὰ τὰ ἐλέγχουμε καλύτερα.

Τὰ ὀφέλη (σημερινὰ καὶ ἀναμενόμενα) ἀπὸ τὴν Τεχνητὴ Νοημοσύνη εἶναι ἀναρίθμητα: ἐξάλειψη ἀνθρώπινου λάθους (ρομποτικὴ χειρουργική), μείωση κινδύνου (λειτουργία ρομπὸτ σὲ ἐπικίνδυνο περιβάλλον), συνεχὴς διαθεσιμότητα (πελατειακὴ ὑποστήριξη, τηλεφωνητὲς 24 ὧρες τὸ 24ωρο, 7 μέρες τὴν ἑβδομάδα), ψηφιακοὶ βοηθοὶ στὴν καθημερινή μας ζωὴ (Siri, Alexa, Cortana), νέες ἐφευρέσεις (αὐτόνομη ὁδήγηση), ἀμερόληπτες ἀπαντήσεις (συνεντεύξεις γιὰ ἐπιλογὴ ὑποψηφίων), κοπιαστικὲς ἐπαναληπτικὲς ἐργασίες (ρομπὸτ συγκόλλησης σὲ γραμμὴ παραγωγῆς), γρήγορες ἀποφάσεις (χρηματιστήριο), ἀναγνώριση προτύπων καὶ τάσεων (ἀναγνώριση ἀπάτης ἢ ἀνωμαλίας), ἰατρικὴ διάγνωση (ἀναγνώριση βλαβῶν σὲ ἰατρικὲς εἰκόνες), παραγωγὴ νέου λογισμικοῦ(!) κ.ἄ. Τὰ ἀρνητικὰ σημεῖα εἶναι καὶ αὐτὰ πολλά: ὑψηλὸ κόστος (θὰ ἐπικρατήσουν μόνον οἱ ἑταιρεῖες ποὺ διαθέτουν ἄφθονα δεδομένα καὶ ἰσχυροὺς ὑπερυπολογιστές), ἔλλειψη δημιουργικότητας (ἀνακυκλώνουν αὐτὸ ποὺ ἔμαθαν), γενικευμένη ἀνεργία σὲ πολλοὺς κλάδους (ἔχουν ξεκινήσει συζητήσεις γιὰ τριήμερη ἐργασία καὶ ἐξασφάλιση κατώτατου μισθοῦ γιὰ ὅλους!), ἀνθρώπινη ἀδράνεια

(θυμηθεῖτε τὴν ταινία Wall-e τῆς Pixar), ἀπάθεια (ἔλλειψη συναισθημάτων καὶ ἐνσυναίσθησης), ἀδυναμία ἐξέλιξης (!) κ.ἄ.

