Ηθική. Περιοδικό φιλοσοφίας

Αρ. 12 (2019)

Ποιος θα φταίει όταν ο HAL σκοτώσει ξανά;

Άλκης Γούναρης

doi: 10.12681/ethiki.22772

Βιβλιογραφική αναφορά:

Γούναρης Ά. (2020). Ποιος θα φταίει όταν ο HAL σκοτώσει ξανά;. *Ηθική. Περιοδικό φιλοσοφία*ς, (12), 4–10. https://doi.org/10.12681/ethiki.22772

Ποιος θα φταίει όταν ο HAL σκοτώσει ξανά;

Αλκης Γουναρης[1]

Οι περισσότεροι από εμάς γνωρίσαμε τον HAL 9000 ως βασικό πρωταγωνιστή της ταινίας του Stanley Kubrick «2001, Οδύσσεια του Διαστήματος», η οποία βασίστηκε στο σενάριο και την ομότιτλη νουβέλα του Arthur Clarke[2]. Ο HAL, ένας Ευρετικά[3] Προγραμματισμένος Αλγοριθμικός Υπολογιστής –μια εξελιγμένη μορφή Τεχνητής Νοημοσύνης (ΤΝ)–, προκειμένου να διασφαλίσει την επιτυχία της αποστολής που έχει αναλάβει, όταν κατάλαβε ότι αυτή κινδυνεύει, αποφάσισε να σκοτώσει το πλήρωμα του διαστημόπλοιου, μέσα στο οποίο ήταν εγκατεστημένος, και να αποκτήσει τον έλεγχο.

Βρισκόμενος στο κατώφλι των καταιγιστικών τεχνολογικών εξελίξεων του 21ου αιώνα, ο Daniel Dennett, ο οποίος συμμετείχε τότε σε ένα ακριβοπληρωμένο αλλά αποτυχημένο –όπως αποδείχθηκε στη συνέχεια– πρόγραμμα κατασκευής ευφυούς ΤΝ[4], δημοσιεύει το δημοφιλές πλέον κείμενό του, σχετικά με τα ηθικά ζητήματα που αφορούν την εξέλιξη της τεχνολογίας των ηλεκτρονικών υπολογιστών, θέτοντας το ερώτημα «When HAL Kills, Who's to Blame?»[5].

Δυο δεκαετίες μετά, επαναδιατυπώνοντας το κλασικό πια ερώτημα του Daniel Dennett, επιχειρώ να τονίσω την ανάγκη για μια επικαιροποιημένη φιλοσοφική απάντηση, υπό το πρίσμα τόσο των σημερινών εξελίξεων στον τομέα της ΤΝ, όσο και εκείνων του άμεσου μέλλοντος.

Η αλήθεια είναι ότι, ενώ το 1997 έμοιαζε τότε αποκαλυπτικό για την έκρηξη της τεχνολογίας των υπολογιστών και της ρομποτικής, σήμερα φαντάζει ως λίθινη για την ΤΝ εποχή, όχι μόνο από πλευράς τεχνολογικών και υπολογιστικών δυνατοτήτων, αλλά και από πλευράς προσβασιμότητας και κόστους[6].

Ξεδιπλώνοντας το επιχείρημά του, ο Dennett επικαλείται τη μνημειώδη πρώτη νίκη στο σκάκι, του υπολογιστή της IBM, Deep Blue, επί του παγκοσμίου πρωταθλητή Gary Kasparov το 1996. Συγκεκριμένα, υποστηρίζει ότι αναγνωρίζουμε και θαυμάζουμε την ικανότητα του υπολογιστή να κερδίζει στο σκάκι και συγχαίρουμε τους προγραμματιστές του για το επίτευγμα, όμως η νίκη ανήκει στον υπολογιστή και όχι στους προγραμματιστές. Οι τελευταίοι, εάν αντιμετώπιζαν τον παγκόσμιο πρωταθλητή, προφανώς θα έχαναν μέσα σε λίγα λεπτά. Η ευθύνη που αναλογεί στους προγραμματιστές για τη νίκη του Deep Blue είναι ισοδύναμη με αυτή που αντιστοιχεί στον προπονητή ή στον δάσκαλο του Kasparov, όμως τελικά υπεύθυνοι για το αποτέλεσμα της αναμέτρησης είναι οι ίδιοι οι παίκτες.

