Παιδαγωγικός Λόγος

Τόμ. 32, Αρ. 1 (2026)

Λόγος περί της Τεχνητής Νοημοσύνης

Τα συστήματα Τεχνητής Νοημοσύνης και η οντολογική προσέγγιση του Karl R. Popper

Ειρήνη Δαρκαδάκη

doi: 10.12681/plogos.33774

Βιβλιογραφική αναφορά:

Δαρκαδάκη Ε. (2026). Τα συστήματα Τεχνητής Νοημοσύνης και η οντολογική προσέγγιση του Karl R. Popper. *Παιδαγωγικός Λόγος*, *32*(1), 11–29. https://doi.org/10.12681/plogos.33774

Ειρήνη Δαρκαδακη

Τα συστήματα Τεχνητής Νοημοσύνης και η οντολογική προσέγγιση του Karl R. Popper

doi:https://doi.org/10.12681/plogos.33774

Εισαγωγή

Η ταχεια αναπτυξη της Τεχνητης Νοημοσυνης (ΤΝ) εχει μεταβάλλει ριζικά τον τρόπο ζωής και αλληλεπίδρασης των ανθρώπων, εγείροντας βαθιά φιλοσοφικά και ηθικά ερωτήματα. Στο κέντρο της συζήτησης βρίσκεται η δυνατότητα των συστημάτων ΤΝ να θεωρηθούν ηθικά υποκείμενα, δηλαδή οντότητες ικανά να λαμβάνουν ηθικές αποφάσεις και να αναλαμβάνουν ευθύνη για τις πράξεις τους. Αυτή η μελέτη εξετάζει το ζήτημα μέσα από το πρίσμα της οντολογικής θεωρίας των τριών κόσμων του Karl R. Popper, η οποία προσφέρει έναν ιδιαίτερα κατατοπιστικό θεωρητικό πλαίσιο για την κατανόηση της φύσης της ΤΝ.

Ο σκοπός αυτής της εργασίας είναι διττός: αφενός, να αναδείξει τις δυνατότητες και τους περιορισμούς της ΤΝ ως ηθικού παράγοντα μέσω της Ποππεριανής προσέγγισης, και αφετέρου, να προτείνει μία μεταηθική θεώρηση που θα επιτρέπει την καλύτερη κατανόηση των ηθικών προκλήσεων που εγείρονται από την αυτονομία της ΤΝ. Με αυτόν τον τρόπο, η μελέτη στοχεύει να συμβάλει στη συνεχιζόμενη φιλοσοφική και ηθική συζήτηση γύρω από τη θέση της ΤΝ στην ανθρώπινη κοινωνία και τους τρόπους με τους οποίους μπορούμε να διαχειριστούμε τις επιπτώσεις της.

Μια μεταηθική θεμελίωση μέσω της οντολογικής προσέγγισης των τριών κόσμων του Karl R. Popper

Οι Κόσμοι του Karl Popper

Οι ιδέες ή αξίες στην συλλογιστική του Popper δεν υπάρχουν a priori της ανθρώπινης ύπαρξης γιατί εάν συνέβαινε κάτι τέτοιο, υποστηρίζει πως δεν θα μπορούσαμε να τις προσεγγίσουμε[1]. Αυτή η θεμελιώδης διευκρίνιση του φιλοσόφου, τον διαχωρίζει από τις αντίστοιχες πλατωνικές θεωρήσεις και κάνουν φανερή τη ρεαλιστική στάση που υιοθετεί.

Η βασική διάκριση που προτείνει στην οντολογική του θεμελίωση της γνώσης είναι αναγκαίο να αναφερθεί εξαρχής προκειμένου τα όσα ειπωθούν στη συνέχεια καθώς και η απόπειρα ηθικής σύνδεσης με τον κόσμο της τεχνητής νοημοσύνης να γίνουν κατανοητά. Διαχωρίζει λοιπόν το σύμπαν (εάν θα μπορούσαμε να το θέσουμε ως τέτοιο) ή αλλιώς τον κόσμο, σε τρία βασικά επίπεδα[2].

Ο Popper διαχωρίζει τον περιβάλλοντα κόσμο σε «Κόσμος 1= φυσικός κόσμος των φυσικών αντικειμένων», «Κόσμος 2= κόσμος των υποκειμενικών εμπειριών (νοητικός κόσμος/ κόσμος των ψυχικών καταστάσεων)» και «Κόσμος 3= κόσμος των προτάσεων καθεαυτών[3] (εδώ εντάσσονται τα σύνολα προτάσεων, των προϊόντων της γλώσσας εν γένει, της επιστήμης, των μαθηματικών και της φιλοσοφίας αλλά και των τεχνών)[4]. Στον Κόσμο 3 σύμφωνα με τον φιλόσοφο βρίσκεται η αντικειμενική γνώση για τα πράγματα[5]. Πρέπει λοιπόν να καταστεί σαφές πως ο Κόσμος 3 συνδέεται με την αντικειμενική γνώση μέσω της αποθήκευσης και της διάδοσης των γνώσεων που μπορούν να είναι ανεξάρτητες από τις υποκειμενικές εμπειρίες και ψυχικές καταστάσεις (Κόσμος 2). Με αυτόν τον τρόπο, ο κόσμος 3 παίζει κρίσιμο ρόλο στην επιστημονική πρόοδο και την αναζήτηση της

---

[1] Στέλιος Βιρβιδάκης, «Ο *Κόσμος 3,* του Καρλ Πόππερ ως ερμηνευτική πρόταση για την κατανόηση των μοντέλων μετριοπαθούς αξιακού ρεαλισμού», *Φιλοσοφία* 49, νο. ΙΙ (2020): 198-200.

[2] Karl R. Popper, *"Objective knowledge: an evolutionary approach"* (New York: Oxford University Press, 1972), 106-107.

[3] ό.π., 154-157 "… there are three worlds: the first is the physical world, the world of physical states; the second is the mental world or the world of mental states; and the third world is the world of intelligibles, or the world of ideas in the objective sense; it is the world of possible objects of thought.."

[4] Στέλιος Βιρβιδάκης, «Ο *Κόσμος 3* του Καρλ Πόππερ», 194-207.

[5] Karl R. Popper, *"Objective knowledge"*, 106- 107.

αντικειμενικής γνώσης. Ο Popper όπως γίνεται φανερό από τα άνωθι, τοποθετείται με την πλευρά της ρεαλιστικής προσέγγισης του ζητήματος και συνεπώς θεωρεί πως υπάρχουν υπερβατικές ηθικές αξίες (αν όχι τουλάχιστον μία ολιστική) επάνω στις οποίες βασίζονται οι ηθικές επιταγές των πρακτικών των ανθρώπων και δύνανται κατά αυτόν τον τρόπο να οριστούν, και αυτές εντάσσονται στον Κ3[6]. Εκ πρώτης όψεως η ανάλυση και προσέγγιση του ζητήματος φαίνεται να ομοιάζει με την πλατωνική θεωρία των ιδεών αλλά κάτι τέτοιο καθίσταται σαφές από τον ίδιο τον φιλόσοφο πως είναι ανυπόστατο[7]. Προκειμένου αυτός ο διαχωρισμός να γίνει κατανοητός αξίζει αρχικώς να αναφερθεί πως ο Popper δεν αναφέρεται σε αναλλοίωτες ιδέες ή αξίες που υπάρχουν σε μία διαφορετική διάσταση όπως γίνεται στη πλατωνική θεώρηση των ιδεών.

Αντιθέτως, αναζητά και αποπειράται να αποδείξει την ενοποίηση αυτών των φαινομενικά ασύνδετων κόσμων, την ενοποίηση δηλαδή της φυσικής και υποκειμενικής γνώσης του φυσικού κόσμου (Κ1 και Κ2) με τον κόσμο της απόλυτης γνώσης (Κ3). Αυτό, το κατορθώνει υποστηρίζοντας ότι στον Κόσμο 1 θεωρούμε πραγματικό ό,τι *ενεργεί* επάνω σε φυσικά πράγματα ή ό,τι υφίσταται κάποιου είδους ενέργεια από αυτά. Τα νοητικά καθαυτά αντικείμενα του Κόσμου 3 (π.χ. θεωρίες) ασκούν επίδραση στον φυσικό κόσμο και τον αλλάζουν, και άρα, αναγκαία πρέπει να θεωρηθούν πραγματικά παρότι δεν είναι εμπειρικώς αντιλήψιμα. Η σύνδεση αυτών των δύο κόσμων στη θεωρία του Popper επικυρώνεται μέσω του Κόσμου 2 που θα μπορούσε να ερμηνευθεί ως «μεταβατικός κόσμος»[8].

