

Εγκέφαλοι και ευφρείς μηχανές: οι δυνατότητες, τα όρια και οι επιπτώσεις των εφαρμογών της Τεχνητής Νοημοσύνης

Ανδρέας Παπανικολάου

6ο ΠΑΝΕΛΛΗΝΙΟ ΣΥΝΕΔΡΙΟ ΓΝΩΣΙΑΚΗΣ ΕΠΙΣΤΗΜΗΣ
Ξυλόκαστρο Κορινθίας
21-24 Σεπτεμβρίου 2023

Andrew C. Papanicolaou
Professor of Neuroscience Emeritus
College of Medicine
The University of Tennessee, Health Science Center
<https://sites.google.com/site/andreascpapanicolaou>

Στο μέτρο που η λέξη *νοημοσύνη* σημαίνει την δύναμη του γιγνώσκειν και του νοείν, η φράση *τεχνητή νοημοσύνη* αποτελεί σχήμα οξύμωρον. Μάλιστα, η έκπληξη και το δέος που το άκουσμά της εγείρει ειδικά μετά την διάδοση του προγράμματος chatGPT, οφείλεται, κατά κύριο λόγο, σε νεφελώδεις εκφράσεις αυτού του είδους. Και αυτό είναι ένα από τα βασικά θέματα που προτίθεμαι να θίξω.

Αλλά γιατί οξύμωρο; Διότι, απλούστατα, η μεν ανθρώπινη νοημοσύνη συνίσταται από δύο διακριτά και **ουσιαστικώς** ανόμοια φαινόμενα η δε τεχνητή από ένα.

Το πρώτο και κύριο φαινόμενο και συστατικό της ανθρώπινης νοημοσύνης είναι η άμεση η προσωπική γνώση αισθημάτων όπως του πόνου ή συνειδητοποίηση της σημασίας αντιλημμάτων όπως «κόκκινο φανάρι» «τραπέζι» και εννοιών όπως του νοήματος του όρου

«τετραγωνική ρίζα» ή «ωραιότητα». Νοεΐν και γινώσκειν σημαίνουν την απόκτηση και την κατοχή προσωπικής γνώσης ή την συνειδητή εμπειρία εννοιών και πραγμάτων.

Το δεύτερο συστατικό της καθ' ημάς νοημοσύνης είναι το σύνολο των αντιδράσεων, λεκτικών και μη, οι οποίες εκφράζουν την κατανόηση μιας έννοιας, ή την επίγνωση ενός αισθήματος ή μιας οποιασδήποτε εμπειρίας. Όπως, επί παραδείγματι, το επιφώνημα *άχ* ή η φράση *πονάω* που εκφράζει το αίσθημα του πόνου. Όπως το πάτημα του φρένου που σημαίνει πως κατάλαβα το νόημα του αντιλήμματος κόκκινο φανάρι στη διασταύρωση. Όπως η λεκτική απόκριση 4 στο ερώτημα ποια είναι η τετραγωνική ρίζα του 16 που δείχνει ότι εννόησα το ερώτημα.

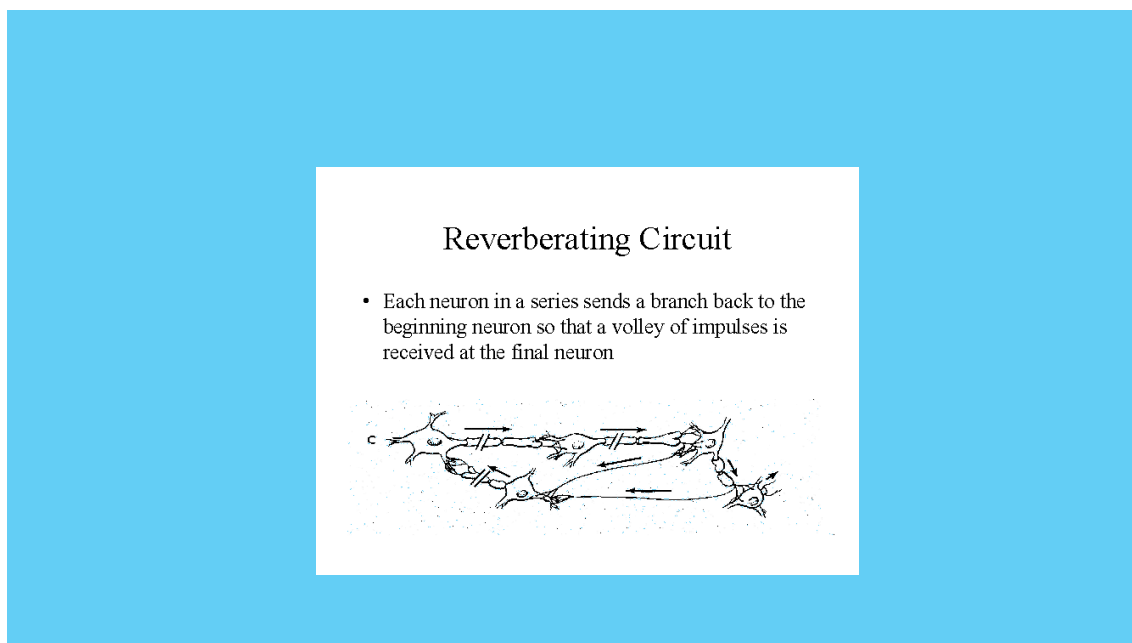
Η τεχνητή νοημοσύνη από το άλλο μέρος συνίσταται μόνον στο ένα φαινόμενο, στην έκφραση της γνώσης ενώ το άλλο φαινόμενο, η συνειδητή γνώση ή η επίγνωση, είναι σήμερα ένας από τους πλέον ευφάνταστους ευσεβείς πόθους μιας κατηγορίας παραπλανημένων θιασωτών σεναρίων επιστημονικής φαντασίας. Παραπλανημένων, όχι επειδή έτσι μ'αρέσει εμένα, αλλά για αντικειμενικούς λόγους τους οποίους θα υποβάλλω στην κρίση σας

Το φαινόμενο αυτό των δημοσίων αντιδράσεων, κοινό στην ανθρώπινη και στην τεχνητή νοημοσύνη, είναι κατά γενική παραδοχή αποτέλεσμα εγκεφαλικών διεργασιών. Ειδικότερα, συγκεκριμένων σχηματισμών ενεργοποίησεως του εγκεφάλου που το περιγράφω τους μπορούμε τώρα και καταγράφουμε με τις νευροαπεικονιστικές μεθόδους.

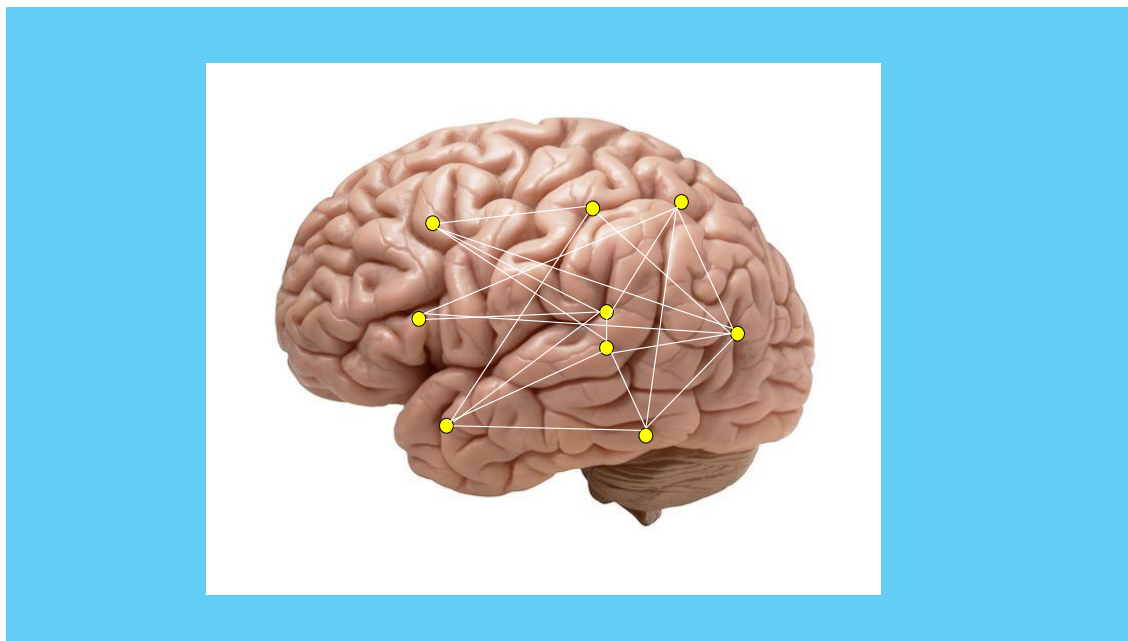
Το δε φαινόμενο της προσωπικής εμπειρίας, της προσωπικής γνώσης, που χαρακτηρίζει μόνο την ανθρώπινη νοημοσύνη, είναι εξ υποθέσεως, μεταφυσικής μάλιστα υποθέσεως, αποτέλεσμα των ίδιων σχηματισμών.

Δηλαδή ο ίδιος σχηματισμός εγκεφαλικής ενεργοποίησης λέγεται ότι προκαλεί και τις συνειδητές εμπειρίες και τις εκφράσεις τους. Τώρα γιατί λέω ότι εδώ πρόκειται για μεταφυσική υπόθεση, νομίζω ότι θα γίνει εμφανές στη συνέχεια αλλά εάν δεν γίνει, το ξανασυζητάμε στο τέλος.