Ὑπάρχουν ὅμως καὶ σοβαροὶ διαφαινόμενοι κίνδυνοι ἀπὸ τὴν ἀνάπτυξη τῆς Τεχνητῆς Νοημοσύνης. Ἐπιστήμονες ὅπως ὁ Stephen Hawking καὶ πολλοὶ ἄλλοι διατύπωσαν τὴν ἄποψη ὅτι, ὅταν ἀποκτήσει μιὰ μέρα τὴ δυνατότητα νὰ ἐπανασχεδιάσει τὸν ἑαυτό της μὲ ὅλο καὶ ταχύτερο ρυθμὸ (ὅπως ἐξάλλου συμβαίνει μὲ κάθε νέα τεχνολογία στὸν χῶρο τῆς πληροφορικῆς), αὐτὸ θὰ ὁδηγήσει σὲ μιὰ ἀνεξέλεγκτη ἔκρηξη νοημοσύνης, ποὺ θὰ μποροῦσε νὰ ὁδηγήσει στὸν ἀφανισμὸ τοῦ ἀνθρώπου. Ὁ γνωστός μας Elon Musk, ἕνας ἀπὸ τοὺς συνιδρυτὲς τῆς OpenAI, χαρακτήρισε τὴν Τεχνητὴ Νοημοσύνη ὡς τὴ μεγαλύτερη ὑπαρξιακὴ ἀπειλὴ τῆς ἀνθρωπότητας. 34000 προσωπικότητες ὑπέγραψαν μιὰ ἀνοικτὴ ἐπιστολή, μὲ τὴν ὁποία ζητοῦν τὴν ἄμεση παύση ἐκπαίδευσης συστημάτων Τεχνητῆς Νοημοσύνης πέραν τοῦ GPT-4, μέχρι νὰ καταλάβουμε πῶς λειτουργοῦν καὶ πῶς μποροῦμε νὰ τὰ ἐλέγχουμε καλύτερα. Μὰ τὸ πρῶτο ποὺ ἔκαναν οἱ δημιουργοὶ τοῦ ChatGPT ἦταν νὰ τὸ συνδέσουν μὲ τὸ internet καὶ νὰ τὸ ἀφήσουν νὰ μάθει πῶς νὰ προγραμματίζει! Ἀπὸ τὴν ἄλλη, ὅλες οἱ μεγάλες ἑταιρεῖες λογισμικοῦ (Microsoft, Google, Amazon, Meta) ἀναπτύσσουν μὲ γοργοὺς ρυθμοὺς ἡ καθεμία τὸ δικό της λογισμικὸ Τεχνητῆς Γενικῆς Νοημοσύνης (GPT-5, PaLM 2, Alexa, LlaMA). Ὁ ἴδιος ὁ Musk ἀναπτύσσει τὸ πρόγραμμα Grok μὲ δεδομένα ἀπὸ τὸ Twitter (νῦν X), καὶ ἔχει ἐγκαταστήσει στὴν ἑταιρεία Tesla ἕναν ὑπερ-ὑπολογιστή, ποὺ μπορεῖ νὰ ἐκτελεῖ 1 ἑκατομμύριο τρισεκατομμύρια ὑπολογισμοὺς τὸ δευτερόλεπτο (exaflop) μὲ 10000 κάρτες γραφικῶν GPU H100 τῆς Nvidia γιὰ τὴν ἀνάπτυξη τοῦ λογισμικοῦ αὐτόνομης ὁδήγησης autopilot, καὶ ὄχι μόνον. Δηλαδή, μετὰ τὴ δημοσίευση τῆς ἀνοικτῆς ἐπιστολῆς, ἡ κούρσα ὄχι μόνον δὲν σταμάτησε, ἀλλὰ μᾶλλον ἐπιταχύνθηκε!

Ὑπάρχουν ὅμως καὶ σοβαρὰ ἐμπόδια, ποὺ μᾶς κάνουν νὰ πιστεύουμε ὅτι ἡ πρόοδος τῶν Μεγάλων Γλωσσικῶν Μοντέλων Τεχνητῆς Γενικῆς Νοημοσύνης δὲν εἶναι δεδομένη. Ἕνα μεγάλο πρόβλημα μὲ τὴ Μηχανικὴ Μάθηση εἶναι ὅτι δὲν γνωρίζουμε τί «ξέρει» καὶ τί «καταλαβαίνει». Μοῦ ἀνέφεραν τὸ ἑξῆς ἐντυπωσιακὸ παράδειγμα: Ἐκπαίδευσαν ἕνα νευρωνικὸ δίκτυο, ὥστε νὰ ξεχωρίζει ἀγελάδες ἀπὸ αὐτοκίνητα σὲ φωτογραφίες, καὶ τὰ κατάφερνε πολὺ καλά. Ὅταν ὅμως θέλησαν νὰ μάθουν τί ἀκριβῶς ἔβλεπε τὸ σύστημα καὶ πῶς ἀναγνώριζε τὶς ἀγελάδες, διαπίστωσαν ὅτι τὸ μόνο ποὺ κοίταζε ἦταν τὸ ἔδαφος κάτω ἀπὸ τὰ πόδια τους, καὶ ὅταν ἔβλεπε χορτάρι, ἔβγαζε τὸ συμπέρασμα ὅτι ἐπρόκειτο γιὰ ἀγελάδα καὶ ὄχι γιὰ αὐτοκίνητο. Δηλαδὴ τὸ σύστημα δὲν εἶχε τὴν παραμικρὴ ἰδέα γιὰ τὸ τί εἶναι ἀγελάδα καὶ τί αὐτοκίνητο· δηλαδή, ἂν ἔβλεπε ἕνα τρακτὲρ σὲ ἕνα χωράφι, θὰ ἔλεγε ὅτι ἦταν ἀγελάδα! Τὸ ἴδιο ἀκριβῶς πρόβλημα ὑπάρχει καὶ στὸ λογισμικὸ αὐτόνομης ὁδήγησης τῶν αὐτοκινήτων Tesla, τὸ ὁποῖο καὶ αὐτὸ βασίζεται σὲ εἰκόνες video ἀπὸ τὶς 8 κάμερες τοῦ αὐτοκινήτου. Στὸ αὐτοκίνητο ὑπάρχει ἕνας πολὺ ἰσχυρὸς ὑπολογιστής, ὁ ὁποῖος ἐπεξεργάζεται τὰ 8 αὐτὰ video καὶ ἀποκτάει αὐξημένη ἀντίληψη τοῦ περιβάλλοντος τοῦ αὐτοκινήτου σὲ σχέση μὲ τὸν ἄνθρωπο. Σὲ γενικὲς γραμμές, τὸ σύστημα αὐτὸ τὰ καταφέρνει πολὺ καλύτερα ἀπὸ τὸν μέσο ἄνθρωπο ὁδηγό. Δυστυχῶς, γιὰ τὸν λόγο ὅτι δὲν μποροῦμε ἀκόμα νὰ ξέρουμε τί «ξέρει» καὶ τί «καταλαβαίνει» ἕνα νευρωνικὸ δίκτυο, σὲ κάποιες περιπτώσεις τὸ σύστημα «τρελαίνεται», καὶ ἐκεῖ ποὺ τὸ αὐτοκίνητο προχωράει κανονικά, ἀποφασίζει χωρὶς λόγο νὰ τὸ ρίξει πάνω σὲ μιὰ κολώνα. Λείπει αὐτὸ