Πιο αναλυτικά, προκειμένου να επιτευχθεί μία ενοποίηση αυτών των τριών ειδών της πραγματικότητας, ο Popper τοποθετείται με σαφήνεια για την ποιότητα και το περιεχόμενο του τρίτου κόσμου που ορίζει. Αναφέρει πως ο Κόσμος 3 ομοιάζει με την αντίστοιχη θεωρία αντικειμενικών ποιοτήτων που περιέχονται στην φιλοσοφική θεώρηση του Frege[9] και άρα θέλει να αποδείξει με αυτή του την απόπειρα πως η επιστημολογία όπως εφαρμόζεται έως σήμερα βασίζει τα πορίσματά της σε υποκειμενικές θεωρήσεις του κόσμου καθώς μελετά συνιστώσες που ο ίδιος εντοπίζει πως βρίσκονται μόνο στον Κ2 και όχι στον Κ3. Ο Κ2 αποτελεί επομένως το σημερινό πεδίο ερευνών της επιστήμης, κάτι το οποίο σύμφωνα με τον φιλόσοφο αποτελεί λανθασμένη προσέγγιση της γνώσης καθότι μόνο στον

[6] ό.π., 158-161.
[7] ό.π., 106/ 154.
[8] Βιρβιδάκης, «Ο Κόσμος 3», 198-199.
[9] Karl R. Popper, *"Objective knowledge"*, 106-107.

Κ3 εμπεριέχονται οι αλήθειες και οι αντικειμενικές θεωρήσεις. Προκειμένου να αιτιολογήσει αυτόν του το διαχωρισμό προχωράει σε μία επιχειρηματολογία βασιζόμενη σε τρεις προκείμενες:

(Π1) : Η παραδοσιακή επιστημολογία κρίνεται ανεπαρκής στην εξήγηση των φυσικών φαινομένων και αληθειών καθώς κατά την μελέτη φυσικών φαινομένων στον Κ1 μεταβαίνουν στον Κ2 για να τα αποδείξουν και όχι στον Κ3 όπου βρίσκεται η αληθινή επιστημονική γνώση[10].

(Π2) : Είναι αναγκαία και υποχρεωτική η μελέτη του Κ3 για την αντικειμενική ανεύρεση των αληθειών που διέπουν τον Κ1 καθώς ο Κ3 αποτελεί έναν αυτόνομο κόσμο εντελώς εκτός της φυσικής πραγματικότητας, όχι όμως με την πλατωνική σημασία της θεωρίας των ιδεών[11].

Για την Π2 διευκρινίζει πως οι επιστήμονες πράττουν βασιζόμενοι σε πεποιθήσεις ή υποθέσεις εις άτοπον απαγωγής οι οποίες όμως εδράζονται σε μία υποκειμενική θεώρηση της πραγματικότητας από τους ίδιους, και όχι στην αντικειμενική ποιότητα των αληθειών που θα τους παρέχονταν αν μελετούσαν τα ίδια αντικείμενα μέσω του Κ3[12].

(Π3) : Η μελέτη του Κ3 αποτελεί κρίσιμης σημασίας ζήτημα καθώς μπορεί να παρέχει μία αντικειμενική σκοπιά θέασης η οποία θα έχει ως αποτέλεσμα να διαφωτίσει πλευρές του κόσμου οι οποίες μάς είναι εντελώς άγνωστες και προς το παρόν δεν έχει ανευρεθεί κάποιος τρόπος εξήγησής τους. Ταυτόχρονα, ο Κ3 θα αιτιολογήσει την υποκειμενική σκοπιά θεώρησης που υιοθετούν οι επιστήμονες (δηλαδή τον Κ2) προσφέροντας μία πιο αντικειμενική προσέγγιση η οποία κατ' επέκταση οδηγεί σε μία πιο έγκυρη εφαρμογή της μεθοδολογίας των επιστημόνων[13]. Η έλευση στον Κ3 μέσω του Κ2, αποτελεί λανθασμένη πεποίθηση και ορίζεται ως ανέφικτη[14].

(Σ) : Η μελέτη του Κ3 θα έπρεπε να αποτελεί αποκλειστικά το κέντρο

[10] Karl R. Popper, *"Objective knowledge"*, "Scientific knowledge", 111.

[11] Στο ίδιο, 1-10/ 106.

[12] Το ερώτημα που προκύπτει από τις άνωθι προκείμενες και αφορά το ζήτημα της παρούσας εργασίας είναι το πώς κανείς αποκτά πρόσβαση στον Κ3 και εάν, τα συστήματα τεχνητής νοημοσύνης μπορούν όπως ακριβώς και στη περίπτωση του ανθρώπου, να έχουν πρόσβαση σε αυτόν ή τουλάχιστον κάποιου είδους σύνδεσης με αυτόν.

[13] Επομένως η αναγκαία συνεπαγωγή που προκύπτει ορίζει επειδή Κ3= Τ (Τ ορίζεται ως True δηλαδή τιμή αληθείας της συνεπαγωγής), τότε ισχύει ότι Κ3→Κ2 και όχι ότι Κ2→Κ3 γιατί η Κ2 προκειμένου να ισχύει προϋποτίθεται να πηγάζει από την Κ3 και όχι το αντίστροφο.

[14] Karl R. Popper, *"Objective knowledge"*, 107.

μελέτης όλων των επιστημόνων και οποιοσδήποτε άλλος τρόπος προσέγγισης αυτού οδηγεί σε πλάνες και υποκειμενικές θεωρήσεις που μας απομακρύνουν από την αλήθεια[15].

Προκειμένου να ενισχύσει τις άνωθι προκείμενες ο Karl Popper κάνει τρεις βασικές επισημάνσεις που αφορούν τον Κ3:

1. Ο Κόσμος 3 αποτελεί παράγωγο της ανθρώπινης οντότητας, το οποίο το παρομοιάζει με τον ιστό της αράχνης[16]. Όπως γύρω από την αράχνη, η οποία βρίσκεται στο κέντρο του ιστού της, ξεδιπλώνονται όλες οι διαστάσεις του ιστού της σε ευρεία κλίμακα και ποικιλία πάχους, έτσι και στη περίπτωση του Κ3, ο άνθρωπος βρίσκεται στο κέντρο και περιβάλλεται από ένα σύμπαν το οποίο έχει παραχθεί από τον ίδιο ακόμη και ο ίδιος ο άνθρωπος δεν έχει επίγνωση αυτού[17].
2. Ο Κόσμος 3 αποτελεί ως επί το πλείστον μία αυτόνομη προέκταση της ύλης η οποία επιδρά στον φυσικό κόσμο και στον άνθρωπο και αντίστοιχα ο φυσικός κόσμος, και κατ' επέκταση ο άνθρωπος, επηρεάζεται και πράττει με κινητήριο δύναμη αυτόν τον ίδιο[18] και μάλιστα στη δεύτερη περίπτωση, ασκεί καταλυτική επίδραση στον

[15] Προκειμένου να υποστηρίξει το συμπέρασμά του και την επιδραστικότητα αλλά και σύνδεση του Κ3 στον Κ1 προχωρεί στην εξής εγκυροποίηση: "In our attempts to solve these other problems we may invent new theories. These theories, again, are produced by us: they are the product of our critical and creative thinking (K2), in which we are greatly helped by other existing third-world theories. Yet the moment we have produced these theories, they create new, unintended and unexpected problems, autonomous problems, problems to be discovered." Karl R. Popper, *"Objective knowledge"*, 161.

[16] Στο ίδιο, 112.

[17] Η ανάλυση που έχει γίνει μέχρι αυτό το σημείο γεννά το βασικό ερώτημα κατά πόσον η οντολογική θεώρηση του Popper προσφέρει ένα εύφορο έδαφος προκειμένου να ανακαλυφθεί μία αντίστοιχης υφής σύνδεση των συστημάτων ΤΝ και του Κ3. Εκ πρώτης όψεως η θεώρηση αυτή φαίνεται ατελέσφορη για αυτήν την απόπειρα, όμως δημιουργεί ένα έδαφος αναλογικού συλλογισμού πάνω στο οποίο μπορούμε να στηρίξουμε μία θεωρία αντίστοιχης φύσεως που να αφορά τα συστήματα ΤΝ αλλά να μη σχετίζεται με τον Κ3, να δρα όμως παράλληλα και σε αντιδιαστολή με αυτόν. Αυτό είναι και το δεύτερο σημείο εστίασης της παρούσας εργασίας. Ο λόγος λοιπόν για τον οποίο χρησιμοποιείται αυτή η θεώρηση στη παρούσα εργασία είναι για να αναδείξει την διαφορετική λειτουργία των συστημάτων ΤΝ από τους ανθρώπους και να χρησιμοποιηθεί με τρόπο αναλογικό ώστε να δημιουργηθεί ένα νέο οντολογικό γίγνεσθαι στο οποίο θα μπορούν να αποδοθούν ηθικές ποιότητες σε αυτά τα συστήματα, αντίστοιχες με αυτές του ανθρώπου.