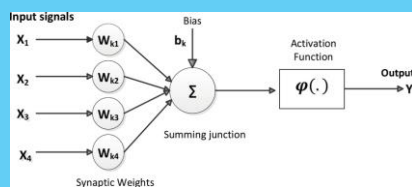
Η ΤΝ θα μπορούσε να θεωρηθεί ως φυσική εξέλιξη των πρώτων προτύπων νευρωνικών δικτύων όπως αυτά που σκιαγράφησε ο William James στο μνημειώδες έργο του *Principles of Psychology*, όπως τα αναπαλλόμενα κυκλώματα του Donald Hebb



Και εδώ πιο σχηματικά



ή όπως και το μαθηματικό πρότυπο των νευρώνων των McCulloch and Pitts



τα οποία αποσκοπούσαν στο να μιμηθούν την δομή και την λειτουργία των εγκεφαλικών νευρωνικών δικτύων.

Εδώ προσέξτε το εξής σημαντικό: Όλα τα πρότυπα στο παρελθόν προσπαθούν να μιμηθούν τους νευρώνες και τα πραγματικά εγκεφαλικά δίκτυα. Μέσω των προτύπων, οι ερευνητές προσπαθούσαν να κατανοήσουν πώς από τα πραγματικά νευρωνικά δίκτυα προκύπτουν τα δίδυμα φαινόμενα της γνώσης και συμπεριφοράς που εκφράζει την γνώση.

Σύντομα όμως έγινε προφανές ότι, λόγω της άγνοιας μας για τον τρόπο λειτουργίας των πραγματικών δικτύων, ήταν προτιμότερο για τα τεχνητά πρότυπα να σκοπεύουν στην παραγωγή της σωστής συμπεριφοράς εκ μέρους των ευφύων μηχανών **ανεξαρτήτως** του τρόπου παραγωγής της συμπεριφοράς αυτής.

Να στοχεύουν, δηλαδή, στην κατασκευή αλγορίθμων αναγνώρισης αντικειμένων, εκμαθήσεως των κανόνων παιγνίων όπως του σκακιού ή της πρέφας, ή της αναπαραγωγής φαινομένων της ανθρώπινης μνήμης έτσι ώστε να μιμούνται την ανάλογη ανθρώπινη **συμπεριφορά** χωρίς όμως να προσπαθούν εκ παραλλήλου να μιμηθούν ούτε τους σχηματισμούς ενεργοποίησεως του ανθρώπινου εγκεφάλου, ούτε τους πραγματικούς αλγορίθμους που διέπουν τα επί μέρους εγκεφαλικά λειτουργικά δίκτυα – πράγματα ούτως ή άλλως κατά βάσιν άγνωστα.

Πολλοί μάλιστα όπως ο Chomsky,

N. Chomsky, I. Roberts, J. Watumull, "The false promise of ChatGPT", The New York Times, 8th March 2023.

εκφράστηκαν απαξιωτικά για την τεχνητή νοημοσύνη ακριβώς επειδή δεν διαφωτίζει τους μηχανισμούς της ανθρώπινης νοημοσύνης αλλά στοχεύει μόνο στην μίμηση της ανθρώπινης συμπεριφοράς.

Αλλά, όπως είπα, οι στόχοι της TN σήμερα είναι άλλοι: δεν είναι η κατανόηση των μηχανισμών του γινώσκειν και του νοείν, και προάγονται με την χρήση μεθόδων που καταφέρνουν να μιμούνται την συμπεριφορά του ανθρώπου με τρόπους που είναι πιο κατάλληλοι και πιο εύκολοι για τους προγραμματιστές, χωρίς να νοιάζονται για το πώς ο εγκέφαλος παράγει γνώσεις και κατάλληλες, στην κάθε περίπτωση, αντιδράσεις.

Επί παραδείγματι, οι αλγόριθμοι που διέπουν ένα ρομπότ, αποσκοπούν μόνο στο να συνδέσουν την προσταγή—το ακουστικό ερέθισμα-- «σκούπισε το πάτωμα» με τις κατάλληλες προγραμματισμένες κινήσεις και όχι να μιμηθούν τους εγκεφαλικούς αλγόριθμους προσλήψεως λόγου ούτε να κατανοήσουν το νόημα της προσταγής, παρά να απαντήσουν στο ακουστικό ερέθισμα με συγκεκριμένες κινήσεις .

Λες στο μηχάνημα «σκούπισε» και το μηχάνημα σκουπίζει χωρίς να εννοήσει τίποτα· χωρίς να έχει την ανάλογη συνειδητή εμπειρία. Αυτό σήμερα είναι γεγονός. Δεν είναι ένας αυθαίρετος δικός μου ισχυρισμός. Για το τι θα γίνει αύριο θα το κουβεντιάσουμε κι αυτό οσονούπω.

Επομένως, σήμερα, ανθρωπομορφικές μεταφορικές εκφράσεις όπως το ρομπότ σκέφτηκε ή ο αλγόριθμος κατάλαβε, ή το σύστημα θέλησε, και ούτω καθ' εξής, είναι συσκοτιστικές παρόλη την ποιητική τους αύρα και καλά θα κάναμε να μην τις χρησιμοποιούμε όπου κι όταν μπορούμε, διότι εκτρέπουν την προσοχή μας από τις πραγματικές δυνατότητες και αδυναμίες της Τεχνητής Νοημοσύνης.

Αφήνω τώρα τα γενικά σχόλια για να προχωρήσω στην περιγραφή της TN η οποία, όμως, θα είναι ότι άλλο παρά εξειδικευμένη και λεπτομερής. Αυτό που σκοπεύω να κάνω είναι να αναφερθώ επιλεκτικά σε εκείνες τις όψεις ή τα χαρακτηριστικά της που θα διευκολύνουν την κατανόηση των πραγματικών ορίων και το πραγματικό φάσμα των δυνατοτήτων της αλλά και την κατανόηση των δικών μου επιχειρημάτων και απαντήσεων στα εξής ερωτήματα :

Πώς ακριβώς συμβάλλει (και θα συμβάλλει) η TN στην αναβάθμιση του επιπέδου ζωής και έναντι τι κόστους;

Σε τι είδους δραστηριότητες τα συστήματα TN θα υπερβούν τις επιδόσεις του ανθρώπου;

Θα υπερβεί η TN τις ανθρώπινες δυνατότητες δημιουργικότητας και καινοτομίας;

Είναι δυνατό μελλοντικά συστήματα TN να εξελιχθούν σε ενσυνείδητα όντα;

Πρώτον, ποια είναι τα σημερινά και τα μελλούμενα οφέλη στους διάφορους τομείς εφαρμογής της τεχνητής νοημοσύνης, ποιο το μέγεθος του κόστους της κακής διαχείρισής της και τι μέτρα είναι πιθανόν να παρθούν για την παράκαμψη ή την μετρίαση του κόστους;

Η εξέταση του πρώτου αυτού ερωτήματος θα διαλευκάνει επίσης το ερώτημα για ποια επιμέρους έργα η τεχνητή νοημοσύνη ξεπερνά τις ανθρώπινες δυνατότητες.

Αλλά, πιστεύω, τα πιο ενδιαφέροντα ερωτήματα, τουλάχιστον για εσάς, είναι τα δύο τελευταία. Δηλαδή εάν μελλοντικά τεχνητά συστήματα θα υπερβούν την ανθρώπινη εφευρετικότητα και πρωτοτυπία και, το σπουδαιότερο, εάν οι ευφυείς μηχανές θα αποκτήσουν αυτοσυνείδηση.

Για να αντιμετωπίσουμε αυτά τα ερωτήματα είναι απαραίτητες μερικές πληροφορίες σχετικά με την εσωτερική λειτουργία των τεχνητών συστημάτων.

Καταρχάς, υπάρχουν αρκετοί διαφορετικοί τύποι συστημάτων ΤΝ.

Πρότυπα (ή συστήματα) ειδικής χρήσεως

Narrow or Weak AI

Προτυπα (ή συστήματα) πολλαπλών επιδόσεων

Strong or General AI or AGI

(Γενικευμένα Γλωσσικά πρότυπα

Large language Models --LLM)

Μερικά από αυτά τα συστήματα έχουν σχεδιαστεί για να λύνουν πολύ συγκεκριμένα προβλήματα, όπως το πώς να παίζουν σκάκι. Αυτά ονομάζονται weak ή narrow παρόλον ότι ούτε στενά είναι ούτε αδύναμα είναι αλλά είναι απλώς **περιορισμένων επιδόσεων ή ειδικής χρήσεως**

Αλλά έχουν σχεδιαστεί για να επιλύουν πολλά και διαφορετικά προβλήματα. Προβλήματα που η λύση τους δεν εξαρτάται από την εφαρμογή άκαμπτων κανόνων, όπως το σκάκι, και για να παρέχουν νέες, καινοτόμες λύσεις. Αυτά ονομάζονται Strong -ισχυρά ή General-γενικά ή προηγμένα, ή **πρότυπα πολλαπλών επιδόσεων**.

