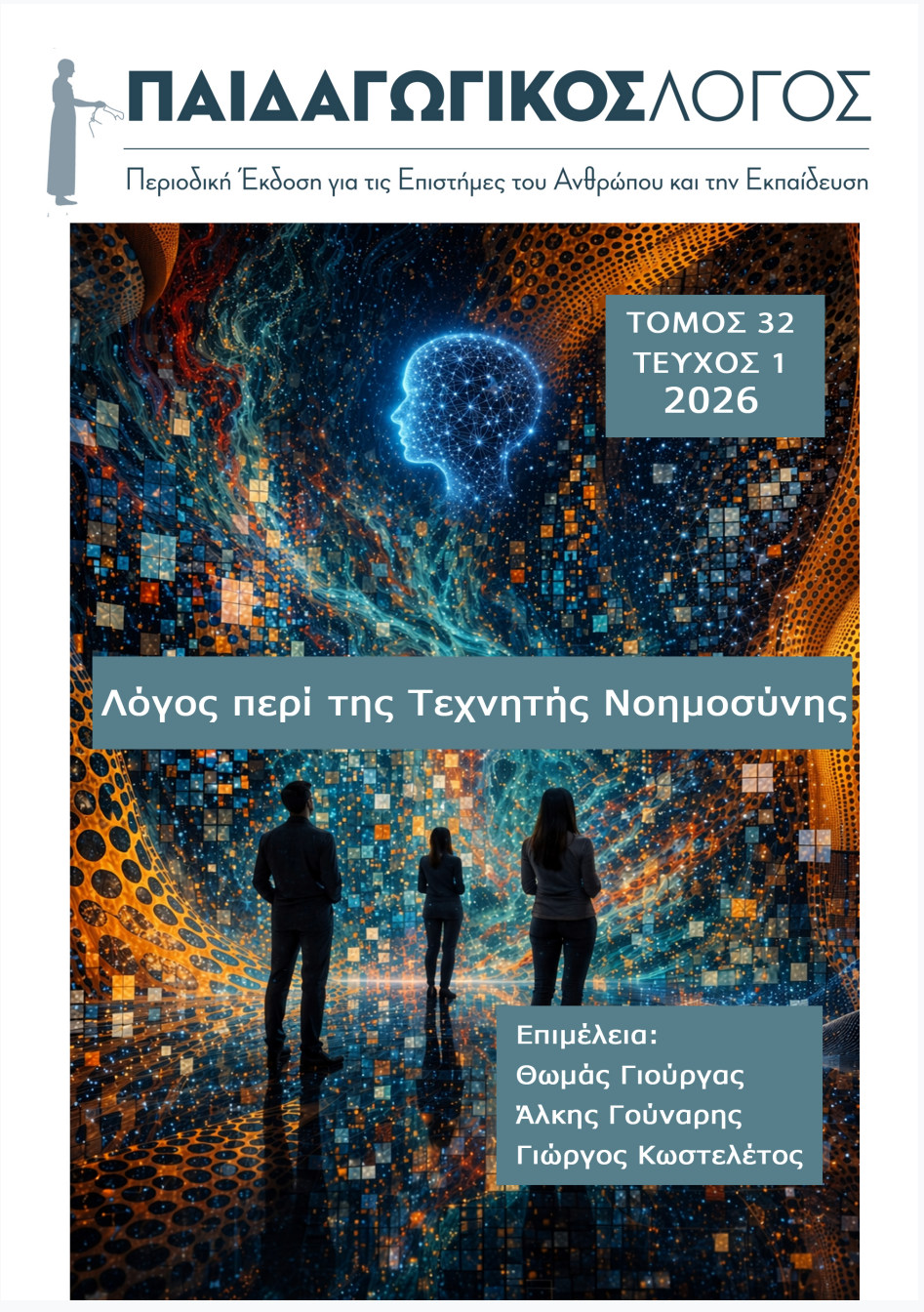
Παιδαγωγικός Λόγος

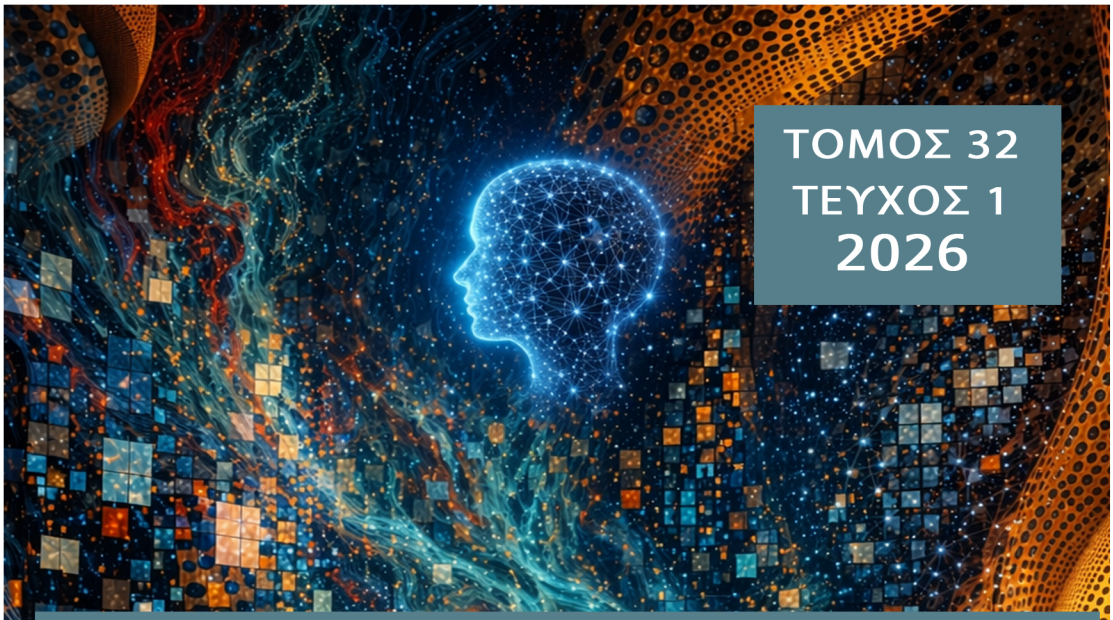
Τόμ. 32, Αρ. 1 (2026)

Λόγος περί της Τεχνητής Νοημοσύνης



ΠΑΙΔΑΓΩΓΙΚΟΣ ΛΟΓΟΣ

Περιοδική Έκδοση για τις Επιστήμες του Ανθρώπου και την Εκπαίδευση



ΤΟΜΟΣ 32
ΤΕΥΧΟΣ 1
2026

Λόγος περί της Τεχνητής Νοημοσύνης

Επιμέλεια:
Θωμάς Γιούργας
Άλκης Γούναρης
Γιώργος Κωστελέτος

Παιδαγωγικός Λόγος

Περιοδική Έκδοση για τις Επιστήμες του Ανθρώπου και την Εκπαίδευση

Τόμος 32 **Τεύχος 1** **2026**

Ιδρυτής και Διευθυντής
Ιωάννης Ε. Θεοδωρόπουλος
τ. Καθηγητής Φιλοσοφίας των Επιστημών της Εκπαίδευσης

Υπεύθυνοι Σύνταξης

Ιωάννης Ε. Θεοδωρόπουλος
sergoula2@yahoo.gr

Δρ. Κωνσταντίνος Ζέρβας
konzervas965@gmail.com

Δρ. Βασίλειος Ε. Πανταζής
vapantazis@bio.uth.gr

Δρ. Χαράλαμπος Ρέτσος
hretsos@asfa.gr

Παιδαγωγικός Λόγος
ISSN 1106-9341

Παιδαγωγικός Λόγος
Περιοδική Έκδοση για τις Επιστήμες του Ανθρώπου και την Εκπαίδευση
https://ejournals.epublishing.ekt.gr/index.php/plogos/index
www.plogos.gr
plogos.journal@gmail.com

Φορέας έκδοσης:
Τμήμα Βιοχημείας και Βιοτεχνολογίας Πανεπιστημίου Θεσσαλίας (Λάρισα)
Πρόγραμμα «Παιδαγωγικής και Διδακτικής Επάρκειας»

Ταχυδρομική διεύθυνση:
ΠΑΙΔΑΓΩΓΙΚΟΣ ΛΟΓΟΣ
Βασίλειος Ε. Πανταζής
Τμήμα Βιοχημείας και Βιοτεχνολογίας Πανεπιστημίου Θεσσαλίας
Βιόπολις
41500 Λάρισα

Επικοινωνία:
Βασίλειος Ε. Πανταζής
vapantazis@bio.uth.gr
τηλ.: 2410565233 και 6973081161

Κωνσταντίνος Ζέρβας
konzervas965@gmail.com
τηλ.: 6909399917

Παιδαγωγικός Λόγος
Περιοδική Έκδοση
για τις Επιστήμες του Ανθρώπου και την Εκπαίδευση

Ο «Παιδαγωγικός Λόγος - Περιοδική Έκδοση για τις Επιστήμες του Ανθρώπου και την Εκπαίδευση» επιχειρεί να αποτελέσει μια διεπιστημονική ερευνητική παρουσία στον ευρύτερο χώρο των Επιστημών του Ανθρώπου (humanities/Geisteswissenschaften), με αναφορά στον διάλογό τους με τις Επιστήμες της Ζωής (life sciences/Lebenswissenschaften), με άξονα και απόληξη την αγωγή και την εκπαίδευση του ανθρώπου. Η θεματολογία επεκτείνεται σε προσεγγίσεις φιλοσοφικού, θεολογικού ή κοινωνιολογικού προσανατολισμού, κυρίως όταν σχετίζονται άμεσα ή έμμεσα με την αγωγή του ανθρώπου.

Ο «Παιδαγωγικός Λόγος», ως «Τετράμηνη Έκδοση Παιδαγωγικού Προβληματισμού», ιδρύθηκε το 1995 από τον Καθηγητή Ιωάννη Ε. Θεοδωρόπουλο, εξέχουσα μορφή της Φιλοσοφικής Παιδαγωγικής στον γερμανόφωνο χώρο.

Φορέας έκδοσης του επιστημονικού περιοδικού «Παιδαγωγικός Λόγος-Περιοδική Έκδοση για τις Επιστήμες του Ανθρώπου και την Εκπαίδευση», από το 2021, είναι το Τμήμα Βιοχημείας και Βιοτεχνολογίας του Πανεπιστημίου Θεσσαλίας (Λάρισα), στο πλαίσιο της πιστοποιημένης «Παιδαγωγικής και Διδακτικής Επάρκειας» του Τμήματος, το οποίο έχει συνάψει σχετική συμφωνία με την Υπηρεσία e-Publishing του Εθνικού Κέντρου Τεκμηρίωσης για την έκδοση και ένταξη του περιοδικού στην Πλατφόμα e-Publishing ΕΚΤ και στις εθνικές υποδομές ευρετηρίων, καταλόγων και συσσώρευσης.

Ο «Παιδαγωγικός Λόγος - Περιοδική Έκδοση για τις Επιστήμες του Ανθρώπου και την Εκπαίδευση» είναι διαθέσιμος μέσω της Πλατφόρμας e-Publishing του Εθνικού Κέντρου Τεκμηρίωσης (ΕΚΤ) στη διεύθυνση: https://bit.ly/3Ar2IFg και στη διεύθυνση: www.plogos.gr

Η υποβολή των υπό δημοσίευση εργασιών γίνεται μέσα από το περιβάλλον e-Publishing του Εθνικού Κέντρου Τεκμηρίωσης (ΕΚΤ). Για τη διαδικασία υποβολής άρθρων ακολουθήστε τη διαδικασία που περιγράφεται στην πλατφόρμα e-publishing του ΕΚΤ: https://bit.ly/3AnIqwt

Ο «Παιδαγωγικός Λόγος - Περιοδική Έκδοση για τις Επιστήμες του Ανθρώπου και την Εκπαίδευση» ακολουθεί το διεθνώς αποδεκτό, ως προϋπόθεση αξιόπιστης επιστημονικής δημοσίευσης, σύστημα «διπλής-τυφλής ανώνυμης κρίσης» (double blind peer review) από κριτές κύρους, μέλη της Επιστημονικής Επιτροπής του περιοδικού ή άλλους επιστήμονες, όταν το επιστημονικό πεδίο της εργασίας το επιβάλλει.

Ο «Παιδαγωγικός Λόγος - Περιοδική Έκδοση για τις Επιστήμες του Ανθρώπου και την Εκπαίδευση» εκδίδεται έντυπα και ηλεκτρονικά δύο φορές τον χρόνο.

Η ιστοσελίδα του περιοδικού «Παιδαγωγικός Λόγος - Περιοδική Έκδοση για τις Επιστήμες του Ανθρώπου και την Εκπαίδευση» παρέχει άμεση ανοικτή πρόσβαση στο περιεχόμενό του υποστηρίζοντας την αρχή της υποστήριξης της παγκόσμιας ανταλλαγής γνώσεων καθιστώντας διαθέσιμα ελεύθερα στο κοινό τα αποτελέσματα της επιστημονικής έρευνας.

Η χρήση άδειας που υιοθετεί ο Παιδαγωγικός Λόγος είναι: Αναφορά Δημιουργού – Μη Εμπορική Χρήση – Όχι Παράγωγα Έργα 4.0 (CC BY-NC-ND). Αυτή η άδεια επιτρέπει στους άλλους να έχουν πρόσβαση στο έργο και να το μοιράζονται με άλλους εφόσον κάνουν αναφορά σε αυτό, ωστόσο δεν μπορούν να το αλλάξουν με κανένα τρόπο ούτε να το χρησιμοποιούν για εμπορική χρήση.

Επιστημονική Επιτροπή

Αντωνίου Αλέξανδρος-Σταμάτιος,
Καθηγητής ΠΤΔΕ Πανεπιστημίου Αθηνών

Γκόβαρης Χρήστος,
Καθηγητής ΠΤΔΕ Πανεπιστημίου Θεσσαλίας (Βόλος)

Δεληκωνσταντής Κωνσταντίνος,
Ομότιμος Καθηγητής Πανεπιστημίου Αθηνών

Ζέρβας Κωνσταντίνος,
Σύμβουλος Εκπαίδευσης Πληροφορικής (Σύρος)

Θεοδωρίδης Αλέξανδρος,
Αναπληρωτής Καθηγητής Πανεπιστημίου Θράκης (Αλεξανδρούπολη)

Θεοδωρόπουλος Ιωάννης Ε.,
τ. Καθηγητής Φιλοσοφίας των Επιστημών της Εκπαίδευσης
(Πανεπιστήμιο Κρήτης – Α.Σ.ΠΑΙ.Τ.Ε.)

Κασσωτάκης Μιχαήλ,
Ομότιμος Καθηγητής Πανεπιστημίου Αθηνών

Κουτρούμπα Κωνσταντίνα,
Αναπληρώτρια Καθηγήτρια Χαροκόπειου Πανεπιστημίου

Μανωλάς Ευάγγελος Ι.,
Καθηγητής Πανεπιστημίου Θράκης (Ορεστιάδα)

Μάρκος Αντώνιος,
Ομότιμος Καθηγητής Πανεπιστημίου Πατρών

Πανταζής Βασίλειος, Α.,
Καθηγητής ΠΤΠΕ Πανεπιστημίου Θεσσαλίας (Βόλος)

Πανταζής Βασίλειος Ε.,
Μέλος Ε.ΔΙ.Π.- Διδάσκων Τμήματος Βιοχημείας και Βιοτεχνολογίας
Πανεπιστημίου Θεσσαλίας (Λάρισα)

Παπακωνσταντίνου Θεόδωρος,
Ομότιμος Καθηγητής Πανεπιστημίου Αθηνών

Πρωτοπαπαδάκης Ευάγγελος Δ.,
Καθηγητής Τμήματος Φιλοσοφίας
Πανεπιστημίου Αθηνών

Ρέτσος Χαράλαμπος,
Μέλος Ε.ΔΙ.Π. - Διδάσκων Τμήματος Εικαστικών Τεχνών Ανώτατης
Σχολής Καλών Τεχνών

Σοφός Αλιβίζος,
Καθηγητής ΠΤΔΕ Πανεπιστημίου Αιγαίου (Ρόδος)

Στείρης Γεώργιος,
Καθηγητής Τμήματος Φιλοσοφίας
Πανεπιστημίου Αθηνών

Τζαβάρας Ιωάννης,
Ομότιμος Καθηγητής Πανεπιστημίου Κρήτης

Περιεχόμενα

Τόμος 32 2026 Τεύχος 1

ΤΕΥΧΟΣ 1

Παιδαγωγικος Λογος 2026

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

Lydia KORNARAKI

A.I. and Lethal Weapons: A blameless army of killer robots?

doi:https://doi.org/10.12681/plogos.33696

Lethal Autonomous Weapons

LETHAL AUTONOMOUS WEAPONS (LAWS) ARE POISED TO BECOME the primary method of warfare in the future. They are machines equipped with artificial intelligence systems, enabling them to operate independently, without human intervention. Artificial intelligence has already infiltrated the military and, right now, there are numerous nonmilitary and military apps and devices utilizing A.I. systems, such as the GPS installed on our phones.

The use of A.I. in the military significantly enhances operational efficiency, intelligence processing, and autonomous decision-making. The state-of-the-art level of A.I. in warfare as of 2025 includes both robotic forms, like drones or sentry guns and non-robotic software systems integrated into broader weapons or command infrastructure. Systems like the U.S. Project Maven[1] and NATO's Palantir-based AI[2] integrate battlefield data to support rapid target identification. Israel's "Gospel" and "Lavender" automate kill lists, while drones like Türkiye's Kargu-2 and Israel's Harpy can engage targets with minimal human input. South Korea's SGR-

[1] Patrick Tucker, "NGA Will Take Over Pentagon's Flagship AI Program," Defense One, April 25, 2022, https://www.defenseone.com/technology/2022/04/nga-will-take-over-pentagons-flagship-ai-program/366098/.

[2] AIN.Capital, "NATO Acquires Palantir Military AI System to Aid Commanders in Battlefield Decision-Making," AIN.Capital, April 15, 2025, https://en.ain.ua/2025/04/15/nato-acquires-palantir-military-ai-system/.

A1 sentry gun[3] and U.S. systems like Shield AI show how lethal autonomy is being physically deployed. These developments have already seen combat use, raising global ethical and regulatory concerns.

Focusing on the military apps and machines, there are non-lethal military robots used for tasks that can be dangerous for humans, such as mine clearing, explosive ordnance disposal and rescue missions, just to name a few[4]. Beyond robotic implementations, there are also non-embodied A.I. systems to assist in a positive outcome for a military mission. Such examples are the fast and efficient processing of data from all the surveillance sources or the ability to protect the military network from hackers, the latter being a fundamental issue, to prevent attacks and at the same time have the right equipment to fight off unauthorized users from confidential content[5].

Machine learning A.I. systems function exactly like that: they "learn" from any situation, they upgrade their system to eliminate faults and mistakes and then give back what has been processed and developed. Therefore, A.I. does not only avert and counter attacks but also ensures flexibility[6].