Επεκτείνοντας τον συλλογισμό αυτόν σε ένα πιο εξελιγμένο σύστημα ΤΝ, όπως ο HAL, το οποίο διαθέτει απείρως μεγαλύτερες υπολογιστικές δυνατότητες από τον Deep Blue, λειτουργεί «αυτόνομα» και προβαίνει σε πράξεις ζωής και θανάτου,

θα λέγαμε ότι ο HAL βεβαίως έχει την κύρια ευθύνη των πράξεών του. Η άποψη του Dennett επί του θέματος είναι σαφής· Εκτιμά πως η νοημοσύνη που επιδεικνύει ο HAL συνεπάγεται ηθική συμπεριφορά[7]. Το γεγονός ότι ο HAL κατορθώνει να εξαπατήσει τον Dave, τον κυβερνήτη του διαστημόπλοιου, και, τελικά, να τον σκοτώσει, αποτελεί για τον Dennett (ο οποίος επικαλείται τον Nietzsche)[8] χαρακτηριστικό τόσο της υψηλής νοημοσύνης του όσο και της ηθικής συμπεριφοράς του. *Η μοχθηρία αποτελεί βασικό χαρακτηριστικό του ηθικού ζώου*[9].

Εν τέλει, ο Dennett αποδίδει στον HAL χαρακτηριστικά ηθικού προσώπου, ανεξάρτητα με το αν ο HAL μετανιώνει, αισθάνεται τύψεις, συναισθάνεται ή κατανοεί τι σημαίνει να είναι κανείς ηθικό πρόσωπο[10].

Ωστόσο, υπάρχουν τρεις βασικές έννοιες -οι οποίες σχετίζονται με τα ηθικά προβλήματα που ανακύπτουν από τις πράξεις της «αυτόνομης» ΤΝ- που ο Dennett απέφυγε να αντιμετωπίσει. Πρόκειται για τις έννοιες της Αυτονομίας, του Ελέγχου και του Κενού Ευθύνης και για τον τρόπο με τον οποίο αυτές συμπλέκονται μεταξύ τους. Κάθε φορά που αναζητάμε την ηθική ευθύνη σε πολύπλοκες καταστάσεις ή συστήματα, στην εξέλιξη των οποίων εμπλέκονται διάφορες διαδικασίες, όπως στην αυτόνομη ΤΝ που οι δημιουργοί της δεν μπορούν να προβλέψουν τις τελικές πράξεις της, ερχόμαστε αντιμέτωποι με αυτό που στην ηθική ονομάζουμε Responsibility Gap[11]. Αυτό το «κενό ευθύνης» το συναντάμε επίσης στον κόσμο της οικονομίας και των επιχειρήσεων, στον πόλεμο, στις διεθνείς σχέσεις και γενικά σε σύνθετες καταστάσεις όπου η κρινόμενη πράξη, ενώ προϋποθέτει τη συμμετοχή πολλών ανθρώπων ή φορέων σε προγενέστερο από την πράξη στάδιο, τελικώς δεν μπορεί να προβλεφθεί ή να ελεγχθεί με ακρίβεια στα προηγούμενα στάδια. Στην αυτόνομη ΤΝ, εν προκειμένω, θα πρέπει να μας προβληματίζει τι μερίδιο ευθύνης αντιστοιχεί –και αν αντιστοιχεί– στους προγραμματιστές, στους developers, στους designers, στους χρηματοδότες της έρευνας, στην εταιρεία που κατασκεύασε το σύστημα ΤΝ κτλ., ακόμα και στους τελικούς χρήστες. Μια σχετικά έγκυρη μέθοδος, για να αποφανθούμε ποιος φέρει τελικά την ευθύνη, είναι να διακριβώσουμε ποιος έχει τον έλεγχο και ποιος τον έχει παραμετροποιήσει· ποιος, δηλαδή, παίρνει την απόφαση και σε ποιο βαθμό για τη συγκεκριμένη πράξη; Για να πούμε ότι κάποιος *αποφασίζει* και επιλέγει τι θα πράξει, είναι απαραίτητο να πληρούνται ορισμένα κριτήρια[12]. Θα μπορούσαμε, χάριν οικονομίας, να θεωρήσουμε ότι το κριτήριο, που οδηγεί αυτόν που αποφασίζει αν θα επιλέξει το Α και όχι το Β, είναι η συνθήκη ή η αρχή εκείνη που θα καθορίζει επί της ουσίας την επιλογή του.

