

ΑΝΕΞΑΡΤΗΤΗ

ΟΜΑΔΑ ΕΜΠΕΙΡΟΓΝΩΜΟΝΩΝ ΥΨΗΛΟΥ ΕΠΙΠΕΔΟΥ

ΓΙΑ ΤΗΝ ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ

ΣΥΣΤΑΘΕΙΣΑ ΑΠΟ ΤΗΝ ΕΥΡΩΠΑΪΚΗ ΕΠΙΤΡΟΠΗ ΤΟΝ ΙΟΥΝΙΟ ΤΟΥ 2018



ΟΡΙΣΜΟΣ ΤΗΣ ΤΕΧΝΗΤΗΣ ΝΟΗΜΟΣΥΝΗΣ:
ΚΥΡΙΕΣ ΔΥΝΑΤΟΤΗΤΕΣ ΚΑΙ ΕΠΙΣΤΗΜΟΝΙΚΑ
ΠΕΔΙΑ

Ορισμός της τεχνητής νοημοσύνης: κύριες δυνατότητες και επιστημονικά πεδία

Ομάδα εμπειρογνομόνων υψηλού επιπέδου για την τεχνητή νοημοσύνη

Αποποίηση ευθύνης και χρήση του παρόντος εγγράφου: Η περιγραφή και ο ορισμός των δυνατοτήτων και των ερευνητικών πεδίων της τεχνητής νοημοσύνης (TN) που επιχειρείται στη συνέχεια αποτελεί μια εξαιρετικά αδρή υπεραπλούστευση των εξελίξεων της τεχνολογίας. Σκοπός του παρόντος εγγράφου δεν είναι ο επακριβής και ολοκληρωμένος ορισμός όλων των τεχνικών και των δυνατοτήτων της TN, αλλά η συνοπτική περιγραφή της κοινής θεώρησης του εν λόγω επιστημονικού πεδίου, στην οποία η ομάδα εμπειρογνομόνων υψηλού επιπέδου βασίζει τα παραδοτέα της. Ωστόσο, ελπίζουμε ότι το παρόν έγγραφο μπορεί να αποτελέσει επίσης ένα χρήσιμο εκπαιδευτικό εφαλτήριο για τους μη ειδικούς στον τομέα της TN, ώστε στη συνέχεια να εντρυφήσουν πιο επισταμένα στην TN και να αποκτήσουν ακριβέστερες γνώσεις για το επιστημονικό πεδίο και την τεχνολογία της TN.

Η ομάδα εμπειρογνομόνων υψηλού επιπέδου για την τεχνητή νοημοσύνη (OEYE για την TN) είναι μια ανεξάρτητη ομάδα εμπειρογνομόνων που συστάθηκε από την Ευρωπαϊκή Επιτροπή τον Ιούνιο του 2018.

Επικοινωνία Nathalie Smuha - Συντονίστρια της OEYE για την TN
E-mail CNECT-HLG-AI@ec.europa.eu

European Commission
B-1049 Brussels

Το έγγραφο δημοσιεύθηκε στις 18 Απριλίου 2019.

Το πρώτο σχέδιο του παρόντος εγγράφου δημοσιεύτηκε στις 18 Δεκεμβρίου 2018, μαζί με το πρώτο σχέδιο κατευθυντήριων γραμμών δεοντολογίας για αξιόπιστη τεχνητή νοημοσύνη της OEYE για την TN, και αναθεωρήθηκε βάσει των παρατηρήσεων που ελήφθησαν μέσω της Ευρωπαϊκής Συμμαχίας για την TN και της ανοικτής διαβούλευσης για το σχέδιο των κατευθυντήριων γραμμών. Θα θέλαμε να ευχαριστήσουμε ιδιαίτερα και θερμώς όσους διατύπωσαν τις απόψεις τους επί του πρώτου σχεδίου του εγγράφου.

Ούτε η Ευρωπαϊκή Επιτροπή ούτε οποιοδήποτε άλλο πρόσωπο ενεργεί εξ ονόματός της φέρει ευθύνη για την ενδεχόμενη χρήση των κάτωθι πληροφοριών. Για τα περιεχόμενα του παρόντος εγγράφου εργασίας αποκλειστικά υπεύθυνη είναι η ομάδα εμπειρογνομόνων υψηλού επιπέδου για την τεχνητή νοημοσύνη (OEYE για την TN). Παρόλο που μέλη του προσωπικού της Επιτροπής διευκόλυναν την εκπόνηση των κατευθυντήριων γραμμών, οι απόψεις που διατυπώνονται στο παρόν έγγραφο

εκφράζουν την άποψη της ΟΕΥΕ για την ΤΝ και δεν μπορούν σε καμία περίπτωση να εκληφθούν ως επίσημη θέση της Ευρωπαϊκής Επιτροπής.

Περισσότερες πληροφορίες σχετικά με την ομάδα εμπειρογνομόνων υψηλού επιπέδου για την τεχνητή νοημοσύνη διατίθενται στο διαδίκτυο (<https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence>).

Η πολιτική για την περαιτέρω χρήση εγγράφων της Ευρωπαϊκής Επιτροπής καθορίζεται στην απόφαση 2011/833/ΕΕ (ΕΕ L 330 της 14.12.2011, σ. 39). Για κάθε χρήση ή αναπαραγωγή φωτογραφιών ή άλλου υλικού που δεν υπόκειται στους κανόνες της ΕΕ για τα δικαιώματα πνευματικής ιδιοκτησίας πρέπει να ζητείται άδεια απευθείας από τους κατόχους δικαιωμάτων πνευματικής ιδιοκτησίας.