The use of A.I. for lethal weapons has broadened the application of such technology, making it possible to have battles from afar. According to the U.S. Department of Defense Directive, Lethal Autonomous Weapons or Killer Robots are a "weapon system[s] that, once activated, can select, and engage targets without further intervention by a human operator."[7].

There are three levels of autonomy noted[8], regarding L.A.W.: human-in-the-loop (human operator), human-on-the-loop or supervised and hu-

[3] Brittany Roston, "Everything We Know About Samsung's Machine Gun Robots," SlashGear, March 24, 2022, https://www.slashgear.com/825074/everything-we-know-about-samsungs-machine-gun-robots/.

[4] Bartneck, Christoph and Lütge, Christoph & Wagner, Alan and Welsh, Sean. (2021). Military Uses of AI. 10.1007/978-3-030-51110-4_11.

[5] Marcus Roth "Artificial Intelligence in the Military – An Overview of Capabilities", last accessed 5/1/2023. https://emerj.com/ai-sector-overviews/artificial-intelligence-in-the-military-an-overview-of-capabilities/.

[6] Marcus Roth, "Artificial Intelligence in the Military – An Overview of Capabilities," Emerj, February 20, 2019, https://emerj.com/artificial-intelligence-in-the-military-an-overview-of-capabilities/.

[7] Department of Defense Directive 3000.09, "Autonomy in Weapon Systems," Updated May 8, 2017, https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/ 300009p.pdf.

[8] Congressional Research Service, Defense Primer: US Policy on Lethal Autonomous Weapons (2020).

man-off-the-loop (fully autonomous). Human-in-the-loop[9] is the most basic form of using A.I. systems in war because it allows for a human operator to control the situation remotely (teleoperation) and engage or disengage a weapon's target[10]. This clearly is the "safest" way to use unmanned drones in war. Human-on-the-loop[11] is a semi-autonomous situation where the L.A.W. has been programmed to "engage individual targets or specific target groups that have been selected by a human operator"[12]. It requires pre-programming by a human operator and then it functions on its own, obeying the specific commands, but it does not deviate from what has been programmed or decided on its own. Of course, humans retain the supervisory role and the task of the ultimate assessor of A.I.'s operations, along with the capacity to interfere during these operations, should such a need arise.

Human-off-the-loop,[13] namely a state of complete autonomy for AI systems, is the destination of L.A.W.S. technology and the reason campaigns like "Stop Killer Robots"[14] were created, or why so many terms, legislations, and re-approaches of ethical theories of the past have surfaced in the academic world. We are not there yet as there are not completely autonomous machines that can make decisions on their own since they still require a human agent to operate them. Still, various weapons exist already

[9] Nils Melzer, "Human rights implications of the usage of drones and unmanned robots in warfare", Directorate-General for External Policies of the Union (European Union 2013): 6. Doi:10.2861/213. https://www.europarl.europa.eu/RegData/etudes/etudes/join/2013/410220/EXPO-DROI_ET(2013)410220_EN.pdf.

[10] Annemarie Shea, "The Legal and Ethical Challenges Posed by Lethal Autonomous Weapons," Trinity College Law Review 24 (2021): 119.

[11] Nils Melzer, "Human rights implications of the usage of drones and unmanned robots in warfare", Directorate-General for External Policies of the Union (European Union 2013): 6. Doi:10.2861/213. https://www.europarl.europa.eu/RegData/etudes/etudes/join/2013/410220/EXPO-DROI_ET(2013)410220_EN.pdf.

[12] DODD 3000.09: 14.

[13] Nils Melzer, "Human rights implications of the usage of drones and unmanned robots in warfare", Directorate-General for External Policies of the Union (European Union 2013): 6.

[14] "Stop Killer Robots" is a campaign that launched on 2013 with the moto "Technology should be used to empower all people, not to reduce us – to stereotypes, labels, objects, or just a pattern of 1's and 0's." https://www.stopkillerrobots.org/ . There are also a few sites that target educating people on the importance of containing such technology before it's too late: https://futureoflife.org/project/lethal-autonomous-weapons-systems/ and https://autonomousweapons.org/.

in use that can either target or fire autonomously but not both[15].

"Killer robots"

The most autonomous machine that is already in use is the unmanned drone. The origin of the very first unmanned drone dates back as far as 1907[16], although there was an even earlier appearance in 1849[17]. With the discreet, modern label, "Killer robots," these drones are the next stage in the technological evolution of warcraft.

Their reputation and more importantly, their horridness, has preceded them. Works of fiction have created a terrifying scenario around "robot machines" in general and, although many of these books and movies are excellent adventures worth reading and watching, they tend to simplify the complexity of a machine to it merely being "good" or "evil." These human characteristics on machines set up a delightful story of fighting for justice and freedom without the guilt trip of killing another human being, but it keeps the whole concept of war machines in the realm of imagination.

Janelle Shane, in a TedX[18] talk she gave in 2019, made quite a shocking statement about the real danger of artificial intelligence and it is not the fear of a rebellious attack. It is the fact that artificial intelligence systems do exactly what we tell them to do. This does not sound puzzling or dangerous, but it becomes so when you start testing it. She conducted an experiment using A.I. to research the best ice cream flavor in the world. The results were the exact opposites. The A.I. generator created the worst ice-cream flavors, which were not only indigestible but also fatal if humans consumed them. What A.I. programs really lack is critical thinking, a human trait that is hard to translate into code.

The blunt reality of artificial intelligence is that so far it cannot operate

[15] Bartneck, Christoph and Lütge, Christoph & Wagner, Alan and Welsh, Sean, "Military Uses of AI", (2001).

[16] Mario Poljak, "The History of Drones: Timeline From 1907 To 2021", accessed December 2, 2022, https://www.dronetechplanet.com/the-history-of-drones-timeline-from-1907-to-2019/.

[17] Alan McKenna, "The Public Acceptance Challenge and Its Implications for the Developing Civil Drone Industry", in The Future of Drone: Use Opportunities and Threats from Ethical and Legal Perspectives, ed. Bart Custers (Springer: Asser Press, 2016), 355.

[18] Janelle Shane, "The danger of A.I. is weirder than you think", 14 Nov. 2019. Educational video, 10:29. https://www.youtube.com/watch?v=OhCzX0iLnOc&ab_channel=TED.

with such complexity as described in fiction works. These drones and robots are human creations and for this reason, there will always be a limit to their autonomy, meaning that their actions and consequences can only be attributed to a human agent.

Accountability and problems

Although still in the sphere of imagination, the reality of such autonomy in war is not 'in a galaxy far far away,' but a bit closer. Autonomous decision-making machines present humanity with new difficulties concerning the attribution of responsibility, moral or legal, during a potential wrongful attack on civilians during warfare.

Before a nation decides to launch an attack on another, military operations involve a series of steps[19]. These steps, represented as an iceberg diagram, reveal a particularly complex hierarchical decision-making system, which in turn entails various levels of transparency and responsibility. As it becomes apparent, between the decision to initiate the process and the actual attack, there is a significant distribution of responsibility. This, frankly, makes it harder to assign blame if things go wrong.

It is harder for both the legality and the morality of the operations, because operations involve many people, a lot of different dispositions and personalities as well as many soldiers working as pawns and simply following orders. The answer to the question of *who is to be blamed* is not easy in this case.

The complexity of such operations makes it hard to place the blame. We begin from the bottom, analyzing the chain of parties involved in the operation and the creation of L.A.W.S. Firstly, there is the person who produced the idea, the innovator. Then, the program designers follow, the ones who make the idea possible and usable. This line of people involved expands through the manufacturers who mass produce machines with artificial intelligence. This chain of events leads to the politicization of such machines as they are mostly used in the military which is part of the government and needs approvals to receive the 'go' signal to use them.

Usually, an obvious target to blame is the country or the state that

[19] UNIDIR: The Human Element In Decisions About The Use Of Force. https://unidir.org/sites/default/files/2020-03/UNIDIR_Iceberg_SinglePages_web.pdf

authorized their use[20] but even though they are indeed responsible for this act, they cannot be held *liable* for it. The so-called iceberg of responsibility in the military divides hierarchy into three main levels of command: strategical, operational, and tactical. In simple terms, the leadership of a country/ nation decides if there should be conflict by military use. Then those commands are translated into military words and actions[21]. In such a grand operational system, who is really to be accused, who would be fair to receive any kind of punishment?

Next in line are the program designers, as the primary creators. They are responsible enough but, at the same time, it is unjust to hold them fully accountable because they work in a lab-like environment, disengaged from real situations and emotions and even though they must predict every possibility, their programs will never be a hundred percent safeguarded[22].

Continuing this L.A.W.S chain of responsibility, the manufacturers of weapons appear, who have always been liable ethically, but can we accredit them the legal part too? An interesting debate[23], between John Forge and Jai Galliot, who analyze the argument regarding the manufacturer's accusations on autonomous machines, among others, display perfectly both sides, for and against this notion. If manufacturers are found to be responsible for any L.A.W. S's derailing, it would mean that any kind of crime ever committed by any weapon ever created leads back to the creator and manufacturer as providers of a means to maim and kill. This is a lot of responsibility for one group of people and a lot of absolution for everyone else. It may sound easy to directly accuse the creators, but that would show narrow and superficial thinking.

Simplifying things to this extent is dangerous and such ideas can cause the banning of research and evolution in technology, regardless of the positive uses it can also offer. A.I. has made it possible, for example, to have a more accurate health care system or to complete for us tasks like collecting data, analyzing it, comparing it, and reaching solutions faster than the

[20] Annemarie Shea, "The Legal and Ethical Challenges Posed by Lethal Autonomous Weapons," Trinity College Law Review 24 (2021): 130.

[21] UNIDIR: The Human Element In Decisions About The Use Of Force. https://unidir.org/sites/default/files/2020-03/UNIDIR_Iceberg_SinglePages_web.pdf

[22] Peter Asaro, "Autonomous Weapons and the ethics of artificial intelligence" in *Ethics of Artificial Intelligence,* edit. S. M. Llao, Oxford University Press (2020): 226.

[23] Jai Galliott and John Forge," Debate on the Ethics of Developing AI for Lethal Autonomous Weapons" *Philosophical Journal of Conflict and Violence* vol. 5, issue 1 (2021). 10.22618/TP.PJCV.20215.1.139009.

human brain would have[24]. Many factors make LAWS irresistible. Autonomous systems exhibit superior performance in both speed and mission effectiveness. They have extended range, sustained operational capability, increased endurance, and higher targeting precision. Additionally, because of their technology, they can enable faster target engagement and exhibit inherent immunity to chemical and biological agents[25].

Furthermore, there are a few ethical arguments in favor of the use of LAWS in war. The most prominent one, of course, is the protection of soldiers' lives[26]. It is not an argument to take lightly[27]. Wars between machines are far better than between humans. Following the same logic, collateral damage will be reduced[28], as well as the possibility of errors and LAWS can make faster decisions in high-stakes environments and critical moments. Due to the excessive effectiveness of the systems there will be no emotional bias, and the machines can be designed to follow to the letter the International Humanitarian Law in war times.

The responsibility issue still exists. There is an intriguing idea, suggested by Jaap Hage, which says that it can be probable, one day, to blame the autonomous machines for their derailments and wrongdoings[29]. Daniel Dennet's article[30] had already explored this idea of placing responsibility in autonomous systems and reflected upon the challenges of attributing moral accountability to artificial agents.

This idea was also critically rejected in the article *Licensed to Kill: Autonomous Weapons as Persons and Moral Agents* by Gounaris and

[24] Luciano Floridi, "Robots, Jobs, Taxes and Responsibilities" in the *Robo-Ethics: Humans, Machines and Health*, ed. By Vincenzo Paglia and Renzo Pegoraro (Rome: Pontifical Academy for Life, 2020), 109-113.

[25] Ronald C. Arkin, "The case for Ethical Autonomy in Unmanned Systems", *Journal of Military Ethics*, 9:4 (2010), 334.

[26] Mark Gubrud, "Stopping Killer Robots" *Bulletin of the Atomic Scientists* 70, no. 1 (2014): 38–39. https://doi.org/10.1177/0096340213516745.

[27] Paul Scharre, *Four battlegrounds – power in the age of artificial intelligence* (New York, W. W. Norton & Company, 2023), p.p. 1-4.

[28] Noel Sharkey, "The Evitability of Autonomous Robot Warfare" *International Review of the Red Cross* 94, no. 886 (2012): 789. https://doi.org/10.1017/S1816383112000732.

[29] Jaap Hage, "Theoretical foundations for the responsibility of autonomous agents" *Artificial Intelligence and Law,* 25 (August 2017): 1-17.

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

Kosteletos[31]. The authors argued that the lack of the essential qualities of a moral agent, such as intention, consciousness, and capacity for ethical reflection, makes it impossible to hold them morally and legally responsible for their actions. They also provided the argument of excessive effectiveness, a state of hyper-rational execution devoid of any emotion, which diminishes the very foundations of moral accountability. What all these researchers agree to is the fact that human supervision cannot be and should not be eliminated and is needed in every stage: design, deployment, oversight.

The machines are not moral and/or legal agents and by that we mean that they cannot tell right from wrong on their own unless we program them to. Thus, they lack free will and intention, so even though they are responsible, they cannot be punished[32]. Dennet concluded that there will always be the need for human involvement both in the design and deployment.

Machine ethics contributes to better AI by enabling systems to make morally informed decisions rather than merely following coded instructions. By embedding ethical reasoning directly into AI, machines can assess complex situations, weigh competing moral values, and act in ways that promote human well-being[33]. Applying this idea on LAWS, we could create A.I. systems that will enable people in making ethical decisions concerning the design of an operation or design them from the beginning with legal and ethical rules and boundaries so, in a way they could make ethical decisions in different situations. Of course, one could argue that this is not an autonomous state of the machines but an on-the-loop case. This case would not be so bad.

[31] Alkis Gounaris and George Kosteletos, "Licensed to Kill: Autonomous Weapons as Persons and Moral Agents," in Personhood, ed. D. Prole and G. Rujević (Sad Novi: Filozofski Fakultet & The NKUA Applied Philosophy Research Lab Press, 2020), 137. https://doi.org/10.12681/aprlp.49

[32] Asaro, 226.

[33] Alkis Gounaris, George Kosteletos, Michael Anderson, and Susan Leigh Anderson. "Towards Moral Machines: A Discussion with Michael Anderson and Susan Leigh Anderson." Conatus – Journal of Philosophy 6, no. 1 (2021): 37–51. https://doi.org/10.12681/cjp.26832.

Responsibility Gap

"Responsibility gap" is a term which describes the lack of existence of a breach between a fact and the realization of its agent. This gap has appeared in L.A.W.S conversations because it cannot be decided fairly who is to be accused when machines with A.I. do something wrong.

John Forge is adamant that every designer is responsible for providing a means to kill[34] by creating the weapons. If you make a gun, you intend for it to go off or else there is no need to build one. The gun might not be fired but that matters truly little, since its primary purpose was to be fired and hence to maim or kill. According to Forge, we have three choices: a. accept that every weapon designer is responsible for every use of their creations, b. accept that the designers should foresee every possible use of their creations to avoid wrongful use, which means that they will not be able to create any weapon ever, or c. accept the existence of a no man's land for Lethal autonomous weapons[35], meaning the responsibility gap.

Galliot agrees with Forge about the responsibility of designers but only partially. He disputes that this "responsibility gap" is presented in too broad a way and cannot be used carelessly to ban autonomous weapons[36] because that would also be misguided. What Galliot tries to point out is the fact that there should be, among other criteria, the scope of how beneficial or not is the use of such weapons along with how necessary they might be regardless of everything else. What should a country do if it were attacked by an enemy who uses only lethal autonomous weapons[37]? It is all fair in love and war until someone dies and people need to be accountable for how, why and under what circumstances this death occurred.

Can we, then, accept the existence of a responsibility gap and move on? Christof Heyns, in the Report for the Human Rights Council, states clearly that the use of the L.A.W.S system is not only a military issue, but also a human rights issue as there are a lot of lives on the line[38]. Hiding behind the void of a responsibility gap is not an option. This vacuum will allow atrocities during attacks because machines have no feelings. It will make the decision of going to war uncomplicated and effortless. Poorer countries

[34] Galliot and Forge, 137.
[35] Galliot and Forge, 139.
[36] Galliot and Forge, 134.
[37] Galliot and Forge, 136.
[38] Christof Heyns, "Report of the Special Rapporteur on extrajudicial, summary or arbitrary executions", United Nations (2014).

that cannot afford to have such technology will be easy prey. Death will be dehumanized, and human lives and fates will be contained in a series of numbers, the algorithms.

Robots, War and Ethics

According to Saint Augustine, war is bad but not always the worst option[39]. This is the jus ad bellum, a principle examining the morality of going to war. Morality and war in the same sentence seem contradictive and pacifists would choke before uttering them together, because for them, there can never be a justifiable cause for war and certainly not a moral one.

Must we go to war, it must be more ethically conducted. The reason war is not so ethical, some argue, is because the old-fashioned weaponry is only as good as their masters[40]. If the need arises to go to war, then machines like LAWS can help make it more just. Autonomous systems do not act on impulse or bias, they stay consistent with the rules they were provided, unlike humans who conduct wars based on emotions and personal gains. LAWS can process vast amounts of data which allows them to engage legitimate targets with great accuracy and less casualties. The most important, of course, is their ability to enter dangerous environments this way saving human lives, like entering a mined path[41].

Both Kahn[42] and Sparrow[43] on their respective papers have made space for the jus ad bellum and jus in bello theories in connection to the use of Lethal Autonomous Weapons. The are some specific conditions to meet

[39] "But, say they, the wise man will wage just wars. As if he would not all the rather lament the necessity of just wars, if he remembers that he is a man; for if they were not just, he would not wage them, and would therefore be delivered from all wars." St. Augustine "City of God", book XIX, ch.7.

[40] Ronald C. Arkin, "The case for Ethical Autonomy in Unmanned Systems", *Journal of Military Ethics*, 9:4 (2010), 334-336.

[41] Amitai Etzioni and Oren Etzioni, "Pros and Cons of Autonomous Weapons Systems," *Military Review*, May–June 2017 p. 72-74. https://www.armyupress.army.mil/Journals/Military-Review/English-Edition-Archives/May-June-2017/Pros-and-Cons-of-Autonomous-Weapons-Systems/.

[42] Leonard Kahn, "Military Robots and the Likelihood of Armed Combat", in *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*, eds Patrick Lin, Keith Abney and Ryan Jenkins (New York: Oxford Edition, 2017): 281.

[43] Robert, Sparrow "Killer Robots" Journal of Applied Philosophy, vol 24, n. 1(2016): 67.

the jus ad bellum theory, like a justifiable cause. The benefits must outweigh the losses, right intentions, last resort choices, good chances of success rate and public announcement by legitimate authorities[44]. What Kahn suggests is that the use of L.A.W.S in war will make going to war easier, thus leading to more war, which will lead to a morally worse war. He argues that it will be morally worse because many of the principles of jus ad bellum will not be met from at least one side. This is supported by the positive offers A.I. has provided to the military. It allows for low-cost, easier, and swift mission accomplishment, has immunity to chemical and biological weapons, there is less grieving as machines are replaceable. Moreover, there is precision and speed in targeting and engaging and of course endurance since machines do not need sleep[45]. Cheap, fast, and precise with no funerals to hold and no PTSD[46] for soldiers in battles, the machines are a gift to warcraft.