Ποια είναι η αρχή, για παράδειγμα, στο γνωστό trolley problem[13], που κάνει κάποιον να στρίψει ή να μη στρίψει το ακυβέρνητο όχημα; Να επιλέξει, προδήλως, να σκοτώσει τον έναν για να σώσει τους πέντε ανθρώπους ή το αντίθετο; Πρόσφατα η Mercedes ανακοίνωσε ότι κατασκευάζει το πρώτο αυτόνομο όχημα με κριτήριο, σε περίπτωση κινδύνου, να βάζει τη ζωή των επιβατών πάντα σε προτεραιότητα[14]. Ποιος έχει θέσει αυτό το κριτήριο και τι μέρος της ευθύνης τελικά του αναλογεί; Τι σημαίνει «σώζω τους επιβάτες», αλλά, ενδεχομένως, σε ένα ακραίο σενάριο μιας επικείμενης σύγκρουσης, σκοτώνω έναν πεζό ή έναν ποδηλάτη; Η αναζήτηση αυτού του κριτηρίου, αυτής της αρχής, μας φέρνει αντιμέτωπους με το σύστημα λήψης

ηθικών αποφάσεων, την ηθική θεωρία που έχουμε υιοθετήσει, και βεβαίως με την έννοια της αυτονομίας.

Για να αποδοθεί ευθύνη σε κάποιον, θεωρούμε ότι ο πράττων είναι αυτόνομος ή εν πάση περιπτώσει δεν τελεί υπό το κράτος εξαναγκασμού. Με άλλα λόγια, για να είναι κάποιος υπεύθυνος για μια πράξη, θα πρέπει να είναι ελεύθερος από εξωγενείς παράγοντες που του επιβάλουν να πράξει με ορισμένο τρόπο (φερ' ειπείν να μην απειλείται με το πιστόλι στον κρόταφο) ή να μην περιορίζεται από μη ελεγχόμενους εσωτερικούς παράγοντες που καθορίζουν την απόφασή του (για παράδειγμα να μη βρίσκεται υπό την επήρεια κάποιου φαρμάκου ή σε κάποια μη ελεγχόμενη νοητική κατάσταση). Η απόφαση, δηλαδή, που οδηγεί κάποιον σε μια συγκεκριμένη πράξη, θα πρέπει να καθορίζεται από τον ίδιο με έλλογο τρόπο[15].

Στην περίπτωση του HAL, ο Dennett ξεπερνάει το ζήτημα της ενδεχόμενης ετερονομίας ενός προγραμματισμένου υπολογιστή, συγκρίνοντάς τον με τον γενετικά ή εμπειρικά «προγραμματισμένο» ηθικό δράστη. Αν ο γενετικός προγραμματισμός και οι ανθρώπινες εμπειρίες απαλλάσσουν τον άνθρωπο από την ηθική του ευθύνη, τότε απαλλάσσουν και τον HAL. Παραβλέπει, όμως, μια σημαντική παράμετρο, η οποία καθιστά τη χρήση του όρου «αυτονομία» στην ΤΝ «κυριολεκτικά» μεταφορική.

Σε αντίθεση με τον άνθρωπο, κάθε αυτόνομη ΤΝ εμπεριέχει έναν δεδομένο σκοπό – μια δεδομένη αποστολή. Δεν είναι, λόγου χάριν, δυνατόν να αναστείλει προσωρινά την αποστολή της και να καθίσει σε ένα καφέ να διαβάσει ένα βιβλίο φιλοσοφίας. Κάθε αποστολή της ΤΝ είναι δεδομένη, αναπόδραστη και έξωθεν ορισμένη με τέτοιο τρόπο ώστε κάθε έννοια αυτονομίας καταργείται. Κι αυτό συμβαίνει όχι γιατί απλώς είναι προγραμματισμένη με κάποιον τρόπο, αλλά γιατί ο σκοπός για τον οποίο υπάρχει εμπεριέχεται στην «ουσία» της. Κάθε μηχανή είναι μια μηχανή «για να» – έχει δηλαδή κατασκευαστεί για να επιτελέσει μια ορισμένη λειτουργία και για να επιτύχει κάποιους στόχους, ανεξάρτητα από την πολυπλοκότητα των στόχων αυτών[16].