ΟΡΙΣΜΟΣ ΤΗΣ ΤΕΧΝΗΤΗΣ ΝΟΗΜΟΣΥΝΗΣ:

ΚΥΡΙΕΣ ΔΥΝΑΤΟΤΗΤΕΣ ΚΑΙ ΕΠΙΣΤΗΜΟΝΙΚΑ ΠΕΔΙΑ

Ας ξεκινήσουμε με τον ακόλουθο ορισμό της τεχνητής νοημοσύνης (TN), όπως προτείνεται στην ανακοίνωση της Ευρωπαϊκής Επιτροπής για την TN¹:

«Η τεχνητή νοημοσύνη (TN) αναφέρεται σε συστήματα που χαρακτηρίζονται από ευφυή συμπεριφορά, αναλύοντας το περιβάλλον τους και ενεργώντας —με κάποιο βαθμό αυτονομίας— για την επίτευξη συγκεκριμένων στόχων.

Τα συστήματα που λειτουργούν βάσει τεχνητής νοημοσύνης μπορούν να βασίζονται αποκλειστικά σε λογισμικό, ενεργώντας στον εικονικό κόσμο (π.χ. βοηθοί φωνής, λογισμικό ανάλυσης εικόνας, μηχανές αναζήτησης, συστήματα αναγνώρισης ομιλίας και προσώπου) ή η τεχνητή νοημοσύνη μπορεί να ενσωματωθεί σε συσκευές υλισμικού (π.χ. προηγμένα ρομπότ, αυτόνομα αυτοκίνητα, δρόνοι ή εφαρμογές του Διαδικτύου των Πραγμάτων).»

Στο παρόν έγγραφο αναπτύσσουμε αυτόν τον ορισμό ώστε να αποσαφηνιστούν ορισμένες πτυχές της TN ως επιστημονικού πεδίου και ως τεχνολογίας. Στόχος είναι να αποφευχθούν οι παρερμηνείες, να δημιουργηθεί κοινό πλαίσιο γνώσεων για την TN που θα είναι δυνατό να αξιοποιούνται και από μη ειδικούς στον τομέα της TN, και να παρασχεθούν χρήσιμα λεπτομερή στοιχεία που θα είναι δυνατό να χρησιμοποιούνται στο πλαίσιο του διαλόγου τόσο για τις κατευθυντήριες γραμμές δεοντολογίας όσο και για τις συστάσεις πολιτικής στον τομέα της TN.

1. Συστήματα TN

Ο όρος TN περιέχει ρητή αναφορά στην έννοια της νοημοσύνης. Δεδομένου, ωστόσο, ότι η νοημοσύνη (τόσο στις μηχανές όσο και στους ανθρώπους) είναι μια αόριστη έννοια, παρά το γεγονός ότι έχει μελετηθεί εκτενώς από ψυχολόγους, βιολόγους και νευροεπιστήμονες, οι ερευνητές της TN χρησιμοποιούν κατά κύριο λόγο την έννοια της ορθολογικότητας (rationality). Ως ορθολογικότητα νοείται η ικανότητα επιλογής της βέλτιστης ενέργειας για την επίτευξη ενός συγκεκριμένου στόχου, με βάση, αφενός, συγκεκριμένα κριτήρια προς βελτιστοποίηση και, αφετέρου, τους διαθέσιμους πόρους. Η ορθολογικότητα δεν είναι βέβαια η μοναδική παράμετρος της έννοιας της νοημοσύνης, αποτελεί όμως σημαντική συνιστώσα της.

Ακολουθώντας, με τον όρο *σύστημα TN* θα νοείται κάθε μεμονωμένο στοιχείο, λογισμικό και/ή υλισμικό που βασίζεται στην TN. Μάλιστα, τα συστήματα TN συνήθως δεν αποτελούν αυτόνομα συστήματα αλλά είναι *ενσωματωμένα* ως μεμονωμένα στοιχεία σε μεγαλύτερα συστήματα.

Επομένως, κάθε σύστημα TN είναι πρώτα και κύρια ορθολογικό, σύμφωνα και με ένα από τα πιο ευρέως χρησιμοποιούμενα συγγράμματα για την TN². Πώς όμως ένα σύστημα TN επιτυγχάνει την ορθολογικότητα; Όπως επισημαίνεται στην πρώτη περίοδο του παραπάνω ορισμού εργασίας για την TN, την επιτυγχάνει με το να αντιλαμβάνεται το περιβάλλον στο οποίο είναι ενσωματωμένο μέσω ορισμένων αισθητήρων, και, κατ' αυτόν τον τρόπο, με το να συλλέγει και να ερμηνεύει δεδομένα, να προβαίνει σε συλλογισμούς σχετικά με τα όσα αντιλαμβάνεται, ή να επεξεργάζεται τις πληροφορίες που εξάγονται από αυτά τα δεδομένα και να αποφασίζει ποια είναι η βέλτιστη ενέργεια και, εν συνεχεία, να ενεργεί αναλόγως, μέσω ορισμένων ενεργοποιητών, τροποποιώντας έτσι, ενδεχομένως, το περιβάλλον. Τα συστήματα TN μπορεί είτε να χρησιμοποιούν συμβολικούς κανόνες είτε να μαθαίνουν ένα αριθμητικό μοντέλο, και μπορεί επίσης να προσαρμόζουν τη συμπεριφορά τους με το να αναλύουν

¹ Ανακοίνωση της Επιτροπής προς το Ευρωπαϊκό Κοινοβούλιο, το Ευρωπαϊκό Συμβούλιο, το Συμβούλιο, την Ευρωπαϊκή Οικονομική και Κοινωνική Επιτροπή και την Επιτροπή των Περιφερειών - Τεχνητή νοημοσύνη για την Ευρώπη, Βρυξέλλες, 25.4.2018 COM(2018) 237 final.

² "Artificial Intelligence: A Modern Approach", S. Russell and P. Norvig, Prentice Hall, 3rd edition, 2009. «Artificial Intelligence: A Modern Approach», S. Russell και P. Norvig, Prentice Hall, 3η έκδοση 2009.