[18] Karl R. Popper, *"Objective knowledge"*, 112-113.

φυσικό κόσμο (Κ1) ώστε να μπορεί να υφίσταται με τον τρόπο ακριβώς που συμβαίνει.

Ενώ φαίνεται από τη θεώρηση του Popper πως ο Κ3 παράγεται από τον άνθρωπο, τον ορίζει ως αυτόνομο, και η αιτία αυτού, εδράζεται στο γεγονός ότι ο Κ3 περιέχει το πεδίο της καθαρής γνώσης κάτι που του δίνει το χαρακτηριστικό της αυτονομίας. Προς απόδειξη αυτού του του επιχειρήματος αναφέρει τα βιβλία και την ανάγνωσή τους από διαφορετικούς ανθρώπους[19]. Ο κάθε άνθρωπος διαβάζοντας ένα βιβλίο θα δώσει μία διαφορετική ερμηνευτική προσέγγιση σε αυτό που διάβασε αυτό όμως δεν αναιρεί την καθαρή πληροφορία γνώσης στην οποία βασίζεται το βιβλίο. Όπως στην περίπτωση των βιβλίων, έτσι και στη περίπτωση του τρίτου κόσμου, μπορεί να αποτελεί παραγόμενο προϊόν της ανθρώπινης ύπαρξης αλλά εντός αυτού υπάρχουν θεωρίες, αλήθειες, γνώση εν γένει οι οποίες δεν βασίζονται στην παραγωγή τους από τον άνθρωπο και μπορούν ποτέ να μην γίνουν αντιληπτές ή κατανοητές από αυτόν. Ένα ακόμη τέτοιο παράδειγμα απόδειξης της αυτονομίας του Κ3 αποτελεί η ίδια η γλώσσα καθαυτή ή η θεωρία των φυσικών αριθμών.

3. Μόνο μέσω της αλληλεπίδρασής των ανθρώπων με τον Κ3 μπορεί η αντικειμενική γνώση να ευδοκιμήσει ενώ ταυτόχρονα μέσω αυτής δύναται να αναπτύσσεται και ο Κ1. Επί παραδείγματι, αυτό εντοπίζεται στη βιολογική εξέλιξη των φυτών και των ζώων[20].

Αυτή η τρίτη διευκρίνιση δημιουργεί ένα πρόσφορο έδαφος ώστε να εξεταστεί το εάν ένα σύστημα τεχνητής νοημοσύνης μπορεί να αλληλεπιδρά και να επηρεαστεί από τον Κ3. Ο φιλόσοφος σε μία απόπειρα να στηρίξει την αυτονομία του Κ3 προβαίνει σε μία επιχειρηματολογία βασιζόμενος στην επιστήμη της βιολογίας. Αρχικώς αναφέρει πως η επιστήμη της βιολογίας μελετά κυρίως α) την συμπεριφορά και φυσική δομή των έμβιων όντων και β) κάποιες από τις μη έμβιες οντότητες που προκύπτουν από τα φυτά και τα ζώα όπως τις φωλιές τους ή τα μονοπάτια που χαράζουν στο δάσος[21]. Η οπτική του Popper εστιάζει στην ανάδειξη της ανάπτυξης της

[19] Στο ίδιο, 116-117.
[20] Στο ίδιο, 112-113.
[21] ό.π.

γνώσης στο πλαίσιο της εξέλιξης της έμβιας ζωής[22]. Τα προβλήματα που εγείρονται από αυτόν τον τρόπο προσέγγισης του αντικειμένου μελέτης των βιολόγων, είναι κυρίως δύο:

Α) Τα ζητήματα που προκύπτουν από τα αποτελέσματα των πράξεων των έμβιων όντων και

Β) Τα ζητήματα που προκύπτουν εάν ληφθούν υπόψη οι δομές των έμβιων όντων καθαυτές[23].

Για την ανθρωπότητα τα Α) και Β) μπορούν αναλογικά να τεθούν στην γλώσσα και την επιστήμη. Αυτά τα ζητήματα ο Popper υποστηρίζει πως είναι αληθή και εντοπίζονται και για τους ανθρώπους: η ανθρωπότητα επίσης έχει δημιουργήσει νέα είδη που εμφανίζονται στον K1 ή στον K2 και τα ορίζει ως «πνευματικά προϊόντα» (intellectual products) τα οποία δομούν το περιβάλλον του ανθρώπου. Τέτοια ορίζονται ως οι μύθοι, τα λογοτεχνικά βιβλία, τα έργα τέχνης ή επιστημονικές θεωρίες που έχουν διατυπωθεί και ισχύουν για τον εμπειρικό κόσμο. Όταν αναφερόμαστε σε μία θεωρία εξελικτικού περιεχομένου αυτά τα πνευματικά προϊόντα *πρέπει* να λαμβάνονται ως αντικείμενα μίας πραγματικότητας *έξω* από εμάς, και μαζί με αυτά εντοπίζεται και η αληθινή γνώση (knowledge)[24]. Επομένως

[22] "…Η προσέγγιση του συγκεκριμένου ζητήματος είναι να τοποθετήσει τον τρόπο ανάπτυξης της καθαρής γνώσης στο πλαίσιο της εξέλιξης των ζώων και του ανθρώπου", Stephen, Thornton "Karl Popper" στο *The Stanford Encyclopedia of Philosophy,* επιμ.: Edward N. Zalta & Uri Nodelman, Winter, 2022, δική μου μετάφραση. https://plato.stanford.edu/archives/win2022/entries/popper/.

[23] Σύμφωνα με τη συλλογιστική του Karl Popper τα ζητήματα του Α) μπορούμε να πούμε ότι εντάσσονται στον K1, ενώ οι δομές που αναφέρονται στο Β) εντάσσονται στον K2. Παρόλα αυτά, είναι δυνατόν να εντοπίσει κανείς πως οι δομές που αναφέρονται στο Β) αποτελούν προϊόντα του εμπειρικού κόσμου (π.χ. η δομή και η λειτουργία ενός κυττάρου) και άρα με βεβαιότητα θα μπορούσε κάποιος να συμπεράνει πως εντάσσονται στον K1. Τα πορίσματα όμως που βγαίνουν μέσω της παρατήρησης των λειτουργιών των δομών αυτών καθώς και οι γενικεύσεις που προκύπτουν μέσω αυτών είναι αναμφιβόλως προϊόντα του K2 και όχι του K1. Άρα εν τέλει μπορούμε να συμπεράνουμε πως το Β) είναι μία παράμετρος που μπορεί να ενταχθεί και στον K1 αλλά και στον K2 ίσως και ταυτόχρονα.

[24] "Αυτό, ο Popper αναφέρει, είναι αληθές και για την περίπτωση των ανθρώπων: εμείς επίσης έχουμε δημιουργήσει νέα είδη προϊόντων, «πνευματικά προϊόντα», τα οποία δομούν το περιβάλλον μας. Τέτοια είναι οι μύθοι, οι ιδέες μας, τα προϊόντα της τέχνης μας, και οι επιστημονικές μας θεωρήσεις για τον κόσμο στον οποίο ζούμε. Όταν αυτή η σκέψη τοποθετείται στο εξελικτικό πλαίσιο, ο Popper προτείνει αυτού του είδους τα προϊόντα πως είναι αναγκαίο να συλλαμβάνονται οργανικά, ως εξωσωματικά κατασκευαστικά προϊόντα. Η ενοποιός δύναμή τους είναι η γνώση", Stephen, Thornton "Karl", δική μου μετάφραση.

τα αποκυήματα της φαντασίας του ανθρώπου ο Popper τα κατατάσσει στον Κ3 και τα οποία μέσω του Κ2 συνδέονται με μία σχέση αποβλεπτικότητας[25] στον Κ1.

