

# φυσική & τεχνητή ΝΟΗΜΟΣΥΝΗ



Κώστας Παγωνδιώτης

## η δοκιμασία του **Turing** το κινέζικο δωμάτιο & η νοημοσύνη

Τα τελευταία χρόνια όλο και πιο συχνά βομβαρδιζόμαστε με διαφημίσεις για διάφορες οικιακές συσκευές «με νοημοσύνη», όπως ψυγεία, πλυντήρια ή κλιματιστικά. Έχουν, άραγε, αυτές οι συσκευές όντως νοημοσύνη; Συνήθως θεωρούμε αυτούς τους ισχυρισμούς διαφημιστικές υπερβολές. Ωστόσο, οι διαισθήσεις μας παύουν να είναι τόσο ξεκάθαρες όταν ερχόμαστε αντιμέτωποι με υπολογιστικά συστήματα που εξαγουν συμπεράσματα από γραπτές ιστορίες, αποδεικνύουν θεωρήματα, παίζουν σκάκι ή αλληλεπιδρούν αυτόνομα με το φυσικό περιβάλλον. Από ποιο σημείο και μετά οφείλουμε να δεχθούμε ότι ένα ον έχει νοημοσύνη;

Επιμέλεια Αφιερώματος  
Άρης Αραγεώργης  
Κώστας Παγωνδιώτης  
Επιμέλεια Εικαστικών  
Ελένη Φιλιππάκη  
Μετάφραση κειμένων  
Τερέζα Μπούκη  
Σπύρος Πετρονάκος

Όταν μιλάμε για φυσική νοημοσύνη εννοούμε κατά κύριο λόγο την ανθρώπινη νοημοσύνη. Συχνά, ωστόσο, στην προσπάθεια να περιχαρακώσουμε και να προσδιορίσουμε τη νοημοσύνη μας, την αντιδιαστέλλουμε προς άλλες μορφές φυσικής νοημοσύνης, όπως είναι η νοημοσύνη των ζώων, των (ενδεχομένων) εξωγήινων ή του Θεού. Τα τελευταία, περίπου, πενήντα χρόνια αυτή η προσπάθεια κατανόησης της ανθρώπινης νοημοσύνης πέρασε σε μια νέα, ιδιαίτερα συναρπαστική, φάση με την επινόηση των υπολογιστών και την ανάπτυξη της τεχνητής νοημοσύνης. Τώρα πλέον, μια κυρίαρχη φιλοσοφική αντίληψη για τη φύση της ανθρώπινης νοημοσύνης μπορεί, όπως θα δούμε, να δοκιμαστεί στην πράξη, με την έννοια ότι έχει ανοίξει ο δρόμος για την κατασκευή μηχανών οι οποίες λειτουργούν με τον τρόπο που σκιαγραφεί η συγκεκριμένη αντίληψη.

Η τεχνητή νοημοσύνη (ΤΝ) είναι ο υποκλάδος της επιστήμης των υπολογιστών που προσφέρει ακριβώς αυτή τη δυνατότητα. Στόχος ενός κυρίαρχου ρεύματος μέσα στην ΤΝ –το οποίο οι φιλόσοφοι ονομάζουν «σκληρή ΤΝ»– είναι να κατασκευάσει σκεπτόμενες μηχανές. Η νοημοσύνη αυτών των μηχανών χαρακτηρίζεται ‘τεχνητή’ επειδή, αν επιτευχθεί, θα είναι προϊόν κατασκευής και όχι αποτέλεσμα φυσικής επιλογής –όπως ορισμένοι πιστεύουμε ότι είναι η ανθρώπινη και η ζωική νοημοσύνη. Η ΤΝ δεν έχει στόχο να μας εξαπατήσει με κάποια έξυπνα τρικ αλλά να δημιουργήσει μηχανές με πραγματική νοημοσύνη, δηλαδή σκεπτόμενες μηχανές. Γι’ αυτό θα ήταν πιο δόκιμο, ίσως, να μιλούσαμε για ‘συνθετική’ νοημοσύνη, κατ’ αναλογία, για παράδειγμα, με τη συνθετική ινσουλίνη.<sup>1</sup>

Η αποδοχή της άποψης ότι με αφετηρία τους υπολογιστές είναι δυνατή η κατασκευή σκεπτόμενων μηχανών προϋποθέτει την παραδοχή ότι υπάρχει κάποια ουσιώδης ομοιότητα ανάμεσα στους υπολογιστές και τον ανθρώπινο νου. Διαφορετικά δεν θα διανοούμαστε καν να ξεκινήσουμε το εγχείρημα κατασκευής σκεπτόμενων μηχανών στηριζόμενοι στις *υπολογιστικές* μηχανές. Μάλιστα, σε παλαιότερες εποχές, είχαν προταθεί άλλου τύπου μηχανισμοί ως μοντέλα για τον τρόπο λειτουργίας του ανθρώπινου νου, όπως, για παράδειγμα, ρολόγια, ατμομηχανές και τηλεφωνικά κέντρα.