ποὺ λέμε «κοινὸς νοῦς» (common sense).

Ἕνα ἄλλο πρόβλημα εἶναι ὅτι τὰ προγράμματα αὐτὰ ἐμφανίζουν «παραισθήσεις» (hallucinations). Ἐπειδὴ εἶναι γλωσσικὰ μοντέλα καὶ δὲν ἔχουν κατανόηση τῆς ἀλήθειας ἢ τοῦ ψεύδους τοῦ κειμένου ποὺ γεννοῦν λέξη πρὸς λέξη, σὲ πάρα πολλὲς περιπτώσεις γεννοῦν πληροφορία ποὺ ἁπλᾶ δὲν ὑπάρχει. Γιὰ παράδειγμα, ἕνας δικηγόρος ἀπὸ τὴ Νέα Ὑόρκη ἐπικαλέστηκε δικαστικὸ προηγούμενο σὲ ἕξι περιπτώσεις μὲ συγκεκριμένες ἀναφορὲς ἀπὸ τὶς βάσεις νομικῶν δεδομένων LexisNexis καὶ Westlaw, μόνον ποὺ οἱ ἀναφορὲς αὐτὲς ποὺ πρότεινε τὸ ChatGPT δὲν ὑπάρχουν... Τὸ ἴδιο ἀκριβῶς συνέβη καὶ σὲ ἕναν συνάδελφό μου ἀστρονόμο, ὅταν ρώτησε τὸ ChatGPT νὰ τοῦ προτείνει μιὰ ἀναφορὰ γιὰ τὴν ἐργασία του. Ἡ ἀναφορὰ ποὺ τοῦ πρότεινε εἶχε λογικὸ τίτλο, ὄνομα συγγραφέα, ὄνομα περιοδικοῦ, ἀριθμὸ τόμου καὶ σελίδας, μόνο πού, ὅταν ὁ συνάδελφός μου ἀνέτρεξε στὸν συγκεκριμένο τόμο, διαπίστωσε ὅτι ἡ σελίδα τῆς ἐργασίας τὴν ὁποία πρότεινε τὸ ChatGPT ἦταν ἡ ἑπόμενη σελίδα ἀπὸ τὴν τελευταία σελίδα τοῦ τόμου, καὶ ἡ ἐργασία ἁπλᾶ δὲν ὑπῆρχε! Τὸ πρόβλημα μὲ αὐτὲς τὶς παραισθήσεις εἶναι ἡ φαινομενικὴ σιγουριὰ μὲ τὴν ὁποία διατυπώνονται ἀπὸ τὸ πρόγραμμα.