Σύμφωνα με τον φιλόσοφο είναι σημαντικότερο να αναζητηθεί η λύση του ζητήματος των δομών καθεαυτές (Β), γιατί εάν εξηγηθεί αυτό, θα έχει δοθεί μία αντικειμενική λύση και στα συμπεριφορικά ζητήματα που αναφέρονται στο Α). Η αναγωγή λοιπόν στη Β περίπτωση πρέπει αναγκαία και ικανά να γίνει στον Κ3 ώστε να μπορούν να εξηγηθούν οι δομές των έμβιων όντων καθαυτές με έναν όσο το δυνατόν πιο αντικειμενικό τρόπο. Αυτή η θέση μπορεί να οριστεί ως το τρίτο ζήτημα (Γ) που κατέχει αντισυμπεριφοριστική χροιά[26].

Επομένως, αυτό που αποτελεί καθοριστικής σημασίας στην αιτιακή σχέση ανάμεσα στους κόσμους του Popper είναι πως με αυτή τη προσέγγιση επιτρέπεται στον φιλόσοφο να αναγάγει την ανάπτυξη και εξέλιξη της ανθρώπινης γνώσης ως μία εξελικτική διαδικασία με εξωσωματικές προσαρμογές[27] οι οποίες τελικώς αποτελούν μία λειτουργική διαδραστική διαδικασία ανάμεσα στη σχέση του Κόσμου 1 (Κ1) και νοητικού κόσμου (Κ2)ενώ ταυτόχρονα επιτυγχάνεται η ίδια διάδραση και με τον κόσμο της αντικειμενικής γνώσης (Κ3) ή αλλιώς του περιεχομένου της σκέψης[28].

---

[25] Ο όρος "αποβλεπτικότητα" αναφέρεται στην ιδιότητα του νου να κατευθύνεται προς "κάτι", πράγματα, αντικείμενα (πραγματικά ή φανταστικά), γεγονότα, καταστάσεις, σχέσεις και ιδέες του κόσμου, τα οποία είναι ως επί το πλείστον εξωτερικά (Κ1), δηλαδή "εκτός" του ίδιου του νοήμονος όντος που νοεί τον κόσμο (Κ2). Για παράδειγμα, όταν σκεφτόμαστε, σκεφτόμαστε κάτι, όταν αποφασίζουμε, αποφασίζουμε κάτι, όταν θέλουμε, θέλουμε κάτι, και ούτω καθεξής. Ο σχηματισμός ενός συγκεκριμένου νοητικού περιεχομένου, ενός νοήματος (Κ3), αποτελεί θεμελιώδες χαρακτηριστικό αυτής της ιδιότητας. Βλέπε: Gounaris, A. (2011). Intentionality and the Emergence of Meaning. Philosophia - Annual Journal of the Research Centre for Greek Philosophy of the Academy of Athens, v.41, pp 319-321, 2011. ISSN 1105-2120

[26] Karl R. Popper, *"Objective knowledge"*, 114-115.

[27] Με τον όρο εξωσωματικές προσαρμογές, ορίζω την αλληλεπίδραση των νοητικών γνώσεων με των εμπειρικών δεδομένων. Τα εμπειρικά δεδομένα, βρίσκονται εκτός της ανθρώπινης σκέψης και νοητικής διαδικασίας όμως παρόλα αυτά, μπορούν να ανατρέψουν θεωρητικά δεδομένα που υπάρχουν στο νου, και να τα αναδομήσουν ή να τα αναπροσαρμόσουν. Επομένως οι εξωσωματικές προσαρμογές αναφέρονται σε διαδικασίες που επιτελούνται εκτός της νοητικής σφαίρας του εαυτού.

[28] "Σε τελική ανάλυση, αυτό που αφορά και αναφέρεται η οντολογική επιστημολογία του Popper, είναι η αιτιακή αλληλεπίδραση ανάμεσα στους κόσμους: αυτή του επιτρέπει να αντικατοπτρίσει την εξέλιξη της ανθρώπινης γνώσης ως μια εξελικτική διαδικασία με εξωσωματικές προσαρμογές η οποία εν τέλει αποτελεί ένα παιχνίδι μεταξύ των σχέσεων

Η ηθική θεμελίωση των συστημάτων Τεχνητής Νοημοσύνης και η κριτική τους μέσω των Κόσμων του Karl R. Popper

Όπως περιγράφεται στην προηγούμενη ενότητα ο τρόπος διαχωρισμού των κόσμων από τον Karl R. Popper, μπορεί να πραγματοποιηθεί και για τις ηθικές αλήθειες πάνω στις οποίες καλούμαστε να ανακαλύψουμε εάν μπορούν να υφίστανται στα συστήματα Τεχνητής Νοημοσύνης (ΤΝ). Τα συστήματα τεχνητής νοημοσύνης καθότι ενεργούν στον (Κ1), έχουν την ικανότητα να δρουν αποκλειστικά με γνώμονα αυτόν. Επομένως, είναι αναγκαίο να αναφερθεί πως σε πρώτο στάδιο πριν τη δημιουργία τους, ήταν νοητικά φαινόμενα (Κ2), και η ύπαρξή τους ως μέρη του φυσικού κόσμου βασίστηκε σε θεωρίες και επιχειρήματα που σύμφωνα με τον Popper πηγάζουν και κατατάσσονται στον Κ3 και άρα αποτελούν προϊόντα αυτού[29].

Οι ηθικές αλήθειες, ενώ ανήκουν στον Κ3, έχουν αντίστοιχη θέση στον φυσικό κόσμο Κ1, καθώς η αποβλεπτικότητά τους μπορεί να εντοπιστεί εκεί. Αρχικώς, ως σύστημα θεωριών, αρχών και νομοθέτησης, αποτελούν μία νοητική διαδικασία που εδράζεται στον Κ2 και η οποία βασίζεται στον αντικειμενικό και καθολικό χαρακτήρα της Κ3[30]. Σε συνάρτηση με τα παραπάνω και με τα εμπειρικά φαινόμενα, η αρμονική συνύπαρξη των έμβιων όντων βασίζεται σε κανονιστικές ηθικές αρχές *έξω* από τον αισθητό κόσμο, οπότε η παρούσα εργασία εξετάζει το ζήτημα των ηθικών αρχών των αυτόνομων οπλικών συστημάτων από μία ηθικώς ρεαλιστική σκοπιά. Επιπρόσθετα σύμφωνα με τον Hage, οι προθέσεις και οι επιθυμίες μας έχουν αναγκαία υπόσταση και ύπαρξη, ανεξαρτήτως της αποβλεπτικότητάς τους στον φυσικό κόσμο[31]. Αυτό τεκμηριώνεται, εάν σκεφτούμε πως οι προθέσεις ή επιθυμίες μας αποτελούν βασικά κίνητρα για τις πράξεις μας ή τη δομή του συναισθηματικού μας κόσμου και της ιδιοσυγκρασίας μας. Μπορεί όμως ένα αυτόνομο οπλικό σύστημα να διαθέτει αυτά τα

ανάμεσα στον φυσικό και νοητικό/πνευματικό κόσμο με τον κόσμο της αντικειμενικής γνώσης ή το περιεχόμενο της σκέψης", Stephen, "Karl", δική μου μετάφραση.

[29] Ο Κ3 περιλαμβάνει τα πολιτιστικά δημιουργήματα, τις γνώσεις, τις ιδέες, τις επιστημονικές θεωρίες, τα μαθηματικά κατασκευάσματα, και άλλα προϊόντα της ανθρώπινης νόησης και διάνοιας τα οποία ενσωματώνονται στα συστήματα ΑΙ. Ένα σύστημα είναι αποτέλεσμα τεχνολογικών και επιστημονικών γνώσεων. Ναι μεν τα συστήματα αυτά σχεδιάζονται και χρησιμοποιούνται για να εκτελούν συγκεκριμένες λειτουργίες στον Κ1, όμως ενσωματώνουν τον Κ3.

[30] Επομένως και για την περίπτωση των συστημάτων τεχνητής νοημοσύνης και για την περίπτωση των ηθικών αληθειών ισχύει η ίδια συνεπαγωγή: επειδή Κ3= Τ τότε ισχύει ότι: Κ3→Κ2→Κ1.

[31] Hage, *Theoretical Foundations*, 259.