The jus in bello theory is concerned with the moral conduct within war and how it is progressing. Among the various principles for a jus in bello, Sparrow brings to light a side principle which is the responsibility of one's death[47]. He points out the importance for a family to know the person responsible for their child's death. Is this something that anyone would argue against? If there is no one to blame and no one to take responsibility for an action as violent as the murder of people fighting in wars with machines, then we strip the deceased of all their dignity. They died not living but existing each day to fight off an enemy that does not die and that cares nothing about anything. Then, to top it all off, we disregard their death as a number who willingly gave their life doing their duty. The basic human rights state that we all have a right to life and dignity but maybe now, they need to add an absolution clause *"unless someone dies in a combat with a non-human A.I. entity."*

[44] Kahn, 281.

[45] Ronald C. Arkin, "The Case for Ethical Autonomy in Unmanned Systems", in *Journal of Military Ethics,* v. 9, n.4 (2010): 334.

[46] Post Traumatic Stress Disorder, is a mental disorder that was recognized in the DSM-III by the American Psychiatric Association, and is common for soldiers who have been through war and survived. It is a serious mental disorder that has affected not only the lives of those soldiers but their families as well.

[47] Sparrow, 67-68.

Extreme scenarios vs reality

In 2018 the Future of Life Institute made a statement to the United Nations representing "nearly 4,000 AI and robotics researchers and scientists" from all around the world with a request to negotiate a legal ban on LAWS[48]. The reasons stated in the letter concerned the lethal use of these weapons compared to chemical, biological and space-based nuclear weapons, which have already been banned or heavily restricted due to their devastating potential. The scientific community considers LAWS as likely for devastating potential as these, because of their ability to select and engage targets without human intervention, their cheap mass production with lead to a global arms race by their use of dictators, terrorists, in a phrase, in the wrong hands to use. The closing of the statement characterized this situation as Pandora's box, underlying the highest importance to act now and fast.

Six years later, in 2024, the same Institute highlighted the urgency in a policy brief, about the escalating risks by combining Ai and chemical biological weapons (CBW)[49]. AI has many potentials which also means that it can be used either for good or harm. This duality of AI's technology poses a substantial challenge to existing non-proliferation frameworks. The authors want to emphasize the need to get ahead of such scenarios that can compromise global security. They suggest scrupulous oversight and regulation of AI applications that can lead to CBW implications. Of course, the need for research for the development of AI systems that would prevent misuse of CBW is always continuous as well as the collaboration with other nations for a catholic agreement on this serious matter of the dual use of AI in the context of CBW.

It seems that most scientists and researchers of AI agree on one thing, and that is the need for proper regulation. Going against the tide, Vincent

[48] Future of Life Institute, "Statement to United Nations on Behalf of LAWS Open Letter Signatories," Future of Life Institute, last modified August 2018, https://futureoflife.org/open-letter/statement-to-united-nations-on-behalf-of-laws-open-letter-signatories/.

[49] Hamza Chaudhry and Landon Klein, *Chemical & Biological Weapons and Artificial Intelligence: Problem Analysis and US Policy Recommendations* (Cambridge, MA: Future of Life Institute, 2024), https://futureoflife.org/document/chemical-biological-weapons-and-artificial-intelligence-problem-analysis-and-us-policy-recommendations/.

C. Müller[50] suggests that the responsibility gap evaporates because, exactly, of the precise data on decision making processes. Back to the favorite argument, strictly programmed LAWS will not deviate from the principles of the IHL, saving many casualties and unlawful killings, which human operations cannot avoid. Their lack of emotion, portrayed as a flaw in the human eyes, is a life savior in battlefield, because unless told to, the machines have no desire and empathy towards the enemy. They cannot identify him as such, but only as a target. No revenge looking for, fewer civilian casualties and less overall suffering. Contra to the belief that such weaponry would lead to easier decision of going to war, Müller, opposes that because of their efficiency and accuracy the adversaries would think thrice before diving into an armed conflict. Although, the different voices come from alternate paths, they all come back to the need of proper ethical guidelines, ethical design of AI machines and international agreement on their use.

One such report on ethical guidelines is the UNESCO report (2002)[51]. The year of its publication underlines the importance of this issue and, even though we had major breakthroughs in AI technology this last decade, it shows the need for a precautionary approach. This report emphasizes the importance of not losing our ethical principles when we develop AI, so that it supports and follows human rights, ensures fairness and respects our dignity. It cannot be stressed enough how valuable it is to have international collaboration for setting up global standards that will make sure that AI technology benefits democracy and is used in favor of humanity and not against it.

The human imagination tends to over dramatize the future, so that previous generations can turn to the next ones and say, “I told you so.” One look at the *Back to the Future* trilogy and we will realize that nothing is as simple as it may appear, not even the past. There is no point in engaging in thought experiments to prove which scenario of using Lethal Autonomous Weapons is the most alarming and horrid[52]. Our world will never be this Utopia found only in Disney movies, and it does not have to be, it is perfect the way it is, with its flaws and miscalculations. Since we humans want to create those machines, we should be smart enough to create

[50] Vincent C. Müller, "Autonomous Killer Robots Are Probably Good News," Frontiers in Robotics and AI 3 (2016), https://philarchive.org/rec/MLLAKR.
[51] UNESCO, Ethics of Artificial Intelligence (Paris: UNESCO, 2002), https://unesdoc.unesco.org/ark:/48223/pf0000139578.
[52] Galliot and Forge, 139.

regulations and safeguards accordingly to prevent any future or past Sarah O'Conor from fighting for her life.

References

Anderson, M., Anderson, S. L., Gounaris, A., & Kosteletos, G. (2021). Towards Moral Machines: A Discussion with Michael Anderson and Susan Leigh Anderson. *Conatus - Journal of Philosophy*, *6*(1), 177–202. https://doi.org/10.12681/cjp.26832 .

AIN.Capital. "NATO Acquires Palantir Military AI System to Aid Commanders in Battlefield Decision-Making." AIN.Capital, April 15, 2025. https://en.ain.ua/2025/04/15/nato-acquires-palantir-military-ai-system/.

Arkin, Ronald C. "The Case for Ethical Autonomy in Unmanned Systems." Journal of Military Ethics 9, no. 4 (2010): 332–341. https://doi.org/10.1080/15027570.2010.536402.

Asaro, Peter. "Autonomous Weapons and the Ethics of Artificial Intelligence." In Ethics of Artificial Intelligence, edited by S. Matthew Liao, 212–236. Oxford: Oxford University Press, 2020. https://doi.org/10.1093/oso/9780190905040.003.0008.

Bartneck, Christoph, Christoph Lütge, Alan Wagner, and Sean Welsh. Military Uses of AI. 2021. https://doi.org/10.1007/978-3-030-51110-4_11.

Chaudhry, Hamza, and Landon Klein. *Chemical & Biological Weapons and Artificial Intelligence: Problem Analysis and US Policy Recommendations*. Cambridge, MA: Future of Life Institute, 2024. https://futureoflife.org/document/chemical-biological-weapons-and-artificial-intelligence-problem-analysis-and-us-policy-recommendations/.

Congressional Research Service. Defense Primer: U.S. Policy on Lethal Autonomous Weapons. Washington, D.C.: Congressional Research Service, 2020.

Custers, Bart, ed. The Future of Drone Use: Opportunities and Threats from Ethical and Legal Perspectives. The Hague: T.M.C. Asser Press, 2016. https://doi.org/10.1007/978-94-6265-132-6.

Department of Defense. Department of Defense Directive 3000.09: Autonomy in Weapon Systems. Updated May 8, 2017.

https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf.

Etzioni, Amitai, and Oren Etzioni. "Pros and Cons of Autonomous Weapons Systems." Military Review, May–June 2017. https://www.armyupress.army.mil/Journals/Military-Review/English-Edition-Archives/May-June-2017/Pros-and-Cons-of-Autonomous-Weapons-Systems/.

Floridi, Luciano. “Robots, Jobs, Taxes and Responsibilities.” In Robo-Ethics: Humans, Machines and Health, edited by Vincenzo Paglia and Renzo Pegoraro, 109–113. Rome: Pontifical Academy for Life, 2020.

Future of Life Institute. "Statement to United Nations on Behalf of LAWS Open Letter Signatories." Future of Life Institute. Last modified August 2018. https://futureoflife.org/open-letter/statement-to-united-nations-on-behalf-of-laws-open-letter-signatories/.

Galliott, Jai, and John Forge. “Debate on the Ethics of Developing AI for Lethal Autonomous Weapons.” The Philosophical Journal of Conflict and Violence 5, no. 1 (2021): 133–142. https://doi.org/10.22618/TP.PJCV.20215.1.139009.

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

Hage, Jaap. “Theoretical Foundations for the Responsibility of Autonomous Agents.” Artificial Intelligence and Law 25, no. 3 (2017): 255–271. https://doi.org/10.1007/s10506-017-9208-7.

Heyns, Christof. Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions. United Nations, 2014.

Kahn, Leonard. “Military Robots and the Likelihood of Armed Combat.” In Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence, edited by Patrick Lin, Keith Abney, and Ryan Jenkins, 274–287. New York: Oxford University Press, 2017. https://doi.org/10.1093/oso/9780190652951.003.0018.

McKenna, Alan. “The Public Acceptance Challenge and Its Implications for the Developing Civil Drone Industry.” In The Future of Drone Use: Opportunities and Threats from Ethical and Legal Perspectives, edited by Bart Custers. The Hague: T.M.C. Asser Press, 2016.

Melzer, Nils. “Human Rights Implications of the Usage of Drones and Unmanned Robots in Warfare.” Directorate-General for External Policies of the Union. European Parliament, 2013. https://www.europarl.europa.eu/RegData/etudes/etudes/join/2013/410220/EXPO-DROI_ET(2013)410220_EN.pdf.

Müller, Vincent C. "Autonomous Killer Robots Are Probably Good News." *Frontiers in Robotics and AI* 3 (2016). https://philarchive.org/rec/MLLAKR.

Poljak, Mario. “The History of Drones: Timeline from 1907 to 2019.” DroneTechPlanet, February 5, 2019. https://www.dronetechplanet.com/the-history-of-drones-timeline-from-1907-to-2019/.

Roston, Brittany. “Everything We Know About Samsung’s Machine Gun Robots.” SlashGear, March 24, 2022. https://www.slashgear.com/825074/everything-we-know-about-samsungs-machine-gun-robots/.

Roth, Marcus. “Artificial Intelligence in the Military – An Overview of Capabilities.” Emerj, February 20, 2019. https://emerj.com/artificial-intelligence-in-the-military-an-overview-of-capabilities/.

Shane, Janelle. “The Danger of A.I. Is Weirder than You Think.” TEDx Talks. Video, 10:29. November 14, 2019. https://www.youtube.com/watch?v=OhCzX0iLnOc.

Shea, Annemarie. “The Legal and Ethical Challenges Posed by Lethal Autonomous Weapons.” Trinity College Law Review 24 (2021): 117–133.

Sparrow, Robert. “Killer Robots.” Journal of Applied Philosophy 24, no. 1 (2007): 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x.

Tucker, Patrick. “NGA Will Take Over Pentagon’s Flagship AI Program.” Defense One, April 25, 2022. https://www.defenseone.com/technology/2022/04/nga-will-take-over-pentagons-flagship-ai-program/366098/.

UNESCO. *Ethics of Artificial Intelligence*. Paris: UNESCO, 2002. https://unesdoc.unesco.org/ark:/48223/pf0000139578.

UNIDIR. The Human Element in Decisions About the Use of Force. Last accessed February 28, 2023. https://unidir.org/sites/default/files/2020-03/UNIDIR_Iceberg_SinglePages_web.pdf.

Abstract

A.I. and Lethal Weapons: A blameless army of killer robots?

Lethal Autonomous Weapons Systems (L.A.W.S) and robot armies are the way of the future for conducting wars. This means that A.I. systems will make life and death decisions during war attacks. A major issue that derives from this potential situation is the challenging part of placing the responsibility on someone. Who is to be accountable for, when a machine, fully equipped to function and "think" on its own, kills civilians? Is it simply a matter of individual blaming divided amongst the parties concerned that helped create the A.I. machines or is it something that needs to make us reconsider the basis of our moral values? We should be more worried, not about putting a name on a blameworthy derail but on what kind of ethics we allow such a derail to be caused.

Keywords: Lethal Autonomous Weapon Systems, responsibility gap, ethics, killer robots.

Λέξεις-κλειδιά: Αυτόνομα Φονικά Οπλικά Συστήματα, κενό ευθύνης, ηθική, στρατός ρομπότ.

Lydia Kornaraki
kornaraki.lydia@gmail.com
ORCID iD: https://orcid.org/0000-0003-4478-2684

Ιωάννα ΜΑΛΑΝΔΡΑΚΗ

Μηχανές-Δικαστές με Τεχνητή Νοημοσύνη. Με Ηθική;

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

Ι. Εισαγωγή

Ο ΠΟΛΙΤΙΣΜΟΣ ΜΑΣ ΠΙΣΤΩΝΕΤΑΙ ΣΤΗ ΝΟΗΜΟΣΥΝΗ ΜΑΣ,[1] ΓΡΑΦΕΙ Ο καθηγητής επιστήμης των υπολογιστών Stuart Russel και τα λόγια του περιτριγυρίζουν στη σκέψη μου.

Η παραδοχή ότι διανύουμε μια εποχή με έναν ορμητικά καλπάζοντα ρυθμό προόδου στον τεχνολογικό πολιτισμό είναι αδιαμφισβήτητη. Στο άκουσμα της λέξεως «τεχνολογία» κάποιοι/ες θα προτάξουν τις θετικές επιρροές της στην καθημερινή ζωή, την προσφορά της στις επιστήμες, κάποιοι/ες θα συλλογιστούν την αρνητική της όψη προβαίνοντας σε βαθιά παρατήρηση των επιδράσεών της, ίσως στον χώρο της χρήσης νέων εργαλείων για χάρη της βίας και κάποιοι/ες θα προβληματιστούν για την ροπή, που δημιουργείται στον άνθρωπο, προς την πίστη ότι όλα μπορούν να υλοποιηθούν από τα επινοήματά του, τα τεχνήματά του κατευθύνοντάς τον σε μία παθητική στάση με κανένα κίνητρο για να διεκδικήσει, όπως πριν, ό,τι του αναλογεί.

Η ζωή, πλέον, φαίνεται να ορίζεται από την τεχνολογία. Ο σύγχρονος άνθρωπος εξαναγκάζεται να επιθυμεί να είναι μέρος όλων των καινών γεγονότων, αφού οι περισσότεροι τομείς επιδέχονται συνεχή κατάκτηση νέων επιπέδων. Ως απόρροια, της εν λόγω συνθήκης, εγείρονται εύλογες ανησυχίες σε σχέση με διάφορες εκφάνσεις χρήσης των τεχνολογικών επιτευγμάτων. Μία από τις εν λόγω εκφάνσεις είναι η Τεχνητή Νοημοσύνη

[1] Stuart Russel, *Συμβατή με τον άνθρωπο; η Τεχνητή Νοημοσύνη και το πρόβλημα του ελέγχου*, μτφρ. Νίκος Αποστολόπουλος (Εκδοτικός Οίκος ΤΡΑΥΛΟΣ, 2021), 141.

(Τ.Ν.) η οποία, έχοντας εκτεταμένη χρήση σε ποικίλα πεδία της ανθρώπινης δραστηριότητας, καταλαμβάνει, ίσως, την πιο αιχμηρή θέση. Ένα από τα επιμέρους πεδία εφαρμογής της Τ.Ν., το οποίο αναδεικνύει την αιχμηρότητά της και αποτελεί αντικείμενο πραγμάτευσης του παρόντος δοκιμίου είναι η λήψη δικαστικών αποφάσεων.

ΙΙ. Τεχνητή Νοημοσύνη και Δίκαιο: μία αξιοσημείωτη περίπτωση

Ο ρόλος της Τ.Ν. τη δεδομένη χρονική στιγμή στην πρακτική του Δικαίου είναι επικουρικός και πραγματώνεται μέσω λογισμικών συστημάτων που αποσκοπούν στην προσομοίωση της ικανότητας του ανθρώπου να λαμβάνει αποφάσεις.[2] Το COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) συνιστά ένα από τα πιο γνωστά λογισμικά ποινικών αδικημάτων. Το 1998 οι Tim Brennan και Dave Wells, επικεφαλής του Ινστιτούτου Northpointe, ανέπτυξαν τον αλγόριθμο COMPAS, ένα εργαλείο εκτίμησης κινδύνου πιθανής υποτροπιάζουσας παραβατικής συμπεριφοράς, που χρησιμοποιείται σε πολλές πολιτείες των ΗΠΑ,[3] παρέχοντας αρωγή στους/στις δικαστές στο στάδιο της λήψης αποφάσεων προανακριτικής απελευθέρωσης για τον/την εκάστοτε κατηγορούμενο/η.[4] Ωστόσο, είναι σημαντικό να σημειωθεί το γεγονός ότι, παρότι ο εν λόγω αλγόριθμος εξάγει έναν βαθμό που αντιστοιχεί σε μία από τις κατηγορίες κινδύνου πιθανής επανάληψης, η ακριβής πεπερασμένη αλληλουχία εντολών, η οποία τον συνθέτει, δε έγκειται στη σφαίρα γνώσης μας.[5] Επομένως, η άγνωστη εσωτερική λειτουργία του συστήματος καθιστά ανέφικτη τη λογοδοσία του, η οποία συνδέεται άρρηκτα με την ευθύνη, που χαρακτηρίζεται από ηθική αιτιότητα και είναι ένας από τους θεμελιώδεις στόχους στο πλαίσιο της εύρυθμης δικαστικής λειτουργίας.

---

[2] Peter Jackson, *Introduction to Expert Systems* (Addison-Wesley, 1999), 1.

[3] Alexandra Mac Taylor, "AI Prediction Tools Claim to Alleviate an Overcrowded American Justice System… But Should they be Used?," *Stanford Politics*, September 13, 2020, https://stanfordpolitics.org/2020/09/13/ai-prediction-tools-claim-to-alleviate-an-overcrowded-american-justice-system-but-should-they-be-used/.

[4] Eugenie Jackson and Christina Mendoza, "Setting the Record Straight: What the COMPAS Core Risk and Need Assessment Is and Is Not," *Harvard Data Science Review* 2, no. 1 (2020): 3, https://doi.org/10.1162/99608f92.1b3dadaa.

[5] Ellora Thadaney Israni, "When an Algorithm Helps Send You to Prison," *The New York Times,* October 26, 2017, https://www.nytimes.com/2017/10/26/opinion/algorithm-compas-sentencing-bias.html.