πώς επηρεάζεται το περιβάλλον από τις προηγούμενες ενέργειές τους. Για να γίνει καλύτερα κατανοητή, η λειτουργία των συστημάτων ΤΝ απεικονίζεται στο σχήμα 1.



Σχήμα 1: Σχηματική απεικόνιση ενός συστήματος ΤΝ.

Αισθητήρες και αντίληψη. Στο σχήμα 1, οι αισθητήρες του συστήματος απεικονίζονται ως σύμβολο ασύρματου δικτύου (Wi-Fi). Στην πράξη μπορεί να είναι κάμερες, μικρόφωνα, ένα πληκτρολόγιο, ένας ιστότοπος ή άλλες μονάδες εισόδου, καθώς και αισθητήρες φυσικών μεγεθών (π.χ. αισθητήρες θερμοκρασίας, πίεσης, απόστασης, δύναμης/ροπής, αφής). Γενικά, το σύστημα ΤΝ πρέπει να είναι εφοδιασμένο με αισθητήρες που μπορούν να αντιλαμβάνονται επαρκώς εκείνα τα δεδομένα που απαντώνται στο περιβάλλον και σχετίζονται με τον στόχο που έχει δοθεί στο σύστημα ΤΝ από τον άνθρωπο που το έχει σχεδιάσει. Για παράδειγμα, εάν θέλουμε να κατασκευάσουμε ένα σύστημα ΤΝ που καθαρίζει αυτόματα το πάτωμα ενός δωματίου όταν είναι βρώμικο, στους αισθητήρες μπορεί να περιλαμβάνονται κάμερες που καταγράφουν εικόνες του πατώματος.

Όσον αφορά τα δεδομένα που συλλέγονται, πολλές φορές είναι χρήσιμο να γίνεται διάκριση μεταξύ δομημένων και αδόμητων δεδομένων. *Δομημένα δεδομένα* είναι τα δεδομένα που έχουν οργανωθεί σύμφωνα με προκαθορισμένα μοντέλα (π.χ. αυτά που περιλαμβάνονται σε μια σχεσιακή βάση δεδομένων), ενώ τα *αδόμητα δεδομένα* δεν έχουν κάποια γνωστή οργάνωση (π.χ. μια εικόνα ή ένα κείμενο).

Συλλογιστική/επεξεργασία πληροφοριών και λήψη αποφάσεων. Στον πυρήνα ενός συστήματος ΤΝ βρίσκεται η λειτουργική μονάδα συλλογιστικής/επεξεργασίας πληροφοριών του, η οποία χρησιμοποιεί ως είσοδο τα δεδομένα που διαβιβάζουν οι αισθητήρες και προτείνει την ενέργεια που πρέπει να εκτελεστεί με βάση τον στόχο προς επίτευξη. Αυτό σημαίνει ότι τα δεδομένα που συλλέγουν οι αισθητήρες πρέπει να μετατραπούν σε πληροφορίες τις οποίες είναι σε θέση να κατανοήσει η λειτουργική μονάδα συλλογιστικής/επεξεργασίας πληροφοριών. Στο προαναφερθέν παράδειγμα του συστήματος ΤΝ που χρησιμοποιείται για τον καθαρισμό του πατώματος, η κάμερα μεταδίδει μια εικόνα του πατώματος στη λειτουργική μονάδα συλλογιστικής/επεξεργασίας πληροφοριών και η συγκεκριμένη μονάδα καλείται να αποφασίσει αν θα καθαρίσει ή όχι το πάτωμα (δηλ. ποια είναι η βέλτιστη

ενέργεια για την επίτευξη του επιθυμητού στόχου). Μπορεί για εμάς τους ανθρώπους η μετάβαση από μια εικόνα του πατώματος στην απόφαση αν χρειάζεται ή όχι καθαρισμό να φαίνεται εύκολη υπόθεση, για μια μηχανή όμως η διαδικασία αυτή δεν είναι εξίσου εύκολη, επειδή η εικόνα είναι απλώς μια αλληλουχία από τα ψηφία 0 και 1. Επομένως, η λειτουργική μονάδα συλλογιστικής/επεξεργασίας πληροφοριών καλείται να κάνει τα εξής:

1. Να ερμηνεύσει την εικόνα για να αποφασίσει αν το πάτωμα είναι καθαρό ή όχι. Σε γενικές γραμμές, αυτό σημαίνει να είναι σε θέση να μετατρέψει τα δεδομένα σε πληροφορίες και να μοντελοποιήσει αυτές τις πληροφορίες με σύντομο και περιεκτικό τρόπο, που θα πρέπει ωστόσο να περιλαμβάνει όλα τα σημαντικά επιμέρους δεδομένα (εν προκειμένω, αν το πάτωμα είναι καθαρό ή όχι).
2. Να προβεί σε συλλογισμούς με βάση αυτή τη γνώση ή να επεξεργαστεί τις πληροφορίες αυτές για να δημιουργήσει ένα αριθμητικό μοντέλο (δηλ. έναν μαθηματικό τύπο) προκειμένου να αποφασίσει ποια είναι η βέλτιστη ενέργεια. Στο συγκεκριμένο παράδειγμα, αν η πληροφορία που εξάγεται από την εικόνα είναι ότι το πάτωμα είναι βρώμικο, η βέλτιστη ενέργεια είναι να ενεργοποιηθεί ο καθαρισμός, διαφορετικά η βέλτιστη ενέργεια είναι η μηχανή να παραμείνει σε αδράνεια.

Επισημαίνεται ότι ο όρος «απόφαση» θα πρέπει να εκλαμβάνεται υπό ευρεία έννοια, ως δηλ. οποιαδήποτε πράξη που συνίσταται στην επιλογή της ενέργειας που θα εκτελεστεί, και δεν σημαίνει κατ' ανάγκη ότι τα συστήματα ΤΝ είναι πλήρως αυτόνομα. Απόφαση μπορεί να είναι επίσης η επιλογή μιας σύστασης που θα τεθεί υπόψη ενός ανθρώπου, ο οποίος θα είναι εκείνος που θα λάβει την τελική απόφαση.