χαρακτηριστικά; Είναι λοιπόν δυνατό, να υποθέσουμε πως τα συστήματα ΤΝ τα οποία έχουν αναπτυχθεί και στα οποία δεν υπάρχει ανθρώπινη παρέμβαση για τη λειτουργία τους (ή έστω είναι σε ελάχιστο βαθμό), διαθέτουν ή μπορούν να αναπτύξουν την ιδιότητα της ελεύθερης βούλησης και σύμφωνα με αυτή να επιτελέσουν ηθικές πράξεις και αποφάσεις[32]; Εάν υποθέσουμε κάτι τέτοιο, προκύπτει πως τα συστήματα ΤΝ θεωρούνται από τους ανθρώπους ως έλλογα όντα με ηθικό καθεστώς και επομένως διαθέτουν τη δυνατότητα αξιολόγησης και ενδελεχούς επεξεργασίας των προθέσεών τους προτού πράξουν. Άρα είναι ικανά να αξιολογούν ηθικά και να επιλέγουν σύμφωνα με τον Ηθικό Νόμο το πώς θα πράξουν. Το συμπέρασμα αυτό προκύπτει μέσω διαφόρων πειραμάτων που έχουν πραγματοποιηθεί ως σήμερα, και αποδεικνύουν πως τα συστήματα ΤΝ διαθέτουν την δυνατότητα να επιλέξουν έναν διαφορετικό τρόπο συμπεριφορικής λειτουργίας από αυτόν που ήταν προγραμματισμένα από τους επιστήμονες να τελέσουν[33].

Προς επίρρωση του συγκεκριμένου επιχειρήματος απόδοσης ηθικής ευθύνης στα συστήματα ΤΝ, επικαλούμαι μία αναγκαία συνεπαγωγή όπως ορίστηκε από τον αναλυτικό φιλόσοφο Thomas Nagel, εάν συνδυάσουμε την αντικειμενική σκοπιά θεώρησης με την πράξη. Σύμφωνα με τη θεωρία του Thomas Nagel, εάν αποδεχθώ ότι ισχύει ότι *A*, τότε αναγκαία αποδέχομαι ότι ισχύει και *B*[34]. Στην περίπτωση λοιπόν των συστημάτων ΤΝ, εάν

---

[32] Κάτι τέτοιο είναι λογικό να μπορούμε να το υποθέσουμε μέσω των πειραμάτων που τελούνται με συστήματα τεχνητής νοημοσύνης και της αποβλεπτικότητας αυτών στον εμπειρικό κόσμο.

[33] Ένα παράδειγμα για αυτές τις περιπτώσεις αποτελούν τα παιχνίδια στρατηγικής (σκάκι ή AlphaGo-κορεάτικο παιχνίδι στρατηγικής παρόμοιο με το σκάκι-) και τα πειράματα που έχουν γίνει με την χρήση συστημάτων ΤΝ Περισσότερες πληροφορίες για αυτά προκύπτουν αν κάποιος ανατρέξει στο ντοκιμαντέρ AlphaGo: https://www.youtube.com/watch?v=WXuK6gekU1Y .

[34] Thomas Nagel, *Η Θέα από το πουθενά*, (Αθήνα: Κριτική, 2002), 234-235. Ο Nagel σε αυτό του το βιβλίο προσπαθεί να αποδείξει τη σύνδεση της υποκειμενικής θεώρησης με την αντικειμενική θεώρηση της πραγματικότητας και πραγματεύεται ποικίλους στοχασμούς είτε από την υποκειμενική θεώρηση του κόσμου είτε προσπαθώντας να προσεγγίσει μία αντικειμενική σκοπιά των πραγμάτων. Κατά αυτόν τον τρόπο, δημιουργεί την αναφερθείσα αναγκαία (σύμφωνα με το συλλογισμό του) συνεπαγωγή, με σκοπό στο συγκεκριμένο κεφάλαιο του βιβλίου του να αποδείξει την ταυτότητα του εαυτού. Τα Α και Β επομένως μπορούν να εφαρμοστούν σε οτιδήποτε έχει αποβλεπτικότητα στον φυσικό κόσμο και παρατηρείται μία αναγκαία αλληλεπίδραση μεταξύ τους η οποία δεν θα μπορούσε να είναι αλλιώς. Προς περεταίρω επεξήγηση μπορούν να αναφερθούν παραδείγματα από τα μαθηματικά όπως ότι αν δεχθώ ότι ισχύει ότι 2+2 ισούται με 4 τότε το άθροισμα αυτών των δύο φυσικών αριθμών δεν μπορεί να μας δώσει τίποτα άλλο πέρα

αποδεχθώ πως διαθέτουν ηθικό καθεστώς (*A*), τότε αποδέχομαι και το γεγονός της απόδοσης ηθικής ευθύνης σε αυτά για τις πράξεις τους (*B*)[35]. Εάν δεχθώ το παραπάνω, το πραγματώνω είτε α) για να είμαι έτοιμος ως έλλογο ον να δεχθώ τις συνέπειες των πράξεων που αυτά θα επιτελέσουν σε περίπτωση που αυτό κρίνεται αναγκαίο, είτε β) για να δεχθώ πως τα συστήματα αυτά που φαίνεται να διαθέτουν την ιδιότητα της ελεύθερης βούλησης, θα αναλάβουν την πλήρη αποδοχή των συνεπειών των πράξεών τους, επιβλαβών ή μη. Στη β) περίπτωση τα θεωρώ έλλογα όντα με αναφορά στον ηθικό νόμο για πράξη και άρα υπεύθυνα για τις πράξεις που αυτά επιτελούν.

Σε αυτή τη περίπτωση (β), στην οποία θεωρούνται *άξια* ανάληψης ευθυνών, αναγκαία θεωρούνται και *ικανά* για λήψη αποφάσεων με ηθικό καταλογισμό. Και στις δύο όμως περιπτώσεις (α και β) θα πρέπει να αναλογιστούμε και να είμαστε έτοιμοι για όλες τις αρνητικές εκβάσεις των πράξεών τους ακόμη και για εκείνες που δεν δυνάμεθα να φανταστούμε ότι θα πραγματώσουν σε μελλοντικό χρόνο. Το συγκεκριμένο επιχείρημα καθιστά σαφή τη σημαντικότητα του χρόνου σε αυτή τη μεταηθική θεμελίωση των αυτόνομων οπλικών συστημάτων, καθώς επίσης και το γεγονός πως η συνιστώσα του χρόνου[36] θα πρέπει να λαμβάνεται υπόψη ως παράγοντας επηρεασμού λήψης μίας απόφασης ενός συστήματος ΤΝ.

από 4. Μία τέτοια θεωρία μπορεί να επιβεβαιωθεί μέσω παραδειγμάτων στον φυσικό κόσμο είτε βρίσκομαι στην υποκειμενική, είτε στην αντικειμενική σκοπιά, με αποτέλεσμα να συμπεράνω πως είναι ένα παράδειγμα απόδειξης μίας θεωρίας που ισχύει καθολικά.

[35] ό.π.

[36] Όταν αναφέρομαι στην έννοια του χρόνου εννοώ το χρονικό διάστημα στο οποίο το σύστημα τεχνητής νοημοσύνης αποφασίζει a priori το πώς θα πράξει, την χρονική στιγμή της τέλεσης της πράξης, τις επιπτώσεις που προκύπτουν από αυτή -βραχυπρόθεσμες ή μακροπρόθεσμες- ακόμη και το πότε το ίδιο το σύστημα τεχνητής νοημοσύνης δημιουργήθηκε κάτι το οποίο ορίζει το πόσο εξελιγμένο είναι. Η παρούσα εργασία γράφεται αυτή τη χρονική περίοδο, κατά την οποία η τεχνολογική εξέλιξη αυτών των συστημάτων διαρκώς μεταβάλλεται και εξελίσσεται. Παράλληλα με αυτό μεταβάλλεται και εξελίσσεται ο τρόπος με τον οποίο μπορεί να αποδοθεί ηθικός καταμερισμός σε αυτά τα συστήματα. Επομένως, τίποτα δεν είναι απόλυτο και σίγουρο από τη στιγμή που υπάρχει μία διαρκής και συνεχόμενη μεταβολή και για αυτό διευκρινίζεται πως ο αστάθμητος παράγοντας του χρόνου δύναται να μεταποιήσει τις ηθικές συνιστώσες που παρουσιάζονται εδώ, σε σχέση με τα συστήματα τεχνητής νοημοσύνης.

Το ηθικό καθεστώς των συστημάτων Τεχνητής Νοημοσύνης

Στο παρόν κεφάλαιο θα πραγματοποιηθεί ένας ορισμός της ηθικής ευθύνης όπως αυτή εμφανίζεται στις ανθρώπινες κοινωνίες, καθώς επίσης και το εάν τα συστήματα Τεχνητής Νοημοσύνης (ΤΝ) μπορούν να θεωρηθούν όντα με ηθικό καταλογισμό. Βασική προϋπόθεση για αυτή την απόπειρα αποτελεί μία στοιχειώδης αναφορά στους βαθμούς ανθρώπινης παρέμβασης των συστημάτων ΤΝ.