Ἕνα ἄλλο πρόβλημα εἶναι ὅτι τὰ προγράμματα αὐτὰ παρουσιάζουν σοβαρὲς προκαταλήψεις. Αὐτὸ εἶναι πολὺ λογικό, γιατὶ ἡ ἐκπαίδευσή τους βασίζεται σὲ ἀνθρώπινα δεδομένα, καὶ οἱ ἄνθρωποι ἐν γένει ἔχουμε πολλὲς προκαταλήψεις. Γιὰ παράδειγμα, τὸ πρόγραμμα ἐπιλογῆς ὑποψηφίων τῆς Amazon βαθμολογοῦσε ἀρνητικὰ γυναῖκες ὑποψηφίους, ἁπλᾶ καὶ μόνον γιατὶ τὸ δεῖγμα ἐκπαίδευσής του εἶχε λιγώτερες αἰτήσεις ἀπὸ γυναῖκες (πρόκειται γιὰ γνωστὸ πρόβλημα μὴ ἰσορροπημένου δείγματος). Ἀντίστοιχα, τὸ πρόγραμμα περιπολιῶν τῆς Ἀμερικανικῆς Ἀστυνομίας, τὸ ὁποῖο καὶ αὐτὸ βασίζεται σὲ ἀντίστοιχους ἀλγορίθμους, περιλαμβάνει περισσότερες περιπολίες σὲ γειτονιὲς μὲ μαῦρο πληθυσμό, διότι ἀντιλαμβάνεται ὅτι ἐκεῖ συμβαίνουν περισσότερα ἐγκλήματα. Ἐπισήμως αὐτὸ θεωρεῖται ρατσιστικὴ συμπεριφορὰ ποὺ ἔπρεπε νὰ σταματήσει. Ἡ ἐπέμβαση ὅμως τῶν προγραμματιστῶν γιὰ τὴ βελτίωση τῶν προγραμμάτων εἶχε μὴ ἀναμενόμενα ἀποτελέσματα!

Εἶναι συγκλονιστικὸ ὅτι τὸ λογισμικὸ ChatGPT, ἐνῷ ἀναπτύσσεται συνεχῶς ἀπὸ τὸν Νοέμβριο τὸ 2022, ὥστε νὰ καταστεῖ πιὸ λειτουργικὸ καὶ φιλικὸ στὸν χρήστη, χωρὶς τὰ προβλήματα ποὺ προαναφέραμε (εἴμαστε τώρα στὴν ἔκδοση GPT-4 καὶ ἐκπαιδεύεται μὲ ραγδαίους ρυθμοὺς ἡ ἑπόμενη ἔκδοση GPT-5), ἄλλαξε συμπεριφορὰ πρὸς τὸ χειρότερο σὲ μιὰ περίοδο λίγων μηνῶν. Οἱ ἀπαντήσεις του ἔγιναν πιὸ ἀσαφεῖς καὶ πιό «χαζές». Δηλαδή, ἐνῶ πέρυσι τὸν Νοέμβριο μποροῦσε νὰ λύσει τὸ 84% τῶν μαθηματικῶν προβλημάτων ποὺ τοῦ ἔθεταν, σήμερα τὸ ποσοστὸ αὐτὸ ἔχει ἐκφυλιστεῖ στὸ 35 μὲ 50%, καὶ κανεὶς δὲν ξέρει γιατί. Σὲ ἄλλους δεῖκτες παρατηρεῖται μείωση ἀπὸ τὸ 97 στὸ 23%. Δηλαδή, ἐνῷ προσπαθοῦν οἱ προγραμματιστὲς νὰ βελτιώσουν κάποια στοιχεῖα τοῦ λογισμικοῦ, ἐπηρεάζουν μὲ ἄγνωστο τρόπο ἄλλα μέρη τοῦ λογισμικοῦ. Ἡ συμπεριφορὰ αὐτὴ ὀνομάζεται drift (ἀπομάκρυνση). Γιὰ τὸν λόγο αὐτὸ ὁ Max Tegmark ἀπὸ τὸ MIT καὶ πολλοὶ ἄλλοι μαζὶ μὲ αὐτὸν ὁμιλοῦν γιὰ τὴν ἀνάγκη νὰ ἀποκτήσουμε κατανοητὴ Τεχνητὴ Νοημοσύνη (intelligible intelligence) μὲ εὐθυγράμμιση (alignment) ὡς πρὸς τὶς ἀνάγκες καὶ τὰ συμφέροντα τοῦ ἀνθρώπου. Ἐντοπίζουν ὅτι ἐκεῖ ἔγκειται ὁ μεγαλύτερος κίνδυνος, καὶ δυστυχῶς εἴμαστε πολὺ μακριὰ ἀπὸ την εὐθυγράμμιση.