Ενεργοποίηση. Αφού αποφασιστεί η ενέργεια, το σύστημα ΤΝ είναι έτοιμο να την εκτελέσει μέσω των ενεργοποιητών που έχει στη διάθεσή του. Στην παραπάνω εικόνα, οι ενεργοποιητές απεικονίζονται ως αρθρωτοί βραχίονες, στην πραγματικότητα όμως δεν είναι απαραίτητο να έχουν υλική υπόσταση, αλλά μπορεί να έχουν και τη μορφή λογισμικού. Στο παράδειγμα του καθαρισμού, το σύστημα ΤΝ μπορεί να εκπέμπει σήμα που θα ενεργοποιεί την ηλεκτρική σκούπα αν η ενέργεια συνίσταται στον καθαρισμό του πατώματος. Ένα άλλο παράδειγμα είναι τα συστήματα συνομιλίας (δηλ. τα διαλογορομπότ/chatbot), τα οποία, ανταποκρινόμενα σε φράσεις που διατυπώνουν οι χρήστες, ενεργούν με τη δημιουργία κειμένων.

Η εκτελούμενη ενέργεια είναι πιθανό να επιφέρει τροποποίηση του περιβάλλοντος, οπότε την επόμενη φορά το σύστημα θα χρειαστεί να χρησιμοποιήσει και πάλι τους αισθητήρες του προκειμένου να αντιληφθεί ενδεχομένως διαφορετικές πληροφορίες από το τροποποιημένο περιβάλλον.

Τα ορθολογικά συστήματα ΤΝ δεν επιλέγουν πάντοτε τη βέλτιστη ενέργεια για τον στόχο τους, οπότε επιτυγχάνουν μόνο *περιορισμένη ορθολογικότητα* (bounded rationality), λόγω περιορισμών σε επίπεδο πόρων όπως ο χρόνος ή η υπολογιστική ισχύς.

Τα *ορθολογικά συστήματα ΤΝ* είναι μια πολύ στοιχειώδης εκδοχή των συστημάτων ΤΝ. Τροποποιούν το περιβάλλον, αλλά δεν προσαρμόζουν τη συμπεριφορά τους με την πάροδο του χρόνου ώστε να πετυχαίνουν καλύτερα τον στόχο τους. Τα *ορθολογικά συστήματα με δυνατότητες μάθησης* είναι ορθολογικά συστήματα τα οποία, αφού ενεργήσουν, αξιολογούν τη νέα κατάσταση του περιβάλλοντος (μέσω της αντίληψης) για να κρίνουν πόσο επιτυχημένη ήταν η ενέργειά τους και εν συνεχεία προσαρμόζουν τους κανόνες συλλογιστικής και τις μεθόδους λήψης των αποφάσεών τους.

2. Η ΤΝ ως επιστημονικό πεδίο

Η παραπάνω περιγραφή είναι μια πολύ απλή και αφηρημένη περιγραφή ενός συστήματος ΤΝ βάσει τριών κύριων δυνατοτήτων: της αντίληψης, της συλλογιστικής/λήψης αποφάσεων και της ενεργοποίησης. Ωστόσο, αρκεί για να μας δώσει τη δυνατότητα να παρουσιάσουμε και να κατανοήσουμε την πλειονότητα των τεχνικών ΤΝ και των επιμέρους επιστημονικών πεδίων ΤΝ που χρησιμοποιούνται σήμερα για την κατασκευή συστημάτων ΤΝ, δεδομένου ότι σχετίζονται στο σύνολό τους με τις διάφορες δυνατότητες των συστημάτων. Σε γενικές γραμμές, όλες αυτές οι τεχνικές είναι δυνατό να ομαδοποιηθούν σε δύο κύριες κατηγορίες που σχετίζονται με τις δυνατότητες της *συλλογιστικής* και της *μάθησης*. Η ρομποτική είναι άλλο ένα πολύ συναφές επιστημονικό πεδίο.

Συλλογιστική και λήψη αποφάσεων. Αυτή η κατηγορία τεχνικών περιλαμβάνει την αναπαράσταση γνώσης και τη συλλογιστική, τον σχεδιασμό, τον προγραμματισμό, την αναζήτηση και τη βελτιστοποίηση. Οι τεχνικές αυτές καθιστούν δυνατή την εκτέλεση συλλογισμών με βάση τα δεδομένα που διαβιβάζουν οι αισθητήρες. Για να γίνει αυτό, τα δεδομένα πρέπει να μετατραπούν σε γνώση. Ως εκ τούτου, ένας τομέας της ΤΝ έχει να κάνει με την αναζήτηση των καλύτερων δυνατών τρόπων για τη μοντελοποίηση αυτής της γνώσης (*αναπαράσταση γνώσης*). Αφού γίνει η μοντελοποίηση της γνώσης, τα επόμενα βήματα είναι η εκτέλεση συλλογισμών με βάση τη γνώση (*συλλογιστική της γνώσης*), που περιλαμβάνει την εξαγωγή συμπερασμάτων μέσω συμβολικών κανόνων, οι δραστηριότητες *σχεδιασμού* και *προγραμματισμού*, η *αναζήτηση* σε ένα μεγάλο σύνολο λύσεων, και η *βελτιστοποίηση* με βάση όλες τις πιθανές λύσεις σε ένα πρόβλημα. Το τελευταίο βήμα είναι να αποφασιστεί η ενέργεια που θα εκτελεστεί. Το σκέλος της συλλογιστικής/λήψης αποφάσεων ενός συστήματος ΤΝ είναι συνήθως εξαιρετικά πολύπλοκο και προϋποθέτει συνδυασμό αρκετών από τις προαναφερθείσες τεχνικές.