Βαθμοί ανθρώπινης παρέμβασης και τα συστήματα ΤΝ

Τα αυτόνομα συστήματα επιδέχονται βαθμούς ανθρώπινης παρέμβασης με τρεις τρόπους: human in the loop[37], human on the loop[38] και human off the loop[39]. Εκτός από την περίπτωση human off the loop, υπάρχει η σκέψη και η δυνατότητα απόδοσης ηθικής ευθύνης σε ένα αυτόνομο σύστημα και στη περίπτωση human on the loop[40]. Θεωρείται πιθανό και ίσως πιο εύλογο να μπορεί να αποδοθεί η ευθύνη στα αυτόνομα οπλικά

[37] Human in the loop= η οποιαδήποτε εντολή επιδιώξει η μηχανή λόγω του αυτόνομου συστήματος που διαθέτει, να πράξει, χρειάζεται την έγκριση του ανθρώπου πριν τη πραγματώσει. Ο άνθρωπος επιτελεί την ενεργό δρώσα οντότητα, και η μηχανή απλώς εκτελεί τις εντολές του

[38] Human on the loop= μεγαλύτερη αυτονομία της μηχανής, ο άνθρωπος δίνει την πρωταρχική εντολή και επιβλέπει.

[39] Human off the loop= Οι αξιολογήσεις, εντολές και οι εκτελέσεις, βασίζονται αποκλειστικά στη μηχανή η οποία αποτελεί τη δρώσα οντότητα, και ο άνθρωπος είναι εντελώς αποσυνδεδεμένος από οποιαδήποτε λειτουργία ή απόφασή της.

[40] Ποιος αναλαμβάνει την ευθύνη μιας λάθος εκτίμησης ενός αυτόνομου οπλικού συστήματος; Μπορεί να αποδοθεί ευθύνη στο ίδιο το σύστημα; Ή θα πρέπει να αποδοθεί στον κατασκευαστή, στον προγραμματιστή, ή ακόμη και στον στρατιωτικό που επέλεξε να θέσει αυτό το όπλο στη μάχη; Ακόμη και στη περίπτωση του human on the loop, είναι ηθικώς θεμιτό να προσδοθεί ηθικός καταμερισμός της ευθύνης στον εντολέα (άνθρωπος) ο οποίος όμως δεν έχει ευθύνη για τις μη ορθές σύμφωνα με μία κανονιστική ηθική θεωρία πρακτικές του αυτόνομου οπλικού συστήματος που χρησιμοποίησε, καθότι το δεύτερο, λειτούργησε με βάση τις προσλαμβάνουσες που αξιολόγησε με έναν δικό του τρόπο (black box paradox); Για το συγκεκριμένο πρόβλημα βλέπε: Γούναρης Ά., & Κωστελέτος Γ. (2024). Γράφοντας τον αλγόριθμο του Καλού: Η Τεχνητή Νοημοσύνη ως μηχανή απόδοσης Δικαιοσύνης. Ηθική. Περιοδικό φιλοσοφίας, (19). https://doi.org/10.12681/ethiki.39654
Επίσης: Gounaris, A., Kosteletos, G. (2020). Licensed to Kill: Autonomous Weapons as Persons and Moral Agents. In Prole, D. and Rujević, G. (ed.). Personhood. Novi Sad, Filozofski Fakultet & The NKUA Applied Philosophy Research Lab Press. DOI: https://doi.org/10.12681/aprlp.49

συστήματα και στις δύο περιπτώσεις καθώς, είτε οι κατασκευαστές, είτε οι προγραμματιστές, κρίνεται αδύνατο και απολύτως παράλογο να επωμιστούν εξολοκλήρου την ευθύνη για πράξεις που έχουν παρθεί εξολοκλήρου από το ίδιο το σύστημα[41].

Σε αυτή τη περίπτωση τα συστήματα τεχνητής νοημοσύνης που χρησιμοποιούνται σε πολεμικές συρράξεις κρίνεται αναγκαίο να θεωρηθούν έλλογα όντα καθώς μόνο μέσω αυτής της κατηγοριοποίησης θα είναι δυνατό να αναλάβουν ευθύνη και να επωμιστούν τις συνέπειες για τις πράξεις τους. Πώς όμως μπορεί να τους αποδοθεί ηθική ευθύνη[42]; Η φύση του ανθρώπου και ο τρόπος λειτουργίας του, τον καθιστούν ικανό για ηθικές πράξεις ή μη, σχεδόν *a priori*, βασιζόμενες στις κοινωνικές του πρακτικές και αλληλεπιδράσεις με τους άλλους ανθρώπους[43]. Αυτό, έχει ως αποτέλεσμα, η ηθικές του αποφάσεις και τα θεμέλια της ηθικής του σύμφωνα με τα οποία πράττει, να διαφοροποιούνται, ή (σε πρώτο στάδιο) να δομούνται και εξαιτίας εξωγενών παραγόντων.

Το ηθικό καθεστώς στις ανθρώπινες κοινωνίες και τα συστήματα ΤΝ

Οι άνθρωποι θεωρούνται υπεύθυνοι και τους προσδίδεται ένα ηθικό καθεστώς εξαιτίας του γεγονότος ότι αποκτούν ένα συγκεκριμένο ρόλο κοινωνικά. Η αιτιακή ευθύνη αποτελεί μία έκφανση αυτής, και συνδέεται άρρηκτα με το αν αποτελεί ο ίδιος ο πράττων την αιτία ενός αποτελέσματος. Σε αυτή τη περίπτωση υπάρχει άμεση σύνδεση αιτίου-αποτελέσματος. Μια πράξη ανατίθεται σε έναν αυτουργό, ο οποίος μετέπειτα θα αναλάβει την ευθύνη του αποτελέσματος αυτής του της πράξης. Το άτομο, αποτελεί υψίστης σημασίας καθώς είναι η αιτία του παραχθέντος αποτελέσματος, και πρέπει να είναι αξιόπιστο σε δύο τομείς: α) στο να μπορεί να επιτελέσει την πράξη με αξιοπιστία, και β) να μπορεί να αναλάβει την ευθύνη για το

[41] Jaap Hage, "Theoretical foundations for the responsibility of autonomous agents", *Springer* (August 2017): 255.

[42] "...παραμένει αδιευκρίνιστο σε τι βαθμό η αυτονομία ενός οπλικού συστήματος είναι ή μπορεί να θεωρηθεί πως πράττει σύμφωνα με τις πληροφορίες που του έχουν εμφυτεύσει. Μία αιτία αυτής της έλλειψης γνώσης από την οποία προκύπτει αυτό το ερώτημα, εδράζεται στο βαθμό αυτονομίας, κάτι που μπορούμε να ισχυριστούμε και για τους ανθρώπους εάν σκεφτούμε την έλλειψη κινήτρων, εδράζεται στη φύση και στο μέγεθος της αυτονομίας κάτι το οποίο αποτελεί από μόνο του ένα αμφιλεγόμενο και δυσκόλως κατανοητό ζήτημα." Robert Sparrow, "Killer Robots", *Journal of Applied Philosophy Vol. 24,* no. 1 (2007): 65, δική μου μετάφραση.

[43] Hage, "Theoretical foundations", 256.

αποτέλεσμα της πράξης του[44].

Η ικανότητα ανάληψης της ευθύνης του αποτελέσματος μίας πράξης, εγείρει το θεμελιώδες ερώτημα για το αν τα συστήματα ΤΝ είναι δυνατόν να διαθέτουν αυτή την ικανότητα (έστω στο μέλλον αν όχι τώρα) όπως αντιστοίχως συμβαίνει στις ανθρώπινες κοινωνίες. Δηλαδή, εάν έχουν ή μπορούν να αποκτήσουν ηθικό καθεστώς. Φυσικά, και στην περίπτωση των έλλογων όντων και στη περίπτωση των συστημάτων ΤΝ, το πότε και το εάν μία πράξη ή η αποδοχή της ευθύνης αυτής θεωρείται ηθικώς θεμιτή, μπορεί να εξεταστεί και αξιολογηθεί από εντελώς διαφορετικές οπτικές κάτι το οποίο μπορεί να αλλάξει την έκβαση μίας απόφασης.