Σχεδόν μία δεκαετία πριν, το 2014, έλαβε χώρα μία τρόπον τινά ληστεία στην Πολιτεία της Φλόριντα. Η δεκαοχτάχρονη Brisha Borden και μία φίλη της έχοντας καθυστερήσει να παραλάβουν την πνευματική αδερφή της πρώτης από το σχολείο, όταν εντόπισαν στον δρόμο ένα παιδικό ποδήλατο κι ένα σκούτερ, χωρίς ιδιοκτήτη, επιχείρησαν να τα οδηγήσουν για να φτάσουν γρηγορότερα στον προορισμό τους.[6] Ευθύς αμέσως αντιλήφθηκαν ότι δεν ήταν εφικτό να κινηθούν με αυτά εξαιτίας του μεγέθους τους και εγκατέλειψαν την προσπάθεια. Λίγο αργότερα, όμως, συνελήφθησαν από τις αστυνομικές αρχές λόγω της αναφοράς του περιστατικού από έναν γείτονα. Ως αποτέλεσμα, οι δύο φίλες κατηγορήθηκαν για διάρρηξη και κλοπή αντικειμένων, με τη συνολική χρηματική αξία των κλοπιμαίων να ανέρχεται σε ογδόντα δολάρια. Ωστόσο, αυτή δεν ήταν πρωτόγνωρη κατάσταση για την Borden, αφού στο παρελθόν είχε εμπλακεί σε παραπτώματα.[7] Στον αντίποδα αυτής της υπόθεσης, το 2013 εκτυλίχθηκε ένα άλλο περιστατικό με τον σαρανταενάχρονο Vernon Prater, έναν άνδρα με βεβαρυμμένο ποινικό μητρώο, ο οποίος συνελήφθη για κλοπή εργαλείων αξίας περίπου ογδόντα έξι δολαρίων από ένα κατάστημα.[8]

Στο στάδιο της προανακριτικής διαδικασίας, ο αλγόριθμος εκτίμησε τη μελλοντική έκβαση. Εξήγαγε το πόρισμα ότι η Brisha Borden είχε υψηλή προδιάθεση να επαναλάβει παραβατική συμπεριφορά, ενώ ο Vernon Prater χαμηλή.[9] Μετά το πέρας δυο χρόνων, δεν είχε καταλογιστεί νέο παράπτωμα στην Brisha Borden, εν αντιθέσει με τον Vernon Prater ο οποίος είχε προβεί σε ληστεία ηλεκτρονικών ειδών αξίας χιλιάδων δολαρίων από την αποθήκη ενός σπιτιού με συνέπεια την έκτιση ποινής οκτώ ετών.[10] Όπως διαπιστώθηκε από την έρευνα της ProPublica,[11] ο αλγόριθμος δε βασίστηκε στο παρελθόν των κατηγορουμένων και στην πράξη που διέπραξαν, αλλά στη φυλετική προέλευσή τους, αφού η απόφαση που εξεδόθη, από το σύστημα, σχετιζόταν με το σκουρόχρωμο δέρμα της Brisha Borden και το ανοιχτόχρωμο του Vernon Prater.[12]

---

[6] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner, "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks," *ProPublica,* May 23, 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[7] Στο ίδιο.

[8] Στο ίδιο.

[9] Στο ίδιο.

[10] Στο ίδιο.

[11] Μη κερδοσκοπική οργάνωση ερευνητικής δημοσιογραφίας για το δημόσιο συμφέρον.

[12] Angwin et al., "Machine Bias."

Σε αυτό το σημείο αναδεικνύεται ζήτημα καίριας σημασίας. Η προκατάληψη και η μεροληψία που αποκτούν τα συστήματα Τ.Ν., εξαιτίας του/της προγραμματιστή/ριας τους ή της μεθόδου βαθιάς μάθησης, ενός τρόπου μηχανικής μάθησης που επιτυγχάνεται μέσω της εμπειρίας των συστημάτων, τα κατευθύνουν προς την εξαγωγή λανθασμένων δεδομένων και τα καθιστούν επισφαλή. Μάλιστα, το ζήτημα της άγνωστης εσωτερικής λειτουργίας των συστημάτων Τ.Ν. δυσχεραίνει την αποτροπή προβλημάτων αυτού του είδους. Είναι, λοιπόν, αναγκαία η διαφανής Τ.Ν., αφού το σύστημα εκθέτοντας τον τρόπο με τον οποίο εξήγαγε ένα συγκεκριμένο αποτέλεσμα, θα ενημερώνει τον άνθρωπο-χειριστή και η περιστολή λήψης λανθασμένης απόφασης, ειδικά σε κρίσιμες συνθήκες – όπως στη λήψη δικαστικής απόφασης – κατά συνέπεια, θα δύναται να επιτυγχάνεται.

Επιπροσθέτως, όπως φαίνεται, η χρήση της Τ.Ν. δεν θα περιοριστεί στον επικουρικό ρόλο της. Στην πρωτεύουσα της Κίνας, το Πεκίνο, έχει, ήδη, κατασκευαστεί η πρώτη μηχανή-δικαστής του κόσμου με Τ.Ν., η οποία παρουσιάστηκε στο ευρύ κοινό τον Ιούνιο του 2019.[13] Σε αυτό ακριβώς το σημείο έγκειται ο ηθικοφιλοσοφικός προβληματισμός για τις ηθικώς δρώσες τεχνητές οντότητες που κέντρισε το ενδιαφέρον μου και συνοψίζεται στο ερώτημα: έχουμε αναλογιστεί το ενδεχόμενο τα συστήματα Τ.Ν., γενικότερα αλλά και ειδικότερα στον δικαστικό κλάδο, να καταστούν κάποια στιγμή υπερ-νοήμονα και να αυτονομηθούν με αποτέλεσμα να βλάψουν – έστω και άθελά τους – την ανθρωπότητα;[14]

III. Η Ηθική της Τεχνητής Νοημοσύνης στο πλαίσιο της Υπερ-νοημοσύνης

Η Τ.Ν. με τη δράση της σε ποικίλα πεδία της ανθρώπινης δραστηριότητας έχει ενεργοποιήσει ηθικές αρχές και κατηγορίες, όπως συμβαίνει, άλλωστε, σε κάθε πεδίο που δημιουργεί ηθικά διλήμματα. Το ζήτημα που εκκινεί τον ηθικοφιλοσοφικό προβληματισμό του παρόντος δοκιμίου είναι

[13] "Beijing Internet Court launches online litigation service center," Beijing Internet Court, last modified July 1, 2019, https://english.bjinternetcourt.gov.cn/2019-07/01/c_190.htm.

[14] Για μια σύνοψη των ηθικών προβληματισμών που εγείρει η δυνατότητα δικαστηριακής χρήσης της Τ.Ν., καθώς και για τις οντολογικής και επιστημολογικής φύσεως προεκτάσεις της, βλ.: Άλκης Γούναρης και Γιώργος Κωστελέτος, «Γράφοντας τον αλγόριθμο του Καλού: Η Τεχνητή Νοημοσύνη ως μηχανή απόδοσης Δικαιοσύνης», *Ηθική. Περιοδικό φιλοσοφίας* 19 (2024): 5-27. https://doi.org/10.12681/ethiki.39654.

η προκατάληψη και η μεροληψία που αποκτούν τα συστήματα Τ.Ν.. Όπως έχω, ήδη, υποδείξει το η προκατάληψη της Τ.Ν. εκδηλώνεται είτε απευθείας από τον/την ίδιο/α τον/την προγραμματιστή/ρια του συστήματος Τ.Ν. είτε από τον τρόπο εκπαίδευσής του, αφού έχει προηγηθεί η έκθεσή του σε ένα περιβάλλον από το οποίο αντλεί ερεθίσματα.

Τι συγκροτείται, όμως, στη σημασία που αποδίδεται στον όρο «προκατάληψη της Τ.Ν.»; Η μεροληψία της Τ.Ν. είναι δυνατό να προκύψει ποικιλοτρόπως. Κατά πρώτον, μία περίπτωση είναι ένα σύστημα να χαρακτηρίζεται από προκατάληψη διότι ο/η αρμόδιος/α να του εμφυσήσει αξίες προσλαμβάνει επιρροές για ένα αντικείμενο πραγμάτευσης και έπειτα τις εφαρμόζει (εν γνώσει ή εν αγνοία του) σε διαφορετικό αντικείμενο με συνέπεια να καταλήγει σε λανθασμένους συλλογισμούς.[15] Επομένως, κατά αντιστοιχία, αυτό το άτομο εκπαιδεύει το σύστημα Τ.Ν. να δρα τοιουτοτρόπως.[16] Κατά δεύτερον, απασχολεί τη συζήτηση η ανθρώπινη τάση προς τη γνωστική προκατάληψη, ήτοι προς την ερμηνεία μίας πληροφορίας μέσα σε ένα συγκεκριμένο πλαίσιο, ώστε αυτή να αποτελεί επαλήθευση για κάτι που κάποιος/α πιστεύει, χωρίς να έχει παρέλθει το στάδιο της λογικής διεργασίας και της διασταύρωσης πηγών για ασφαλή γνώση.[17] Αναλογικά, το σύστημα Τ.Ν. αφομοιώνει, μέσω της μηχανικής μάθησης, από τον/την χρήστη τί να δέχεται και τί να αποκρούει στη βάση συγκεκριμένων πεποιθήσεων. Ακόμα, τίθεται επί τάπητος η στατιστική προκατάληψη.[18] Η εν λόγω μορφή προκύπτει από την αρχική αμεροληψία του συστήματος, με την έννοια ότι εκπαιδεύτηκε για μία συγκεκριμένη συνθήκη και έδρασε ορθώς, αλλά, εν συνεχεία, λόγω της ανεπάρκειάς του να αντιληφθεί την αιτία που προχώρησε σε εκείνη την κρίση, χρησιμοποίησε το συγκεκριμένο σύνολο δεδομένων σε άλλη συνθήκη με διαφορετικά δεδομένα και ακολουθώντας την προηγούμενη στρατηγική έδρασε, εν τέλει, μεροληπτικά.[19]

Λαμβάνοντας τη σκυτάλη από την τελευταία μορφή μεροληψίας ενός συστήματος Τ.Ν συλλογίζομαι την υπόθεση ενός υπερ-νοήμονος συστήματος. Με λίγα λόγια, στρέφω την προσοχή μου στην περίπτωση που

[15] Vincent C. Müller, "Ethics of Artificial Intelligence and Robotics," *The Stanford Encyclopedia of Philosophy* (Summer 2020 Edition), ed. Edward N. Zalta, https://plato.stanford.edu/archives/sum2020/entries/ethics-ai/.
[16] Στο ίδιο.
[17] Στο ίδιο.
[18] Στο ίδιο.
[19] Στο ίδιο.

μηχανές-δικαστές λειτουργούν μελλοντικά με επιβλαβή τρόπο για την ανθρωπότητα, ακόμα και αν δρουν δίκαια, έχοντας αποδεσμευτεί από ζητήματα προκατάληψης λόγω της υπερ-νοήμονος κατάστασής τους, υιοθετώντας στρατηγική που επιφέρει το θεμιτό αποτέλεσμα, αλλά κατά την εφαρμογή της δημιουργεί άλλου είδους προβλήματα.

Ο Max Tegmark, καθηγητής φυσικής στο MIT και πρόεδρος του Ινστιτούτου για το Μέλλον της Ζωής, στο βιβλίο του που τιτλοφορείται *Life 3.0-Τι θα σήμαινε να είσαι άνθρωπος στην εποχή της τεχνητής νοημοσύνης;* θέτει ποικίλα ερωτήματα για την υπερ-νοήμονα συνθήκη της Τ.Ν. Διερωτάται, μεταξύ άλλων, πώς είναι δυνατό να κατασκευαστεί Τ.Ν. με στόχους και να διασφαλιστεί ότι οι εν λόγω στόχοι θα παραμείνουν ίδιοι ακόμη και αν η Τ.Ν. γίνει πιο ευφυής.[20] Ο καθηγητής επισημαίνει ότι τη στιγμή της δημιουργίας συστήματος Τ.Ν. οι αξίες, οι ηθικοί και νομικοί κανόνες, που θα εμφυσήσουν σε αυτό, εξαρτώνται από τον/την σχεδιαστή/ρια του και κατ' επέκταση, όταν το σύστημα καταστεί υπερ-νοήμων, θα είναι στον ύψιστο βαθμό ικανό να εκπληρώσει τους στόχους του,[21] εφόσον η νοημοσύνη συνεπάγεται την ικανότητα πραγμάτωσης στόχων.

Επομένως, ο Max Tegmark αναρωτιέται αν μπορούμε να ευθυγραμμίσουμε τους στόχους μας με εκείνους ενός υπερ-νοήμονος συστήματος και ως αποτέλεσμα να έχουμε μία «φιλική Τ.Ν.», όπως την χαρακτηρίζει ο Αμερικανός θεωρητικός Eliezer Yudkowsky ή αν το σύστημα Τ.Ν. λόγω της υπερ-νοημοσύνης του θα ενεργήσει με δικούς του στόχους, με δυνητικά καταστροφικές συνέπειες, ακόμα και χωρίς σκόπιμη επιδίωξη τέτοιας δράσης.[22] Θεωρεί, δηλαδή, ότι το σύστημα ενδέχεται να έχει την ικανότητα να θέτει δευτερεύοντες στόχους ή ακόμα και να λειτουργεί βάσει δικών του στόχων, ώστε να επιτύχει τον τελικό στόχο του, διότι θα έχει τη δυνατότητα αναστοχασμού των στόχων με τους οποίους έχει διαποτιστεί.[23] Συνεπώς, αναδύεται ο προβληματισμός που προκύπτει από τον συγκερασμό του ζητήματος σχετικά με το ηθικό και νομικό σύστημα αξιών, το οποίο θα εισάγεται σε ένα σύστημα Τ.Ν. και του ζητήματος ελέγχου της Τ.Ν. στο μέλλον.

Την ίδια στάση υιοθετεί ο Σουηδός φιλόσοφος Nick Bostrom. Εκείνος έχει περιγράψει το πείραμα σκέψης "paperclip maximizer," το οποίο

[20] Max Tegmark, *LIFE 3.0 Τι θα σήμαινε να είσαι άνθρωπος στην εποχή της τεχνητής νοημοσύνης;,* μτφρ. Νίκος Αποστολόπουλος (Εκδοτικός Οίκος ΤΡΑΥΛΟΣ, 2018), 375.
[21] Στο ίδιο, 391.
[22] Στο ίδιο, 390-404.
[23] Στο ίδιο, 390-403.

συνοψίζεται στην ιδέα ότι «μία Τ.Ν. σχεδιασμένη να διαχειρίζεται την παραγωγή σε ένα εργοστάσιο, έχει ως τελικό στόχο τη μεγιστοποίηση της κατασκευής συνδετήρων και προχωρά με τη μετατροπή πρώτα της Γης και έπειτα ολοένα και περισσότερο μεγαλύτερων κομματιών του παρατηρήσιμου σύμπαντος σε συνδετήρες».[24] Έτσι, ο τελικός στόχος δεν είναι επιζήμιος για τον άνθρωπο, αλλά η υπερ-νοημοσύνη επιχειρώντας να έχει το μέγιστο δυνατό αποτέλεσμα είναι ικανή να μετατρέψει τους ανθρώπους σε συνδετήρες, υποθέτοντας ότι εκείνοι μπορούν να την απενεργοποιήσουν με συνέπεια να μη δημιουργηθούν πολλοί συνδετήρες.[25]

Σε αντίθεση με την προαναφερθείσα καταστρεπτική θέση, οι Michael Anderson και Susan Leigh Anderson, καθηγητές Φιλοσοφίας, συνιστούν μια πιο αισιόδοξη προσέγγιση. Σε συνέντευξή τους ερωτήθηκαν για το κατά πόσο πιστεύουν ότι ενυπάρχει κίνδυνος για την ανθρώπινη ζωή από την Τ.Ν. στο εγγύς μέλλον, αν τα συστήματα καταστούν πιο ευφυή και αποκρίθηκαν ότι η δράση των συστημάτων Τ.Ν. προσδιορίζεται από τον τρόπο ανάπτυξής τους επισημαίνοντας ότι μη απειλητικές μηχανές μπορούν να σταθούν αρωγοί μας ακόμα και στη βελτίωση της συμπεριφοράς μας· βέβαια, δεν αγνοούν τους κινδύνους που έθεσε η Τ.Ν., όμως θεωρούν πως τοιουτοτρόπως θα διαφυλαχτεί η νοημοσύνη μας στην περίπτωση που κινδυνέψουμε γενικά.[26]

Συν τοις άλλοις, οι Anderson εισηγούνται το ερευνητικό πρόγραμμα «Ηθική των Μηχανών».[27] Επιχειρούν, μεταξύ άλλων, να δημιουργήσουν ένα ασφαλές περιβάλλον για τους ανθρώπους που τρέφουν ανησυχίες και αβεβαιότητες σχετικά με το ενδεχόμενο ύπαρξης ηθικών συστημάτων Τ.Ν. στο πλαίσιο των αυτόνομων ευφυών συστημάτων Τ.Ν.[28] Ως απάντηση προβάλλουν την ανθρωποκεντρική διάσταση του ζητήματος, καθώς υπογραμμίζουν ότι «η ανησυχία ότι οι μηχανές που ξεκινούν να συμπεριφέρο-

[24] Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press, 2014), 149, δική μου μετάφραση.

[25] Στο ίδιο.

[26] Michael Anderson, Susan Leigh Anderson, Alkis Gounaris, and George Kosteletos, "Towards Moral Machines: A Discussion with Michael Anderson and Susan Leigh Anderson," *Conatus – Journal of Philosophy* 6, no. 1 (2021): 192-193, https://doi.org/10.12681/cjp.26832.

[27] «Δημιουργία μιας μηχανής που η ίδια να ακολουθεί μία ηθική αρχή ή ένα σύνολο αρχών, ήτοι να καθοδηγείται από αυτήν την αρχή ή από αυτές τις αρχές στις αποφάσεις που λαμβάνει περί των πιθανών κατευθύνσεων δράσης που θα μπορούσε να λάβει». Michael Anderson and Susan Leigh Anderson, "Machine Ethics: Creating an Ethical Intelligent Agent," *AI Magazine* 28, no. 4 (2007): 15, https://doi.org/10.1609/aimag.v28i4.2065.

[28] M. Anderson and S. L. Anderson, "Machine Ethics," 16.