Το εγχείρημα της ΤΝ στηρίζεται σε μια συγκεκριμένη φιλοσοφική αντίληψη για τη φύση της ίδιας της ανθρώπινης νοημοσύνης. Πρόκειται για μια αντίληψη που άρχισε να διαμορφώνεται στη νεότερη δυτική φιλοσοφία, σύμφωνα με την οποία η σκέψη είναι υπολογισμός και, ειδικότερα, υπολογιστικός χειρισμός νοητικών συμβόλων. Με αφετηρία αυτή την ιδέα και μέσα από μια μακρά και επίπονη πορεία θεωρητικών επεξεργασιών και εννοιολογικών μεταπολίσεων, φτάσαμε τελικά στην ίδια την επινόηση των ψηφιακών υπολογιστών, δηλαδή των μηχανών που η λειτουργία τους βασίζεται στην υπολογιστική επεξεργασία συμβόλων. Αυτό είναι ένα είδος επεξεργασίας που δεν στηρίζεται στη σημασία

των συμβόλων αλλά αποκλειστικά στη σύνταξη τους, στον τρόπο, δηλαδή, που οι *μορφές* των συμβόλων σχετίζονται μεταξύ τους. Γι’ αυτό και οι ψηφιακοί υπολογιστές έχουν χαρακτηρισίσιμη και ως συντακτικές μηχανές.

Το πιο σημαντικό ίσως χαρακτηριστικό αυτών των μηχανών, το οποίο ενίσχυσε την αντίληψη ότι ο ίδιος ο ανθρώπινος νους αποτελεί έναν ψηφιακό υπολογιστή, είναι ότι μπορούν να αλλάζουν τη λειτουργία τους χωρίς να αλλάζουν τα μηχανικά τους μέρη. Αυτό το επιτυγχάνουν χάρη στη δυνατότητα προγραμματισμού με βάση τον οποίο μπορούν κάθε φορά να εκτελούν αυτόματα ένα νέο σύνολο εντολών κωδικοποιημένων σε μια τυπική γλώσσα. Γι’ αυτό για να αλλάξουμε τη λειτουργία του υπολογιστή μας δεν χρειάζεται να του αλλάξουμε κάποιο εξάρτημα, αρκεί να του ‘φορτώσουμε’ ένα καινούριο πρόγραμμα. Αυτή η αυτονομία που παρουσιάζει το επίπεδο του προγράμματος από το υλικό υπόστρωμα του ψηφιακού υπολογιστή θεωρήθηκε από πολλούς ερευνητές εντελώς ανάλογη προς την αυτονομία που παρουσιάζει η ανθρώπινη σκέψη από το σώμα μέσα στο οποίο πραγματοποιείται. Έτσι, από τα μέσα του 20ου αιώνα άρχισε να θεωρείται όλο και πιο εύλογη η ιδέα ότι ο ίδιος ο νους δεν είναι τίποτα άλλο παρά ένας (πολύ σύνθετος) ψηφιακός υπολογιστής. Η θέση αυτή αποκρυσταλλώθηκε και αναπτύχθηκε σε φιλοσοφική θεωρία με την έλευση του Λειτουργισμού, μια προσέγγιση που έχει ως βασικό σλόγκαν ότι ο νους σε σχέση με το σώμα είναι ό,τι το λογισμικό (το πρόγραμμα) σε σχέση με το υλισμικό, δηλαδή το υλικό από το οποίο αποτελείται ο υπολογιστής.

Το εγχείρημα της ΤΝ εγείρει διάφορα ηθικά ζητήματα που σχετίζονται με το πώς πρέπει να συμπεριφερόμαστε σε όντα με τεχνητή νοημοσύνη και με το εάν πρέπει να κατασκευάσουμε τέτοια όντα. Για παράδειγμα, αν αύριο ανακαλύψω ότι ο καλύτερος φίλος μου κάτω από το δέρμα του δεν έχει ανθρώπινους ιστούς και ανθρώπινα οστά αλλά ένα σύνολο από μικροεπεξεργαστές και άλλους μεταλλικούς μηχανισμούς, πρέπει άραγε να αλλάξω τον τρόπο που τον αντιμετωπίζω; Τα ηθικά ζητήματα που εγείρει το εγχείρημα της ΤΝ γίνονται όλο και πιο πιεστικά λόγω της αλματώδους προόδου της υπολογιστικής τεχνολογίας και της αυξανόμενης εξάρτησης των ανθρώπινων κοινωνιών από τους υπολογιστές. Ωστόσο, προτού κανείς αρχίσει να διερευνά τέτοιου τύπου ερωτήματα είναι ίσως σκόπιμο να εξετάσει τι είναι η ανθρώπινη νοημοσύνη και αν όντως αυτή έγκειται στον υπολογισμό. Αν, όπως ορισμένοι φιλόσοφοι υποστηρίζουν, ο ανθρώπινος νους δεν είναι τελικά ένας ψηφιακός υπολογιστής, τότε οι υπολογιστικές μηχανές που κατασκευάζει η ΤΝ ενδεχομένως να μην αποκτήσουν ποτέ νοημοσύνη, τουλάχιστον ανθρώπινη νοημοσύνη, και ίσως θα πρέπει να τις αντιμετωπίζουμε απλώς ως χρήσιμα εργαλεία που επεκτείνουν της δυνατότητές μας. Σε αυτό το αφιέρωμα θα επικεντρωθούμε στη συζήτηση γύρω από το εάν ο νους είναι υπολογιστής και δεν θα ασχο-

ληθούμε με τα ηθικά ζητήματα που εγείρει το εγχείρημα της ΤΝ.