Η βασική αιτία για αυτό, εντοπίζεται στο γεγονός πως οι άνθρωποι πράττουν εκ προθέσεως κινητοποιημένοι από τις προσταγές της ελεύθερης βουλήσεώς τους να πράξουν κατά το δοκούν[45]. Αντίθετα, τα συστήματα ΤΝ φαίνεται πως δεν διαθέτουν προδιαθέσεις για πράξεις όμοιες με αυτές του ανθρώπου και εξαιτίας αυτού, σε καμία περίπτωση προς το παρόν, δε μπορεί να θεωρηθεί πως διαθέτουν ελεύθερη βούληση[46] ακόμη και εάν προγραμματιστούν με σκοπό να αναπτύξουν μία που να ομοιάζει με ελεύθερη βούληση[47]. Η συμπεριφορά αυτών, βασίζεται στους ίδιους κανόνες με αυτούς των ανθρώπων σύμφωνα με την αποβλεπτικότητα αυτής στον εμπειρικό κόσμο. Ένα σύστημα τεχνητής νοημοσύνης το οποίο είναι προγραμματισμένο να επιτελεί εξολοκλήρου αυτόνομα τις πράξεις του, είναι δυνατόν εκ των αποτελεσμάτων αυτών να φανεί πως πράττει ορθά, και στις περισσότερες περιπτώσεις να μην είναι διακριτό το εάν πρόκειται για αποτελέσματα ανθρώπινων πράξεων ή ενός συστήματος τεχνητής νοημοσύνης[48].

Τα συστήματα ΤΝ δομούν τις πράξεις τους σύμφωνα με τις αλγοριθμικές αλληλουχίες που επιτελούν και οι οποίες βασίζονται στον τρόπο με τον οποίο ο άνθρωπος έχει κατασκευάσει/προγραμματίσει αυτά τα

[44] Hage, "Theoretical foundations", 257.

[45] οι οποίες για να θεωρηθούν ηθικές θα πρέπει να είναι σύμφωνες με τον ηθικό νόμο.

[46] Hage, "Theoretical foundations", 258.

[47] Αιτία αυτού είναι το γεγονός πως δεν μπορούμε με απόλυτη βεβαιότητα να γνωρίζουμε εάν διαθέτουν μία εσωτερική προδιάθεση για πράξη η οποία να εδράζεται σε έναν στιβαρό και αναλλοίωτο Ηθικό Νόμο όπως αυτός ορίζεται στη καντιανή θεωρία, και όπως αυτός βλέπουμε να ισχύει για τους ανθρώπους -τους οποίους ορίζουμε ως έλλογα όντα που πράττουν σύμφωνα με τον Ηθικό Νόμο-.

[48] Luciano Floridi, "On the Morality of Artificial Intelligence", *Minds and Machines*, (August 2014): 5
DOI: 10.1023/B:MIND.0000035461.63578.9d

συστήματα. Αντιθέτως, ο άνθρωπος, κινητοποιημένος από τον Ηθικό Νόμο[49] (περιεχόμενο του Κ3) αποφασίζει και πράττει σύμφωνα με την ελεύθερή του βούληση. Εξαιτίας αυτής της σημαντικής διαφοράς ανάμεσα σε έναν άνθρωπο και σε μία μηχανή ΤΝ, εγείρεται το σημαντικό ερώτημα εάν θα πρέπει να υπάρχει μία διαφορετικού τύπου μεταχείριση των συστημάτων αυτών, από αυτή που υφίσταται ήδη για τους ανθρώπους όσον αφορά το ζήτημα του ηθικού καθεστώτος[50]. Ποια είναι η ηθική επιταγή σύμφωνα με την οποία αποφασίζει ένα σύστημα ΤΝ να επιτελέσει μία του πράξη; Προς το παρόν, μας είναι άγνωστο.

Το πρόβλημα λοιπόν παραμένει. Μπορεί η ανθρωπότητα να διακρίνει εάν ένα τέτοιο σύστημα ΤΝ αντλεί τις αρχές της ηθικής από τον Κ3 προκειμένου να προβεί σε μία πράξη; Εάν ναι, μήπως έτσι κατορθώνει να συνδέεται στον Κ2 που έως τώρα θεωρούσαμε πως όχι; Εάν όχι, τότε τι είναι αυτό που ισχύει για την ηθική και τα συστήματα ΤΝ;

Συμπεράσματα

Η μεταηθική προσέγγιση αναδεικνύει ερωτήματα και προβλήματα που ενδεχομένως προκύπτουν από τη χρήση συστημάτων τεχνητής νοημοσύνης που ως τώρα μπορεί να μην είχαν γίνει αντιληπτά. Ο μετασχηματισμός της οντολογικής προσέγγισης του Popper δημιουργεί ένα γόνιμο έδαφος που μας οδηγεί να υποθέσουμε πως ο Κ3 μπορεί να δημιουργείται ή προσεγγίζεται και από άλλες οντότητες. Είναι γεγονός πως εντοπίζεται μία τάση ή και επιθυμία εξίσωσης της ανθρώπινης φύσης με τις λειτουργίες και πρακτικές μίας αυτόνομης μηχανής σε ηθικό επίπεδο, στις περιπτώσεις που αυτό κρίνεται αναγκαίο ώστε να μπορούν να αποδοθούν ευθύνες ειδικά στην περίπτωση λάθος πρακτικών.

Η Ποππεριανή προσέγγιση του ζητήματος όμως αναδεικνύει μία ειδοποιό διαφορά ανάμεσα στα συστήματα τεχνητής νοημοσύνης και στις ανθρώπινες οντότητες. Στα συστήματα Τ.Ν. η μετάβαση πραγματοποιείται από τον Κ3 που διαθέτουν πρόσβαση, στη λειτουργία και πραγμάτωση αυτού στο Κ1 χωρίς τον ενδιάμεσο κόσμο Κ2. Αντιθέτως, ο άνθρωπος διαθέτει πρόσβαση και στον Κ2 στον οποίο δημιουργεί Κ3 και αποβλέπει στον Κ1. Η μετάβαση των συστημάτων ΤΝ στον Κ2 από τον Κ3 και

[49] Ιμμάνουελ Καντ, *Θεμελίωση της Μεταφυσικής των Ηθών*, επιμ. Κώστας Ανδρουλιδάκης, (Πανεπιστημιακές εκδόσεις Κρήτης: 2017), 75-79.
[50] Hage, "Theoretical foundations, 256.

κατόπιν στον Κ1 (από τον ένα δηλαδή κόσμο του Popper στον άλλο) είναι αβέβαιη και ενδεχομένως να περιορίζεται στην αποβλεπτικότητά τους μέσω των αντανακλάσεων από τον Κ1 στον Κ1 ή όπως εξηγήθηκε από τον Κ3 στον Κ1 εξαιτίας του τρόπου δημιουργίας τους. Το ερώτημα που προκύπτει λοιπόν είναι: Μπορεί να υπάρξει ηθική για τα συστήματα ΤΝ εάν απουσιάζει από αυτά ο Κ2; Εν αντιθέσει με τις ηθικές αλήθειες που αναγνωρίζουν οι άνθρωποι και οι οποίες *πηγάζουν* από τον Κ3 και αντανακλούν στον Κ1, τα συστήματα Τεχνητής Νοημοσύνης είναι αβέβαιο αν μπορούν να επιτελέσουν μία τέτοιου είδους αναγωγή και αντ' αυτού είναι σαν να δημιουργούν ένα καινούργιο επίπεδο το οποίο ακόμη μας είναι άγνωστο.

Στα συστήματα τεχνητής νοημοσύνης θα μπορούσαμε να πούμε με όσα έχουν αναφερθεί στην παρούσα εργασία, πως ενδέχεται *μόνο* η δημιουργία τους να εντοπίζεται και να πηγάζει από τον Κ3 και να αντανακλά στον Κ1 εάν όπως αναλύθηκε στο πρώτο κεφάλαιο οι επιστήμονες έχουν ως αντικείμενο έρευνας τα όσα βρίσκονται στον Κ3 και όχι στον Κ2 όπως φαίνεται να κάνουν έως σήμερα. Σκοπός όμως αυτής της οντολογικής προσέγγισης του ζητήματος, είναι να εντοπιστεί η ίδια συλλογιστική αλληλουχία *και* στις ηθικές συνιστώσες τέτοιων συστημάτων. Επομένως ανοίγει ένα νέο πεδίο έρευνας το οποίο εξετάζει ένα νέο κόσμο, τον νοητικό κόσμο των συστημάτων ΤΝ, στον οποίο απουσιάζει ο Κ2 και που ενδεχομένως λειτουργεί με το δικό του αξιακό σύστημα, ενέχει αποβλεπτικότητας στον Κ1, αλλά αποτελεί μόνο κατά ένα μέρος του αντικείμενο του φυσικού κόσμου. Τα ερωτήματα παραμένουν, και μαζί με αυτά εγείρονται καινούργια, καθώς και καινούργιες προβληματικές που ενδέχεται να προκύψουν εξαιτίας του αστάθμητου παράγοντα του χρόνου.