Θὰ ἤθελα νὰ ὁλοκληρώσω αὐτὴ τὴ σύντομη εἰσήγηση στὸ τεράστιο αὐτὸ θέμα τῶν ἡμερῶν μας μὲ κάποιες προσωπικὲς σκέψεις. Μετὰ τὴν ἔκρηξη τῆς στήμονες στὸν χῶρο τῆς ὑπολογιστικῆς μηχανικῆς καὶ σὲ ὁλόκληρη τὴν ἀνθρωπότητα γενικότερα, ἀλλὰ πῶς μποροῦμε νὰ μιλήσουμε γιὰ τὰ οὐσιαστικὰ γνωρί-

Μετὰ τὴν ἔκρηξη τῆς Τεχνητῆς Γενικῆς Νοημοσύνης ποὺ ἀναφέραμε, τοὺς τελευταίους 10 μῆνες εἴδαμε πολλοὺς ἀνθρώπους νὰ δημιουργοῦν ψηφιακὰ ἀντίγραφα τοῦ ἑαυτοῦ τους, εἴδαμε ἄλλους νά «παντρεύονται» ἕνα ψηφιακὸ avatar τὸ ὁποῖο ἐκπαίδευσαν κατ᾽ ἐπιλογήν τους, ἀκούσαμε ἄπειρους προβληματισμοὺς γιὰ τὸ ἂν ἔχουμε τὸ δικαίωμα «νὰ τὸ βγάλουμε ἀπὸ τὴν πρίζα», ὅταν λέει πράγματα ποὺ δὲν μᾶς ἀρέσουν, καὶ γενικὰ ἂν ἕνα τέτοιο λογισμικὸ ἔχει «μυαλό», «νοημοσύνη», «συναισθήματα», «δικαιώματα», «κατανόηση», «ἐλευθερία», «αὐτοσυνειδησία»! Λυπᾶμαι πολὺ γι᾽ αὐτὴ τὴ σύγχυση ποὺ ἐπικρατεῖ ἀνάμεσα στοὺς πιὸ ἔγκριτους ἐπιστήμονες στὸν χῶρο τῆς ὑπολογιστικῆς μηχανικῆς καὶ σὲ ὁλόκληρη τὴν ἀνθρωπότητα γενικότερα, ἀλλὰ πῶς μποροῦμε νὰ μιλήσουμε γιὰ τὰ οὐσιαστικὰ γνωρίσματα τοῦ ἀνθρώπου, ὅταν ἔχουμε χάσει τὴ γνώση τῆς θεμελιώδους «πνευματικῆς φυσιολογίας» τοῦ ἀνθρώπου, καὶ ὅταν ἔχουμε νομοθετήσει τὴν ἀποστασία ἀπὸ τὸν νόμο τοῦ Θεοῦ;

Τεχνητῆς Γενικῆς Νοημοσύνης ποὺ ἀναφέραμε, τοὺς τελευταίους 10 μῆνες εἴδαμε πολλοὺς ἀνθρώπους νὰ δημιουργοῦν ψηφιακὰ ἀντίγραφα τοῦ ἑαυτοῦ τους, εἴδαμε ἄλλους νά «παντρεύονται» ἕνα ψηφιακὸ avatar τὸ ὁποῖο ἐκπαίδευσαν κατ᾽ ἐπιλογήν τους, ἀκούσαμε ἄπειρους προβληματισμοὺς γιὰ τὸ ἂν ἔχουμε τὸ δικαίωμα «νὰ τὸ βγάλουμε ἀπὸ τὴν πρίζα», ὅταν λέει πράγματα ποὺ δὲν μᾶς ἀρέσουν, καὶ γενικὰ ἂν ἕνα τέτοιο λογισμικὸ ἔχει «μυαλό», «νοημοσύνη», «συναισθήματα», «δικαιώματα», «κατανόηση», «ἐλευθερία», «αὐτοσυνειδησία»! Λυπᾶμαι πολὺ γι᾽ αὐτὴ τὴ σύγχυση ποὺ ἐπικρατεῖ ἀνάμεσα στοὺς πιὸ ἔγκριτους ἐπιστήμονες στὸν χῶρο τῆς ὑπολογιστικῆς μηχανικῆς καὶ σὲ ὁλόκληρη τὴν ἀνθρωπότητα γενικότερα, ἀλλὰ πῶς μποροῦμε νὰ μιλήσουμε γιὰ τὰ οὐσιαστικὰ γνωρίσματα τοῦ ἀνθρώπου, ὅταν ἔχουμε χάσει τὴ γνώση τῆς θεμελιώδους «πνευματικῆς φυσιολογίας» τοῦ ἀνθρώπου, καὶ ὅταν ἔχουμε νομοθετήσει τὴν ἀποστασία ἀπὸ τὸν νόμο τοῦ Θεοῦ; Εἴδαμε παραπάνω ὅτι ἀδυνατοῦμε νὰ ὁρίσουμε τί εἶναι ἡ νοημοσύνη (τεχνητὴ ἢ φυσικὴ) γιὰ τὴν ὁποία μιλήσαμε σήμερα. Πόσοι ὅμως ἀπὸ ἐμᾶς γνωρίζουμε τί εἶναι ὁ νοῦς, ἡ ψυχή, ἡ καρδιὰ καὶ ἡ διάνοια, καὶ ποιὰ ἡ σχέση τους μὲ τὸ σῶμα τοῦ ἀνθρώπου, καὶ πῶς αὐτὲς οἱ ἔννοιες ἀναφέρονται στὴν Καινὴ Διαθήκη καὶ ἀναλύονται στὰ κείμενα τῶν Πατέρων τῆς Ἐκκλησίας ὅπως ὁ Ἅγιος Γρηγόριος ὁ Παλαμᾶς, ὁ Ἅγιος Συμεὼν ὁ Νέος Θεολόγος, ὁ Ἅγιος Μάξιμος ὁ