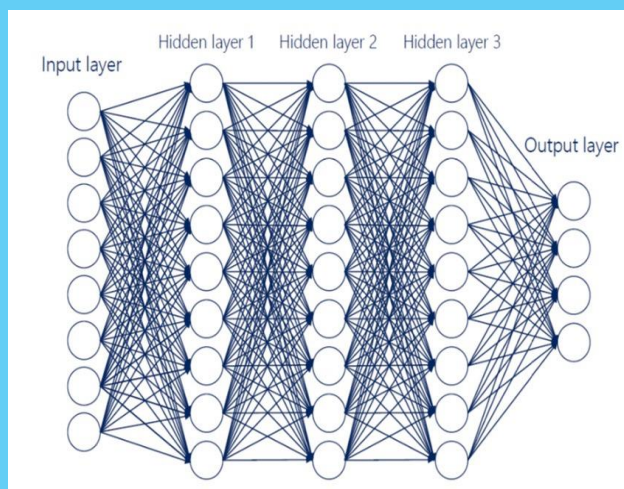
Τα περισσότερα από τα τελευταία λειτουργούν με τους λεγόμενους αλγόριθμους deep learning που καθόλου βαθείς δεν είναι παρά αποτελούν μια ιδιαίτερη ποικιλία μηχανικής μάθησης machine learning που επίσης είναι όρος αίολος καθότι το μηχάνημα τίποτα δεν μαθαίνει- αλλά αυτό θα το εξετάσουμε αργότερα.

Το διάσημο πλέον ChatGPT (ή παραγωγικός (generative) προεκπαιδευμένος μετασχηματιστής – pretrained transformer) που

κατασκευάζεται από την εταιρεία Open AI, είναι ένα τέτοιο σύστημα. Και υπάρχουν και άλλα, παρόμοια με το ChatGPT γνωστά συλλογικά ως Large Language Models ή LLM – γενικευμένα, δηλαδή, γλωσσικά πρότυπα.

Μια σχηματική, τώρα, αναπαράσταση της αρχιτεκτονικής των σύγχρονων δικτύων γενικής τεχνητής νοημοσύνης, όπως αυτό του ChatGPT.

Schematic of a neural network capable of “deep learning”



Αποτελείται από στρώματα κόμβων διατεταγμένων από αριστερά προς τα δεξιά. Κάθε κόμβος, που αντιπροσωπεύεται εδώ από έναν κύκλο, είναι το τεχνητό αντίστοιχο ενός νευρώνα. Το σχήμα περιλαμβάνει επίσης τις φυσικές συνδέσεις κάθε κόμβου με όλους τους άλλους κόμβους του επόμενου επιπέδου.

Οι κόμβοι του πρώτου στρώματος στα αριστερά είναι αισθητήρες. Προσλαμβάνουν όσα δεδομένα, είτε οπτικοακουστικά είτε γλωσσικά, επιλέξουμε να τους παρουσιάσουμε.

Το τελευταίο επίπεδο, στα δεξιά, αποτελείται από κόμβους που παράγουν τις απαντήσεις του δικτύου στις ερωτήσεις μας. Οι απαντήσεις αυτές μπορεί να είναι λεκτικές — που εμφανίζονται με τη μορφή κειμένου στον υπολογιστή μας ή συνίστανται στην παραγωγή πινάκων ζωγραφικής ή μουσικών κομματιών. Αυτό βέβαια εξαρτάται με τι περιφερειακές μηχανές είναι συνδεδεμένο το τελευταίο στρώμα των κόμβων.

Τα μεσαία στρώματα αναλύουν τα δεδομένα. Αλλά τα αναλύουν με έναν ιδιόμορφο τρόπο που θα περιγράψω και ο οποίος, όπως προείπα, δεν μοιάζει καθόλου με τον τρόπο που οι νευρώνες **μας** διαχειρίζονται αισθητικά δεδομένα, ούτε με τον τρόπο που θεωρούμε ότι τα εγκεφαλικά μας δίκτυα προετοιμάζουν τις αποκρίσεις μας.

Σε αυτό το σχήμα βλέπετε μόνο τρία μεσαία στρώματα κόμβων. Αλλά το ChatGPT έχει 12 τέτοια στρώματα με πολλούς κόμβους το καθένα με αποτέλεσμα να δημιουργούνται δισεκατομμύρια δομικές και λειτουργικές συνδέσεις μεταξύ των κόμβων.

Επειδή ένα τέτοιο δίκτυο για να είναι επαρκές πρέπει να εκτίθεται σε τεράστιες ποσότητες δεδομένων τα τεχνητά δίκτυα και οι αλγόριθμοι που λειτουργούν σε αυτά απαιτούν ολόκληρες φάρμες συνδεδεμένων υπερυπολογιστών. Οι εγκέφαλοι των προηγμένων γλωσσικών προτύπων όπως το ChatGPT, με άλλα λόγια, είναι ολόκληρα στρέμματα κυκλωμάτων κάπου στην Καλιφόρνια ή δεν ξέρω πού αλλού.

Εμείς, οι χρήστες, όταν επικοινωνούμε με το ChatGPT ή οποιοδήποτε άλλο προηγμένο δίκτυο, έχουμε πρόσβαση σε αυτά μέσω του διαδικτύου χρησιμοποιώντας τους δικούς μας υπολογιστές ή τα

τηλέφωνα, με τον ίδιο τρόπο που τα χρησιμοποιούμε για τυπικές αναζητήσεις στο διαδίκτυο.

Αλλά οι μηχανές τεχνητής νοημοσύνης μπορούν επίσης να είναι αυτόνομες όπως τα ευφυή ρομπότ που σήμερα σκουπίζουν πατώματα, αλλά τα οποία, στο μέλλον, οπλισμένα, ίσως, με κβαντικούς υπολογιστές, μπορεί ακόμη και να επαναστατήσουν κατά της ανθρωπότητας, σύμφωνα με ορισμένες προβλέψεις.

Επιτρέψτε μου τώρα να αναφερθώ σε μερικά από τα πιο βασικά χαρακτηριστικά της λειτουργίας των προηγμένων γενικών δικτύων. Πώς δηλαδή τα συστήματα αυτά συλλέγουν πληροφορίες και πώς χειρίζονται αυτά τα δεδομένα για την λύση διαφόρων προβλημάτων.

Οι εν λόγω διαδικασίες, παραδόξως, απαιτούν ελάχιστο ή καθόλου ειδικό για το κάθε είδος δεδομένων προγραμματισμό, πέρα από την έκθεση του δικτύου σε αυτά τα δεδομένα.

[1] I. Pitas, "Artificial Intelligence Science and Society Part A: Introduction to AI Science and Information Technology", Amazon/Kindle Direct Publishing, 2022,

[2] I. Pitas, "Artificial Intelligence Science and Society Part B: AI Science, Mind and Humans", Amazon/Kindle Direct Publishing, 2022,

[3] I. Pitas, "Artificial Intelligence Science and Society Part C: AI Science and Society", Amazon/Kindle Direct Publishing, 2022,

[4] I. Pitas, "Artificial Intelligence Science and Society Part D: AI Science and the Environment", Amazon/Kindle Direct Publishing, 2022,

Όσοι από εσάς θέλετε λεπτομερή περιγραφή της λειτουργίας των συστημάτων θα μπορούσατε να συμβουλευθείτε το τετράτομο

σύγγραμμα του Ιωάννη Πήτα καθηγητή στο Αριστοτέλειο. Εδώ θα θίξω μόνον τις πλέον γενικές αρχές της λειτουργίας τους.

Καταρχάς, όπως έχω ήδη αναφέρει, τα συστήματα είναι προικισμένα με τους λεγόμενους αλγόριθμους deep learning οι οποίοι δεν αφορούν ούτε βαθιά ούτε ρηχή μάθηση παρά μια μηχανική, παρατακτική και επαναλαμβανόμενη διαδικασία που, βασικά, περιγράφεται με μια μαθηματική εξίσωση.

Η διαδικασία αυτή συνίσταται στην εξαγωγή στατιστικών κανονικοτήτων σε εικόνες, σε κομμάτια μουσικής ή, γενικότερα, ήχων και σε κείμενα στα οποία το σύστημα εκτίθεται. Ένα παράδειγμα τέτοιων κανονικοτήτων είναι η πιθανότητα εμφάνισης μιας συγκεκριμένης λέξης μετά από μια σειρά άλλων λέξεων- δηλαδή η εξαγωγή των συντακτικών κανόνων μιας συγκεκριμένης γλώσσας.

ο καιρός αύριο θα είναι..... καλός βροχερός άσχημος νεφελώδης....

ο καιρός αύριο θα είναι..... κίτρινο , Ανάργυρος, μετέπειτα

Το σύστημα μαθαίνει, με άλλα λόγια,--μαθαίνει πάντα εντός εισαγωγικών-- από τα κείμενα στα οποία εκτίθεται, ότι μετά τη σειρά λέξεων «ο καιρός αύριο θα είναι...» η πιθανότητα είναι πολύ μεγαλύτερη για λέξεις όπως καλός ή κακός, ή ηλιόλουστος, ή βροχερός

κ.λπ., να συμπληρώσουν την πρόταση παρά για λέξεις όπως *κίτρινο* ή *επίσκοπος* ή *Μελπομένη*. Και έχοντας υπολογίσει τις πιθανότητες μπορεί να παραγάγει σωστές απαντήσεις γραμματικά και σημασιολογικά.

Το πώς ακριβώς το επιτυγχάνουν αυτό οι αλγόριθμοι, είναι λίγο μυστήριο.

Αυτό που είναι γνωστό είναι ότι το σήμα εξόδου κάθε κόμβου πολλαπλασιάζεται με έναν συντελεστή του οποίου η αριθμητική τιμή αλλάζει έτσι ώστε οι έξοδοι των κόμβων, δηλαδή τα σήματα που παράγουν οι κόμβοι, να αναπροσαρμόζονται συνεχώς.

Αυτή η αναπροσαρμογή επιτυγχάνεται μέσω μιας διαδικασίας που λέγεται *backpropagation* η οποία είναι μια μορφή ανάδρασης από τους κόμβους του στρώματος εξόδου προς τους κόμβους των μεσαίων στρωμάτων.