Η μεταηθική θεμελίωση ανοίγει καινούργια μονοπάτια σκέψης στον τρόπο ηθικής θεμελίωσης των συστημάτων τεχνητής νοημοσύνης. Μήπως θα ήταν καλύτερο να δομήσουμε ένα εξολοκλήρου νέο ηθικό σύστημα από εντελώς διαφορετική σκοπιά, για τις ηθικές συνιστώσες και πρακτικές (που επιτελούν ή θα επιτελέσουν στο μέλλον) αυτών των συστημάτων ή να ερευνήσουμε περισσότερο τα ερωτήματα «είναι οι οντότητες ΤΝ ισότιμες με τον άνθρωπο;» ή αν έστω «είναι οι οντότητες της τεχνητής νοημοσύνης έλλογα όντα;» ώστε να καταλήξουμε στο «εάν είναι δυνατόν οι οντότητες της ΤΝ να παράγουν έναν διαφορετικό τον Κ2 που οδηγεί και στον Κ3»; Το σίγουρο είναι, πως δεν μπορούμε να πούμε με βεβαιότητα ότι προσφέρεται μία οριστική και άμεση λύση στο ζήτημα.

Βιβλιογραφία

Ελληνική

Βιρβιδάκης, Στέλιος. «Ο *Κόσμος 3* του Καρλ Πόππερ ως ερμηνευτική πρόταση για την κατανόηση των μοντέλων μετριοπαθούς αξιακού ρεαλισμού», *Φιλοσοφία 49,* νο. ΙΙ (2020): 194-207.

Γούναρης Ά., & Κωστελέτος Γ. (2024). Γράφοντας τον αλγόριθμο του Καλού: Η Τεχνητή Νοημοσύνη ως μηχανή απόδοσης Δικαιοσύνης. Ηθική. Περιοδικό φιλοσοφίας, (19). https://doi.org/10.12681/ethiki.39654

Καντ, Ιμμάνουελ. *Θεμελίωση της Μεταφυσικής των Ηθών*, επιμ. Κώστας Ανδρουλιδάκης, Πανεπιστημιακές εκδόσεις Κρήτης: 2017.

Nagel, Thomas. *Η Θέα από το πουθενά*, Αθήνα: Κριτική, 2002.

Ξενόγλωσση

Arkin, Ronald C. "The Case of Ethical Autonomy in Unmanned Systems", *Journal of Military Ethics Vol. 9*, no. 4, 2010.

Asaro, Peter. "THE NEAR FUTURE OF ARTIFICIAL INTELLIGENCE", *Autonomous Weapons and the Ethics of A.I.,* 2020.

Floridi, Luciano. "On the Morality of Artificial Intelligence", *Minds and Machines*, August: 2014. DOI: 10.1023/B:MIND.0000035461.63578.9d

Gounaris, A. (2011). Intentionality and the Emergence of Meaning. Philosophia - Annual Journal of the Research Centre for Greek Philosophy of the Academy of Athens, v.41, pp 319-321, 2011. ISSN 1105-2120

Gounaris, A., Kosteletos, G. (2020). Licensed to Kill: Autonomous Weapons as Persons and Moral Agents. In Prole, D. and Rujević, G. (ed.). Personhood. Novi Sad, Filozofski Fakultet & The NKUA Applied Philosophy Research Lab Press. DOI: https://doi.org/10.12681/aprlp.49

Hage, Jaap. "Theoretical foundations for the responsibility of autonomous agents", *Springer*: August 2017.

Kim, Jaegwon, *Η Φιλοσοφία του Νου*, Leader Books, Αθήνα: 2005.

Popper, Karl R. *Objective knowledge: an evolutionary approach* , New York: Oxford University Press, 1972.

Robinson, Howard. "Dualism", The Stanford Encyclopedia of Philosophy (Fall 2020 Edition), επιμέλεια: Edward N. Zalta.

https://plato.stanford.edu/archives/fall2020/entries/dualism/
Sparrow, Robert. “Killer Robots”, *Journal of Applied Philosophy Vol. 24,* no. 1, 2007.
Thornton, Stephen. “Karl Popper”, *The Stanford Encyclopedia of Philosophy* (Winter 2022 Edition), επιμέλεια: Edward N. Zalta & Uri Nodelman.
https://plato.stanford.edu/archives/win2022/entries/popper/

Διαδικτυακοί ιστότοποι

DeepMind. AlphaGo - The Movie | Full award-winning documentary. Youtube, 2020. https://www.youtube.com/watch?v=WXuK6gekU1Y
Bronstein, Michael. AMMI Course "Geometric Deep Learning" - Lecture 1 (Introduction). Youtube, 2021.
https://www.youtube.com/watch?v=PtA0lg_e5nA

Περίληψη

Η ραγδαία ανάπτυξη των συστημάτων Τεχνητής Νοημοσύνης (ΤΝ ή A.I.-Artificial Intelligence-) αποτελεί ένα βασικό παράγοντα βελτίωσης της ποιότητας ζωής των ανθρώπων και της κοινωνίας εν γένει σε όλους τους τομείς. Από τα συστήματα παραγωγικής Τεχνητής Νοημοσύνης έως τα αυτοματοποιημένα μηχανήματα έκδοσης εισιτηρίων και τα αυτοματοποιημένα οχήματα, ο τρόπος διαβίωσης των ανθρώπινων όντων αναδιαμορφώνεται. Σαφέστατα, κρίνεται πως είναι πολύ νωρίς ώστε να προκύψουν συνολικά συμπεράσματα για το αν αυτή η εισβολή της τεχνολογίας αποτελεί κάτι κοινωνικά ωφέλιμο ή όχι. Η σκέψη αυτή προβληματίζει την επιστημονική κοινότητα όχι τόσο για τις περιπτώσεις ημιαυτόνομης λειτουργίας τέτοιων συστημάτων[51], όσο για τις περιπτώσεις πλήρους αυτονομίας και «αυτενεργείας» τους[52]. Σε αυτές ο άνθρωπος έχει ελάχιστη ή και μηδαμινή συμμετοχή. Καραδοκεί θα μπορούσαμε να πούμε, ένας μόνιμος φόβος ή

[51] Human in the loop
[52] Human on the loop ή human off the loop

αλλιώς αμφιβολία, για το αν τέτοια συστήματα μπορούν να δράσουν αυτόνομα με τέτοιο τρόπο ώστε οι ενέργειες αλλά και τα κίνητρά τους να ομοιάζουν με αυτές των ανθρώπων, εάν δηλαδή διαθέτουν ηθικό καθεστώς. Επομένως το βασικό ερώτημα που προκύπτει είναι εάν τέτοιου είδους συστήματα έχουν ή μπορούν να αποκτήσουν ηθικό καθεστώς ή και κατ' επέκταση εάν μπορούν να θεωρηθούν ηθικά υποκείμενα. Ποια είναι τα ηθικά θεμέλια πάνω στα οποία αναπτύσσονται και λειτουργούν τα συστήματα Τεχνητής Νοημοσύνης; Υπάρχουν; Είναι αναγκαίο να βασίζονται σε τέτοιου είδους θεμέλια; Ενδεχομένως τα παραπάνω ερωτήματα να βρουν απαντήσεις στη μεταηθική προσέγγιση του ζητήματος υπό το πρίσμα της οντολογικής θεωρίας των τριών κόσμων του Karl R. Popper.

Λέξεις-κλειδιά: Τεχνητή Νοημοσύνη, απόδοση ηθικής ευθύνης, ηθική ευθύνη, ηθικό καθεστώς, Οντολογία, μεταηθική, οι Κόσμοι του Karl R. Popper, Αποβλεπτικότητα

Ειρήνη Δαρκαδάκη
Τμήμα Φιλοσοφίας, ΕΚΠΑ
Ηλεκτρονική Διεύθυνση: iomalan@philosophy.uoa.gr
ORCID iD: https://orcid.org/0000-0001-9942-7124