Για πολλούς φιλοσόφους το ερώτημα για τη φύση της ανθρώπινης νοημοσύνης είναι άμεσα συναρτημένο με ένα γνωσιολογικό ερώτημα: ποια είναι τα κριτήρια που θα μας επιτρέψουν να διακρίνουμε αν ένα όν έχει νοημοσύνη; Έτσι, ειδικότερα, ακόμα και αν δεχθούμε ότι η ‘ουσία’ της ανθρώπινης νοημοσύνης έγκειται στον υπολογισμό, με ποιο τρόπο θα μπορέσουμε να αποφασίσουμε αν



μια υπολογιστική μηχανή που κατασκευάσαμε έχει νοημοσύνη; Τα τελευταία χρόνια, για παράδειγμα, όλο και πιο συχνά βομβαρδιζόμαστε με διαφημίσεις για διάφορες οικιακές συσκευές «με νοημοσύνη», όπως ψυγεία, πλυντήρια ή κλιματιστικά. Έχουν, άραγε, αυτές οι συσκευές όντως νοημοσύνη; Συνήθως θεωρούμε αυτούς τους ισχυρισμούς διαφημιστικές υπερβολές. Ωστόσο, οι διαισθήσεις μας παύουν να είναι τόσο ξεκάθαρες όταν ερχόμαστε αντιμέτωποι με υπολογιστικά συστήματα που εξάγουν συμπεράσματα από γραπτές ιστορίες,

αποδεικνύουν θεωρήματα, παίζουν σκάκι ή αλληλεπιδρούν αυτόνομα με το φυσικό περιβάλλον. Από ποιο σημείο και μετά οφείλουμε να δεχθούμε ότι ένα όν έχει νοημοσύνη;

Είναι ίσως εύλογο να υποθέσουμε ότι αν ένα όν έχει νοημοσύνη, τότε αυτή θα πρέπει κάπως να εκδηλώνεται και μάλιστα με τέτοιο τρόπο ώστε να μπορούμε να τη διαπιστώσουμε. Χρειαζόμαστε κάποιο δημόσια διαπιστώσιμο κριτήριο προκειμένου να προσδιορίσουμε με αντικειμενικό τρόπο την ύπαρξη νοημοσύ-

νης. Διαφορετικά φαίνεται να κινδυνεύουμε να εμπλακούμε σε μια ατέρμονη συζήτηση χωρίς τη δυνατότητα συμφωνίας. Ένα τέτοιο κριτήριο πρότεινε ο μεγάλος Βρετανός μαθηματικός Alan Turing, ο οποίος ήταν ένας από τους ανθρώπους που επινόησαν τη βασική αρχή λειτουργίας των ψηφιακών υπολογιστών. Ο Turing θεώρησε ότι το ερώτημα «μπορούν οι μηχανές να σκέφτονται;» είναι εντελώς χωρίς νόημα για να αξίζει εξέτασης. Για τον λόγο αυτό θέλησε να το αντικαταστήσει με ένα πιο συγκεκριμένο ερώτημα το οποίο αφορά το εάν ένας

ψηφιακός υπολογιστής μπορεί να κερδίσει σε μια δοκιμασία που ο Turing ονόμασε «παιχνίδι της μίμησης». Η δοκιμασία του Turing βασίζεται σε ένα παιχνίδι προσποίησης με τρεις παίκτες. Οι δύο παίκτες είναι άνθρωποι και ο τρίτος υπολογιστής. Ο ένας από τους δύο ανθρώπους είναι ο ανακριτής και έχει ως αποστολή του να θέτει ερωτήματα μέσω ενός τερματικού στους άλλους δύο παίκτες προκειμένου να διακρίνει ποιος από τους δύο είναι ο υπολογιστής. Οι δύο ανακρινόμενοι παίκτες μπορούν με τη σειρά τους να απαντούν στα ερωτήματα γραπτώς μέσω του τερματικού. Αποστολή του υπολογιστή είναι να μιμηθεί όσο το δυνατόν καλύτερα την ανθρώπινη γλωσσική συμπεριφορά προκειμένου να παραπλανήσει τον ανακριτή, ενώ, αντίθετα, αποστολή του ανακρινόμενου ανθρώπου είναι να λείπει την αλήθεια προκειμένου να βοηθήσει τον ανακριτή στο έργο του. Ο Turing ισχυρίζεται ότι αν σε αυτό το παιχνίδι ο ανακριτής φτάσει να κάνει λάθος τόσο συχνά όσο θα έκανε αν αποστολή του ήταν να διακρίνει υπό τις ίδιες συνθήκες έναν άνδρα από μια γυναίκα, τότε πρέπει να δεχθούμε ότι ο υπολογιστής έχει περάσει επιτυχώς τη δοκιμασία και ότι άρα έχει νοημοσύνη.