Ο σκοπός αυτός μετατρέπει την, κατά τα άλλα ευφυή, μηχανή σε αυτό που ο Bostrom[17] ονομάζει «Πολλαπλασιαστή Συνδετήρων». Πρόκειται για ένα σενάριο που αναδεικνύει την «ηλιθιότητα» των «ευφυϊών» των μηχανών, οι οποίες παραμένουν άκαμπτα και αναπόδραστα προσκολλημένες στον σκοπό για τον οποίο υπάρχουν. Η υποθετική μηχανή του Bostrom έχει την αποστολή να πολλαπλασιάσει τους συνδετήρες γραφείου. Η συγκεκριμένη -ευφυέστατη κατά τα λοιπά- ΤΝ, δρώντας «αυτόνομα», θα πρέπει να βρίσκει και να συλλέγει συνδετήρες ή και να μαζεύει χρήματα για να αγοράσει συνδετήρες ή και να φτιάξει μονάδα ή μονάδες παραγωγής συνδετήρων και γενικά να μπορεί, μόνη της, να βελτιώνει συνεχώς τη νοημοσύνη της και να βελτιστοποιεί την ικανότητά της πολλαπλασιάζοντας τους συνδετήρες με κάθε τρόπο, ξεπερνώντας τελικά κατά πολύ την ανθρώπινη ικανότητα. Λόγω της προσήλωσης στον τελικό της σκοπό, η ΤΝ αγνοεί άλλες, πέραν του σκοπού της, αξίες, όπως η ζωή, η γνώση, η φιλία, η ειλικρίνεια, η εντιμότητα, ή αξίες που συνδέονται με μικρές απολαύσεις της ζωής. Σύμφωνα με τον Bostrom, η μηχανή, δεν θα δύναται να γνωρίσει άλλες αξίες, αφού κάθε αξία πέρα από τον εγγενή της σκοπό θα υπονομεύει την ίδια της την αποστολή – ακόμα και την ίδια της την ύπαρξη.

Στο σενάριο αυτό, ο Bostrom καταφέρνει επιτυγχάνει να δείξει ότι μια ΤΝ μπορεί να πραγματοποιήσει εξαιρετικά σημαντικές επιδόσεις σε έναν συγκεκριμένο στόχο, χωρίς να είναι όμως σε θέση να γνωρίσει ή να κατανοήσει τις καθοριστικές για το ανθρώπινο είδος αξίες, όπως αυτές έχουν αναπτυχθεί στην εξελικτική μας πορεία. Χωρίς την κατανόηση (ή την ενσωμάτωση) των αξιών αυτών, η στοχοπροσηλωμένη ΤΝ –ακόμα κι αν δεν καταστεί όπως ο HAL, με τον έναν ή τον άλλο τρόπο, απειλή για το ανθρώπινο είδος– θα απέχει πολύ από το να χαρακτηριστεί ηθικό πρόσωπο.

Σε αυτό το σημείο είναι απαραίτητη, θεωρώ, μια σημαντική διευκρίνιση, η οποία μπορεί να βάλει, κατά κάποιο τρόπο, σε τάξη τη σκέψη εκείνων που θέτουν ερωτήματα σχετικά με την ηθική της ΤΝ. Είναι διαφορετικό πράγμα να επιδεικνύουν οι μηχανές συμπεριφορά ή να προβαίνουν σε πράξεις οι οποίες μπορούν να αξιολογηθούν ηθικά, και διαφορετικό πράγμα να έχουν οι μηχανές ηθική ευθύνη για τις πράξεις αυτές[18]. Πιο συγκεκριμένα, είναι διαφορετικό να αξιολογούμε το αποτέλεσμα μιας επιχείρησης μη επανδρωμένων οπλικών συστημάτων και διαφορετικό να αποδίδουμε ηθική ευθύνη στα ίδια τα συστήματα. Για να καταλάβουμε τη διαφορά, αρκεί να σκεφτούμε ένα ευγενικό ηχογραφημένο μήνυμα στον τηλεφωνητή, το οποίο, όταν βρισκόμαστε σε αναμονή, λέει: *Παρακαλώ περιμένετε*. Πόσο απέχει η μετάδοση αυτού του ευγενικού μηνύματος, από το να θεωρήσουμε υπεύθυνο τον τηλεφωνητή για την ευγενική του συμπεριφορά;