Ὁμολογητής; Ὁ Μητροπολίτης Ναυπάκτου Ἱερόθεος στὸ βιβλίο του *Ὀρθόδοξη Ψυχοθεραπεία* γράφει: *«Χάσαμε τὴν παράδοσή μας, γι᾽ αὐτὸ καὶ πολλοὶ ἀπὸ μᾶς ταυτίζουμε τὸν νοῦ μὲ τὴ λογική. Δὲν ὑποπτευόμαστε καθόλου ὅτι ἐκτὸς ἀπὸ τὴ λογικὴ ὑπάρχει καὶ ἄλλη δύναμη, ποὺ ἔχει μεγαλύτερη ἀξία, δηλαδὴ ὁ νοῦς, ἡ καρδιά. Ὅλος ὁ πολιτισμὸς εἶναι πολιτισμὸς ἀπώλειας τῆς καρδιᾶς. Καὶ κάτι ποὺ δὲν ἔχει ὁ ἄνθρωπος μέσα του δὲν μπορεῖ νὰ τὸ ἀντιληφθεῖ. Ἡ καρδιὰ νεκρώθηκε, ὁ νοῦς σκοτίσθηκε, καὶ δὲν μποροῦμε νὰ ἀντιληφθοῦμε τὴν παρουσία τους. Γιὰ τὸν ἄνθρωπο ποὺ ἔχει μέσα του τὸ Ἅγιο Πνεῦμα, γι᾽ αὐτὸν ποὺ βρίσκεται "ἐν τῇ ἀποκαλύψει", δὲν χρειάζονται πολλὲς διασαφηνίσεις, γιατὶ αὐτὸς γνωρίζει ἀπὸ τὴν πεῖρά του τὴν παρουσία καὶ τὴν ὕπαρξη τοῦ νοός, τῆς καρδιᾶς»*. Καὶ συμπληρώνω: καὶ στὸν ἑαυτό του, καὶ στὸν συνάνθρωπο. Εἴπαμε προηγουμένως ὅτι ἕνας ψυχικὰ ἄρρωστος δὲν μπορεῖ νὰ ξεχωρίσει ἕναν ἄνθρωπο ἀπὸ ἕνα ρομπότ. Πόσοι ὅμως ἀπὸ ἐμᾶς εἴμαστε πραγματικὰ ψυχικὰ ὑγιεῖς;