Από τη δομή αυτής της δοκιμασίας φαίνεται ότι ο Turing θεωρεί τη γλωσσική συμπεριφορά καθοριστικό κριτήριο για την ύπαρξη νοημοσύνης. Μάλιστα, η γλωσσική συμπεριφορά για την οποία κάνει λόγο δεν είναι κάποια μεμονωμένη και αποσπασματική συμπεριφορά αλλά γλωσσική συμπεριφορά που προσδιάζει σε διάλογο. Αρκετοί θεωρούν προάγγελο αυτής της ιδέας τον ίδιο τον Καρτέσιο, ο οποίος, προκειμένου να προσάψει τη θέση του ότι τα ζώα είναι απλώς μηχανικά αυτόματα χωρίς την ικανότητα σκέψης, υποστήριξε ότι, σε αντίθεση με τους ανθρώπους, τα ζώα δεν έχουν την ικανότητα να χρησιμοποιούν τη γλώσσα. Ο Καρτέσιος, όπως φαίνεται και από μια επιστολή του που μεταφράσαμε για τις ανάγκες του αφιερώματος, υποστήριξε πιο συγκεκριμένα ότι η γλωσσική έκφραση αποτελεί το μόνο εξωτερικό σημάδι που μπορεί να βεβαιώσει την ύπαρξη σκέψης, αλλά, σε αντίθεση με τους υποστηρικτές της ΤΝ, αρνιόταν ότι οι αυτοκινούμενες μηχανές (και τα ζώα) μπορούν να εκδηλώσουν αυτό το σημάδι: «...καμία από τις εξωτερικές πράξεις μας δεν μπορεί να βεβαιώσει κάποιον που τις εξετάζει ότι το σώμα μας δεν είναι απλώς μια αυτοκινούμενη μηχανή αλλά διαθέτει και ψυχή με σκέψεις, εκτός από τις λέξεις ή άλλα σήματα που αναφέρονται σε ζητήματα τα οποία παρουσιάζονται χωρίς να σχετίζονται με κάποιο πάθος».<sup>2</sup>

Η δοκιμασία του Turing ουσιαστικά μας λέει ότι προκειμένου να διαπιστώσουμε αν ένα πλάσμα σκέφτεται πρέπει να προσδιορίσουμε, μέσω της γλωσσικής συμπεριφοράς του, το *περιεχόμενο* των μηνυμάτων του. Αυτή είναι μια πολύ απαιτητική δοκιμασία αν τη συγκρίνουμε με δοκιμασίες που θα μας επέτρεπαν να διαπιστώσουμε αν ένα πλάσμα σκέφτεται χωρίς να χρειάζεται να προσδιορίσουμε τι σκέφτεται. Φανταστείτε, για παράδειγμα, μια δοκιμασία

στην οποία ο εξεταστής θέτει ερωτήματα και ελέγχει απλώς κάποιες (μη γλωσσικές) αντιδράσεις του πλάσματος, όπως, λόγω χάρην, την πίεση του αίματός ή τον βαθμό συστολής της κόρης του ματιού του (αν υποθέσουμε προς στιγμήν ότι το εξεταζόμενο πλάσμα έχει αίμα και μάτια). Σε μια τέτοια δοκιμασία δεν θα χρειαζόταν να προσδιορίσουμε τι σκέφτεται προκειμένου να αποφανθούμε για το αν σκέφτεται. Υπό αυτή την έννοια λοιπόν, η δοκιμασία του Turing είναι πολύ πιο απαιτητική γιατί ο εξεταστής χρειάζεται να λειτουργήσει ως *ερμηνευτής* γλωσσικών σημείων.

Παρόλα αυτά η συγκεκριμένη δοκιμασία δεν φαίνεται τελικά να προσφέρει ένα επαρκές κριτήριο για τη διαπίστωση της ύπαρξης νοημοσύνης, γιατί υπάρχει περίπτωση το εξεταζόμενο πλάσμα να καταφέρνει να πληκτρολογεί κάποια σχήματα τα οποία ερμηνεύονται ως αποδεκτή γλωσσική συμπεριφορά από τον εξεταστή, χωρίς ωστόσο το ίδιο το πλάσμα να κατανοεί τι σημαίνουν αυτά τα σχήματα. Θα μπορούσε ενδεχομένως να υποστηριχθεί ότι αυτή η ανεπάρκεια της δοκιμασίας του Turing οφείλεται στο γεγονός ότι είναι αποκλειστικά μια συμπεριφοριστική δοκιμασία, γιατί το μόνο που εξετάζει ο ανακριτής είναι οι (γλωσσικές) αντιδράσεις των παικτών στα (γλωσσικά) ερεθίσματα που δέχονται. Ο ανακριτής αδιαφορεί εντελώς για τον τρόπο με τον οποίο οι παίκτες παράγουν αυτές τις αντιδράσεις. Ίσως λοιπόν θα μπορούσε να συμπληρωθεί αυτή η δοκιμασία με κάποιο έλεγχο της εσωτερικής λειτουργίας των παικτών. Μάλιστα, αν δεχθούμε τη βασική υπόθεση της σκληρής ΤΝ ότι η σκέψη είναι υπολογισμός, ο κατάλληλος τρόπος για να συμπληρώσουμε τη δοκιμασία θα ήταν να ελέγχουμε παράλληλα αν το εξεταζόμενο πλάσμα εκτελεί κάποιο υπολογιστικό πρόγραμμα.