Σήμερα, μισό αιώνα μετά από τον μυθιστορηματικό HAL, υπάρχει η τεχνολογική υποδομή για τη μεγάλη ανάπτυξη της αυτόνομης ΤΝ. Απλές ή λιγότερο απλές μορφές αυτόνομων υπολογιστικών συστημάτων αποτελούν ήδη μέρος της καθημερινότητάς μας. Τέτοια συστήματα ρυθμίζουν την εναέρια κυκλοφορία, μας υποδεικνύουν τη συντομότερη διαδρομή όταν κινούμαστε με το αυτοκίνητο, εξυπηρετούν τις αναζητήσεις μας στο διαδίκτυο, επιλέγουν για εμάς τις κατάλληλες διαφημίσεις – και όχι μόνο, αφού ακόμα δημιουργούν[19] και αξιολογούν διαφημίσεις[20], γράφουν μουσική, βαθμολογούν σε ρόλο κριτή τους αθλητές της ενόργανης γυμναστικής[21] και των καταδύσεων, υποστηρίζουν διαστημικές αποστολές και άλλα πολλά. Μπορεί τα παραπάνω να μη μοιάζουν τόσο εντυπωσιακά γιατί δεν τα βλέπουμε να κυκλοφορούν ανάμεσά μας, όμως τα αυτόνομα οχήματα εντός των επομένων δεκαετιών θα κυκλοφορούν και αναμένεται να υποκαταστήσουν σε μεγάλο βαθμό την παραδοσιακή οδήγηση. Η μεγαλύτερη, όμως, έρευνα και ανάπτυξη στον τομέα αυτό γίνεται στην πολεμική βιομηχανία. Ήδη αυτόνομες πολεμικές μηχανές, τα μη επανδρωμένα drones κυρίως, επιχειρούν για αμυντικούς ή επιθετικούς σκοπούς και δοκιμάζονται σε πραγματικές εμπόλεμες καταστάσεις[22]. Τα εξοπλιστικά προγράμματα των κρατών έχουν ήδη ξεκινήσει έναν αγώνα δρόμου για την απόκτηση του ανταγωνιστικού πλεονεκτήματος.

Το 2015 οι Max Tegmark (MIT) και Stuart Russell (Berkeley) δημοσίευσαν μια ανοιχτή επιστολή[23] προς τους ερευνητές ΤΝ και Ρομποτικής, με σκοπό την απαγόρευση κατασκευής αυτόνομων οπλικών συστημάτων και φονικών ρομπότ. Μεταξύ των άλλων, αναφέρουν ότι τα αυτόνομα οπλικά συστήματα αποτελούν σήμερα την τρίτη επανάσταση στις πολεμικές επιχειρήσεις (μετά την πυρίτιδα και τα πυρηνικά) και, λόγω του χαμηλού σχετικά κόστους και της ευκολίας

κατασκευής τους, αναμένεται να διαδοθούν ευρέως και να παραχθούν μαζικά, με κίνδυνο να χρησιμοποιηθούν *για τρομοκρατικές ενέργειες, εθνοκαθάρσεις, δολοφονίες προσωπικοτήτων, αποσταθεροποίηση εθνών, υποδούλωση πληθυσμών και επιλεκτική εξόντωση εθνικών ή κοινωνικών ομάδων*. Για τον λόγο αυτόν, καλούν τους ερευνητές ΤΝ να αρνηθούν να συμμετάσχουν στην έρευνα και την κατασκευή τέτοιων οπλικών συστημάτων, όπως αντίστοιχα οι βιολόγοι, οι χημικοί και οι φυσικοί υποστηρίζουν ευρέως ανάλογες διεθνείς συμφωνίες για την απαγόρευση χημικών και βιολογικών όπλων ή και όπλων εφοδιασμένων με laser.

Το ερώτημα, για το ποιος φταίει κάθε φορά που ο HAL σκοτώνει ξανά, είναι πέρα για πέρα επίκαιρο και οι φιλόσοφοι θα πρέπει να δώσουν νέες απαντήσεις, που θα συμβάλλουν καθοριστικά στις πρακτικές, νομικές, οικονομικές, πολιτικές και κατασκευαστικές διαστάσεις του ζητήματος.

Είναι αλήθεια –και για αυτό ίσως θα πρέπει να αισθανόμαστε τυχεροί– ότι βρισκόμαστε μπροστά σε ένα εξαιρετικά ενδιαφέρον μέλλον, όπου πανίσχυρες υπολογιστικές μηχανές πρόκειται να αλλάξουν τη ζωή μας, τις συνήθειές μας και τον τρόπο που αντιλαμβανόμαστε τα πράγματα. Είναι βέβαιο ότι νοήμονες μηχανές θα προσομοιώνουν τις ανθρώπινες πράξεις και θα κάνουν απείρως ταχύτερα και καλύτερα τις δουλειές μας, καθώς και νέες δουλειές που σήμερα δεν μπορούμε καν να φανταστούμε. Είναι στο χέρι μας αν αυτή η προοπτική θα κάνει την ανθρώπινη ζωή καλύτερη ή αν θα τη θέσει σε κίνδυνο.