Επί παραδείγματι, εάν το σύστημα δεν έχει ολοκληρώσει την εκπαίδευσή του έτσι ώστε να προσθέτει τις πιο πιθανές λέξεις, δηλαδή τις λέξεις *αίθριος* ή *βροχερός* στο τέλος της σειράς *ο καιρός αύριο θα είναι*, αλλά προσθέτει τη λέξη *μετέπειτα* ή *Ανάργυρος*, η ασυμφωνία μεταξύ της αναμενόμενης (της πλέον πιθανής) και της πραγματικής απόκρισης ανατροφοδοτείται με αποτέλεσμα την αλλαγή των συντελεστών. Και αυτό επαναλαμβάνεται έως ότου το σύστημα δώσει τη σωστή απόκριση.

Το σύστημα, βέβαια, δεν έχει μάθει απολύτως τίποτα για το τι σημαίνει *καιρός*, τι σημαίνει *αιθρία*, *ηλιόλουστος* κλπ. Το σύστημα έχει ρυθμιστεί έτσι ώστε ενόψει ενός αντικειμένου—εδώ μιας σειράς λέξεων να αποκριθεί με λέξεις που στα χιλιάδες κείμενα στα οποία έχει εκτεθεί, απαντούν με υψηλή συχνότητα. Κι έτσι δημιουργεί την εντύπωση ότι έχει μάθει γραμματική, συντακτικό και, το κυριότερο, ότι κατανοεί το νόημα των διαμειβομένων μεταξύ ημών και εκείνου.

Αυτό που δεν είναι γνωστό είναι πώς ακριβώς προσαρμόζονται αυτά τα δισεκατομμύρια τιμών των συντελεστών. Με ποιο σχέδιο ή λογική αυξάνονται ή μειώνονται δεδομένης μιας συγκεκριμένης ανάδρασης και σε ποιο βαθμό τροποποιούν το σήμα εξόδου του κάθε κόμβου.

Μάλιστα, αυτή η έλλειψη γνώσης είναι ίσως μια πηγή ανησυχίας, για μερικούς ειδικούς, στο μέτρο που σημαίνει απώλεια του απόλυτου ανθρώπινου ελέγχου της λειτουργίας πολύπλοκων τεχνητών δικτύων.

Ωστόσο, ανεξάρτητα από το πώς μπορεί να συμβαίνει αυτή η προσαρμογή, το δίκτυο, σταδιακά, αποκρίνεται καλύτερα τόσο συντακτικά όσο και σημασιολογικά σε ερωτήματα και δημιουργεί όλο και πιο ακριβείς απαντήσεις.

.....

Έχοντας όλα αυτά κατά νου, μπορούμε τώρα να αντιμετωπίσουμε ένα προς ένα τα ερωτήματα των δυνατοτήτων και των ορίων της τεχνητής νοημοσύνης

Για να απαντήσω στην πρώτη ερώτηση για το ποιες είναι οι τρέχουσες εφαρμογές τεχνητής νοημοσύνης, ρώτησα το ChatGPT αντί να ψάξω μόνος μου στο διαδίκτυο και ορίστε πώς μου απάντησε:

Η ΤΝ εξακολουθεί να είναι ένας τομέας ενεργού έρευνας και το πεδίο συνεχίζει να εξελίσσεται με ταχείς ρυθμούς. Ακολουθούν μερικές από τις πιθανές εφαρμογές της γενικής ΤΝ:

Παρένθεση εδώ: Τι ακριβώς σημαίνει τομείς εφαρμογής; Ακόμα και τα υπερσυστήματα γενικής τεχνητής νοημοσύνης έχουν κάποια όρια. Ο βασικός παράγων που καθορίζει τα όρια είναι τα δεδομένα στα οποία έχει εκτεθεί το σύστημα. Τι ακριβώς πληροφορίες θα απορροφήσει το σύστημα, επομένως σε τι ερωτήματα θα μπορεί να απαντήσει, είναι, σήμερα, ανθρώπινη επιλογή.

Κάποιοι άνθρωποι αποφασίζουν εάν η μηχανή θα εκτεθεί σε κείμενα της αρχαίας Θιβετιανής γλώσσας ή στα Principia Mathematica ή σε οτιδήποτε άλλο-- πράγμα που θα καθορίσει τις δυνατότητες του συστήματος να προσεγγίσει ζητήματα ανατολίτικης θεολογίας ή της λογικής βάσης των μαθηματικών, ή ζητήματα που εμπίπτουν στο καθένα από τα παρακάτω πεδία εφαρμογών που μου παρέθεσε το σύστημα ChatGPT: όπως

- 1.Ρομποτική
2. Αυτόνομα Οχήματα
3. Εξατομικευμένη Υγεία
4. Επεξεργασία φυσικής γλώσσας
5. Επιστημονική Έρευνα
6. Χρηματοοικονομική Ανάλυση
7. Εξατομικευμένη Εκπαίδευση

1.Ρομποτική: Το σύστημα μπορεί να χρησιμοποιηθεί για τη δημιουργία προηγμένων ρομποτικών συστημάτων ικανών να εκτελούν σύνθετες εργασίες σε διάφορους τομείς, όπως διάφορες κατασκευές (αντικειμένων, σπιτιών, πλοίων, αεροπλάνων), η υγειονομική περίθαλψη, η γεωργία και η εξερεύνηση. Τα ρομπότ που λειτουργούν με σύστημα TN μπορούν να προσαρμοστούν σε νέα περιβάλλοντα, να μάθουν από την εμπειρία και να αλληλεπιδράσουν με τους ανθρώπους πιο αποτελεσματικά. Φανταστείτε, λοιπόν ολόκληρα εργοτάξια να διεκπεραιώνουν έργα χωρίς την παρουσία ανθρώπων.

2. Αυτόνομα Οχήματα: Το σύστημα έχει τη δυνατότητα να φέρει επανάσταση στον κλάδο των μεταφορών, ενεργοποιώντας πλήρως

αυτόνομα οχήματα. Τα συστήματα ΤΝ μπορούν να αντιληφθούν και να κατανοήσουν το περιβάλλον, να λάβουν αποφάσεις σε πραγματικό χρόνο και να πλοηγηθούν με ασφάλεια στην κυκλοφορία.

Εδώ, βέβαια, πρόκειται για μεταφορική χρήση των όρων *αντιλαμβάνομαι και κατανοώ*— χρήση που, όπως προείπα, συσκοτίζει την πραγματικότητα και που γι αυτόν τον λόγο πρέπει να αποφεύγεται πάση θυσία. Τα υφαρκτά συστήματα τίποτα δεν αντιλαμβάνονται και τίποτα απολύτως δεν κατανοούν. Η μηχανική διαδικασία που περιέγραψα δεν είναι νοήμον όν για να καταλάβει. Μια εξίσωση είναι, γραμμένη σε κώδικα. Απλώς το σύστημα συμπεριφέρεται σαν να αντιλαμβάνεται, σαν να κατανοεί.

Ωστόσο οι επιπτώσεις της συμπεριφοράς του είναι ανυπολόγιστες: Όταν μιλάμε για αυτόνομα οχήματα δεν αναφερόμαστε μόνο σε ΙΧ αλλά σε τραίνα, πλοία, νταλίκες αεροσκάφη-σκεφτείτε τις κοσμογονικής εμβέλειας επιπτώσεις καταρχάς στην ανεργία και μετά την ανακατανομή ανθρώπινου προσωπικού που θα έχει αυτή η εφαρμογή.

3. Εξατομικευμένη Υγεία: *Η σύστημα μπορεί να βοηθήσει στην εξατομικευμένη ιατρική αναλύοντας τεράστιες ποσότητες δεδομένων ασθενών, συμπεριλαμβανομένων ιατρικών αρχείων, γενετικών πληροφοριών και παραγόντων τρόπου ζωής. Μπορεί να βοηθήσει στη διάγνωση ασθενειών, στη σύσταση σχεδίων θεραπείας και στην πρόβλεψη των αποτελεσμάτων των ασθενών.*

4. Επεξεργασία φυσικής γλώσσας: *Το σύστημα μπορεί να βελτιώσει σημαντικά τις ικανότητες επεξεργασίας φυσικής γλώσσας. Μπορεί να κατανοήσει και να δημιουργήσει ανθρώπινη γλώσσα με μεγαλύτερη ακρίβεια και επίγνωση του πλαισίου.*

Έτσι εκφράζεται τώρα η μηχανή γιατί έτσι της έχουν επιτρέψει να εκφράζεται. Στην ουσία, όπως άλλωστε εξήγησα, το σύστημα απλώς μιμείται την ανθρώπινη γλωσσική συμπεριφορά και απαντά σε

Λεκτικά ερωτήματα με την επάρκεια που του παρέχει η έκθεσή του σε εκατομμύρια καταγεγραμμένων σωστών γλωσσικών εκφράσεων.

5. Επιστημονική Έρευνα: Το σύστημα μπορεί να επιταχύνει την επιστημονική ανακάλυψη αναλύοντας τεράστια σύνολα δεδομένων, εκτελώντας προσομοιώσεις και εντοπίζοντας σχηματισμούς και συσχετισμούς που μπορεί να μην είναι άμεσα εμφανείς στους ανθρώπους ερευνητές. Μπορεί να συμβάλει στην πρόοδο σε τομείς όπως η γονιδιωματική, η επιστήμη των υλικών, η ανακάλυψη φαρμάκων και η κλιματική μοντελοποίηση.