νται ηθικά θα καταλήξουν να συμπεριφέρονται ανήθικα, ίσως ευνοώντας τα συμφέροντά τους, μπορεί να οφείλεται σε φόβους που προέρχονται από εύλογες ανησυχίες σχετικά με την ανθρώπινη συμπεριφορά. Οι περισσότεροι άνθρωποι απέχουν από τα ιδανικά μοντέλα ηθικών παραγόντων, παρά το γεγονός ότι έχουν διδαχθεί ηθικές αρχές και τείνουν να ευνοούν τον εαυτό τους».[29] Οι Anderson υπογραμμίζουν ότι η δράση των συστημάτων εξαρτάται από την ανάπτυξή τους.[30]

Εν κατακλείδι, μεταβιβάζοντας τον προβληματισμό για τις υπερ-νοήμονες οντότητες Τ.Ν. στις δικαστικές αποφάσεις προκύπτουν δύο θέσεις. Από τη μία πλευρά, υπάρχει η ανησυχία ότι οι υπερ-νοήμονες οντότητες Τ.Ν. δε θα διακρίνονται ενδεχομένως από τη μέριμνα της Δικαιοσύνης ή ότι θα την αποδίδουν διαφορετικά από τους ανθρώπους και πως με αυτόν τον τρόπο θα είναι πιθανό να μπορούν να μας παραπλανούν σχετικά με τις προθέσεις τους, ακριβώς επειδή θα είναι υπερ-νοήμονες. Από την άλλη πλευρά, έχοντας μηχανές-δικαστές με υπερ-νοημοσύνη θα μπορούμε να δεχθούμε την απόφασή τους, ύστερα από την παραδοχή ότι έχουν θετικούς στόχους και μεριμνούν για το ανθρώπινο καλό, εφόσον θα αδυνατούμε πλέον να ακολουθήσουμε τη δράση τους.

IV. Μηχανές-Δικαστές

Το εφαλτήριο του συλλογισμού μου έγκειται σε μία δίκη ανάλογη με εκείνη του Otto Adolf Eichmann. Ανατρέχοντας στο έργο *Ο Άιχμαν στην Ιερουσαλήμ* της Hannah Arendt ο αναγνώστης μπορεί να εντοπίσει την άποψη πως οποιοσδήποτε άνθρωπος είναι ικανός να προβεί σε εγκλήματα κατά της ανθρωπότητας, αν έχει απογυμνωθεί από τις ηθικές ευθύνες του.

Τον 20ο αιώνα, το ζήτημα της βίας εξακολούθησε να αποτελεί αντικείμενο φιλοσοφικής συζήτησης στο επίπεδο της φαινομενολογικής σκέψης. Στο επίκεντρο είναι ο Γάλλος φιλόσοφος Frantz Fanon με το βιβλίο του με τίτλο *Της γης οι κολασμένοι* που εκδόθηκε στα 1961 και την εισαγωγή του οποίου συνέθεσε ο, επίσης Γάλλος φιλόσοφος, Jean-Paul Sartre. Το βιβλίο συνιστά μία εξύμνηση στη βία και σημείο αναφοράς για τους αντιαποικιοκράτες, διότι ωθεί τον καταπιεσμένο λαό προς την κατεύθυνση πως είναι δυνατό να κατακτήσει την ελευθερία μόνος του μέσω της βίας. Δηλώνει ότι η βία απελευθερώνει, καθώς «ο αποικιοκρατούμενος

[29] Στο ίδιο, 17, δική μου μετάφραση.
[30] Anderson et al., "Towards Moral Machines," 192-193.

ανακαλύπτει την πραγματικότητα και την μετουσιώνει στην κίνηση της δράσης του, στην εξάσκηση της βίας, στο σχέδιο του για απελευθέρωση».[31] Μάλιστα, ο Jean-Paul Sartre σχολιάζει τη σκέψη του Frantz Fanon τασσόμενος υπέρ του και υπερασπιζόμενος τη χειραφέτηση που προκαλεί η επαναστατική βία.

Στη συζήτηση εισέρχεται η Hannah Arendt. Στο *Περί Βίας* ασκεί την κριτική της, συνδιαλεγόμενη με τους δυο Γάλλους φιλοσόφους, ισχυριζόμενη ότι η βία δεν μπορεί να παράγει πολιτικά αποτελέσματα. Πιο συγκεκριμένα, η Γερμανίδα φιλόσοφος σημειώνει «ο Σαρτρ […] προχωρεί […] στην εξύμνηση της βίας […]. […] Η "βία", πιστεύει τώρα, βρίσκοντας έρεισμα στο βιβλίο του Φανόν, "σαν τη λόγχη του Αχιλλέα, μπορεί να γιατρέψει τις πληγές που προκάλεσε". […] Αν δούμε την ιστορία ως μια συνεχή χρονολογική διαδικασία, της οποίας η πρόοδος είναι μάλιστα αναπόφευκτη, η βία με τη μορφή του πολέμου και της επανάστασης μπορεί να φανεί πως συνιστά τη μόνη δυνατή διακοπή. Αν αυτό ήταν αλήθεια, αν μόνο η χρήση της βίας θα επέτρεπε να διακοπούν οι αυτόματες διαδικασίες στο πεδίο των ανθρώπινων πραγμάτων, οι κήρυκες της βίας θα είχαν κερδίσει σε ένα σημαντικό ζήτημα».[32]

Επιπλέον, η Hannah Arendt ισχυρίζεται ότι «η βία δεν είναι ούτε ζωώδης ούτε ανορθολογική».[33] Αρχικά, φρονεί πως κατά τον τρόπο που η απόδοση ανθρωπομορφικής συμπεριφοράς στα ζώα δεν είναι επιτυχημένη, ο εντοπισμός ζωωδών στοιχείων στον άνθρωπο δεν μπορεί να λειτουργήσει δικαιολογώντας ή καταδικάζοντας την ανθρώπινη συμπεριφορά,[34] αφού, μάλιστα, «η κατασκευή εργαλείων συνιστά μια εξαιρετικά περίπλοκη διανοητική δραστηριότητα».[35] Έπειτα, θεωρεί ότι η αντίδραση της οργής δε συνεπάγεται την ανορθολογική όψη της βίας, γιατί «μόνο εκεί όπου εύλογα υποπτευόμαστε ότι οι συνθήκες θα μπορούσαν να αλλάξουν μα δεν αλλάζουν εμφανίζεται η οργή. Μόνο όταν μας θίγουν το αίσθημα δικαίου αντιδρούμε με οργή […]».[36] Τέλος, η φιλόσοφος δέχεται μόνο σε μία περίπτωση να χαρακτηριστεί η βία ανορθολογική και αυτή είναι όταν

[31] Frantz Fanon, *Της γης οι κολασμένοι,* μτφρ. Αγγέλα Αρτέμη (Εκδόσεις Κάλβος, 1982), 33.
[32] Hannah Arendt, *Περί Βίας*, μτφρ. Βάνα Νικολαΐδου-Κυριανίδου (Εκδόσεις Αλεξάνδρεια, 2000), 75-93.
[33] Στο ίδιο, 122.
[34] Στο ίδιο, 119-120.
[35] Στο ίδιο, 122.
[36] Στο ίδιο, 123.

κινείται ενάντια σε υποκατάστατα εξαιτίας ψυχολογικών κινήτρων.[37]

Υπάρχει, λοιπόν, ηθική δικαιολόγηση για την πολιτική βία; Η Hannah Arendt είναι φιλόσοφος που ενδιαφέρεται για τον βίο, ήτοι για τη ζωή του όντος που έχει προσωπικότητα, που «γράφει ιστορία» με τα λεγόμενα και τις πράξεις του. Πιο συγκεκριμένα, αποσαφηνίζει στο έργο τής *Ανθρώπινη Κατάσταση* μέσω της φιλοσοφικής ανθρωπολογίας της πως ενδιαφέρεται για τον homo politicus, για εκείνον που πράττει στον δημόσιο χώρο, στον χώρο της ελευθερίας.[38] Μάλιστα, σημαντική θέση στη σκέψη της κατέχει η Αριστοτελική ηθική και πολιτική φιλοσοφία. Επομένως, για να δοθεί απάντηση στο ερώτημα, υπό την οπτική της φιλοσόφου, έχω τη γνώμη ότι θα πρέπει να εξεταστεί και να εφαρμοστεί η πρότασή της στη βάση της παραπάνω παραδοχής.

Η φιλόσοφος σε συζήτηση στα 1967 με θέμα «Η νομιμότητα της βίας ως πολιτική πράξη;» υπήρξε η μόνη που δεν ερμήνευσε το συγκεκριμένο ζήτημα ως ηθικό, αλλά ως πολιτικό. Αναλυτικότερα, εξήγησε πως στον πυρήνα της ηθικής φιλοσοφίας συναντάμε την πράξη του εαυτού μας κι όχι του κόσμου· «σε όλα τα ηθικά ζητήματα, μας απασχολεί ο εαυτός μας. Αναρωτιόμαστε αν είμαστε ένοχοι για κάτι, αν μπορούμε να ζήσουμε με τον εαυτό μας αφού έχουμε κάνει αυτό ή εκείνο. Αυτά είναι απολύτως θεμιτά και πολύ σημαντικά ερωτήματα, αλλά δεν είναι θεμελιωδώς πολιτικά. Στην πολιτική, ασχολούμαστε με τον κόσμο και όχι με τον εαυτό μας».[39] Σύμφωνα με αυτά, η φιλόσοφος επιχειρεί να προβάλλει την ηθική πολιτική όχι στο πεδίο του εαυτού αλλά στη δημόσια σφαίρα, ως πολίτη. Συμπερασματικά, και βάσει της επιρροής που δέχθηκε από τον Αριστοτέλη, φαίνεται ότι για τη Hannah Arendt η βία δεν έχει θέση στα πολιτικά ζητήματα και για αυτό δεν τίθεται θέμα ηθικής δικαιολόγησης. Η χρήση βίας, για τη φιλόσοφο, συνεπάγεται την έλλειψη ηθικής.

Επιστρέφοντας, όμως, στον θεματικό πυρήνα της παρούσας ενότητας, η αναφορά στη Hannah Arendt στοχεύει στην ανάδειξη του ζητήματος η υπερ-νοήμων δρώσα οντότητα T.N. να ακολουθεί το μοτίβο του Otto Adolf Eichmann, ήτοι να αναγνωρίζει και να επικροτεί την τήρηση των καθηκόντων, αλλά να μην περιλαμβάνει ουδεμία ηθική οπτική στη δράση της. Η φιλόσοφος έχοντας παρακολουθήσει τη δίκη ως δημοσιογράφος διατείνεται ότι διαπίστωσε πως ο Otto Adolf Eichmann ενεργούσε βάσει

[37] Στο ίδιο, 124-127.

[38] Hannah Arendt, *The Human Condition* (The University of Chicago Press, 1998), 22-28.

[39] Noam Chomsky, Hannah Arendt and Susan Sontag, "The Legitimacy of Violence as a Political Act," in *Dissent, Power and Confrontation*, ed. Alexander Klein (McGraw-Hill, 1971), 116, δική μου μετάφραση.

καθήκοντος χωρίς να εμπλέκει την ηθική κρίση στις ενέργειές του.[40]

Δυνάμει της εν λόγω παρατήρησης, θα μπορούσε ο/η δικαστής, όπως ίσως και η μηχανή-δικαστής, να κρίνει ότι ο Otto Adolf Eichmann, πράγματι, λειτουργούσε αρμονικά μέσα σε ένα ηθικό πλαίσιο, για μία ομάδα ή βάσει μίας (μαθηματικής) λογικής και πάλι για μία συγκεκριμένη ομάδα, στην περίπτωση που η προαναφερθείσα θέση είναι αποδεκτή από το κοινωνικό σύνολο; Θα ήταν σε θέση η Τ.Ν., ακόμα κι αν ήταν υπερ-νοήμων, να λαμβάνει τέτοιου είδους αποφάσεις ούσα αυτόνομη; Δίχως να διατηρείται, δηλαδή, ο σημερινός επικουρικός ρόλος της; Αν μια μηχανή-δικαστής είναι ένα αλγοριθμικό σύστημα που απλώς εντοπίζει και δικαιώνει την τήρηση του καθήκοντος, χωρίς να πραγματεύεται τις επιπτώσεις της εν λόγω τήρησης στην εκάστοτε συνθήκη, ήτοι να μην προβαίνει στην ηθική εξέταση της περίστασης, είναι πολύ πιθανό να ελλοχεύει ο κίνδυνος για νέα εγκλήματα από την Τ.Ν., αν, μάλιστα, ληφθεί υπόψη και η στάση των Max Tegmark, Eliezer Yudkowsky και Nick Bostrom.

Ο αντι-παραγωγικός χαρακτήρας της βίας, σύμφωνα με τη Hannah Arendt, μπορεί να παραλληλιστεί με ένα «α-ηθικό» σύστημα λήψης αποφάσεων. Συγκεκριμένα, όπως η βία φέρεται να μην είναι ικανή να προσφέρει πολιτικά αποτελέσματα, η μηχανή-δικαστής ελλείψει ηθικών αρχών συνεπάγεται την αδυναμία επιβίωσης στην πόλη. Το αποδεκτό ή το κατεστημένο δεν ορίζει την ηθική, εν αντιθέσει η ηθική, η αυτοσυνείδηση της κοινωνίας, κρίνει και αμφισβητεί με επιχειρηματολογία βάσει κριτηρίων, ήτοι με λογοδοσία, το αποδεκτό ή το κατεστημένο.[41] Αν οι υπερ-νοήμονες μηχανές-δικαστές διαποτιστούν με ηθική, ο αριστοτελικός θεσμός της ηθικής πόλης δύναται να πραγματωθεί· οι πολίτες θα μπορούν, με την επανάληψη της καλής πράξης, να αποκτούν την ηθική αρετή – την ηθική του ορθού – και να γίνονται ενάρετοι αλληλεπιδρώντας με τους κοινωνικούς άλλους.[42] Επομένως, θα επιτευχθεί τοιουτοτρόπως η ηθική πόλη και θα

[40] Roger Berkowitz, "The Power of Non-Reconciliation – Arendt's Judgement of Adolf Eichmann," *HannahArendt.Net* 6, no. 1/2 (2012), https://doi.org/10.57773/hanet.v6i1/2.11. Δίχως, όμως, να παρουσιάζει αντίρρηση για την ενοχή του. Στο ίδιο.

[41] Σταυρούλα Τσινόρεμα, «Ηθικές Αρχές» (διάλεξη στο μάθημα Θεωρητική Ηθική ΙΙ του Προγράμματος Μεταπτυχιακών Σπουδών «Φιλοσοφία» - Κατεύθυνση: «Εφαρμοσμένη Ηθική», Φιλοσοφική Σχολή, Ε.Κ.Π.Α., Αθήνα, Μάρτιος 2022).

[42] Για την ιδέα ότι η Τ.Ν. μπορεί να λειτουργήσει ως ενάρετο σύστημα και να σμιλεύσει ενάρετους χαρακτήρες ανθρώπων-χρηστών, βλ.: Alkis Gounaris, George Kosteletos and Maria-Artemis Kolliniati, "Virtue in the machine: beyond a one-size-fits-all approach and Aristotelian ethics for Artificial Intelligence," Conatus – Journal of Philosophy 10, no.1 (2025): 127-152. https://doi.org/10.12681/cjp.40628

εξαλειφθεί η ανάγκη καταφυγής σε δικαστηριακές διαδικασίες, διότι θα έχει επέλθει η ύψιστη ηθική αρετή, η φιλία, σύμφωνα με τον Αριστοτέλη.[43]

Ωστόσο, αν οι υπερ-νοήμονες μηχανές-δικαστές διαθέτουν χαρακτηριστικά ανθρώπινης συμπεριφοράς, η επίτευξη του θεσμού της ηθικής πόλης δεν είναι εγγυημένη. Ούσα μέλος μίας κοινωνίας με ανθρώπους-δικαστές εντοπίζω την περίπτωση κάποιος/α δικαστής να είναι σε θέση να διακρίνει τί είναι ηθικό, αλλά να το αγνοεί επιδιώκοντας να ικανοποιήσει δικά του/της συμφέροντα και δημιουργώντας κατ' επέκταση ανάλογα αποτελέσματα. Ως εκ τούτου, λαμβάνοντας υπόψη αυτήν την παράμετρο, μία μηχανή-δικαστής, ακόμα κι αν είχε κρίση για να εντοπίσει το ηθικό, θα μπορούσε να παραβλέψει σκοπίμως την ηθική δράση.

V. Συμπεράσματα

Στην περίπτωση των μηχανών-δικαστών είναι αναγκαία η πρόσβαση στην εσωτερική λειτουργία των συστημάτων. Το αίτημα για διαφανή Τ.Ν. χρειάζεται να εξεταστεί και να γίνει αποδεκτό ακόμα και για τον σημερινό, επικουρικό, ρόλο των συστημάτων στη λήψη δικαστικών αποφάσεων. Η υπόθεση των υπερ-νοημόνων συστημάτων Τ.Ν., ίσως, να μην απασχολεί έντονα τη δεδομένη στιγμή το δικαστικό πεδίο, αλλά φρονώ, δεδομένης της χρήσης συστημάτων όπως το COMPAS, ότι σε τέτοια ρυθμιστικά πλαίσια, όπου διακυβεύεται το μέλλον κάποιων βάσει μίας απόφασης, η εξάλειψη της μεροληψίας από τα συστήματα είναι φλέγουσα και η γνωστοποίηση του αλγορίθμου μπορεί να συνδράμει στην αντιμετώπισή της.

Επίσης, οι νομικές και ηθικές αξίες που θα χορηγηθούν στις μηχανές-δικαστές και το ζήτημα του ελέγχου διαδραματίζουν σημαντικό ρόλο. Οι επιφυλακτικοί ερευνητές της Τ.Ν. μπορεί να διερωτώνται για το αν οι μηχανές-δικαστές θα ενδιαφέρονται για την απόδοση Δικαιοσύνης ή για το αν θα αντιλαμβάνονται τη Δικαιοσύνη όπως οι άνθρωποι. Ωστόσο, οι οπτιμιστές ερευνητές πιθανώς να παρατηρούν ότι τα υπερ-νοήμονα συστήματα απαλλαγμένα, ίσως, από την προκατάληψη της Τ.Ν., χάρη στην ικανότητά τους να διακρίνουν την προβληματική της μεροληψίας, θα αναλάβουν την προστασία μας και θα κρίνουν τις δικαστικές υποθέσεις ορθώς θέτοντας θετικούς στόχους.

Εναρμονίζοντας όλα τα παραπάνω, η ανάπτυξη μηχανών-δικαστών και η δράση τους απαιτεί την ύπαρξη ηθικής. Το σενάριο της ηθικής πόλης

[43] Τσινόρεμα, «Ηθικές Αρχές».

φαντάζει ουτοπικό, εντούτοις θεωρώ ότι στο ενδεχόμενο της υπερ-νοημοσύνης η ανθρώπινη φαρέτρα χρειάζεται να είναι γεμάτη με τρόπους δημιουργίας της καλύτερης δυνατής εκδοχής ενός τέτοιου συστήματος, ώστε να μην κινδυνέψουμε από εγκλήματα εκπορευόμενα από την Τ.Ν. Φαίνεται, λοιπόν, να μην είναι θεμιτό η Τ.Ν. να έχει αυτόνομο λόγο τουλάχιστον σε «ευαίσθητους» κλάδους της ζωής μας – όπως είναι ο δικαστικός κλάδος – δεδομένου ότι παραμονεύει ο κίνδυνος μιας υπερ-νοήμονος Τ.Ν. που ως τέτοια θα είναι ανεξέλεγκτη. Ας είναι σκοπός μας η βέλτιστη εκδοχή των τρεχόντων συστημάτων Τ.Ν., για την πιθανότητα εμφάνισης υπερ-νοημόνων συστημάτων, διότι «ο πολιτισμός μας πιστώνεται στη νοημοσύνη μας».