Ωστόσο, ούτε αυτή η συμπλήρωση φαίνεται να είναι επαρκής. Σκεπτείτε την ακόλουθη περίπτωση που επινόησε ο John Searle (1980/1993): έστω ότι έχουμε έναν Βρετανό, ο οποίος δεν γνωρίζει κινέζικα, και ο οποίος είναι κλεισμένος σε ένα δωμάτιο με επικοινωνεί με τον εξωτερικό κόσμο μέσω γραπτών μηνυμάτων. Τα μηνύματα που δέχεται είναι γραμμένα στα κινέζικα, το ίδιο και τα μηνύματα που στέλνει. Αυτό το επιτυγχάνει χάρη σε ένα υπολογιστικό πρόγραμμα, δηλαδή ένα σύνολο συντακτικών κανόνων, το οποίο είναι γραμμένο στα αγγλικά και προσδιορίζει πώς να χειρίζεται συντακτικά τα εισερχόμενα μηνύματα καθώς και μια βάση δεδομένων (η οποία είναι επίσης στα κινέζικα) προκειμένου να βρίσκει τα κινέζικα σύμβολα που γράφει στα εξερχόμενα μηνύματα. Από τη σκοπιά ενός εξωτερικού παρατηρητή που γνωρίζει κινέζικα, τα μηνύματα που στέλνονται στο δωμάτιο είναι ερωτήματα, τα δε μηνύματα που εξέρχονται από αυτό είναι απαντήσεις στα συγκεκριμένα ερωτήματα. Στον εξωτερικό παρατηρητή, λοιπόν, δίδεται η εντύπωση ότι μέσα στο δωμάτιο βρίσκεται κάποιο πλάσμα που επιτυγχάνει στη δοκιμασία του Turing και το οποίο καταλαβαίνει κινέζικα. Ωστόσο, αυτή η εντύπωση είναι προφανώς ψευδής από την

οπτική του ανθρώπου που βρίσκεται μέσα στο δωμάτιο, διότι ο ίδιος δεν γνωρίζει κινέζικα. Το μόνο που κάνει ο συγκεκριμένος άνθρωπος είναι να μιμείται τον τρόπο λειτουργίας των υπολογιστών, δηλαδή να χειρίζεται σύμβολα στηριζόμενος αποκλειστικά στη μορφή τους. Το συμπέρασμα, λοιπόν, στο οποίο φτάνει ο Searle είναι ότι ο συντακτικός χειρισμός συμβόλων δεν επαρκεί για την κατανόηση των συμβόλων. Από τη σύνταξη δεν μπορούμε να πάρουμε σημασιολογία. Εφόσον ο άνθρωπος μέσα στο δωμάτιο δεν καταλαβαίνει κινέζικα όταν εκτελεί το υπολογιστικό πρόγραμμα, ούτε και ένας υπολογιστής μπορεί να καταλαβαίνει κινέζικα όταν εκτελεί το ίδιο υπολογιστικό πρόγραμμα.

Το ενδιαφέρον με το επιχείρημα του Searle είναι ότι μας καλεί να εξετάσουμε το πρόβλημα της νοημοσύνης από μια διαφορετική οπτική, την οπτική του πρώτου προσώπου. Σε αντίθεση με τη δοκιμασία του Turing, όπου καλούμαστε να αποφασίσουμε από την οπτική του τρίτου προσώπου –την οπτική του ανακριτή– για το εάν ένα άλλο ον έχει νοημοσύνη, ο Searle μάς ζητεί να λειτουργήσουμε οι ίδιοι ως υπολογιστές και να εξετάσουμε από «πρώτο χέρι» αν αυτός ο τρόπος λειτουργίας μάς αποφέρει κατανόηση της κινεζικής γλώσσας.

Ωστόσο, ο Searle μέσα σε αυτή την οπτική του πρώτου προσώπου φαίνεται ότι μας τοποθετεί σε μια λανθασμένη θέση, δηλαδή μας

τοποθετεί στη θέση του εγκεφαλικού μηχανισμού ενός Κινέζου και όχι στη θέση του ίδιου του Κινέζου. Κατά συνέπεια, το μόνο που δικαιούται ο Searle να συναγάγει από το νοπτικό του πείραμα είναι ότι ο εγκεφαλικός μηχανισμός, ο οποίος υποστηρίζει την ικανότητα ενός Κινέζου να διαβάζει και να γράφει στη γλώσσα του, δεν καταλαβαίνει κινέζικα. Η σωστή θέση μέσα από την οπτική του πρώτου προσώπου είναι εκείνη που, εκτός από τον επεξεργαστή, περιλαμβάνει επίσης το υπολογιστικό πρόγραμμα και τη βάση δεδομένων. Δηλαδή, αν υπάρχει κάτι που καταλαβαίνει κινέζικα, αυτό δεν θα έπρεπε να είναι ο άνθρωπος μέσα στο δωμάτιο, αλλά ολόκληρο το δωμάτιο.