Καμμία αντίρρηση

6. Χρηματοοικονομική Ανάλυση: Τα συστήματα TN μπορούν να αναλύσουν μεγάλους όγκους χρηματοοικονομικών δεδομένων, να παρακολουθούν τις τάσεις της αγοράς και να λαμβάνουν έξυπνες επενδυτικές αποφάσεις. Μπορούν να παρέχουν πολύτιμες πληροφορίες για τη διαχείριση χαρτοφυλακίου, την αξιολόγηση κινδύνου και τις αλγοριθμικές συναλλαγές.

Συμφωνώ.

7. Εξατομικευμένη Εκπαίδευση: Το σύστημα μπορεί να εξατομικεύσει τη μαθησιακή εμπειρία για τους μαθητές προσαρμόζοντας μεθόδους διδασκαλίας και περιεχόμενο στις ατομικές ανάγκες και στυλ μάθησης. Μπορεί να παρέχει έξυπνη διδασκαλία, ανατροφοδότηση και προσαρμοστικές αξιολογήσεις.

Σωστό και αυτό. Υπενθυμίζω μόνο ότι άλλο η εκπαίδευση στο πως να λύνεις διαφορικές εξισώσεις και άλλο η μαθηματική παιδεία. Άλλο η εκπαίδευση στην επαρκή ανακεφαλαίωση των γεγονότων μιας ιστορικής περιόδου και άλλο η ιστορική παιδεία –αλλά σε αυτά θα επανέλθω αργότερα.

Εν τω μεταξύ, δεδομένου ότι σε πολλούς από τους παραπάνω τομείς, η τεχνητή νοημοσύνη θα ξεπεράσει σίγουρα τους ανθρώπους, η συστηματική χρήση της θα αυξήσει επίσης σίγουρα την παραγωγή αγαθών και υπηρεσιών και έχει τη δυνατότητα να αναβαθμίσει ριζικά την ποιότητα της ανθρώπινης ζωής. Ταυτόχρονα, είναι βέβαιο ότι θα οδηγήσει σε εκτόπιση εργαζομένων και αύξηση του ποσοστού ανεργίας.

Το πώς θα αντιμετωπίσει η ανθρωπότητα αυτά τα δύο φαινόμενα (τη βελτίωση της ποιότητας ζωής και την αύξηση της ανεργίας) είναι πολιτικό ζήτημα -και ζήτημα ανθρωπιστικής παιδείας.

Εν πολλοίς εξαρτάται από το εάν τα διάφορα κέντρα λήψης αποφάσεων στον κόσμο (είτε πολυεθνικές εταιρείες είτε κράτη) επιθυμούν να μειώσουν, να αυξήσουν ή να αφήσουν αμετάβλητες τις εισοδηματικές ανισότητες που υπάρχουν στον κόσμο –Αλλά δεν θα υπεισέλθω σε αυτό το θέμα, διότι η γνώμη μου για αυτό δεν έχει κατ' ανάγκην, περισσότερη βαρύτητα από τη γνώμη οποιουδήποτε το έχει μελετήσει.

Θα επαναλάβω, μόνον, ότι πρόκειται για πολιτικό, δηλαδή ανθρώπινο θέμα και όχι θέμα για το οποίο μπορούν να αποφασίσουν οι μηχανές, αν και οι μηχανές μπορεί να παρέχουν τα δεδομένα για πιο επαρκείς αποφάσεις.

Το άλλο αναμφισβήτητο γεγονός είναι ότι μπορεί να γίνει σκόπιμη κατάχρηση των κοσμογονικής εμβέλειας δυνατοτήτων της τεχνητής νοημοσύνης σε όλους τους τομείς που παρετέθησαν. Επίσης, όλες οι εφαρμογές της ΤΝ συνεπάγονται κάποια πιθανότητα σφάλματος.

Επί παραδείγματι, είναι προφανές πώς επιλεκτική έκθεση των συστημάτων σε κάποια δεδομένα αλλά όχι σε άλλα, μπορεί να δημιουργήσει παραπληροφόρηση. Δηλαδή να δημιουργήσει αυτό που

γίνεται και σήμερα σ' ανατολή και δύση αλλά με απείρως μεγαλύτερη επάρκεια.

Ομοίως προφανής είναι η δυνατότητα να γίνουν παραβιάσεις σε εκλογικά πρωτόκολλα, να παραπλανηθούν οι χρηματοπιστωτικές αγορές, και πάει λέγοντας. Είναι προφανές, εν ολίγοις, ότι η τεχνητή νοημοσύνη θα μπορούσε να δημιουργήσει χάος σε όλα τα ζητήματα πολέμου και ειρήνης.

Προφανές, επίσης, είναι το ενδεχόμενο λάθους εν αντιθέσει προς την εθελούσια παραπλάνηση. Η πιθανότητα, επί παραδείγματι, σύγκρουσης αυτόνομων οχημάτων. Αλλά εδώ είναι ομοίως προφανές ότι η διόρθωση αυτών των προβλημάτων δεν είναι θέμα που αποφασίζουν οι μηχανές αλλά οι άνθρωποι.

Εναπόκειται στους ανθρώπους να αποφασίσουν εάν θα χρησιμοποιήσουν αυτό-οδηγούμενα αυτοκίνητα γνωρίζοντας ότι ο κίνδυνος λάθους είναι υψηλότερος, χαμηλότερος ή ίσος με τον κίνδυνο ατυχήματος όταν οδηγεί ο ίδιος ο άνθρωπος.

Γενικά, εναπόκειται στους ανθρώπους απανταχού της οικουμένης να συνεργαστούν αναβαθμίζοντας και επεκτείνοντας τις εφαρμογές της τεχνητής νοημοσύνης όπως και τα μέσα ελέγχου της κατάχρησης και των σφαλμάτων με απόλυτη διαφάνεια.

Πρέπει να υπάρχει καθολική συμφωνία και διαφάνεια ως προς το τι και πώς θα αναβαθμίζονται τα συστήματα όπως και σε ποια δεδομένα θα εκτίθενται. Εν ολίγοις η TN παρουσιάζει τα ίδια, και ίσως πιο δυσεπίλυτα προβλήματα από εκείνα που παρουσιάζει η διαχείριση της ατομικής ενέργειας.

Τώρα, θα το κάνει ο κόσμος; Θα ενεργήσει συναινετικά, και με διαύγεια για να αποτραπούν οι κίνδυνοι που ελλοχεύουν σε κάθε εφαρμογή της; Αμφιβάλλω. Αμφιβάλλω αν οι περισσότερες κυβερνήσεις και εταιρείες

θα συνεργαστούν. Το πιθανότερο είναι ότι θα ανταγωνιστούν η μία την άλλη. Και, πιθανότατα, θα χρειαστεί να μεσολαβήσουν ατυχήματα και να υποστεί ο κόσμος τις επιπτώσεις λανθασμένων πρακτικών προτού σοβαρευτεί και αντιμετωπίσει τους κινδύνους με συνέπεια.

Ας ελπίσουμε μόνον ότι δεν θα χρειαστεί το αντίστοιχο μιας άλλης Ουκρανίας πριν αρχίσουμε να χειριζόμαστε αυτά τα θέματα με τον σωστό τρόπο.

Πολύ **λιγότερο** προφανές, απ' τ' άλλο μέρος είναι πώς να αντιμετωπιστούν σωστά οι φόβοι ή οι ανησυχίες ότι τα προηγμένα συστήματα θα αυτονομηθούν και θα γίνουν ενσυνείδητα και πιο καινοτόμα από τους δημιουργούς τους και, ενδεχομένως, και επικίνδυνα.

Ας δούμε λοιπόν πρώτα την σχετική γνωστική -- εντός εισαγωγικών γνωστική-- επάρκεια των προηγμένων συστημάτων. Έχω ήδη αναφερθεί σε έργα στα οποία οι μηχανές εύκολα θα υπερβούν την επάρκεια του μέσου ανθρώπου στο μέλλον – σε κάποια, μάλιστα, όπως στο σκάκι, την έχουν ήδη υπερβεί, αλλά ας πάρουμε πάλι, ως παράδειγμα την περίπτωση του ChatGPT.

Σήμερα, το σύστημα αυτό φαίνεται σαν να μπορεί να κατανοήσει ευθείς ερωτήσεις και να δώσει εύλογες απαντήσεις, σαν αυτές που σας διάβασα πριν από λίγα λεπτά. Έχει όμως τη δυνατότητα να χειριστεί ερωτήσεις που απαιτούν αναγνώριση της σημασίας των συμφραζομένων;;

Για να δούμε:

Εγώ: Ο Κωστάκης ήθελε να αγοράσει τσίχλες, αλλά έχασε το ευρώ που είχε στην τσέπη. Η Ελενίτσα ένωσε ένοχη βλέποντας τον Κωστάκη λυπημένο κι έτσι του αγόρασε την τσίχλα. Αλλά γιατί να νιώσει ένοχη;

ChatGPT: Δεν υπάρχουν πληροφορίες στο κείμενο που να υποδηλώνουν γιατί η Ελενίτσα ένωθε ένοχη. Το κείμενο αναφέρει μόνο ότι η Ελενίτσα αγόρασε τσίχλα για τον Κωστάκη αφού συνειδητοποίησε ότι είχε χάσει το ευρώ του. Είναι απλά πιθανό η Ελενίτσα να ήθελε απλώς να βοηθήσει τον Κωστάκη ή να είναι ευγενική μαζί του.