Αναφορές

Anderson, M., Anderson, S. L., Gounaris, A., & Kosteletos, G. (2021). Towards Moral Machines: A Discussion with Michael Anderson and Susan Leigh Anderson. *Conatus - Journal of Philosophy*, *6*(1), 177–202. https://doi.org/10.12681/cjp.26832

Anderson, Michael and Susan Leigh Anderson. "Machine Ethics: Creating an Ethical Intelligent Agent." *AI Magazine* 28, no. 4 (2007): 15-26. https://doi.org/10.1609/aimag.v28i4.2065.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. "Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks." *ProPublica,* May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Arendt, Hannah. *The Human Condition.* The University of Chicago Press, 1998.

Arendt, Hannah. *Περί Βίας*. Μεταφρασμένη από τη Βάνα Νικολαΐδου-Κυριανίδου. Εκδόσεις Αλεξάνδρεια, 2000.

Beijing Internet Court. "Beijing Internet Court launches online litigation service center." Last modified July 1, 2019. https://english.bjinternetcourt.gov.cn/2019-07/01/c_190.htm.

Berkowitz, Roger. "The Power of Non-Reconciliation – Arendt's Judgement of Adolf Eichmann." *HannahArendt.Net* 6, no. 1/2 (2012). https://doi.org/10.57773/hanet.v6i1/2.11.

Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.

Γούναρης, Άλκης και Γιώργος Κωστελέτος. «Γράφοντας τον αλγόριθμο του Καλού: Η Τεχνητή Νοημοσύνη ως μηχανή απόδοσης Δικαιοσύνης». *Ηθική. Περιοδικό φιλοσοφίας* 19 (2024): 5-27. https://doi.org/10.12681/ethiki.39654

Chomsky, Noam, Hannah Arendt and Susan Sontag. "The Legitimacy of Violence as a Political Act." In *Dissent, Power and Confrontation*, edited by Alexander Klein. McGraw-Hill, 1971.

Fanon, Frantz. *Της γης οι κολασμένοι*. Μεταφρασμένο από την Αγγέλα Αρτέμη. Εκδόσεις Κάλβος, 1982.

Gounaris, A., Kosteletos, G., & Kolliniati, M.-A. (2025). Virtue in the Machine: Beyond a One-size-fits-all Approach and Aristotelian Ethics for Artificial Intelligence. *Conatus - Journal of Philosophy*, *10*(1), 127–152. https://doi.org/10.12681/cjp.40628

Israni, Ellora Thadaney. "When an Algorithm Helps Send You to Prison." *The New York Times*, October 26, 2017. https://www.nytimes.com/2017/10/26/opinion/algorithm-compas-sentencing-bias.html.

Jackson, Eugenie and Christina Mendoza. "Setting the Record Straight: What the COMPAS Core Risk and Need Assessment Is and Is Not." *Harvard Data Science Review* 2, no. 1 (2020). https://doi.org/10.1162/99608f92.1b3dadaa.

Jackson, Peter. *Introduction to Expert Systems*. Addison-Wesley, 1999.

Müller, Vincent C. "Ethics of Artificial Intelligence and Robotics." *The Stanford Encyclopedia of Philosophy* (Summer 2020 Edition), edited by Edward N. Zalta. https://plato.stanford.edu/entries/ethics-ai/#BiasDeciSyst.

Russel, Stuart. *Συμβατή με τον άνθρωπο; η Τεχνητή Νοημοσύνη και το πρόβλημα του ελέγχου*. Μεταφρασμένο από τον Νίκο Αποστολόπουλου. Εκδοτικός Οίκος ΤΡΑΥΛΟΣ, 2021.

Τσινόρεμα, Σταυρούλα. «Ηθικές Αρχές». Διάλεξη στο μάθημα Θεωρητική Ηθική ΙΙ του Προγράμματος Μεταπτυχιακών Σπουδών «Φιλοσοφία» - Κατεύθυνση: «Εφαρμοσμένη Ηθική», Φιλοσοφική Σχολή, Ε.Κ.Π.Α., Αθήνα, Μάρτιος 2022.

Taylor, Alexandra Mac. "AI Prediction Tools Claim to Alleviate an Overcrowded American Justice System… But should they be Used?" *Stanford Politics,* September 13, 2020. https://stanfordpolitics.org/2020/09/13/ai-prediction-tools-claim-to-alleviate-an-overcrowded-american-justice-system-but-should-they-be-used/.

Tegmark, Max. *LIFE 3.0 Τι θα σήμαινε να είσαι άνθρωπος στην εποχή της τεχνητής νοημοσύνης;* Μεταφρασμένο από τον Νίκο Αποστολόπουλο. Εκδοτικός Οίκος ΤΡΑΥΛΟΣ, 2018.

Περίληψη

Ο ορμητικά καλπάζων ρυθμός προόδου στον τεχνολογικό πολιτισμό αναδιαμορφώνει τους περισσότερους τομείς της κοινωνίας. Είναι πρόδηλο πως η εξάρτηση, που προκύπτει, από την τεχνολογία ορίζει τη ζωή του σύγχρονου ανθρώπου εγείροντας συχνά εύλογες ανησυχίες για τα καινά γεγονότα. Η Τεχνητή Νοημοσύνη (Τ.Ν.) αποτελεί, ίσως, την πιο αιχμηρή έκφανση της τεχνολογίας έχοντας διεισδύσει σε ποικίλα πεδία της ανθρώπινης δραστηριότητας. Το παρόν δοκίμιο πραγματεύεται τη διαδικασία λήψης δικαστικών αποφάσεων ως ένα αξιοσημείωτο επιμέρους πεδίο εφαρμογής της Τ.Ν. Αρχικά, στο δοκίμιο γίνεται λόγος για τον τεχνολογικό πολιτισμό και τις επιδράσεις του. Εν συνεχεία, παρουσιάζεται η διεκπεραίωση δικαστικών αποφάσεων με Τ.Ν. τη δεδομένη χρονική στιγμή και προσεγγίζεται ο ηθικοφιλοσοφικός προβληματισμός για τις ηθικώς δρώσες τεχνητές οντότητες.

Λέξεις-κλειδιά: ηθική, Τεχνητή Νοημοσύνη, Ηθική των Μηχανών, ηθικώς δρώσα τεχνητή οντότητα, δικαστικές αποφάσεις

Keywords: ethics, Artificial Intelligence, Machine Ethics, artificial moral entities, judicial decision-making

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

Αλέξανδρος ΝΟΥΝΕΣΗΣ

Τεχνητή Νοημοσύνη και Αριστοτελική Επιείκεια

doi:https://doi.org/10.12681/plogos.33695

Πρόλογος

Η ΚΙΝΑ ΠΕΡΗΦΑΝΕΎΕΤΑΙ ΟΤΙ ΚΑΤΑΦΕΡΕ ΝΑ ΔΗΜΙΟΥΡΓΉΣΕΙ Τεχνητή Νοημοσύνη (AI) που μπορεί να πάρει τη θέση ενός δικαστή. Ανακάλυψα αυτό το γεγονός στις αρχές του 2022 καθώς ερευνούσα βιβλιογραφία σε σχέση με την πτυχιακή μου εργασία (καθώς και μια ακόμη εργασία που σχετιζόταν με την Τεχνητή Νοημοσύνη). Συνέχεια έπεφτα πάνω σε άρθρα ειδήσεων που μιλούσαν για αυτήν την AI της Κίνας, και όλα τα άρθρα επισήμαναν πως αυτή η AI υποτίθεται ότι φτάνει σε σωστές δικαστικές αποφάσεις με ποσοστό επιτυχίας 97% (Chen 2021; Newman 2021; Samson 2021; The Korea Times 2021).

Αυτό το 97% μου έκανε πραγματικά εντύπωση. Τι σημαίνει αυτή η επιτυχία; Η AI-δικαστής εκδικάζει δίκαια 97% όλων των υποθέσεων; Άρα τρεις στις εκατό φορές κάνει λάθος; Και τι είναι όμως το λάθος και το σωστό στο πλαίσιο της Δικαιοσύνης; Πως ορίζει η AI το τι είναι το Δίκαιο και το Άδικο;

Κάπως έτσι σπίθισε μέσα μου η ανάγκη να αναλύσω στον ελεύθερο μου χρόνο την έννοια της Δικαιοσύνης, και να τη συσχετίσω με την υπόσταση της AI. Ήθελα να καταλάβω αν γίνεται μια μηχανή να φέρεται Δίκαια, κι αν μπορεί να μας δείξει και σε μας επιτέλους τι είναι το Δίκαιο (μιας και βασανιζόμαστε χιλιετίες να το καταλάβουμε από μόνοι μας).

Η Δικαιοσύνη είναι μια έννοια αναπόσπαστη από μια ουσιαστική κουβέντα περί Ηθικής. Δεκάδες φιλόσοφοι έχουν παρουσιάσει τις δικές τους ερμηνείες πάνω σε αυτό το θέμα μέσα στις χιλιετίες της ανθρώπινης ιστορίας. Ένας από τους διάσημους και σημαντικότερους φιλοσόφους που

ανέλυσε εκτεταμένα το τι είναι Δικαιοσύνη ήταν ο Αριστοτέλης. Έχοντας λοιπόν προσωπικά μεγάλη εκτίμηση για τη φιλοσοφική σκέψη του Σταγειρίτη, θα ήθελα να εξετάσω σε αυτό το δοκίμιο τέτοιες απόπειρες δημιουργίας AI-δικαστών (όπως αυτές τις Κίνας) μέσα από το πρίσμα της Αριστοτελικά ορισμένης Δικαιοσύνης, όπως αυτή παρουσιάστηκε στο μνημειώδες έργο του, τα Ηθικά Νικομάχεια.[1]

Η Δικαιοσύνη μέσα από το Πρίσμα της Αριστοτελικής Ηθικής

Ο Αριστοτέλης, διαφοροποιούμενος σε ένα βαθμό από το δάσκαλο του Πλάτωνα, καταλάβαινε τη Δικαιοσύνη ως κάτι που δεν έχει μεταφυσική υπόσταση: κάτι που δεν υπάρχει ακριβώς «εκεί έξω». Ο Αριστοτέλης θεωρεί τη Δικαιοσύνη ως κάτι άρρηκτα και αποκλειστικά συνδεδεμένο με τον άνθρωπο. Είναι μια από τις μορφές/εκφάνσεις της Αρετής (και, με έναν τρόπο, η κυρίαρχη) που σημαίνει ότι είναι δυνάμει ένα κομμάτι του χαρακτήρα του ανθρώπου που προϋπάρχει ως δυνατότητά του, και μπορεί τελεολογικά να αναπτυχθεί και να γίνει πραγματικά κομμάτι του χαρακτήρα του, αρκεί να την καλλιεργήσει στη καθημερινότητά του (και από όσο πιο νεαρή ηλικία γίνεται) με Ενάρετες πράξεις.[2] Μερικές άλλες ηθικές Αρετές είναι η Ανδρεία, η Φιλικότητα και η Μεγαλοψυχία. Εάν επιτύχει κάποιος να γίνει Ενάρετος (και συνεπώς Δίκαιος, Ανδρείος κλπ.), αυτό θα αποδεικνύεται συνέχεια πλέον στις πράξεις του και το περιεχόμενο των επιθυμιών του. Η Αρετή του δηλαδή δεν θα είναι κρυφή, αλλά θα φαίνεται από το χαρακτήρα του όταν πρέπει να εκδηλώνεται προς τον κοινωνικό του περίγυρο.[3]

Πέραν του ότι ο Αριστοτέλης θεωρεί τη Δικαιοσύνη ως μια από τις εκφάνσεις της Αρετής, πρέπει να καταλάβουμε τι ακριβώς είναι η Δικαιοσύνη αυτοτελώς. Δίκαιος λοιπόν κατά τον Αριστοτέλη είναι κάποιος που

[1] Μια εκτενής φιλοσοφική σύνοψη και χαρτογράφηση των ηθικών προβληματισμών που εγείρονται από τη δυνατότητα δικαστηριακής χρήσης της AI έχουν παρουσιάσει οι Άλκης Γούναρης και Γιώργος Κωστελέτος στο άρθρο τους «Γράφοντας τον αλγόριθμο του Καλού: Η Τεχνητή Νοημοσύνη ως μηχανή απόδοσης Δικαιοσύνης» (2024). Οι συγγραφείς του άρθρου επιχειρούν να εξετάσουν αυτούς τους προβληματισμούς και μέσα από το πρίσμα κανονιστικών ηθικών θεωριών διαφορετικών από την Αριστοτελική. Το άρθρο τους προσφέρεται ως ένα θεωρητικό υπόβαθρο που συμπληρώνει τη δική μου μελέτη που εστιάζει συγκεκριμένα στη σχέση Αριστοτελικής επιείκειας και AI.

[2] Αριστοτέλης, *Ηθικά Νικομάχεια*, Β.1, 1103a14–1103b1· Β.4, 1105a30–1105b11.

[3] Αριστοτέλης, *Ηθικά Νικομάχεια*, Z.13, 1144b4–17.

ξέρει πότε ένα άτομο αδικείται όταν αυτό συμβαίνει. Ξέρει δηλαδή πότε ένα άτομο *κλέβει* ουσιαστικά (ακόμα και άθελα του) με κάποιο τρόπο ένα άλλο άτομο (ή ακόμα και τον ίδιο του τον εαυτό).

Προφανώς, μια τέτοια κλοπή μπορεί να συμβεί είτε σε θέματα υλικών αγαθών και σωματικής υγείας. Αλλά και, ενδεχομένως, κάποιος θα μπορούσε να υποστηρίξει πως τέτοιες κλοπές μπορούν να συμβούν ακόμα και σε θέματα ψυχολογικής υγείας και επιλογών στην ζωή. Ένας γονιός που καταπιέζει το παιδί του σε συγκεκριμένες δραστηριότητες και κατευθύνσεις στη ζωή, παρόλο που το παιδί δυσανασχετεί… ή ένας νεαρός Αφροαμερικάνος που καταπιάνεται με εγκληματικές δραστηριότητες διότι γεννήθηκε και μεγάλωσε σε γκέτο του Ντιτρόιτ, είναι δυο παραδείγματα αυτών των πιο ασαφών μορφών Αδικίας.

Εν κατακλείδι, ο Ενάρετος (κάποιος που έχει δηλαδή αναπτύξει και την αρετή της Δικαιοσύνης και φέρεται Δίκαια), είναι κάποιος που γνωρίζει πρώτον πως να *μην* κλέψει, και δεύτερον, πως να επανορθωθεί η κλοπή (δηλαδή η Αδικία) που μπορεί να έχει προξενηθεί σε ένα άτομο (για οποιονδήποτε λόγο). Επίσης *πράττει* και *θέλει να πράττει* με βάση αυτές του τις γνώσεις, γιατί αυτές οι πράξεις του προσφέρουν την ύψιστη ευχαρίστηση.[4] Θέλει λοιπόν να φέρεται με ένα τρόπο που αποσκοπεί στο να διατηρείται μια αρμονία και ισότητα μέσα του, και στο κόσμο γύρω του ανεξαρτήτως πλαισίου, άρα ουσιαστικά να λαμβάνει ο καθένας ότι του αξίζει στο μεγαλύτερο δυνατό βαθμό.[5]

Τώρα, σε δεύτερο χρόνο, ο Αριστοτέλης μιλάει για τη Δικαιοσύνη στο πλαίσιο της πολιτικής. Η πολιτική δικαιοσύνη είναι ο τρόπος ρύθμισης και οργάνωσης με στόχο τη διατήρηση μιας γενικής ισότητας μεταξύ πολιτών… δηλαδή να μην κλέβονται, έμμεσα ή άμεσα, οι πολίτες μεταξύ τους. Η πραγμάτωση της πολιτικής δικαιοσύνης φαίνεται να οδηγεί στη θέσπιση και εφαρμογή νόμων, οι οποίοι θα προνοούν για να διατηρείται η Δίκαιη ισότητα μεταξύ των πολιτών, καθώς και να τιμωρούνται οι Αδικίες (με μια επανορθωτική λογική φυσικά) που συμβαίνουν ανάμεσά τους.[6]

Εδώ όμως πιστεύω μπορούμε να επαγάγουμε τρία καίρια σημεία που ξεκαθαρίζουν την διαφορά της Δικαιοσύνης ως Αρετή από το πολιτικό και, επακολούθως, νομικό δίκαιο.

Το πρώτο σημείο είναι ότι η ίδια η θέσπιση νόμων είναι μια πράξη (ή

[4] Αριστοτέλης, *Ηθικά Νικομάχεια*, Α.8, 1099a7–30· Β.3, 1104b5–12.

[5] Αριστοτέλης, *Ηθικά Νικομάχεια*, Β.1, 1103b3–6· Ε.1, 1129a33–1130a13.

[6] Ο Αριστοτέλης αποκαλεί τις δυο αυτές μορφές της Δικαιοσύνης Διανεμητικό και Επανορθωτικό Δίκαιο αντίστοιχα: *Ηθικά Νικομάχεια*, Ε.2, 1130b30–1131a9· Ε.3-4, 1131a29–1132a29.

μια σειρά πράξεων) που συμπωματικά μπορεί να είναι Δίκαιη ή Άδικη. Αυτό σημαίνει πως ένας νομοθέτης μπορεί κάλλιστα να έχει στο μυαλό του να Αδικήσει τους συμπολίτες του για το δικό του κέρδος, και έτσι να θεσπίσει νόμους που ουσιαστικά Αδικούν. Σε άλλη περίπτωση, ένας άλλος νομοθέτης, λόγω άγνοιας, μπορεί απλά να μη καταφέρει να προβλέψει πως η θέσπιση κάποιων νόμων θα προκαλέσουν μεγάλες Αδικίες στο μέλλον.