Ωστόσο, ο Searle θεωρεί ότι ακόμα κι αν βάλουμε κάποιον στη θέση ολόκληρου του δωματίου, αυτός θα εξακολουθεί να μην καταλαβαίνει κινέζικα. Έστω, για παράδειγμα, ότι ο άνθρωπος που ήταν κλεισμένος μέσα στο δωμάτιο απομνημονεύει τώρα το υπολογιστικό πρόγραμμα και την ‘κινέζικη’ βάση δεδομένων ούτως ώστε να μπορεί να βρίσκει χωρίς εξωτερική βοήθεια τα εκάστοτε σύμβολα που γράφει. Ο Searle θεωρεί διαισθητικά προφανές ότι ούτε αυτή η προσθήκη δεν θα επιτρέψει στον συγκεκριμένο άνθρωπο να καταλαβαίνει κινέζικα. Ωστόσο, άλλοι φιλόσοφοι δεν έχουν τις ίδιες διαισθήσεις.

Όμως μια συζήτηση όταν καταλήγει απλώς σε μια σύγκριση διαισθήσεων παύει να είναι

διαφωτιστική. Εκείνο που παραμένει, παρά όλα αυτά, ενδιαφέρον από την επιχειρηματολογία του Searle είναι το γεγονός ότι εισάγει μια εσωτερική διάσταση σε αυτό που λέμε «νοημοσύνη»: αν προσεγγίσουμε από την οπτική του πρώτου προσώπου τη νοημοσύνη, τότε αυτή εκδηλώνεται ως *συνειδητή* κατανόηση. Η συνείδηση για τον Searle είναι συνθήκη για να έχουμε νοπτικές καταστάσεις και νοημοσύνη. Το πρόβλημα, ωστόσο, με αυτή την πρόταση είναι ότι αν προκειμένου να κατανοήσουμε το τι είναι νοημοσύνη χρειάζεται προηγουμένως να κατανοήσουμε το τι είναι συνείδηση, τότε φαίνεται να οδηγούμαστε σε μια εσωτερική, ιδιωτική σφαίρα –τη σφαίρα της υποκειμενικότητας– η οποία, ακριβώς για αυτό τον λόγο, δεν προσφέρεται, εκ πρώτης όψεως, για διυποκειμενικά ελέγχιμη έρευνα.<sup>3</sup> Το πρόβλημα που αντιμετωπίζει η προσπάθεια επιστημονικής διερεύνησης της συνείδησης παρουσιάζεται αναλυτικότερα στο κείμενο της Φαίης Ζήκα «Μαίρη, η επιστήμων του χρώματος» στη στήλη «De Coloribus», ενώ το άρθρο του Αθανάσιου Ραφτόπουλου «Οπτική συνειδητότητα: άνθρωποι και μηχανές» περιγράφει δύο πρόσφατες επιστημονικές προσεγγίσεις στο πρόβλημα της οπτικής συνείδησης.

Ωστόσο, ακόμα κι αν δεχθούμε ότι η προσέγγιση της νοημοσύνης από την οπτική του πρώτου προσώπου οδηγεί στο πρόβλημα της συνείδησης, εξακολουθεί να παραμένει ανοικτή για μας η οδός της οπτικής του τρίτου

προσώπου. Εξάλλου, στον καθημερινό μας βίο, όταν ‘αποδίδουμε’ νοημοσύνη σε έναν άλλο άνθρωπο, σίγουρα δεν έχουμε άμεση πρόσβαση ούτε σε κάποια υποκειμενική, ιδιωτική του σφαίρα ούτε, βέβαια, στον εγκέφαλό του και στις υποτιθέμενες υπολογιστικές του λειτουργίες. Το μόνο στο οποίο έχουμε άμεση πρόσβαση είναι οι δημόσια διαπιστώσιμες συμπεριφορές του. Ίσως, λοιπόν, το πρόβλημα με τη δοκιμασία του Turing δεν έγκειται στο γεγονός ότι αυτή είναι συμπεριφοριστική αλλά στο γεγονός ότι περιορίζει δραστικά τα συμπεριφοριστικά τεκμήρια που πρέπει να λάβουμε υπόψη για να διαπιστώσουμε αν ένα ον έχει νοημοσύνη.

Πράγματι, ο έλεγχος της *γραπτής* γλωσσικής συμπεριφοράς επιτρέπει, στην καλύτερη περίπτωση, τη διαπίστωση της λογικής συνοχής του παραγόμενου λόγου. Αλλά δεν μας φανερώνει τίποτα για το εάν το εξεταζόμενο πλάσμα εννοεί όντως κάτι με τα γραπτά σύμβολα που παράγει. Με άλλα λόγια, δεν μας φανερώνει το εάν το εξεταζόμενο πλάσμα έχει την ικανότητα να αναφέρεται στον κόσμο. Επιπλέον, η γραπτή γλωσσική συμπεριφορά, από μόνη της, δεν μας φανερώνει ούτε το εάν υπάρχει συνέπεια μεταξύ λόγων και πράξεων. Τα δύο αυτά ζητήματα διαπιστώνονται πρωτογενώς μόνο αν το εξεταζόμενο πλάσμα είναι σωματικά παρόν, ούτως ώστε να μπορούμε να παρατηρούμε αφενός το εάν τα λόγια του ανταποκρίνονται σε όσα αισθητηριακά αντιλαμβάνεται και αφετέρου το εάν οι πράξεις του ανταποκρίνονται σε όσα λέει ότι πιστεύει και επιθυμεί.