Μάθηση. Αυτή η κατηγορία τεχνικών περιλαμβάνει τη μηχανική μάθηση, τα νευρωνικά δίκτυα, τη βαθιά μάθηση, τα δενδροδιαγράμματα αποφάσεων και πολλές άλλες τεχνικές μάθησης. Χάρη στις τεχνικές αυτές, ένα σύστημα ΤΝ μπορεί να μαθαίνει πώς να επιλύει προβλήματα που δεν είναι δυνατό να προσδιοριστούν επακριβώς ή η μέθοδος επίλυσης των οποίων δεν είναι δυνατό να περιγραφεί με συμβολικούς κανόνες συλλογιστικής. Παραδείγματα τέτοιων προβλημάτων είναι όσα σχετίζονται με δυνατότητες αντίληψης όπως η *ομιλία* και η *κατανόηση της γλώσσας*, καθώς και η *όραση υπολογιστή* ή η *πρόβλεψη συμπεριφορών*. Επισημαίνεται ότι, αν τα προβλήματα αυτά φαίνονται εύκολα, είναι γιατί όντως είναι συνήθως εύκολα για τους ανθρώπους. Ωστόσο, δεν είναι εξίσου εύκολα για τα συστήματα ΤΝ, δεδομένου ότι αυτά δεν μπορούν να χρησιμοποιήσουν κοινή λογική (τουλάχιστον όχι ακόμα) και, μάλιστα, είναι ιδιαίτερα δύσκολα όταν ένα σύστημα καλείται να ερμηνεύσει αδόμητα δεδομένα. Εδώ ακριβώς είναι που αποδεικνύονται χρήσιμες οι τεχνικές που υιοθετούν την προσέγγιση της *μηχανικής μάθησης*. Ωστόσο, τεχνικές μηχανικής μάθησης είναι δυνατό να χρησιμοποιούνται εκτός από την αντίληψη και για πολλές ακόμη εργασίες. Οι τεχνικές μηχανικής μάθησης παράγουν ένα αριθμητικό μοντέλο (δηλ. έναν μαθηματικό τύπο) που χρησιμοποιείται για τον υπολογισμό της απόφασης βάσει των δεδομένων.

Η μηχανική μάθηση λαμβάνει διάφορες μορφές. Οι πιο διαδεδομένες προσεγγίσεις είναι αυτές της *επιβλεπόμενης μάθησης*, της *μη επιβλεπόμενης μάθησης*, και της *ενισχυτικής μάθησης*.

Στην επιβλεπόμενη μηχανική μάθηση, αντί να δώσουμε στο σύστημα συμπεριφορικούς κανόνες, του παρέχουμε παραδείγματα συμπεριφορών εισόδου-εξόδου με την ελπίδα ότι θα μπορέσει να κάνει γενικεύσεις βάσει των παραδειγμάτων (που κατά κανόνα περιγράφουν το παρελθόν) και να συμπεριφερθεί ικανοποιητικά και σε καταστάσεις που δεν παρουσιάζονται στα παραδείγματα (οι οποίες θα ήταν δυνατό να αντιμετωπιστούν στο μέλλον). Στην περίπτωση του παραδείγματός μας, θα δίνουμε στο σύστημα πολλά παραδείγματα εικόνων ενός πατώματος και την αντίστοιχη ερμηνεία (δηλ. αν το πάτωμα στην κάθε εικόνα είναι καθαρό ή όχι). Αν δώσουμε αρκετά παραδείγματα, τα οποία να είναι διαφορετικά μεταξύ τους και να περιλαμβάνουν αρκετές περιπτώσεις από τις περισσότερες καταστάσεις, το σύστημα, μέσω του αλγόριθμου μηχανικής μάθησης που διαθέτει, θα είναι σε θέση να κάνει γενικεύσεις ώστε να γνωρίζει επίσης πώς να ερμηνεύει ικανοποιητικά εικόνες πατωμάτων που δεν έχει ξαναδεί. Κάποιες προσεγγίσεις μηχανικής μάθησης υιοθετούν αλγόριθμους που βασίζονται στην έννοια των *νευρωνικών δικτύων*, τα οποία ως έναν βαθμό είναι εμπνευσμένα από τη λειτουργία του ανθρώπινου εγκεφάλου, από την άποψη ότι διαθέτουν ένα δίκτυο μικρών μονάδων επεξεργασίας (κάτι ανάλογο με τους δικούς μας νευρώνες) μεταξύ των οποίων υπάρχουν πολλές σταθμισμένες συνδέσεις. Το νευρωνικό δίκτυο έχει ως είσοδο τα δεδομένα που διαβιβάζουν οι αισθητήρες (στο παράδειγμά μας, την εικόνα του δαπέδου) και ως έξοδο την ερμηνεία της εικόνας (στο παράδειγμά μας, το κατά πόσο το πάτωμα είναι καθαρό ή όχι). Κατά την ανάλυση των παραδειγμάτων (δηλ. τη φάση *εκπαίδευσης* του δικτύου) οι συντελεστές στάθμισης των συνδέσεων προσαρμόζονται ώστε να αντιστοιχούν όσο το δυνατόν περισσότερο σε αυτό που δηλώνουν τα διαθέσιμα παραδείγματα (δηλ. να ελαχιστοποιηθεί το σφάλμα μεταξύ της αναμενόμενης εξόδου και της εξόδου που υπολογίζεται από το σύστημα). Στο τέλος της φάσης εκπαίδευσης, στο πλαίσιο της φάσης δοκιμής της συμπεριφοράς του νευρωνικού δικτύου έναντι καινοφανών παραδειγμάτων, ελέγχεται το αν η μάθηση της εν λόγω εργασίας έχει γίνει ικανοποιητικά.