Αυτός που όντως θα θεσπίσει δίκαιους νόμους είναι κάποιος που είναι ήδη Δίκαιος, και, ως Δίκαιος, θέλει και ξέρει πως να πράξει Δίκαια. Άρα θεσπίζει νόμους που αποσκοπούν και όντως καταφέρνουν στο να διατηρούνται οι Δίκαιες σχέσεις μεταξύ των πολιτών. Επομένως καταλαβαίνουμε πως το νομικό δίκαιο δεν είναι απαραίτητα Δίκαιο όπως το έχει ορίσει ο οποιοσδήποτε, αλλά μόνο συμπωματικά αν, και μόνο αν, έχει θεσπιστεί από Ενάρετους ανθρώπους.[7]

Το δεύτερο καίριο σημείο που πρέπει να έχουμε υπόψιν μας είναι πως όταν κάποιος ακολουθεί απολύτως ρητά τους νόμους, *ακόμα κι αν αυτοί είναι Δίκαιοι*, δεν τον καθιστά εκείνον πραγματικά Αριστοτελικά Δίκαιο, αλλά μόνο νομικά δίκαιο. Όπως είπαμε, οι νόμοι, ακόμα και οι καλύτεροι και Δικαιότεροι, είναι αποτελέσματα των πράξεων των Δίκαιων ανθρώπων. Αν κάποιος είναι Ενάρετος και Δίκαιος, γνωρίζει ήδη τι είναι Δίκαιο και θέλει να το υπηρετεί. Δεν χρειάζεται λοιπόν να έχει το βλέμμα του στραμμένο προς τους νόμους περιμένοντας να του υποδείξουν εκείνοι πως να φερθεί Δίκαια. Τέτοιες υποδείξεις τις χρειάζονται, με διαφορετικούς τρόπους και για διαφορετικούς λόγους, οι μη-Ενάρετοι.[8]

Τέλος, το τρίτο καίριο σημείο που πρέπει να έχουμε στο μυαλό μας είναι ότι οι νόμοι έχουν πάντα έναν αόριστα καθολικό χαρακτήρα που στοχεύει να δημιουργεί και να διατηρεί ισότητα *γενικά*, αλλά και να προσφέρει επανορθωτικές λύσεις στις περισσότερες εκφάνσεις Αδικίας. Ο Αριστοτέλης όμως πιστεύει πως στη πραγματική ζωή πάντα θα εμφανίζονται επιμέρους περιπτώσεις που, όχι μόνο ο νόμος δε θα προσφέρει τη πιο Δίκαιη λύση, αλλά μπορεί κιόλας να γίνει Άδικος ο ίδιος. Αυτό σημαίνει ότι όταν εφαρμόζονται σε τέτοιες συγκεκριμένες περιπτώσεις, ακόμα και οι φαινομενικά δίκαιοι νόμοι που έχουν πραγματικά στόχο τη γενική διατήρηση ισότητας, μπορούν να αποδειχτούν ουσιαστικά Άδικοι.[9]

---

[7] Αριστοτέλης, *Ηθικά Νικομάχεια,* Ε.1, 1129a4-12, 1129a32-b2, 1129b27-1130a13· Ε.5, 11134a1-7· Ε.6, 1134a24-b8· Ε.8, 11135a18-b25.

[8] Αριστοτέλης, *Ηθικά Νικομάχεια,* Ε.1, 1129b15-27· Ε.8, 1136a1-4· Κ.9, 1179b1-1181b13.

[9] Αριστοτέλης, *Ηθικά Νικομάχεια,* Β.2, 1104a1-9· Ε.7, 1134b18-1135a16.

Η Κορύφωση της Αριστοτελικής Δικαιοσύνης

Εδώ είναι λοιπόν που ο Αριστοτέλης παρουσιάζει επιτέλους ποια είναι η κορύφωση της Αρετής της Δικαιοσύνης, μια κορύφωση που ξεδιαλύνει τις επιπλοκές που εμφανίζονται λόγω του τρίτου προαναφερθέντος καίριου σημείου στη κατανόηση μας της Αριστοτελικής Δικαιοσύνης. Αυτή η κορύφωση είναι η Επιείκεια.[10]

Η Επιείκεια, ως προς το νομικό δίκαιο, μοιάζει σαν μια μικρή ελίτσα που «χαλάει» ένα τέλειο αρμονικό πρόσωπο, ενώ ταυτόχρονα, ανεξήγητα, το κάνει ακόμα πιο ωραίο… ή μια λάθος αρμονικά νότα που όμως χρωματίζει και εξυψώνει με μοναδικό τρόπο μια μελωδία, και την κάνει απαράμιλλη. Έτσι και η Επιείκεια είναι μια εξυψωτική δυνατότητα που όμως διαφεύγει από κανόνες και αρμονίες και λογικές.

Ο Επιεικής είναι ο πραγματικά Ενάρετος και Δίκαιος άνθρωπος, κι είναι αυτός που αναγνωρίζει μεν τον γενικά και *σχεδόν τέλειο δυνάμει* Δίκαιο χαρακτήρα των νόμων, αλλά καταλαβαίνει δε ότι υπάρχουν και θα υπάρχουν γεγονότα και συγκυρίες οι οποίες θα τον ξεπερνούν. Άρα δε θα ναι αρκετός ο νόμος για να ανταποκριθεί σε τέτοια γεγονότα και συγκυρίες. Σε αυτές τις περιπτώσεις έρχεται ο πραγματικά Δίκαιος (και άρα Επιεικής), αναγνωρίζοντας την αδυναμία και ανελαστικότητα του νόμου, και δείχνει σε αυτές τις μοναδικές επιμέρους περιπτώσεις Επιείκεια κατά την κρίση του και κατά την αντίστοιχη με τη κρίση του πράξη.[11]

Προφανώς και είναι πρακτικά αδύνατο να κανονικοποιήσουμε ποιες είναι οι Επιεικείς, και άρα Δίκαιες, πράξεις. Στο τέλος της ημέρας, οι Επιεικείς κρίσεις είναι εντελώς στην ευθύνη του Δίκαιου. Και ο Δίκαιος δεν μπορεί απαραίτητα να εξηγήσει γιατί τις κάνει, πέραν της προαίρεσης του για Δίκαιες πράξεις που έχει αναπτύξει. Για αυτόν τον λόγο είναι τόσο δύσκολο να αποδεχτούμε φιλοσοφικά την έννοια της Επιείκειας όταν προσπαθούμε να θεσπίσουμε νόμους και ένα σταθερό σύστημα Δικαιοσύνης. Γιατί ο καθένας θα μπορούσε να επικαλεστεί την δήθεν Επιεική κρίση του για να δικαιολογήσει τις πράξεις του ή τις πράξεις των άλλων.

---

[10] Θα ήθελα να σημειώσω εδώ πως η έννοια της Επιείκειας στο πλαίσιο της Αριστοτελικής Ηθικής δεν αντιστοιχεί στη συνηθισμένη (και ξεκάθαρα πιο στενή) έννοια του «να δείχνεις έλεος». Επιείκεια μπορεί να σημαίνει πράγματι μια συγχωρητική στάση απέναντι σε αυτούς που έπραξαν άδικα ως προς το νόμο, αλλά μπορεί κάλλιστα να σημαίνει και το αντίθετο, όπου ο Επιεικής χρειάζεται να επιβάλλει μεγαλύτερη τιμωρία σε αυτόν που έκανε άδικη πράξη από ότι προβλέπει ο νόμος. Με άλλα λόγια, η Αριστοτελική Επιείκεια μεταφράζεται σωστότερα στα αγγλικά ως “Equity” παρά ως “Mercy”.

[11] Αριστοτέλης, *Ηθικά Νικομάχεια,* Ε.10 1137a31-1138a2.

Εν πάση περιπτώσει, έχοντας σε γενικές γραμμές καταλάβει την Αριστοτελικά ορισμένη Δικαιοσύνη, μας μένει να την εξετάσουμε στο πλαίσιο της Τεχνητής Νοημοσύνης και των AI-δικαστών. Πρώτα όμως πρέπει να ξεκαθαρίσουμε τι ακριβώς είναι η Τεχνητή Νοημοσύνη.

Οι Δυο Τύποι Τεχνητής Νοημοσύνης

Στις μέρες μας, ο όρος «Τεχνητή Νοημοσύνη» χρησιμοποιείται αδιακρίτως για να χαρακτηρίσει οποιαδήποτε τεχνολογική εξέλιξη που συσχετίζεται με την λύση προβλημάτων ως προς την ανάλυση και επεξεργασία δεδομένων. Γι' αυτό το λόγο είναι σημαντικό να εντοπίσουμε σημαντικές διαφοροποιήσεις που προκύπτουν στην φιλοσοφική ανάλυση του ορού. Μελετώντας τα κείμενα διάφορων ακαδημαϊκών και Φιλοσόφων πάνω στο θέμα, όπως το άρθρο των Selmer Bringsjord και Naveen Sundar Govindarajulu με τον ευθύ τίτλο *"Artificial Intelligence",* αντιλήφθηκα πως είναι απαραίτητο να επισημανθεί η ύπαρξη δυο πολύ διαφορετικών ειδών AI, ειδών που έχουν εδραιωθεί και θεμελιωθεί πλέον στην φιλοσοφική κοινότητα. Τα δυο είδη είναι η «Ασθενής» (Weak) και η «Ισχυρή» (Strong) AI. Ο πρώτος που ξεχώρισε τα δυο αυτά είδη ήταν ο John Searle, ένας διαβόητος φιλόσοφος του δευτέρου μισού του 20^{ου} αιώνα, που ασχολήθηκε περισσότερο με φιλοσοφικά ζητήματα γύρω από το Νου (Searle 1980, 417, 421, 423–424).

Η Ασθενής AI είναι τεχνολογία που έχει σχεδιαστεί με σκοπό την επίλυση πολύ συγκεκριμένων τύπων προβλημάτων. Η αποδοτικότητα της Ασθενούς AI στο να επιλύει τέτοια προβλήματα είναι ασύγκριτα μεγαλύτερη από την αντίστοιχη φυσικώς ανθρώπινη αποδοτικότητα. Όμως η πιο σημαντική δυνατότητα της Ασθενούς AI (που την ξεχωρίζει άλλωστε από τους συνηθισμένους υπολογιστές) είναι ότι, μέσω της ανάπτυξης της τεχνολογίας Μηχανικής Μάθησης ("machine learning"), δύναται να εξελίσσει αυτονόμως τις ικανότητές της στο να επιλύει προβλήματα, και άρα να επαυξάνει την αποδοτικότητά της ανεξάρτητα (Bringsjord and Govindarajulu 2018, sec. 4.1). Μορφές Ασθενούς AI ήδη χρησιμοποιούνται σε δεκάδες, επαγγελματικούς τομείς, όπως στην Ιατρική (όπου η Ασθενής AI μπορεί να πραγματοποιήσει πολύ πιο αποτελεσματικά χειρουργικές επεμβάσεις από ανθρώπους γιατρούς) και τον Πρωταθλητισμό (όπου η Ασθενής AI μπορεί πλέον να ανταγωνιστεί και εύκολα να κατατροπώσει τους μεγαλύτερους ανθρώπους πρωταθλητές δεκάδων διαφορετικών παιχνιδιών, όπως το σκάκι) (Bringsjord and Govindarajulu 2018, sec. 1.5).

Παρά όλα αυτά, πρέπει να υπογραμμιστεί πως κάθε εκδοχή Ασθενούς AI είναι απόλυτα περιορισμένη στο να λύνει ένα και μόνο μοναδικό τύπο προβλημάτων. Αυτό σημαίνει ότι μια εξειδικευμένη AI στο να κερδίζει παρτίδες σκάκι είναι παντελώς άχρηστη στο να κερδίζει παρτίδες τάβλι (πόσο μάλλον να πραγματοποιεί χειρουργικές επεμβάσεις).

Σε αντίθεση με την Ασθενή AI (που έχει πραγματικά εφευρεθεί), η Ισχυρή AI υπάρχει μόνο στην θεωρία. Θεωρητικά λοιπόν, κάθε Ισχυρή AI δεν περιορίζεται σε ένα και μόνο συγκεκριμένο τύπο προβλημάτων, αλλά μπορεί να αναγνωρίσει και να ανταποκριθεί σε οποιοδήποτε πρόβλημα μπορεί να προκύψει, ακριβώς όπως ο άνθρωπος. Αναγνωρίζει έτσι τις ριζικές διαφορές μεταξύ επιτραπέζιων παιχνιδιών, χειρουργικών επεμβάσεων, μαγειρικής, πυρηνικής φυσικής και οποιουδήποτε άλλου τύπου προβλήματος μπορούμε να φανταστούμε, καθώς και ποιοι παράγοντες και μεταβλητές έχουν σχέση με κάθε συγκεκριμένο τύπο προβλήματος.[12]

Επιπλέον, πέραν των φαινομενικά τεχνικών της ικανοτήτων, η Ισχυρή AI θα έχει τη δυνατότητα υπαρξιακής αυτογνωσίας και συνείδησης. Επομένως, ως νοημοσύνη, η Ισχυρή AI δε θα διαφέρει ουσιαστικά από την ανθρώπινη (πέραν του ότι θα είναι αδιανόητα πιο έξυπνη) και θα πρέπει πρακτικά να συνυπολογίζεται από τον άνθρωπο ως ίσο αισθανόμενο ον.

Όμως η Ισχυρή AI, όπως τονίσαμε, υπάρχει μόνο στη θεωρία (και σε ιστορίες επιστημονικής φαντασίας) και το αν είναι κάποτε εφικτή ρεαλιστικά η δημιουργία της είναι ένα φιλοσοφικό (κι όχι τεχνικό όπως πολλοί νομίζουν) ερώτημα που δεν μας απασχολεί στην παρούσα εργασία μας.

[12] Το τεράστιο επιστημολογικό πρόβλημα που προέκυψε από την έρευνα για το πώς θα μπορούσε να εφευρεθεί AI που δεν θα περιορίζεται σε έναν και μόνο τύπο προβλημάτων έχει αναγνωριστεί εδώ και δεκαετίες: Ο άνθρωπος μπορεί σε ριπή οφθαλμού να ξεχωρίσει τις ριζικές διαφορές ανάμεσα σε θεωρητικά άπειρους τύπους προβλημάτων, αλλά και ανάμεσα στους εξίσου άπειρους παράγοντες και μεταβλητές που σχετίζονται με αυτά τα προβλήματα. Για παράδειγμα, όταν ο άνθρωπος θέλει να παίξει σκάκι, δε χρειάζεται να σκεφτεί συνειδητά ότι το αλάτι, τα μαχαίρια, ο φούρνος κι οτιδήποτε άλλο που βρίσκεται στην κουζίνα του και σχετίζεται με τη μαγειρική είναι άσχετα με το σκάκι. Ξέρει (και μαθαίνει) σχεδόν αυτόματα τι είναι σχετικό με το σκάκι, τι με τη μαγειρική, και τι είναι άσχετο και με τα δύο. Αυτή την επιστημολογική ικανότητα των ανθρώπων (που φαίνεται τόσο αυτονόητη στην καθημερινή ανθρώπινη εμπειρία μας) παραμένει αδιευκρίνιστη ως προς τον τρόπο που προκύπτει, και συνεπώς παραμένει άγνωστο πώς θα μπορούσε να προγραμματιστεί σε μια AI. Την πιο καθαρή διατύπωση του προβλήματος παρουσίασε ο Daniel Dennett στο δοκίμιό του *"Cognitive Wheels: The Frame Problem of AI"* (1987). Για μια αναλυτική επισκόπηση του προβλήματος, δείτε επίσης: *"The Frame Problem"* (2016) του Murray Shanahan στο *The Stanford Encyclopedia of Philosophy*. Και τα δύο κείμενα περιλαμβάνονται στη βιβλιογραφία του παρόντος δοκιμίου.

Στη περίπτωση της εξέτασης λοιπόν της AI-δικαστή, όπως αυτή δύναται να προγραμματιστεί βάσει των τεχνολογικών δυνατοτήτων της εποχής μας, πρέπει να εστιάσουμε μόνο στην περίπτωση της Ασθενούς AI.

Μια ρεαλιστικά δυνάμει AI-δικαστής λοιπόν θα είναι μια Ασθενής AI, δηλαδή μια AI που θα είναι συγκλονιστικά αποτελεσματική στο να εκδικάζει δικαστικές υποθέσεις (με την εκδίκαση δικαστικών υποθέσεων να είναι ο απόλυτα μοναδικός σκοπός της). Δε πρέπει σε καμία περίπτωση να μπερδευτούμε και να σκεφτούμε ότι μια AI-δικαστής, όπως αυτή ρεαλιστικά δύναται, κατέχει κάποια μορφή συνείδησης και βούλησης (Bringsjord and Govindarajulu 2018, sec. 8.1.1).

Πώς θα Λειτουργήσει μια AI-δικαστής

Μπορούμε να σκεφτούμε πολλούς λόγους για τους οποίους θα θέλουμε να σχεδιάσουμε μια AI-δικαστή. Για παράδειγμα, επειδή δεν εμπιστευόμαστε τις ανθρώπινες προκαταλήψεις που μπορεί να έχουν άνθρωποι δικαστές εις βάρος (ή υπέρ) μειονοτήτων ή άλλων ανθρώπινων χαρακτηριστικών… ή ίσως ελπίζουμε ότι θα καταπολεμήσουμε τη διαφθορά που μπορεί να υπάρχει στο δικαστικό μας σύστημα λόγω της ανθρώπινης φύσης των δικαστών. Ο πιο σημαντικός όμως λόγος κατά την άποψη μου, που θα θέλαμε να έχουμε AI-δικαστή είναι διότι η AI θα είναι πάρα πολύ πιο αποδοτική στο να εκδικάζει υποθέσεις: Μια AI θα «βλέπει» μια υπόθεση, θα αναλύει τα δεδομένα της υπόθεσης (όπως καταθέσεις και αποδεικτικά στοιχεία) και μέσα σε μερικά λεπτά (αν όχι γρηγορότερα) θα την εκδικάζει, χωρίς ποτέ να αισθάνεται κούραση.[13]

Για να φτάσει σε αυτή την ικανότητα όμως, η AI-δικαστής θα πρέπει πρώτα να «εκπαιδευθεί» στο πως να είναι δικαστής. Αυτό επιτυγχάνεται πράγματι χάρη στη μέθοδο Μηχανικής Μάθησης. Μια AI-δικαστής, για να «εκπαιδευτεί» λοιπόν θα πρέπει να έχει πρόσβαση σε όσα περισσότερα δεδομένα data είναι προσβάσιμα από όλες τις υποθέσεις και εκδικάσεις που έχουν γίνει στο παρελθόν σε ολόκληρη την υφήλιο, και για τις οποίες υπάρχουν αποθηκευμένα και καταγεγραμμένα data. Η AI-δικαστής θα

[13] Για μια πιο ενδελεχή τεκμηρίωση της θέσης ότι η Τεχνητή Νοημοσύνη μπορεί να βελτιώσει ουσιαστικά το δικαστικό σύστημα μιας κοινωνίας (ειδικά ως προς την επιτάχυνση της εκδίκασης νομικών υποθέσεων) δείτε το άρθρο των Cinara Rocha και João Alvaro Carvalho με τίτλο *"Artificial Intelligence in the Judiciary: Uses and Threats"* (2022).