Σε κάθε περίπτωση οι μηχανές αυτές δεν θα πληρούν –τουλάχιστον όχι στο ορατό μέλλον– τις φιλοσοφικές εκείνες προϋποθέσεις που θα τις καθιστούν κυριολεκτικά ηθικά πρόσωπα. Ακόμα κι όταν εμπλέκονται σε γεγονότα τα οποία μπορούν να αξιολογηθούν ηθικά, το ηθικό βάρος και η ευθύνη, για τις πράξεις (ή τις παραλήψεις) τους, θα πρέπει να αναζητηθεί στη μακρά αλυσίδα των συντελεστών και των φορέων που εμπλέκονται στα συγκεκριμένα γεγονότα πριν από την τελική δράση της ΤΝ[24]. Καθώς η κατασκευή και η χρήση ενός συστήματος αυτόνομης ΤΝ δεν θα πρέπει να νοείται ως μια αξιολογικά ουδέτερη και καθαρά τεχνική πράξη, όλοι όσοι εργάζονται για τη δημιουργία και τη χρήση της θα πρέπει να έχουν επίγνωση των ηθικών παραμέτρων και των συνεπειών που αυτές συνεπάγονται και να αναλαμβάνουν, εκ των προτέρων και δεσμευτικά, την ευθύνη που τους αναλογεί. Επιπρόσθετα, θα μπορούσαν να αναπτυχθούν συστήματα συνεχούς ελέγχου, ανατροφοδότησης και αναβαθμίσεων, ώστε η αυτόνομη ΤΝ να είναι αυτόνομη υπό όρους, δηλαδή να «εκπαιδεύεται» διαρκώς και σε τακτά χρονικά διαστήματα προς την επιθυμητή συμπεριφορά. Σε κάθε περίπτωση, όπως λέει ο Hume, η ηθική είναι και θα παραμείνει, τουλάχιστον προς το παρόν, ανθρώπινη υπόθεση . Το ίδιο και η απόδοση ευθυνών, ακόμα κι αν μπροστά στις νέες εξελίξεις είναι επιτακτική μια αναθεώρηση ορισμών.

ΣΗΜΕΙΩΣΕΙΣ

1. Ο Δρ. Άλκης Γούναρης είναι επιστημονικός συνεργάτης, ερευνητής και διδάσκων στο ΕΚΠΑ. alkisg@philosophy.uoa.gr, www.alkisgounaris.gr

2. Arthur C. Clarke, 2001: A Space Odyssey, New American Library, NY, 1968.

3. Ευρετικοί ή ευριστικοί μηχανισμοί ονομάζονται οι υπολογιστικές τεχνικές επίλυσης προβλημάτων οι

οποίες για οικονομία χρόνου αξιολογούν και προκρίνουν ενδιάμεσες καταστάσεις απορρίπτοντας τις υπόλοιπες. Στην ΤΝ παρότι οι τεχνικές αυτές κωδικοποιούνται αλγοριθμικά, δεν θεωρούνται «ακριβώς» αλγόριθμοι, καθώς οι αλγόριθμοι οδηγούν πάντα σε ακριβή αποτελέσματα, ενώ οι μηχανισμοί αυτοί προσομοιάζουν περισσότερο την ανθρώπινη «διαισθητική» σκέψη και την «σταθμισμένη εικασία».

4. Hubert L. Dreyfus, «Why Heideggerian AI Failed and How Fixing It Would Require Making it More Heideggerian», *Philosophical Psychology* Vol. 20 No. 2 (2007): σσ. 247-268.

5. Daniel C. Dennett, "When Hal Kills, Who's to Blame? Computer Ethics", *Hal's Legacy: 2001's Computer as Dream and Reality*, D. Stork, (επ.), MA: MIT Press, Cambridge, 1997, σσ. 351-365.

6. Είναι χαρακτηριστικό ότι ένα αποθηκευτικό μέσο χωρητικότητας ψηφιακών δεδομένων μερικών megabyte την εποχή εκείνη, κόστιζε περισσότερο απ' ό,τι κοστίζει σήμερα ένας σκληρός δίσκος ενός terabyte. Με τα ίδια χρήματα δηλαδή το 2019 αγοράζεις έναν δίσκο 100.000 φορές μεγαλύτερης χωρητικότητας απ' ό,τι το 1997 -άρα κάθε μονάδα αποθηκευμένης πληροφορίας κοστίζει 1/100.000 λιγότερο απ' ό,τι πριν από είκοσι χρόνια. Ακόμα μεγαλύτερη, είναι η διαφορά στο κόστος της υπολογιστικής ικανότητας. Ενδεικτικά ένα gigaFLOP υπολογιστικής ισχύος κόστιζε το 1997 $30.000 και το 2017 $0,03 δηλαδή 1/1.000.000 λιγότερα χρήματα μέσα σε δύο δεκαετίες. Οι σχέσεις αυτές είναι ενδεικτικές της αλματώδους εξέλιξης στην αγορά των υπολογιστών τα τελευταία χρόνια. Για τη σχέση κόστους - αποθήκευσης δεδομένων χρονολογικά βλ. https://jcmit.net/memoryprice.htm. Για τη σχέση κόστους - υπολογιστικής ικανότητας χρονολογικά βλ. https://en.wikipedia.org/wiki/FLOPS.