Είναι σημαντικό να επισημανθεί ότι η προσέγγιση αυτή (όπως και όλες οι τεχνικές μηχανικής μάθησης) έχει πάντοτε ένα ορισμένο ποσοστό σφάλματος, το οποίο συνήθως είναι μικρό. Μια πολύ σημαντική έννοια, λοιπόν, είναι η *ακρίβεια*, δηλ. η μέτρηση του ποσοστού σωστών απαντήσεων για να διαπιστωθεί πόσο μεγάλο είναι.

Υπάρχουν αρκετά είδη νευρωνικών δικτύων και προσεγγίσεων μηχανικής μάθησης, εκ των οποίων μία από τις πιο επιτυχημένες είναι σήμερα η *βαθιά μάθηση*. Η εν λόγω προσέγγιση βασίζεται στο γεγονός ότι στο νευρωνικό δίκτυο μεσολαβούν διάφορα επίπεδα μεταξύ της εισόδου και της εξόδου, που καθιστούν δυνατή τη μάθηση της συνολικότερης σχέσης εισόδου-εξόδου σε διαδοχικά βήματα. Αυτό καθιστά τη συνολική προσέγγιση ακριβέστερη και περιορίζει την ανάγκη ανθρώπινης καθοδήγησης.

Τα νευρωνικά δίκτυα είναι ένα μόνο από τα εργαλεία μηχανικής μάθησης —υπάρχουν και πολλά άλλα, με διάφορα χαρακτηριστικά: τυχαία δάση και ενισχυμένα δενδροδιαγράμματα, μέθοδοι ομαδοποίησης, παραγοντοποίηση πινάκων κ.λπ.

Άλλη μία χρήσιμη προσέγγιση μηχανικής μάθησης είναι η *ενισχυτική μάθηση*. Στο πλαίσιο της εν λόγω προσέγγισης αφήνουμε ελεύθερο το σύστημα TN να λαμβάνει τις αποφάσεις του, με την πάροδο του χρόνου, και σε κάθε απόφαση δίνουμε στο σύστημα ένα σήμα ανατροφοδότησης που του λέει αν η απόφαση ήταν καλή ή κακή. Στόχος του συστήματος, σε βάθος χρόνου, είναι η μεγιστοποίηση της θετικής ανατροφοδότησης που λαμβάνει. Η προσέγγιση αυτή χρησιμοποιείται, για παράδειγμα, στα συμβουλευτικά συστήματα (όπως τα διάφορα διαδικτυακά συμβουλευτικά συστήματα που προτείνουν στους χρήστες προϊόντα που ίσως θα ήθελαν να αγοράσουν) ή και στο μάρκετινγκ.

Οι προσεγγίσεις μηχανικής μάθησης είναι χρήσιμες όχι μόνο για την εκτέλεση των εργασιών που σχετίζονται με την αντίληψη, όπως η αναγνώριση εικόνων και η κατανόηση κειμένων, αλλά και όλων εκείνων των εργασιών που δεν είναι εύκολο να οριοθετηθούν και δεν είναι δυνατό να περιγραφούν ολοκληρωμένα με συμπεριφορικούς συμβολικούς κανόνες.

Επισημαίνεται ότι θα πρέπει να γίνεται διάκριση μεταξύ των προσεγγίσεων μηχανικής μάθησης που χρησιμοποιούνται για τη μάθηση μιας εργασίας που δεν είναι δυνατό να περιγραφεί ικανοποιητικά με συμβολικό τρόπο, και των ορθολογικών πρακτόρων με δυνατότητες μάθησης (που αναφέρονται στην προηγούμενη ενότητα) οι οποίοι προσαρμόζουν τη συμπεριφορά τους με την πάροδο του χρόνου ώστε ο εκάστοτε στόχος να επιτυγχάνεται καλύτερα. Οι δύο αυτές τεχνικές μπορεί να επικαλύπτονται ή να συνεργάζονται, χωρίς αναγκαστικά να συμπίπτουν.

Ρομποτική. Η ρομποτική μπορεί να οριστεί ως «η πρακτική εφαρμογή της TN στον υλικό κόσμο» (είναι γνωστή στα αγγλικά και ως «*embodied AI*», δηλ. «ενσάρκωση της TN»). Ρομπότ είναι μια μηχανή με υλική υπόσταση που καλείται να αντεπεξέλθει στη δυναμική, τις αβεβαιότητες και την πολυπλοκότητα του υλικού κόσμου. Η αρχιτεκτονική ελέγχου των ρομποτικών συστημάτων συνήθως ενσωματώνει δυνατότητες αντίληψης, συλλογιστικής, εκτέλεσης ενεργειών, μάθησης, αλλά και αλληλεπίδρασης με άλλα συστήματα. Πέραν της TN, ρόλο στη σχεδίαση και τη λειτουργία των ρομπότ παίζουν και άλλα επιστημονικά πεδία, όπως η μηχανολογία και η θεωρία ελέγχου. Παραδείγματα ρομπότ είναι οι ρομποτικοί χειριστές, τα αυτόνομα οχήματα (π.χ. αυτοκίνητα, δρόνοι, ιπτάμενα ταξί), ανθρωποειδή ρομπότ, ρομποτικές ηλεκτρικές σκούπες κ.λπ.

Στο σχήμα 2 απεικονίζονται τα περισσότερα από τα επιμέρους επιστημονικά πεδία της TN, καθώς και η μεταξύ τους σχέση. Είναι σημαντικό, ωστόσο, να επισημανθεί ότι η TN είναι πολύ πιο πολύπλοκη από ό,τι δείχνει το σχήμα, δεδομένου ότι περιλαμβάνει πολλά ακόμα επιμέρους επιστημονικά πεδία και τεχνικές. Επίσης, όπως επισημάνθηκε παραπάνω, και η ρομποτική βασίζεται σε τεχνικές που ενδεχομένως είναι εκτός της σφαίρας της TN. Σε κάθε περίπτωση, θεωρούμε ότι όσα αναφέρονται εδώ αρκούν για τον εμπλουτισμό της ανταλλαγής απόψεων, της ευαισθητοποίησης και των συζητήσεων σχετικά με την TN, τη δεοντολογία στον τομέα της TN και τις πολιτικές TN που θα πρέπει να γίνουν στους κόλπους της ιδιαίτερα πολυεπιστημονικής και πολυσυμμετοχικής ομάδας εμπειρογνομώνων υψηλού επιπέδου.