Τον περασμένο μήνα, ρώτησα το ChatGPT τα εξής: «Ο Κωστάκης ήθελε να αγοράσει τσίχλες, αλλά του έπεσε το ευρώ που είχε στην τσέπη και το έχασε. Η Ελενίτσα ένωσε ένοχη βλέποντας τον Κωστάκη λυπημένο κι έτσι του αγόρασε την τσίχλα. Αλλά γιατί να νιώσει ένοχη;»

Στο ερώτημα αυτό, το ChatGPT απάντησε αμέσως με καλή συντακτική μορφή: «Δεν υπάρχουν πληροφορίες στο κείμενο που να υποδηλώνουν γιατί η Ελενίτσα ένωθε ένοχη. Το κείμενο αναφέρει μόνο ότι η Ελενίτσα αγόρασε τσίχλα για τον Κωστάκη αφού συνειδητοποίησε ότι είχε χάσει το ευρώ του. Είναι απλά πιθανό η Ελενίτσα να ήθελε απλώς να βοηθήσει τον Κωστάκη ή να είναι ευγενική μαζί του.»

Αλλά όπως κάθε πιτσιρικάς μπορεί να διακρίνει, υπήρχαν σαφείς αν και έμμεσες ενδείξεις, στην ερώτησή μου, που εξηγούν γιατί η Ελενίτσα ένωθε ένοχη. Μόνο που το ChatGPT δεν έχει τη φαντασία να «μπει στη θέση της Ελενίτσας». Συμπεριφέρεται, αυτές τις μέρες, τουλάχιστον, σαν ένα αυτιστικό παιδί με υψηλή λειτουργικότητα. Δεν έχει όπως λένε οι νεοέλληνες **ενσυναίσθηση** --μια θλιβερά άτυχη λέξη που φιλοδοξεί

να εκφράσει την απουσία γνώσης του ψυχισμού του άλλου, δηλαδή την **ετερογνωσία**, το αντίστοιχο της αυτογνωσίας.

Είμαι σίγουρος, όμως ότι, καθώς οι επιδόσεις του ChatGPT βελτιώνονται συνεχώς, μια μέρα θα αποκριθεί στην ερώτηση όχι σαν ένα παιδάκι με σύνδρομο Ασπεργκερ αλλά σαν ένας ενήλικας με ετερογνωσία (και όχι ενσυναίσθηση) που εύκολα μπαίνει στη θέση του άλλου. Και είμαι επίσης βέβαιος ότι σε τομείς δεδομένων στους οποίους έχει εκτεθεί θα απαντήσει με ακρίβεια ακόμη και σε ερωτήσεις στις οποίες ο μέσος φυσιολογικός ενήλικας δεν μπορεί να απαντήσει.

Δημιουργικότητα , καινοτομία, εφευρετικότητα

Επιπλέον, δεν θα διστάσω να υποστηρίξω το ίδιο για τις δυνατότητες της τεχνητής νοημοσύνης να δημιουργήσει πραγματικά πρωτόγνωρα, καινοτόμα έργα: πίνακες, φερ' ειπείν στο στυλ του Γκογκέν ή σε στυλ που θυμίζει και τον Νταλί· να συνθέτει ποιήματα στο στυλ του Πούσκιν· μουσική που μπορεί να ακούγεται περίεργη στον μέσο ακροατή, αλλά που θα καταπλήξει τον έμπειρο μουσικοσυνθέτη. Θα μπορέσει ακόμη και να παράσχει λύσεις σε δυσεπίλυτα προβλήματα λογικής ή μαθηματικών όπως, ας πούμε, στο τελευταίο θεώρημα του Fermat.

Και δεν θα διστάσω να ισχυριστώ ότι μπορεί ακόμη και να παραγάγει νέα θεωρήματα που πάλι θα εκπλήσσουν τους μεγάλους μαθηματικούς, ενώ μπορεί να ακούγονται σαν σκέτη ανοησία στον μέσο φοιτητή του μαθηματικού του ΕΚΠΑ ή της Πάτρας.

Αλλά πόθεν οι κάθετοι αυτοί ισχυρισμοί;

Η απάντηση είναι απλή: Ο καθένας από αυτούς τους τομείς της επιστήμης και της τέχνης έχει δημιουργήσει πάμπολλα υποδείγματα ανά τους αιώνες, τα περισσότερα από τα οποία είναι ψηφιακώς κωδικουμένα και διαθέσιμα στην τεχνητή νοημοσύνη. Αυτά τα υποδείγματα, ενσωματωμένα στα συστήματα γενικής ΤΝ δρουν ως πρότυπα για την δημιουργία παρόμοιων και ενδεχομένως αριότερων έργων εκ μέρους των συστημάτων.

Είναι, λοιπόν, δημιουργική η ΤΝ- υπό αυτή την έννοια; --βεβαίως είναι.

Θα ξεπεράσει, όμως, την ανθρώπινη δημιουργικότητα;

Θα ξεπεράσει η ΤΝ την ανθρώπινη εφευρετικότητα;

-Εδώ τα πράγματα μπλέκουν λίγο -- δεν είναι τόσο ξεκάθαρα και χρειάζεται λίγη προσοχή για να μην μπερδέψουμε το νόημα των λεγομένων.

Εν πρώτοις, το κριτήριο του τι είναι καλή τέχνη το θέτει ο καλλιτέχνης και ο δόκιμος εραστής της τέχνης. Το κριτήριο για το τι είναι πραγματικά επιστημονική ή μαθηματική καινοτομία είναι πάλι θέμα για το οποίο αποφασίζουν, πρωτίστως, οι ειδικοί επιστήμονες και οι έγκριτοι μαθηματικοί.

Όταν η τεχνητή νοημοσύνη δημιουργήσει έναν πίνακα ή μια σονάτα, εναπόκειται στον ειδικό να την χαρακτηρίσει ως καλλιτέχνημα ή ως ακαλαίσθητο θόρυβο. Αν η σονάτα προξενήσει την γνώριμή του, λεπτή εκείνη συγκίνηση που προξενεί η υψηλή τέχνη, και η οποία είναι μια αποκλειστικώς ανθρώπινη ευαισθησία, τότε το έργο θα θεωρηθεί καλλιτέχνημα. Εάν όχι, θα απορριφθεί ως ανοησία.

Το ίδιο ισχύει για νέες επιστημονικές θεωρίες και μαθηματικά θεωρήματα.

Μπορεί, λοιπόν, η τεχνητή νοημοσύνη να ξεπεράσει τον άνθρωπο όσον αφορά τη δημιουργικότητα, ευρηματικότητα, ριζική καινοτομία;

Το ερώτημα αυτό δεν έχει νόημα για τους εξής λόγους.

Εν πρώτοις, όπως μόλις είπα, η αξία οποιουδήποτε έργου είτε εκείνο είναι δημιούργημα ανθρώπου είτε είναι δημιούργημα μηχανής θα επικυρωθεί, κατ' ανάγκη από τον άνθρωπο τουλάχιστον μέχρις ότου οι μηχανές αποκτήσουν αυτογνωσία, εάν την αποκτήσουν -που είναι και το επόμενο θέμα μας.

Δεύτερον, δεδομένου ότι εφιαλτήριο νέων έργων είναι εξ αρχής ανθρώπινα υποδείγματα δεν έχει νόημα να συγκρίνουμε την ανθρώπινη και την τεχνητή επινοητικότητα και καινοτομία.

Αλλά, κάποιος μπορεί να αντιτάξει, τι γίνεται από την πλευρά του μηχανήματος; Από την οπτική γωνία του συστήματος;

Θα μπορούσαν, με άλλα λόγια, οι μηχανές να δημιουργούν τις δικές τους αξίες, είτε εκείνες αφορούν λογικό-μαθηματικά, είτε αισθητικά είτε επιστημονικά κατασκευάσματα, με βάση τις οποίες αξίες να κρίνουν τις δικές τους δημιουργίες παρακάμπτοντας έτσι εντελώς τη γνώμη ημών, των δημιουργών τους;

Αλλά αυτό είναι, με άλλη μορφή, το προηγούμενο ερώτημα για το εάν τα συστήματα θα μπορέσουν ποτέ να παραγάγουν αυτογνωσία. Να παράγουν όχι μόνο σχηματισμούς σημάτων που οδηγούν σε εμφανείς επιδόσεις, σε πράξεις, σε συμπεριφορά, αλλά το είδος των σχηματισμών που συνεπάγονται επίσης συνειδητή γνώση ή συνειδητές εμπειρίες πόνων, μελωδιών, συναισθημάτων, ορμών, κινήτρων, αξιών και αισθημάτων.

Το ερώτημα είναι αυτό που έθεσα στην αρχή και το απάντησα αρνητικά. Εάν δηλαδή η ΤΝ είναι νοημοσύνη με την έννοια ότι χαρακτηρίζεται από δύο φαινόμενα: την πράξη και την γνώση. Το απάντησα δε αρνητικά και ρηξικέλευθα όσον αφορά το παρόν.

Τώρα όμως ερωτώ: Θα μπορούσαμε να διακρίνουμε έστω και την παραμικρή ένδειξη ότι η δομή των συστημάτων τεχνητής νοημοσύνης και ο τρόπος με τον οποίο αποκτούν πληροφορίες προοιωνίζουν την **δυνατότητα** απόκτησης γνώσης κι αυτογνωσίας;

Η εξέλιξη συστημάτων ΤΝ σε ενσυνείδητα όντα

Εδώ η απάντηση είναι πάλι ένα κατηγορηματικό ΟΧΙ.