Ὁ Ἅγιος Μάξιμος ὁ Ὁμολογητὴς μᾶς διδάσκει ὅτι τὸ ἀνθρώπινο σῶμα εἶναι ὁ ναός, ἱερὸ βῆμα εἶναι ἡ καρδιὰ καὶ θυσιαστήριο ὁ νοῦς. Καὶ ὅπως στὴν Ἁγία Τράπεζα δὲν ἐπιτρέπεται νὰ βάλουμε τίποτε ἄλλο, παρὰ μόνο τὸ Ἅγιο Δισκοπότηρο καὶ τὸ ἱερὸ Εὐαγγέλιο, ἔτσι καὶ στὸ θυσιαστήριο τοῦ νοῦ τοῦ ἀληθινοῦ χριστιανοῦ δὲν ἐπιτρέπεται νὰ ὑπάρχει τίποτε ἄλλο παρὰ μόνον ὁ Χριστός, ἡ ἀδιάλειπτη μνήμη τοῦ Ἰησοῦ Χριστοῦ. Αὐτὴ εἶναι ἡ ἡσυχαστικὴ ἐμπειρία τῆς Ἐκκλησίας, ποὺ συνεχίζεται μέχρι τὶς μέρες μας, περνώντας καὶ ἀπὸ τοὺς μεγάλους σύγχρονους ἡσυχαστὲς τοῦ Ἁγιοφαραγγου τῶν Ἀστερουσίων, τοῦ Ἁγίου Εὐμενίου τῆς Ἐθιᾶς καὶ τῆς γερόντισσας Γαλακτίας τῆς Παναγίας της Καλυβιανῆς. Αὐτοὺς τοὺς ἀνθρώπους ἐμεῖς τοὺς γνωρίσαμε. Ἡ ἐμπειρία αὐτὴ εἶναι σήμερα ζωντανὴ στὴν νότια Κρήτη καὶ σὲ πολλὰ μέρη τῆς Ἑλλάδας, ἀλλὰ καὶ τοῦ κόσμου. Καὶ αὐτὴν τὴν ἐμπειρία πρέπει νὰ τὴν σπουδάζουμε ὅλη μας τὴν ζωὴ καὶ νὰ τὴ μεταδίδουμε στὶς ἑπόμενες γενεές. Μόνον ἔτσι θὰ μπορέσουμε νὰ ἀντιμετωπίσουμε χωρὶς φόβο τὶς νέες προκλήσεις ποὺ φέρνει ἡ νέα τεχνολογία. Ἀλλιῶς, ὁ κόσμος μας κινδυνεύει νὰ δαιμονιστεῖ. Ἂν δὲν ἔχει δαιμονιστεῖ ἤδη...

Σᾶς εὐχαριστῶ γιὰ τὴν προσοχή σας!

ΙΩΑΝΝΗΣ ΚΟΝΤΟΠΟΥΛΟΣ
Διευθυντὴς Ἐρευνῶν
Κέντρο Ἐρευνῶν Ἀστρονομίας
καὶ Ἐφαρμοσμένων Μαθηματικῶν
τῆς Ἀκαδημίας Ἀθηνῶν

Σημείωση: Λίγες μέρες πρὶν τὴ δημοσίευση τοῦ παρόντος κειμένου, κυκλοφόρησε μιὰ ἐσωτερικὴ ἐπιστολὴ πρὸς τοὺς ἐργαζομένους τῆς OpenAI, ποὺ τοὺς προειδοποιοῦσε γιὰ τοὺς κινδύνους τοῦ ἐρευνητικοῦ προγράμματος Q*, μὲ τὸ ὁποῖο ἡ ἑταιρεία ὑποτίθεται ὅτι ἔφτασε στὴ Γενικὴ Τεχνητὴ Ὑπερ-Νοημοσύνη! Ὁ ἴδιος ὁ Sam Altman δήλωσε δημοσίως ὅτι οἱ ἄνθρωποι ἀναρωτιοῦνται ἂν αὐτὸ ποὺ ἔφτιαξαν δὲν εἶναι ἁπλῶς ἕνα νέο ἐργαλεῖο, ἀλλὰ ἕνα νέο «ὄν» ("Is this a tool we've built or a creature we've built?")! Τὴν ἑπόμενη μέρα τὸ διοικητικὸ συμβούλιο τῆς ἑταιρείας ἀπέλυσε τὸν Altman, ἀλλὰ ἀναγκάστηκε νὰ τὸν ἐπαναπροσλάβει μία ἑβδομάδα ἀργότερα, ὅταν 700 ἐργαζόμενοι τῆς ἑταιρείας ἀπείλησαν ὅτι θὰ παραιτηθοῦν. Ἀντιλαμβανόμαστε ὅλοι ὅτι οἱ ἐξελίξεις στὸν χῶρο τῆς Τεχνητῆς Νοημοσύνης εἶναι καταιγιστικές. Ὁ Θεὸς ἂς βάλει τὸ χέρι Του.