Σχήμα 2: Απλουστευμένη επισκόπηση των επιμέρους επιστημονικών πεδίων της TN και της μεταξύ τους σχέσης. Τόσο η μηχανική μάθηση όσο και η συλλογιστική περιλαμβάνουν πολλές άλλες τεχνικές, ενώ και η ρομποτική περιλαμβάνει τεχνικές που δεν εμπίπτουν στο αντικείμενο της TN. Το σύνολο της TN υπάγεται στο επιστημονικό πεδίο της επιστήμης υπολογιστών.

3. Άλλες σημαντικές έννοιες και ζητήματα της TN

Περιορισμένη (ή ασθενής) και γενική (ή ισχυρή) TN. Σκοπός των συστημάτων γενικής TN είναι να αποτελούν συστήματα που μπορούν να εκτελούν τις περισσότερες από τις δραστηριότητες τις οποίες μπορούν να εκτελούν οι άνθρωποι. Αντίθετα, τα συστήματα περιορισμένης TN είναι συστήματα που μπορούν να εκτελούν μόνο μία δραστηριότητα ή μικρό αριθμό συγκεκριμένων δραστηριοτήτων. Τα συστήματα TN που βρίσκονται σήμερα σε επιχειρησιακή λειτουργία είναι παραδείγματα περιορισμένης TN. Όταν η TN έκανε ακόμη τα πρώτα της βήματα, οι ερευνητές χρησιμοποιούσαν διαφορετική ορολογία (ασθενής και ισχυρή TN). Σήμερα παραμένουν ανοιχτές πολλές δεοντολογικές, επιστημονικές και τεχνολογικές προκλήσεις στην πορεία για την ανάπτυξη των δυνατοτήτων που απαιτούνται για να γίνει η γενική TN πραγματικότητα, όπως η συλλογιστική που βασίζεται στην κοινή λογική, η αυτεπίγνωση και η ικανότητα της μηχανής να καθορίζει η ίδια τον σκοπό της.

Ζητήματα και μεροληψία δεδομένων. Δεδομένου ότι πολλά συστήματα TN, όπως αυτά που περιλαμβάνουν μεμονωμένα στοιχεία επιβλεπόμενης μηχανικής μάθησης, βασίζονται σε τεράστιες ποσότητες δεδομένων για να λειτουργούν ικανοποιητικά, είναι σημαντικό να γίνουν κατανοητοί οι τρόποι με τους οποίους τα δεδομένα επηρεάζουν τη συμπεριφορά του συστήματος TN. Για παράδειγμα, αν τα δεδομένα εκπαίδευσης είναι μεροληπτικά, αν δηλαδή δεν είναι αρκετά ισορροπημένα ή συμπεριληπτικά, το σύστημα TN που εκπαιδεύεται με βάση αυτά τα δεδομένα δεν θα είναι σε θέση να κάνει ικανοποιητικές γενικεύσεις και ενδεχομένως θα λαμβάνει μεροληπτικές αποφάσεις υπέρ της μίας ή της άλλης κατηγορίας. Προσφάτως, η κοινότητα της TN επεξεργάζεται μεθόδους για την ανίχνευση και τον περιορισμό της μεροληψίας στα σύνολα δεδομένων εκπαίδευσης, καθώς και σε άλλα μέρη ενός συστήματος TN.

«Μαύρο κουτί» της TN και επεξηγησιμότητα. Κάποιες τεχνικές μηχανικής μάθησης, αν και πολύ επιτυχημένες από την άποψη της ακρίβειας, δεν είναι καθόλου διαφωτιστικές από την άποψη της κατανόησης του τρόπου με τον οποίο λαμβάνουν αποφάσεις. Η έννοια του «μαύρου κουτιού» της TN (black-box AI) αναφέρεται σε αυτά ακριβώς

τα σενάρια, όπου δεν είναι δυνατό να εντοπίσει κανείς τους λόγους στους οποίους βασίζονται ορισμένες αποφάσεις. Σε αντιδιαστολή, επεξηγησιμότητα είναι η ιδιότητα των συστημάτων ΤΝ που είναι σε θέση να παρέχουν κάποιο είδος επεξηγήσεων για τις ενέργειές τους.

ΤΝ κατευθυνόμενη από στόχους. Όλα τα σημερινά συστήματα ΤΝ είναι κατευθυνόμενα από στόχους, που σημαίνει ότι λαμβάνουν από τον άνθρωπο τις προδιαγραφές του στόχου που πρέπει να επιτύχουν και χρησιμοποιούν ορισμένες τεχνικές για να τον επιτύχουν. Δεν καθορίζουν τα ίδια τους τους στόχους τους. Ωστόσο, ορισμένα συστήματα ΤΝ (όπως αυτά που βασίζονται σε ορισμένες τεχνικές μηχανικής μάθησης) μπορεί να έχουν περισσότερη ελευθερία να αποφασίσουν ποια πορεία θα ακολουθήσουν για να επιτύχουν τον δεδομένο στόχο.