7. Ορισμένως, η παραδοχή του Dennett ότι η υψηλή νοημοσύνη συνεπάγεται ηθική συμπεριφορά, εκκινεί από την αποδοχή μιας ισοδυναμίας των όρων νόηση (cognition) και νοημοσύνη (intelligence) επί τη βάσει της υπολογιστικής ικανότητας. Πρόκειται για μια παραδοχή που δεν είναι ευρέως αποδεκτή, καθώς η μεν νοημοσύνη μπορεί να οριστεί ως η ικανότητα επίτευξης πολύπλοκων στόχων και είναι συνεπώς άρρηκτα συνδεδεμένη με την υπολογιστική ικανότητα, η δε νόηση ορίζεται ως η ικανότητα του νοήμονος όντος να μαθαίνει, να αντιλαμβάνεται και να κατανοεί, να προβαίνει σε αξιολογικές κρίσεις και να λαμβάνει αποφάσεις, να νηματοδοτεί κτλ., διαδικασίες δηλαδή που δεν συνδέονται απαραίτητα με την υπολογιστική ικανότητα. Για τις μεταφυσικές διαφορές Νόησης (Cognition) και Νοημοσύνης (Intelligence) βλέπε περισσότερα στο: Γούναρης, Άλκης, «Ανθρώπινη Νόηση και Τεχνητή Νοημοσύνη: Αναζητώντας τις θεμελιώδεις διαφορές του νοήματος στα όρια της μεταφυσικής», (2013). https://alkisgounaris.gr/gr/research/human-cognition-artificial-intelligence/ DOI: 10.13140/RG.2.2.17433.67681.

8. Ο Dennett παραπέμπει στη Γενεαλογία της Ηθικής (πρώτη πραγματεία, παράγραφος 6) όπου κατά την «ιερατική» ζωή, αναδεικνύονται η μοχθηρία και η κακία ως χαρακτηριστικά υπεροχής του ανθρώπου έναντι των υπόλοιπων ζώων. Φρειδερίκος Νίτσε, *Η Γενεαλογία της Ηθικής*, Εκδοτική Θεσσαλονίκης, σελ.74.

9. Το επιχείρημα αυτό αναπτύσσεται με πολύ γλαφυρό τρόπο στο Mark Rowlands, *Ο Φιλόσοφος και ο Λύκος,* Εκδόσεις του Εικοστού Πρώτου, 2010. Σύμφωνα με αυτό, στην ιστορία της εξέλιξης του πιθήκου, η ικανότητά του να ξεγελάει τους άλλους πιθήκους αναπτύχθηκε παράλληλα με την αυξανόμενη ικανότητά του να καταλαβαίνει πότε οι άλλοι προσπαθούν να τον ξεγελάσουν (σελ. 79). Αυτό το γεγονός οδήγησε τον σοφό πήθικο στη σύναψη κοινωνικών συμβολαίων και υιοθέτησης ηθικών κανόνων (σσ. 147-149). Ο ανθρώπινος πολιτισμός και η ίδια η ανθρώπινη νοημοσύνη είναι προϊόντα μιας πορείας βασισμένης στο ψέμα και την εξαπάτηση (σελ.150). Αντίθετα ο Λύκος δεν έχει ανάγκη ηθικών κανόνων. Είναι ευθύς, παρορμητικός, εκρηκτικός, αλλά συγχωρεί και ξεχνάει εύκολα.

10. Ο Dennett διευκρινίζει ότι δεν είναι απαραίτητο ο δράστης να αισθάνεται ενοχές ή μεταμέλεια –ή γενικώς να αισθάνεται κάτι– για μια ηθικά μεμπτή πράξη. Ένας επαγγελματίας δολοφόνος αποτελεί ένα καλό παράδειγμα τέτοιας περίπτωσης.

11. Για το κενό ευθύνης στην ΤΝ βλέπε: Matthias, A, "The responsibility gap: Ascribing responsibility for the actions of learning automata", *Ethics and Information Technology,* 6 (2004): σσ. 175–183.