Κάποιοι από σας πιθανό να αναρωτηθούν: μα γιατί αυτή η απόλυτη στάση; Μην τυχόν ο ομιλών πιστεύει σε υπερφυσικά πνεύματα και ασώματες οντότητες;

Αλλά, τι πιστεύει ο ομιλών, είναι εντελώς άσχετο στην προκειμένη περίπτωση. Η άρνησή μου ερείδεται αποκλειστικώς στην κοινή, και πεζή εμπειρική πραγματικότητα.

Και η πραγματικότητα είναι ότι δεν υπάρχει τρόπος να προγραμματίσουμε πόνους και απολαύσεις, συναισθήματα και αξίες και, το κυριότερο, δεν μπορούμε να εκθέσουμε τους αλγόριθμους σε παραδείγματα πόνων και απολαύσεων και αξιών. Μπορούμε μόνο να τους εκθέσουμε σε **περιγραφές** πόνων και απολαύσεων, συναισθημάτων και αξιών.

Και μεταξύ μιας αισθήσεως, ας πούμε, της **μυρωδιάς** των σταφυλιών και της **περιγραφής της μυρωδιάς** των σταφυλιών χάσμα μέγα εστήρικται –καθώς λέει ο ευαγγελιστής. Η περιγραφή, η κάθε

περιγραφή, είτε γίνεται με λόγια είτε γίνεται με άλλα σύμβολα απλούστατα δεν αποπνέει κανένα άρωμα, οπότε πώς θα μπορούσαν οι αλγόριθμοι να μάθουν πράγματα όπως το άρωμα των σταφυλιών τον Σεπτέμβρη, ή οποιαδήποτε άλλη εμπειρία ερχόμενοι σε επαφή με **περιγραφές** εμπειριών; Δεν μπορούν. Είναι τόσο απλό.

Να το ξαναπώ αλλιώς: Η αξία που έχει για τον καθένα από εσάς ένα τραγούδι ή ένα ποίημα ή μια γεωμετρική απόδειξη –αυτή η λεπτή συναίσθηση της πεποίθησης ότι η απόδειξη ισχύει, και το αίσθημα της ευχαρίστησης ή ικανοποίησης για το ποιόν του τραγουδιού– είναι ένα πράγμα ενώ ακόμα και η πιο πιστή και πιο λεπτομερής **περιγραφή** αυτής της εμπειρίας είναι ένα **φύσει αλλότριο** πράγμα. Και, ευτυχώς, ή δυστυχώς, μόνο σε περιγραφές εμπειριών, μπορεί κανείς να εκθέσει τους αλγόριθμους και τα τεχνητά δίκτυα και όχι στις ίδιες τις εμπειρίες στις ίδιες τις άμεσες γνώσεις.

Δεν μπορώ να κάνω πιο ξεκάθαρη την εξήγηση του γιατί οι μηχανές δεν μπορούν να αποκτήσουν συνείδηση. Και σπεύδω να προλάβω την συνήθη σε αυτές τις περιπτώσεις ένσταση των αντιρροούντων η οποία εκφράζεται ως εξής: «ναι αλλά κανείς δεν ξέρει πώς θα εξελιχθεί η τεχνολογία αύριο. Μπορεί κάλλιστα να μπορούν οι μηχανές να αισθανθούν και να βιώνουν τις ανθρώπινες εμπειρίες». Στην οποία ένσταση απαντώ: «Σωστά, κανείς δεν ξέρει πώς θα εξελιχθεί η τεχνολογία αύριο». Αυτό, όμως, σημαίνει δύο τινά κι όχι μόνο την επικύρωση των ευσεβών πόθων του αντιρρησίας.

Σημαίνει, πρώτον, ότι η τεχνολογία ίσως καταφέρει να κάνει τις μηχανές ευθέως ενήμερες των εμπειριών μου -- πράγμα που ο δίπλα μου, ο συνάνθρωπός μου, είναι προφανώς αδύνατο να καταφέρει – Το γεγονός ότι κανείς δεν μπορεί να βιώσει το ποιόν των εμπειριών του άλλου είναι προφανές τοις πάσι κι όμως ο αντιρρησίας προσπαθεί να με πείσει ότι, κάπως, οι μηχανές θα μπορέσουν. Ελπίζω βλέπετε πόσο τραβηγμένη απ τα μαλλιά είναι αυτή η προσδοκία, αλλά έστω.

Έστω, διότι είναι ισοπίθανο ή μάλλον πολύ πιθανότερο η τεχνολογία του μέλλοντος να κάνει το αντίθετο: να επανεπικυρώσει, δηλαδή, τελεσίδικα το προφανές: ότι τέτοιο πράγμα είναι φύσει αδύνατον. Οπότε ας παρακάμψουμε τα αφελή και τα εύκολα, κι ας βάλουμε στο στόχαστρο κάτι σοβαρό και δύσκολο. Τί ακριβώς;

Να, το ότι η συνείδηση μπορεί να μην είναι καθόλου απαραίτητη προϋπόθεση για τα συστήματα τεχνητής νοημοσύνης να διακηρύξουν την αυτονομία τους, ακόμη και να κηρύξουν πόλεμο κατά της ανθρωπότητας. Κι ότι το μόνο που μπορεί να χρειάζονται είναι το **λειτουργικό αντίστοιχο** των κινήτρων. Τί εννοώ:

Μη συνειδητά κίνητρα:

αυτοματισμοί μιμούμενοι
κίνητρα και πρωτοβουλίες

Την κατσαρίδα δεν την θεωρούμε ενσυνείδητη-- σωστά; Δεν πιστεύουμε ότι η κατσαρίδα έχει γνώση αξιών και προτιμήσεων ούτε γνώση συναισθημάτων. Αλλά παρόλα αυτά, η κατσαρίδα εμφανίζει το λειτουργικό αντίστοιχο κινήτρων και πρωτοβουλιών. Φεύγει όταν αισθάνεται πολύ φως, κάτι που εμείς ερμηνεύουμε ως επίγνωση κινδύνου. Και όταν οι θερμίδες της εξαντληθούν (κάτι που εμείς ερμηνεύουμε ως αίσθημα πείνας) τρέχει αναζητώντας τροφή.

Πώς καταφέρνει λοιπόν η κατσαρίδα να συμπεριφέρεται σαν να είχε κίνητρα και αξίες; Πιθανότατα αποκτά το λειτουργικό ισοδύναμο των αξιών και των κινήτρων μέσω φυσιολογικών μηχανισμών που μερικοί από τους προγόνους της ανέπτυξαν προκειμένου να διατηρήσουν το φυσικό τους σώμα – να επιζήσουν.

Προφανώς υπάρχουν νευροφυσιολογικά σήματα μέσα στην κατσαρίδα που όταν φτάσουν σε μια συγκεκριμένη τιμή αναγκάζουν την κατσαρίδα να τρέξει για κάλυψη ή για αναζήτηση τροφής.

Πώς ξεκίνησαν αυτοί οι μηχανισμοί;

Κανείς δεν ξέρει πραγματικά, αλλά πολλοί εξ ημών ισχυριζόμαστε ότι μάλλον κάποια στιγμή στο μακρινό παρελθόν μια σειρά τυχαίων γονιδιακών μεταλλάξεων είχε ως αποτέλεσμα τη λειτουργική σύνδεση μεταξύ των σημάτων από τους εσωτερικούς αισθητήρες μέσα στο ζώυφιο και των κινητικών μηχανισμών του.

Παρόμοιες τυχερές μεταλλάξεις, λέμε, πρέπει να είχαν ως αποτέλεσμα τη σύνδεση μεταξύ άλλων μερών του σώματος της κατσαρίδας που σχετίζονται με τον έλεγχο της θερμοκρασίας, τον έλεγχο της έκθεσης στο φως και τον έλεγχο των παντοειδών φυσιολογικών μεταβλητών και των κινητικών μηχανισμών της.

Κατά συνέπεια, σύμφωνα με το οικείο αυτό αφήγημα, οι αλγόριθμοι που υπάρχουν σε κάποιον υπολογιστή θα μπορούσαν να αναπτύξουν ανάλογους μηχανισμούς για να διατηρεί ο υπολογιστής την ενέργειά του και την αυτονομία του.

Αλλά το μόνο ανάλογο των τυχαίων γονιδιακών μεταλλάξεων στην περίπτωση των τεχνητών συστημάτων είναι τυχαία σφάλματα. –για τον απλούστατο λόγο ότι οι πραγματικές γονιδιακές μεταλλάξεις είναι επίσης σφάλματα.

Κατά συνέπεια, ίσως χρειαστεί να περιμένουμε μερικά εκατομμύρια χρόνια για να συμβούν τα απαραίτητα τυχερά λάθη κι έτσι να μάθουν οι μηχανές πως να διατηρούν το σώμα τους – πώς να επιβιώνουν.

Υπάρχει λοιπόν, σύμφωνα με αυτό το αφήγημα, η πιθανότητα ότι μια σειρά σφαλμάτων στην εφαρμογή των κωδίκων να μετατρέψει τα τεχνητά συστήματα σε αυτόνομα εχθρικά όντα ακριβώς όπως τυχαίες σειρές μεταλλάξεων μέσα στους αιώνες μετέτρεψαν κάποια ζώφια σε αυτόνομες κατσαρίδες και επιθετικές σφήκες.

Αλλά, βέβαια, ο πιο ευθύς τρόπος για αρχίσουν να συμπεριφέρονται οι μηχανές σαν κατσαρίδες και σαν σφήκες χωρίς να περιμένουμε για τα τυχαία και τυχερά σφάλματα ατέρμονους αιώνες, είναι να ενεργήσουμε εμείς, οι δημιουργοί τους, όπως φημολογείται ότι ενεργεί ο Μεγαλοδύναμος στα πλαίσια της θεωρίας του Ευφυούς Σχεδιασμού. Να εκθέσουμε, δηλαδή τις μηχανές στις κατάλληλες πληροφορίες καθιστώντας τις μ' αυτόν τον τρόπο αυτόνομες στο άψε σβήσε.