4. Επικαιροποιημένος ορισμός της ΤΝ

Προτείνουμε να χρησιμοποιείται ο ακόλουθος επικαιροποιημένος ορισμός της ΤΝ:

«Τα συστήματα τεχνητής νοημοσύνης (ΤΝ) είναι συστήματα λογισμικού (ή ενδεχομένως και υλισμικού) που σχεδιάζονται από ανθρώπους³ και, βάσει ενός δεδομένου σύνθετου στόχου, ενεργούν στην υλική ή ψηφιακή διάσταση με το να αντιλαμβάνονται το περιβάλλον τους μέσω της απόκτησης δεδομένων, να ερμηνεύουν τα δομημένα ή αδόμητα δεδομένα που έχουν συλλεχθεί, να προβαίνουν σε συλλογισμούς με βάση τις γνώσεις ή να επεξεργάζονται τις πληροφορίες που εξάγονται από αυτά τα δεδομένα και να αποφασίζουν ποια είναι η βέλτιστη ενέργεια (ή οι βέλτιστες ενέργειες) που θα πρέπει να εκτελέσουν για να επιτύχουν τον δεδομένο στόχο. Τα συστήματα ΤΝ μπορεί είτε να χρησιμοποιούν συμβολικούς κανόνες είτε να μαθαίνουν ένα αριθμητικό μοντέλο, και μπορεί επίσης να προσαρμόζουν τη συμπεριφορά τους με το να αναλύουν πώς επηρεάζεται το περιβάλλον από τις προηγούμενες ενέργειές τους.

Ως επιστημονικό πεδίο, η ΤΝ περιλαμβάνει διάφορες προσεγγίσεις και τεχνικές, όπως η μηχανική μάθηση (συγκεκριμένα παραδείγματα της οποίας είναι η βαθιά μάθηση και η ενισχυτική μάθηση), η μηχανική συλλογιστική (που περιλαμβάνει τον σχεδιασμό, τον προγραμματισμό, την αναπαράσταση και τη συλλογιστική γνώσης, την αναζήτηση και τη βελτιστοποίηση) και η ρομποτική (που περιλαμβάνει έλεγχο, αντίληψη, αισθητήρες και ενεργοποιητές, καθώς και την ενσωμάτωση όλων των άλλων τεχνικών σε κυβερνο-υλικά συστήματα).»

και να γίνεται παραπομπή σε αυτό το έγγραφο ως πηγή συμπληρωματικών πληροφοριών που τεκμηριώνουν τον εν λόγω ορισμό.

³ Humans design AI systems directly, but they may also use AI techniques to optimise their design.

**Το παρόν έγγραφο εκπονήθηκε από τα μέλη της ομάδας εμπειρογνομόνων υψηλού
επιπέδου για την TN**

τα οποία παρατίθενται ακολούθως κατ' αλφαβητική σειρά

Pekka Ala-Pietilä, πρόεδρος OEYE για την TN AI Finland, Huhtamaki, Sanoma	Pierre Lucas Orgalim – Europe's technology industries
Wilhelm Bauer Fraunhofer	Ieva Martinkenaite Telenor
Urs Bergmann Zalando	Thomas Metzinger JGU Mainz & European University Association
Mária Bielíková Slovak University of Technology της Μπρατισλάβα	Catelijne Muller ALLAI Netherlands & EOKE
Cecilia Bonefeld-Dahl DigitalEurope	Markus Noga SAP
Yann Bonnet ANSSI	Barry O'Sullivan, αντιπρόεδρος OEYE για την TN University College Cork
Loubna Bouarfa OKRA	Ursula Pacht Ευρωπαϊκό Γραφείο Ενώσεων Καταναλωτών (ΕΓΕΚ)
Stéphan Brunessaux Airbus	Nicolas Petit Πανεπιστήμιο της Λιέγης
Raja Chatila IEEE Initiative Ethics of Intelligent/Autonomous Systems & πανεπιστήμιο της Σορβόνης	Christoph Peylo Bosch
Mark Coeckelbergh Πανεπιστήμιο της Βιέννης	Iris Plöger BDI
Virginia Dignum Πανεπιστήμιο της Umeå	Stefano Quintarelli Garden Ventures
Luciano Floridi Πανεπιστήμιο της Οξφόρδης	Andrea Renda College of Europe Faculty & Κέντρο Μελετών Ευρωπαϊκής Πολιτικής (CEPS)
Jean-Francois Gagné Element AI	Francesca Rossi* IBM
Chiara Giovannini Ευρωπαϊκή Ένωση για τον Συντονισμό της Εκπροσώπησης των Καταναλωτών στην Τυποποίηση (ANEC)	Cristina San José Ομοσπονδία Ευρωπαϊκών Τραπεζών (EBF)
Joanna Goodey Οργανισμός Θεμελιωδών Δικαιωμάτων	George Sharkov Digital SME Alliance
Sami Haddadin Σχολή ρομποτικής και νοημοσύνης των μηχανών του Μονάχου (MSRM)	Philipp Slusallek Γερμανικό Κέντρο Έρευνας για την TN (DFKI)
Gry Hasselbalch The thinkdotank DataEthics & πανεπιστήμιο της Κοπεγχάγης	Françoise Soulié Fogelman Σύμβουλος σε θέματα TN
Fredrik Heintz Πανεπιστήμιο του Λινσέπινγκ	Saskia Steinacker Bayer
Fanny Hidvegi Access Now	Jaan Tallinn Ambient Sound Investment
Eric Hilgendorf Πανεπιστήμιο του Würzburg	Thierry Tingaud STMicroelectronics
Klaus Höckner Hilfsgemeinschaft der Blinden und Sehschwachen	Jakob Uszkoreit Google
Mari-Noëlle Jégo-Laveissière Orange	Aimee Van Wynsberghe TU Delft
Leo Kärkkäinen Nokia Bell Labs	Thiébaut Weber Ευρωπαϊκή Συνομοσπονδία Συνδικαλιστικών Οργανώσεων (CES)
Sabine Theresia Köszegi TU Wien	Cecile Wendling AXA
Robert Kroplewski Δικηγόρος & σύμβουλος της πολωνικής κυβέρνησης	Karen Yeung Πανεπιστήμιο του Μπέρμιγχαμ
Elisabeth Ling RELX	

*Η Francesca Rossi ήταν εισηγήτρια του παρόντος εγγράφου.