Το αίτημα ότι πρέπει να υπάρχουν αυτά τα δύο είδη ανταπόκρισης μαζί με το αίτημα ότι ο παραγόμενος λόγος πρέπει να έχει λογική συνοχή σημαίνουν ουσιαστικά ότι η νοήμων συμπεριφορά χαρακτηρίζεται από διάφορα είδη κανονιστικότητας. Γι’ αυτόν ακριβώς τον λόγο

όταν μια συμπεριφορά είναι νοήμων επιδέχεται χαρακτηρισμών όπως ‘ορθή/εσφαλμένη’, ‘αρμόζουσα/μη-αρμόζουσα’ κ.ο.κ. Αντίθετα, όταν μια συμπεριφορά δεν είναι νοήμων, όπως για παράδειγμα ένα τικ, δεν έχει νόημα να τη χαρακτηρίσουμε ως ανάρμοστη ή εσφαλμένη. Φυσικά, προκειμένου μια συμπεριφορά να διέπεται όντως από κανονιστικότητα δεν αρκεί να είναι χαρακτηρίσιμη ως τέτοια από την οπτική του τρίτου προσώπου, αλλά θα πρέπει και το ίδιο το ον που εκδηλώνει τη συμπεριφορά να μεριμνά σχετικά με την ορθότητά της. Για το πρόβλημα της κανονιστικότητας ο αναγνώστης μπορεί να ανατρέξει στο κείμενο του Σπύρου Πετρουνάκου «Μπορούν οι υπολογιστές να ακολουθούν κανόνες».

Η συνείδηση και η κανονιστικότητα είναι δύο από τα σημαντικότερα ‘χαρακτηριστικά’ της ανθρώπινης νοημοσύνης. Ωστόσο, φαίνεται ότι βρισκόμαστε ακόμα πολύ μακριά από το να κατασκευάσουμε υπολογιστικές μηχανές που να έχουν συνείδηση και η συμπεριφορά τους να διέπεται από κανονιστικότητα. Η κλασική προσέγγιση στην ΤΝ, η οποία εμπνέεται από την ιδέα ότι η νοημοσύνη έγκειται στον υπολογιστικό χειρισμό εσωτερικών συμβολικών αναπαραστάσεων, δεν έχει αποφέρει ιδιαίτερους καρπούς ως προς αυτό το ζήτημα. Γι’ αυτό τον λόγο, τα τελευταία 20 χρόνια έχουν αρχίσει να αναπτύσσονται δυναμικά ορισμένες αντι-αναπαριστασιακές προσεγγίσεις στην ΤΝ οι οποίες έχουν ως κοινό γνώρισμα ότι μελετούν και προσπαθούν να αναπαράγουν ‘κατώτερα’ είδη νοημοσύνης, όπως είναι η νοημοσύνη των ζώων ή ανθρώπινες πρακτικές δεξιότητες οι οποίες δεν είναι γλωσσικά αρθρωμένες.

Αυτός ο νέος προσανατολισμός της έρευνας υποστηρίζεται φιλοσοφικά, κυρίως, από το έργο του Hubert Dreyfus και του John Haugeland οι οποίοι θεωρούν τον νο

ενσώματο και ενταγμένο [embedded] στο φυσικό και κοινωνικό περιβάλλον. Σύμφωνα με αυτούς τους φιλοσόφους, οι οποίοι εμπνέονται από το έργο του Heidegger και του Merleau-Ponty, το μεγαλύτερο μέρος των γνώσεών μας δεν είναι προτασιακό αλλά βρίσκεται ενσωματωμένο στις διάφορες πρακτικές δεξιότητές μας καθώς και στον φυσικό και κοινωνικό κόσμο. Για παράδειγμα, το μόνο που χρειάζεται να ξέρω για να πάω από την Αθήνα στη Θεσσαλονίκη είναι ποιο δρόμο να πάρω. Αρκεί να καταφέρω να βρω αυτόν τον δρόμο και μετά θα με «οδηγήσει» ο δρόμος, η γνώση είναι «αποθηκευμένη» μέσα σε αυτόν. Σκεφτείτε πόσα άλλα πράγματα θα έπρεπε να γνωρίζω αν έπρεπε να βρω τον προσανατολισμό μου μέσα από μονοπάτια του δάσους.