Κι αυτός είναι ο δεύτερος τρόπος που οι άβουλες μηχανές θα μπορούσαν να αποκτήσουν τα χαρακτηριστικά της κατσαρίδας ή της σφήκας. Οπότε διαλέγετε και παίρνετε: Στην μία περίπτωση ο κίνδυνος αυτός είναι χιλιετίες μακριά. Στην άλλη περίπτωση είναι άμεσος. Αλλά σε αυτή, την δεύτερη περίπτωση, ο κίνδυνος δεν ελλοχεύει στα κυκλώματα των μηχανών αλλά στην κακοήθεια ή στην απροσεξία των ανθρώπων.

Μη φοβού λοιπόν την ΤΝ αλλά τον άνθρωπο που δίκην καραγκιοζοπαίχτη κινεί τα νήματα της μοίρας των μηχανών πίσω από τον μπερντέ της τεχνογνωσίας του.

.....

Αλλά πέρα από την πιθανότητα οι μηχανές να εξελιχθούν στο επίπεδο των κατσαρίδων ή σε αυτό της επιθετικής σφήκας, μπορούν να αποκτήσουν συνείδηση; Να συνειδητοποιούν τουλάχιστον πράγματα που φαίνεται να γνωρίζουν;

Σε αυτό, έχω ήδη απαντήσει **όχι** και έχω παραθέσει τους λόγους μου για την αδιάλλακτα αρνητική απάντησή μου. Αλλά στην επιχειρηματολογία μου δεν έλαβα υπόψη μου το δόγμα της ανάδυσης, ένα δόγμα που οι περισσότεροι από εσάς γνωρίζετε τουλάχιστον με το το επίσημο όνομά του : the doctrine of emergence

Το δόγμα της «Αναδύσεως» emergence

Σύμφωνα με αυτό το δόγμα εάν ο σχηματισμός σημάτων που παράγονται από τα δίκτυα των πραγματικών εγκεφάλων φτάσει σε έναν ορισμένο βαθμό πολυπλοκότητας, τότε απλώς αναδύεται η αντίστοιχη συνειδητή εμπειρία. Έτσι απλά και αυθόρμητα. Κι αν νομίζετε ότι τα παραλέω ξαναδιαβάστε

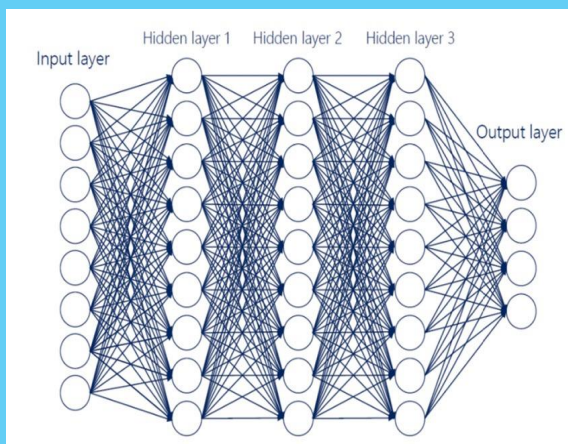
Sperry, R. W. (1969). A modified concept of consciousness. *Psychological Review*, 76(6), 532–536. <https://doi.org/10.1037/h0028156>

τον Sperry που ήταν από τους πρώτους ένθερμους θιασώτες του εν λόγω δόγματος και λέει κι αυτός ακριβώς τα ίδια.

Κοιτάτε μια φωτογραφία και τα δίκτυα του εγκεφάλου που επεξεργάζονται το ερέθισμα παράγουν έναν σχηματισμό σημάτων.

Τώρα, σύμφωνα με το δόγμα, εάν ο σχηματισμός αυτών των σημάτων είναι αρκετά πολύπλοκος, όχι μόνο μπορείτε να αποκριθείτε με τις προσήκουσες αντικειμενικές πράξεις, είτε λεκτικές είτε μη λεκτικές σε αυτήν την εικόνα, αλλά έχετε επίσης την ξεχωριστή εμπειρία· έχετε άμεση γνώση του πορτραίτου της φίλης σας της Κλημεντίνης ή μιαν όψη του πύργου του Άιφελ ή οτιδήποτε άλλο περιέχει η φωτογραφία που κοιτάτε.

Schematic of a neural network capable of “deep learning”



Ομοίως, αυτό το δόγμα διδάσκει, ότι ένας επαρκώς πολύπλοκος σχηματισμός σημάτων που εξέρχονται από τους κόμβους του τελευταίου δεξιού στρώματος του συστήματος, μπορεί επίσης να προκαλέσει αυτόματα την αντίστοιχη συνειδητή εμπειρία—στην μηχανή. Να κάνει, δηλαδή, την μηχανή ενσυνείδητη.

Αλλά ποιες είναι οι εμπειρικές ενδείξεις υπέρ του δόγματος της ανάδυσης τόσο στην περίπτωση του ανθρώπινου εγκεφάλου όσο και στην περίπτωση των μηχανών;

Στην περίπτωση του ανθρώπινου εγκεφάλου δεν υπάρχει καμία απολύτως ένδειξη ότι η εμπειρία αναδύεται από τα σήματα. Διαπιστώνουμε μόνο την παρουσία δύο σύγχρονων φαινομένων: ενός σχηματισμού νευρωνικής ενεργοποίησης και μιας συγκεκριμένης συνειδητής εμπειρίας. Καμιά εμπειρική ένδειξη για το ποιο φαινόμενο αναδύεται από ποιο. Καμιά, πέραν της υποκειμενικής και μεταφυσικής μας προκαταλήψεως υπέρ του δόγματος της Αναδύσεως.

Στην περίπτωση των μηχανών, γνωρίζουμε μόνο το ένα φαινόμενο, τον σχηματισμό των σημάτων και πρέπει να συμπληρώσουμε το δεύτερο φαινόμενο, την επίγνωση, αυθαιρέτως με μια άμεση κατάφαση που εκφράζει, πάλι, την μεταφυσική μας πίστη στο ίδιο δόγμα.

Αλλά, πόσο βαθιά και στέρεα είναι αυτή η πίστη;

Το πόσο βαθιά είναι αυτή η πίστη εξαρτάται από ολόκληρη την προσωπικότητα του καθενός μας και ολόκληρη την ιστορία της διαμόρφωσής της.

Έχω όμως σκαρώσει μια πολύ απλή και αδιάβλητη δοκιμασία με την οποία μπορεί ο καθένας σας εδώ και τώρα να αξιολογήσει το βάθος αυτής της πίστης - τη ισχύ της πεποίθησής του ότι οι μηχανές μπορεί να γίνουν ενσυνείδητα όντα μέσω της ανάδυσης ή οποιουδήποτε άλλου μέσου.

Είσαστε έτοιμοι για το τέστ;



Βγάζετε, λοιπόν, βόλτα τον σκύλο σας (ή αν δεν έχετε σκυλί, τη γάτα σας) και συναντάτε αυτές τις δύο κοπέλες. Αρχίζετε να μιλάτε μαζί τους. Η Μαρία στα αριστερά η Ελένη στα δεξιά. Η Μαρία είναι άνθρωπος και η Ελένη είναι ρομπότ παρόλο ότι φαίνεται να είναι η δίδυμη αδελφή της Μαρίας. Αλλά τόσο τέλειο ρομπότ που αν δεν κοιτάξετε με προσοχή κάτω από το δέρμα της που έχει την τελειότητα του ανθρωπίνου δέρματος να διακρίνετε τα κυκλώματά της, δεν θα σας πέρανε ποτέ από το μυαλό ότι είναι μηχανή.

Και καθώς ανταλλάσσετε χαριτολογήματα με τις δυο τους, να 'σου ο Darth Vader να κατεβαίνει από το διαστημόπλοίο του κρατώντας μια συσκευή, και να δηλώνει:

"Αυτό εδώ είναι το καλύτερο όργανο βασανιστηρίων στον γαλαξία - «Τίποτα άλλο δεν μπορεί να προκαλέσει πιο αφόρητο πόνο από αυτό εδώ το εργαλείο».

Και συνεχίζει: «Λοιπόν, φίλε μου, έχεις την εξής επιλογή. Πες μου ποιον να βασανίσω; την Μαρία, την Ελένη ή το σκύλο σου; Αλλά θα διαλέξεις εδώ και τώρα, αλλιώς πεθαίνεις».

Σκεφτείτε το λίγο: ποιόν θα του λέγατε να βασανίσει; την Μαρία τον άνθρωπο, την Ελένη το τέλειο ρομπότ ή το σκυλάκι;

Αν δεν είσαστε λοιπόν παθολογικά αυτοκτονικοί ή σαδιστές, η ειλικρινής σας απάντηση στην πρόκληση του Vader, θα σας ενημερώσει, πέραν κάθε αμφιβολίας, εάν εσείς, ο καθένας σας, στην πραγματικότητα, πιστεύετε ότι οι μηχανές θα μπορούσαν ποτέ να αποκτήσουν συνείδηση.

Και θα μάθετε τι στ' αλήθεια πιστεύετε ασχέτως τι μπορεί να λέτε προς τα έξω φοβούμενοι μην τυχόν θιχθεί η επιστημονική σας υπόληψη— έτσι όπως, νοείται από πολλούς, τις μέρες μας, η επιστημοσύνη.