12. Στην ψυχολογία του κοινού νου (folk psychology) τα κριτήρια αυτά σχετίζονται με τις πεποιθήσεις, τις επιθυμίες κτλ. – όροι που οδηγούν τους φιλοσόφους σε ατέρμονες συζητήσεις για το τι σημαίνει πεποίθηση, ποιος την έβαλε εκεί, τι σημαίνει επιθυμία, βούληση κ.ο.κ. Αυτός είναι, κατά τη γνώμη μου, ένας ατελέσφορος τρόπος προσέγγισης του προβλήματος.

13. Το κλασικό πρόβλημα του ακυβέρνητου οχήματος αποτελεί ένα νοητικό πείραμα που εισηγήθηκε η

Philippa Foot [Philippa Foot, "The Problem of Abortion and the Doctrine of the Double Effect", *Virtues and Vices,* Basil Blackwell, (επ.), Oxford, 1978, που εμφανίστηκε αρχικά στο *Oxford Review,* Number 5 (1967)] και ανέπτυξε εκτενώς η Judith Thompson (Judith J. Thomson, "Killing, Letting Die, and the Trolley Problem", *The Monist* 59 (1976): σελ. 204). Στο νοητικό αυτό πείραμα αναδεικνύεται η διαφορά συλλογιστικής των δεοντοκρατικών και των ωφελιμιστικών θεωριών στην επίλυση διλημμάτων. Στη σύγχρονη βιβλιογραφία συνδέεται με τα διλήμματα που προκύπτουν από την κατασκευή αυτόνομων οχημάτων και αυτόνομων οπλικών συστημάτων.

14. Βλέπε περισσότερα στο https://www.businessinsider.com/mercedes-benz-self-driving-cars-programmed-save-driver-2016-10.

15. Sarah Buss and Andrea Westlund, "Personal Autonomy", The Stanford Encyclopedia of Philosophy (Spring 2018 Edition), Edward N. Zalta (επ.), URL = <https://plato.stanford.edu/archives/spr2018/entries/personal-autonomy/>.

16. Μοιάζει πραγματικά χωρίς νόημα –και πραγματική πρόκληση– στην έρευνα της ΤΝ, η σύλληψη (πόσο μάλλον η κατασκευή) μιας ευφυούς μηχανής χωρίς κανέναν απολύτως σκοπό – χωρίς αποστολή.

17. Nick Bostrom, Ethical Issues in Advanced Artificial Intelligence, I. Smit, et al ed., *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence,* Vol. 2, Institute of Advanced Studies in Systems Research and Cybernetics, 2003, σσ. 12-17.

18. Michael Anderson and Susan L. Anderson, "Machine Ethics: Creating an Ethical Intelligent Agent", *AI Magazine* 28 (4) (2007): σσ. 15-26.

19. Περισσότερα για τη διαφήμιση του Lexus που γράφτηκε από ΤΝ εδώ: https://www.thedrum.com/news/2018/11/16/lexus-reveals-ad-created-ai-it-gimmick-no-will-it-win-any-awards-probably-not.

20. Περισσότερα για την πρώτη ΤΝ που κρίνει διαφημίσεις εδώ: https://www.adweek.com/creativity/does-pearl-the-first-a-i-ad-awards-juror-know-a-thing-about-creativity/.

21. Περισσότερα για την πρώτη ΤΝ που θα βαθμολογεί αθλητές στους Ολυμπιακούς Αγώνες του 2020 εδώ: http://www.asahi.com/ajw/articles/AJ201812030005.html.

22. Περισσότερα για το μέλλον των πολεμικών επιχειρήσεων με μη επανδρωμένες μηχανές στο: Riza M. Shane, *Killing without Heart: Limits on Robotic Warfare in an Age of Persistent Conflict,* University of Nebraska Press, Potomac Books, 2013, DOI: 10.2307/j.ctt1ddr7mb.

23. Max Tegmark, Life 3.0., Εκδοτικός Οίκος Τραυλός, Αθήνα, 2018, σσ. 177-178.

24. Για μια επισκόπηση της συζήτησης περί της ηθικής ευθύνης στην πληροφορική, βλέπε: Merel Noorman, "Computing and Moral Responsibility", The Stanford Encyclopedia of Philosophy, Edward N. Zalta (επ.), (Spring 2018 Edition), URL = <https://plato.stanford.edu/archives/spr2018/entries/computing-responsibility/>.

25. Μπορούμε να υποστηρίξουμε ότι αναζητώντας ή μιλώντας για ηθική ευθύνη της ΤΝ ή των ρομπότ διαπράττουμε σφάλμα κατηγορίας. Αποδίδουμε ιδιότητες ενός είδους σε ένα άλλο, κάνοντας κακή χρήση της γλώσσας και της λογικής.