Για μια διερεύνηση της ιδέας ότι ο νους είναι ενσώματος και ενταγμένος στο φυσικό και κοινωνικό περιβάλλον ο αναγνώστης μπορεί να ανατρέξει στο κείμενο της Μαρίας Βενιέρη «Είναι η νοημοσύνη ενσώματη; Εγκέφαλοι στη γυάλα και Μάτριξ», ενώ το κείμενο του Vincent Müller «Πενήντα χρόνια τεχνητής νοημοσύνης – γιατί δεν επιτύχαμε ακόμα;» παρουσιάζει τον νέο προσανατολισμό που έχει πάρει η έρευνα στην ΤΝ και κάνει έναν γενικότερο απολογισμό της 50χρονης ιστορίας της.

Το αφιέρωμα στην τεχνητή και τη φυσική νοημοσύνη αρχίζει με τη συνέντευξη που παραχώρησε στο *Cogito* ο φιλόσοφος του Πανεπιστημίου του Σικάγο John Haugeland. Στη συνέντευξη αναπτύσσονται σε βάθος πολλά από τα ζητήματα που θίγονται στα υπόλοιπα κείμενα του αφιερώματος.



#### ΕΝΔΕΙΚΤΙΚΗ ΒΙΒΛΙΟΓΡΑΦΙΑ

- ▶ Churchland, P. (1995/1999): *Η Μηχανή της Λογικής, η Θέση της Ψυχής*. Εκδόσεις Γκοβόστη.
- ▶ Copeland, J. (1993): *Artificial Intelligence – A Philosophical Introduction*. Blackwell.
- ▶ Dreyfus, H. (1992/2001): *Τι δεν μπορούν ακόμα να κάνουν οι υπολογιστές*. Πανεπιστημιακές Εκδόσεις Κρήτης.
- ▶ Glymour, C. (1992/1998): «Γνωσιολογία των Ανδροειδών: Υπολογισμός, Τεχνητή Νοημοσύνη και Φιλοσοφία της Επιστήμης» στο Salmon, H. M. (επιμέλεια) (1998): *Εισαγωγή στη Φιλοσοφία της Επιστήμης*. Πανεπιστημιακές Εκδόσεις Κρήτης.
- ▶ Haugeland, J. (1985/1992): *Τεχνητή Νοημοσύνη*. Εκδόσεις Κάτοπτρο.
- ▶ Preston, J. & Bishop, M. (επιμέλεια) (2002): *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*. Oxford: Clarendon Press.
- ▶ Searle, J. (1980/1993): «Νοήσεις, Εγκέφαλοι και Προγράμματα» στο Hofstadter, D. & Dennett, D. (επιμέλεια): *Το Εγώ της Νόησης – Φαντασίες και Στοχασμοί για τον Εαυτό και την Ψυχή*. Εκδόσεις Κάτοπτρο.
- ▶ Shieber, S. (επιμέλεια) (2004): *The Turing Test: Verbal Behavior as the Hallmark of Intelligence*. The MIT Press.
- ▶ Turing, A.M. (1950/1993): «Υπολογιστικές Μηχανές και Νοημοσύνη» στο Hofstadter, D. & Dennett, D. (επιμέλεια): *Το Εγώ της Νόησης – Φαντασίες και Στοχασμοί για τον Εαυτό και την Ψυχή*. Εκδόσεις Κάτοπτρο.

#### ΕΝΔΕΙΚΤΙΚΕΣ ΔΙΕΥΘΥΝΣΕΙΣ ΣΤΟ ΔΙΑΔΙΚΤΥΟ

- ▶ <http://www.aaai.org/AITopics/html/phil.html> (Στην ιστοσελίδα αυτή της American Association for Artificial Intelligence είναι συγκεντρωμένες οι σπουδαιότερες πηγές που υπάρχουν στο διαδίκτυο σχετικά με τη φιλοσοφία της τεχνητής νοημοσύνης).
- ▶ <http://consc.net/people.html#ai> (Η σελίδα συγκεντρώνει τις διευθύνσεις ιστοσελίδων αρκετών φιλοσόφων που ασχολούνται με τη φιλοσοφία της τεχνητής νοημοσύνης και οι οποίοι προσφέρουν σχετικά άρθρα τους σε ηλεκτρονική μορφή).
- ▶ <http://www.angelfire.com/ego/philosophyradio/#mind> (Στην ιστοσελίδα αυτή μπορείτε να βρείτε και να ακούσετε ραδιοφωνικές εκπομπές που αφορούν τη φιλοσοφία του νου).
- ▶ <http://consc.net/biblio/4.html> (Εδώ θα βρείτε μια σχεδόν εξαντλητική βιβλιογραφία για τη φιλοσοφία της τεχνητής νοημοσύνης).

#### ΣΗΜΕΙΩΣΕΙΣ

<sup>1</sup> Haugeland 1985/1992, σελ. 349.

<sup>2</sup> Βλ., επίσης, το πέμπτο κεφάλαιο από τον *Λόγο περί της Μεθόδου* του Καρτέσιου (Εκδόσεις Παπαζήση, 1976).

<sup>3</sup> Ωστόσο αυτή η κατάληξη δεν είναι υποχρεωτική. Βλέπε, ενδεικτικά, το κείμενό μου «Συνείδηση, Αντίληψη και Τυφλή Όραση» στο δεύτερο τεύχος του *Cogito*.