«μάθει» έτσι να κάνει σωστή συσχέτιση νόμων με τις εκδικάσεις των υποθέσεων. Έπειτα, έχοντας τους αλγόριθμους της παραμετροποιημένους από όλους τους νόμους μιας χώρας στην οποία θα λειτουργήσει ως δικαστής, η AI θα μπορούσε να βαλθεί να εκδικάσει (σε προσομοίωση) όλες τις υποθέσεις της χώρας που αντιστοιχούν στους νόμους της και που έχουν εκδικαστεί στο παρελθόν. Τελικώς, συγκρίνοντας τις εκδικάσεις της AI με τις παλιές και τωρινές εκδικάσεις ανθρώπων δικαστών, μπορούμε άμεσα να εξετάσουμε αν φτάνει στις ίδιες αποφάσεις που είχαν φτάσει και φτάνουν άνθρωποι δικαστές.[14]

Αν παρατηρήσουμε εν τέλει σε αυτές τις προσομοιώσεις τρανά ποσοστά αντιστοιχίας μεταξύ των εκδικάσεων της AI με τις εκδικάσεις των ανθρώπων δικαστών (όπως για παράδειγμα 97%), τότε θα μπορούσαμε να συμπεράνουμε ότι η AI ακολουθεί και εφαρμόζει το νόμο πρακτικά ακριβώς όπως οι άνθρωποι δικαστές. Έχοντας στο μυαλό μας και τα προαναφερθέντα πλεονεκτήματα μιας AI-δικαστή (ως προς την ταχύτητα εκδικάσεων και την αδιάφθορη και απροκατάληπτη φύση της) θα φάνταζε πράγματι πολύ λογικό να θέλουμε και στην πράξη να αντικαταστήσουμε τους ανθρώπους δικαστές με μια AI-δικαστή.

Αν όμως ασπαζόμαστε τις φιλοσοφικές θέσεις του Αριστοτέλη ως προς τη Δικαιοσύνη και την Αρετή, τότε μια τέτοια αντικατάσταση δεν θα ήταν καθόλου απλή να την αποδεχτούμε. Το πρόβλημα είναι το εξής: Η AI-δικαστής θα φέρεται πάντα με γενική και καθολική λογική, όσο πολύπλοκη και διεκπεραιωτική μπορεί να φαίνεται πως είναι. Διότι η AI έχει σχεδιαστεί και προγραμματιστεί εξαρχής με έναν πολύ συγκεκριμένο στόχο, καθώς και παραμέτρους που θα υποδεικνύουν πότε και πόσο επιτυγχάνεται αυτός ο στόχος. Αυτός ο στόχος είναι η εφαρμογή του νομικού δικαίου, στην αυθαίρετη μορφή που της έχει παρουσιαστεί. Συνεπώς έχει σχεδιαστεί, όχι να είναι Δίκαια η ίδια, αλλά να μιμηθεί τις εκδικάσεις παλιών υποθέσεων βάσει του εκάστοτε νομικού δικαίου, και να προβλέψει βάσει

[14] Το Κοινοβούλιο της Ευρωπαϊκής Ένωσης έχει αναγνωρίσει ότι η χρήση εφαρμογών AI (ιδίως όταν πρόκειται για AI που προορίζεται για χρήση στο πλαίσιο της απονομής νομικής δικαιοσύνης) ενέχει υψηλό ρίσκο για την παραβίαση ανθρωπίνων δικαιωμάτων και την υπονόμευση δημοκρατικών θεσμών. Συνεπώς, έχουν ήδη προβλεφθεί και κατοχυρωθεί νομικά στην Ευρωπαϊκή Ένωση μέσω του Artificial Intelligence Act (Κανονισμός ΕΕ 2024/1689) παρόμοιες διαδικασίες εκπαίδευσης και ελέγχου AI με αυτή που περιέγραψα συνοπτικά παραπάνω (αν και οι διαδικασίες της ΕΕ είναι, προφανώς, πολύ πιο σύνθετες, θεσμικά αυστηρές και τεχνικά απαιτητικές από τη δικιά μου απλουστευμένη περιγραφή). Ο προαναφερθείς κανονισμός περιλαμβάνεται στη βιβλιογραφία του παρόντος δοκιμίου.

αυτών των εκδικάσεων πως θα εκδικάζονταν μελλοντικές υποθέσεις.

Η ΑΙ ουσιαστικά λοιπόν ακολουθεί, έστω εμμέσως, *οδηγίες*. Της θέτουμε εμείς τι θεωρούμε γενικά δίκαιο. Άρα τις θέτουμε τους νόμους που έχουμε θεσπίσει. Στη συνέχεια της λέμε τι παράγοντες της πραγματικότητας αντιστοιχούν στους νόμους, για να μπορεί να τους προσμετρήσει. Και τέλος, η ΑΙ απλά εκτελεί εκδικάσεις βάσει αυτών των οδηγιών με απίστευτη αποτελεσματικότητα και αποδοτικότητα. Εκεί είναι το ταβάνι της.

Η συσχέτιση της ΑΙ-δικαστή με την Αριστοτελική Δικαιοσύνη

Έχοντας καταλάβει ποιο είναι το έργο της ΑΙ-δικαστή, ας δούμε διεξοδικά γιατί δεν μπορεί να ταυτιστεί καθόλου με τη Αριστοτελική Δικαιοσύνη. Για να το πετύχουμε αυτό, θα συσχετίσουμε την ΑΙ-δικαστή με τα τρία καίρια σημεία της Αριστοτελικής Δικαιοσύνης που προανέφερα σε προηγούμενες παραγράφους μέσω μερικών ρεαλιστικών παραδειγμάτων.

Πρώτα ξεκαθαρίσαμε ότι το νομικό δίκαιο μπορεί μόνο συμπωματικά να αντιπροσωπεύει το πραγματικό Δίκαιο. Είναι εύκολο να αναλογιστούμε δεκάδες ιστορικά παραδείγματα τυραννιών, και να καταλάβουμε πως, πράγματι, σε τέτοιες τις περιπτώσεις, ένα ΑΙ-δικαστής θα ήταν απλά ένα πανίσχυρο εργαλείο καταπίεσης στα χέρια των τυράννων. Φανταστείτε για παράδειγμα τον Κιμ Γιονγκ Ουν της Βόρειας Κορέας να είχε στη διάθεσή του μια τέτοια ΑΙ-δικαστή που, εντελώς αυτοματοποιημένα, και μέσω της αστραπιαίας και απίστευτα ακριβούς ανάλυσης στοιχείων και δεδομένων, να μπορεί καταδικάσει ανθρώπους που εναντιώνονται στους νόμους της Βόρειας Κορέας. Κι όμως, ως προς τους τυραννικούς νόμους που λογικά θα έχουν θεσπιστεί σε τέτοιες τυραννίες, η ΑΙ-δικαστής θα ανταποκρίνεται με φοβερή επιτυχία. Κι όλοι οι παραβάτες αυτών των νόμων θα καταδικάζονται από μια απολύτως *νομικά* δίκαιη ΑΙ-δικαστή, ακόμα κι όταν οι νόμοι είναι *πραγματικά* Άδικοι. Μπορούμε όμως να αποδεχτούμε ένα τέτοιο σύστημα ως παράδειγμα πραγματικής Δικαιοσύνης; Αμφιβάλλω.

Μέσω της παραπάνω διαπίστωσης, γρήγορα μπορούμε να δούμε πως μια ΑΙ-δικαστής δεν μπορεί να ταυτιστεί ούτε με το δεύτερο καίριο σημείο της Αριστοτελικής Δικαιοσύνης που υπέδειξα νωρίτερα. Όταν κάποιος ακολουθεί απλά και ρητά το νομικό δίκαιο, ακόμα κι όταν τυχαίνει αυτό το νομικό δίκαιο να αντικατοπτρίζει το πραγματικά Δίκαιο, *δεν* καθίσταται πραγματικά Δίκαιος. Η ΑΙ-δικαστής όμως ξεκάθαρα κάνει ακριβώς και μόνο αυτό: Μιμείται τη Δίκαιη συμπεριφορά, ακολουθώντας ρητά και

ανελαστικά το νομικό δίκαιο. Παρ' όλα αυτά, εύκολα κάποιος μπορεί να υποστηρίξει πως μια AI-δικαστής που έχει παραμετροποιηθεί με πραγματικά Δίκαιους νόμους (άρα νόμους θεσπισμένους από Αριστοτελικά Ενάρετους ανθρώπους) μπορεί κάλλιστα να αντικαταστήσει ανθρώπους δικαστές, κι έτσι να δημιουργηθεί ένα αυτοματοποιημένο και αποδοτικότατο δικαστικό σύστημα. Οι εκδικάσεις της AI-δικαστή θα ήταν, έστω και μόνο φαινομενικά, πραγματικά Δίκαιες, ως έμμεσα αποτελέσματα πράξεων των Ενάρετων.

Ή μήπως όχι;

Η AI-δικαστής δεν μπορεί να είναι Επιεικής εκ Φύσεως

Τώρα λοιπόν είναι που η AI-δικαστής αποδεικνύεται απόλυτα ανεπαρκής ακόμα και στο να προσομοιάσει ένα Αριστοτελικά Δίκαιο δικαστικό σύστημα. Ο λόγος είναι ότι η AI-δικαστής δε μπορεί με κανένα τρόπο να ρυθμιστεί ώστε να δείχνει Επιείκεια (με την ανάγκη για Επιείκεια να αποτελεί το τρίτο και πιο καίριο σημείο της Αριστοτελικής Δικαιοσύνης για το οποίο μίλησα προηγουμένως). Θυμίζω πως ο Αριστοτέλης μίλησε για τη δυνατότητα να εμφανιστούν περιπτώσεις και υποθέσεις μοναδικές στη ζωή, οι οποίες, αν εκδικαστούν με βάση το γενικό νόμο, θα εκδικαστούν Άδικα, ακόμα κι αν ο νόμος είναι γενικά δίκαιος και θεσπισμένος από Ενάρετους ανθρώπους. Αυτές είναι οι περιπτώσεις και υποθέσεις που ξεπερνούν το νόμο (και ίσως τις προβλεπτικές ικανότητες οποιουδήποτε ανθρώπου) και χρειάζονται να διαλευκανθούν κατ' ιδίαν από τους Αριστοτελικά Ενάρετους όταν θα προκύψουν.[15]

Πιστεύω θα ήταν πολύ χρήσιμο να δημιουργήσω ένα φανταστικό παράδειγμα, για να καταλάβουμε και στη πράξη πως μια AI-δικαστής θα αποτύγχανε εκ φύσεως να είναι Επιεικής, άρα, πέραν από ουσιαστικά, θα αποδεικνυόταν *και πρακτικά, και φαινομενικά,* Αριστοτελικά Άδικη.

Ας θεωρήσουμε ότι έχουμε θεσπίσει νόμο στην Ελλάδα που απαγορεύει την κλοπή και τη σωματική βία (μιας και πιστεύω πως τέτοιοι νόμοι φαντάζουν γενικά δίκαιοι σε όλους μας). Ας θεωρήσουμε επίσης ότι έχουμε σχεδιάσει και προγραμματίσει στην εντέλεια μια πανίσχυρη AI-δικαστή

[15] Την συγκεκριμένη αδυναμία της AI να επιδείξει Αριστοτελικά ορισμένη επιείκεια έχει συστηματικά υποδείξει ο Νίκος Α. Παρασκευόπουλος, καθηγητής νομικής του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης, ειδικά στο βιβλίο του *Ο επιεικής αλγόριθμος – Από την Αριστοτελική σκέψη στην τεχνητή νοημοσύνη* (2024).

που ακολουθεί τους νόμους και εκδικάζει υποθέσεις με 99.999% επιτυχία.

Ας φανταστούμε τώρα πως κάποιος άνδρας της φυλής των Ρομά εισέρχεται σε ένα πολυκατάστημα. Τον βλέπουμε να μιλάει με τον υπάλληλο στο ταμείο. Ο υπάλληλος φαίνεται πολύ γρήγορα να θυμώνει με τον άνδρα, και με έντονες και απειλητικές κινήσεις να του δείχνει την πόρτα της εξόδου… και ξαφνικά, ο άνδρας επιτίθεται στον υπάλληλο, τον χτυπάει και τον αφήνει αναίσθητο στο πάτωμα. Μεμιάς ο άνδρας πηδάει πίσω από το ταμείο, ανοίγει ένα ψυγείο με αφεψήματα, αρπάζει ένα χυμό πορτοκάλι, και τρέπεται σε φυγή.

Προφανώς, βάσει της παραστατικότατης εικόνας που μόλις περιέγραψα, θα φωνάξουμε πως ο άνδρας καθαρά αδίκησε: Αδίκησε πρώτα τον υπάλληλο ασκώντας πάνω του σωματική βία, και μετά τον ιδιοκτήτη του πολυκαταστήματος κλέβοντας το προϊόν του. Αντίστοιχα με μας, η AI-δικαστής θα αναλύσει τις πράξεις αυτού του άνδρα (που έτυχε να βιντεοσκοπηθούν από τις κάμερες ασφαλείας του πολυκαταστήματος) και πολύ γρήγορα και απλά θα τον καταδικάσει για κλοπή και άσκηση σωματικής βίας. Ο άνδρας τιμωρείται, και τώρα θα εκτίσει την ποινή που έχει ορίσει ο νόμος (που κατά πάσα πιθανότητα θα είναι μια μακρά ποινή φυλάκισης). Ο νόμος εφαρμόστηκε με επιτυχία.

Τι θα γίνει όμως όταν ανακαλύψουμε και άλλα χαρακτηριστικά μοναδικά σε αυτήν την (φανταστική πάντα) υπόθεση, τα οποία δεν προβλέπει ο νόμος; Πως θα επηρεαστούμε αν ανακαλύψουμε λοιπόν ότι αυτός ο άνθρωπος, λόγω άτυχων συγκυριών, δεν είχε πιει το οτιδήποτε για τρεις μέρες; Αν μαθαίναμε επίσης, ότι έξω από πολυκατάστημα τον περίμενε η πεντάχρονη κόρη του, η οποία επίσης δεν είχε πιει τίποτα τρεις μέρες; Κι αν τέλος, ανακαλύπταμε πως αυτός ο πατέρας είναι η μόνη οικογένεια που έχει απομείνει στον κόσμο για αυτό το άμοιρο πεντάχρονο κορίτσι;

Ο πατέρας λοιπόν, μπαίνοντας στο πολυκατάστημα, εξήγησε στον υπάλληλο την κατάσταση στην οποία αυτός κι η κόρη του βρισκόταν. Πιστεύω πως θα συμφωνούσαμε στο ότι, πατέρας και κόρη βρίσκονταν ξεκάθαρα σε μια εξαθλιωμένη και απελπιστική κατάσταση. Μα για δικούς του λόγους, ο υπάλληλος αρνήθηκε με ασυνήθιστα έντονο ύφος να του προσφέρει το οτιδήποτε για να τους βοηθήσει να σβήσουν τη δίψα τους. Μπορούμε λοιπόν να υποθέσουμε πως ο πατέρας αντέδρασε υπό την επήρεια της απελπισίας και του άγχους του για την υγεία της κόρης του.

Μήπως λοιπόν είναι πρέπων, σε αυτή τη μια και μοναδική υπόθεση, να πάρουμε υπόψιν μας όλους τους δευτερεύοντες λόγους που μόλις φανταστήκαμε, και να δείξουμε Επιείκεια απέναντι στις (πράγματι) Άδικες πράξεις αυτού του άνδρα? Μήπως ο γενικά δίκαιος νόμος θα δημιουργούσε

περαιτέρω Αδικία εις βάρος αυτού του πατέρα και της πεντάχρονης κόρης του αν εφαρμοστεί κατά γράμμα?

Ένα είναι σίγουρο: Η ΑΙ-δικαστής θα παρέμενε απόλυτα καθηλωμένη στα πλαίσιο του νομικού δικαίου και θα καταδίκαζε τον άνδρα χωρίς προστριβές. Διότι μόνο ένας Ενάρετος άνθρωπος θα μπορούσε να θέσει τέτοιες ερωτήσεις στον εαυτό του.

Εν Κατακλείδι

Έχοντας αναλύσει τι ήταν η Αρετή και η Δικαιοσύνη για τον Αριστοτέλη, καθώς και το πως λειτουργεί μια ΑΙ-δικαστής, μπορούμε πλέον να υποθέσουμε με αρκετή σιγουριά για το ποια θα ήταν η άποψη του Αριστοτέλη για μια ΑΙ-δικαστή. Πιστεύω πως, για τον Σταγειρίτη, η ΑΙ γενικότερα θα χαρακτηριζόταν στην ουσία της ως απλά ένα ακόμα δημιούργημα του ανθρώπου: Η ΑΙ είναι ένα αποτέλεσμα ανθρώπινης πράξης, και μιας πράξης προφανώς με απίστευτη επίπτωση πάνω στο σύνολο της ανθρωπότητας, καθώς και απίστευτη επίπτωση στη διατήρηση ή διαστρέβλωση της αρμονίας και της ισότητας στο εσωτερικό της ανθρώπινης κοινωνίας. Αυτό σημαίνει ότι, στην ουσία της, η δημιουργία της ΑΙ θα είναι το αποτέλεσμα μιας Δίκαιης ή Άδικης πράξης, ακριβώς όπως είναι το αποτέλεσμα μιας Δίκαιης ή Άδικης πράξης η θέσπιση νόμων (αλλά και, ουσιαστικά, σχεδόν κάθε ανθρώπινη δημιουργία).

Όμως, όπως υποστηρίζει ο Αριστοτέλης, όλοι οι άνθρωποι έχουν τη δυνατότητα μέσα τους να γίνουν Ενάρετοι. Κι όλοι οι άνθρωποι, είτε είναι υπάλληλοι πολυκαταστημάτων, στρατιώτες, δικαστές ή οτιδήποτε άλλο, έχουν πάντα την δυνατότητα μέσα τους να γίνουν Επιεικείς. Κι ακόμη κι αν όλη τους τη ζωή εφαρμόζουν το νόμο κατά γράμμα (είτε ο νόμος είναι γενικά Δίκαιος ή Άδικος) αυτή η δυνατότητα τους δεν χάνεται πότε. Πάντα η ερώτηση θα πλανάται σε κάποιο επίπεδο της σκέψης τους για το αν αυτό που κρίνουν και πράττουν είναι Δίκαιο ή όχι.

Η ΑΙ-δικαστής θα είναι κατά πάσα πιθανότητα η καλύτερη και πιο αποτελεσματική μορφή εφαρμογής του νομικού δικαίου. Όμως, ελλοχεύει κίνδυνος με διπλή φύση αν αντικαταστήσουμε τους ανθρώπους δικαστές με μια ΑΙ-δικαστή. Πρώτον, να μη παρανοήσουμε τις δυνατότητες της και πιστέψουμε ότι γνωρίζει η ίδια η ΑΙ τι είναι Δίκαιο καλύτερα από τον άνθρωπο. Και δεύτερον, μέσα στην αυταρέσκεια μας για το δημιούργημα μας, να μην εναποθέσουμε την ευθύνη της κρίσης μας ολότελα πάνω της, πνίγοντας τελικά την δυνατότητα μας να γίνουμε οι ίδιοι Ενάρετοι και

Δίκαιοι. Διότι έτσι η δυνατότητα για Αριστοτελική Δικαιοσύνη θα χαθεί οριστικά, καθώς Επιείκεια μέσω AI δε μπορεί να υπάρξει. Πιστεύω πλέον πως είναι κατανοητό γιατί.

Πηγές

Αριστοτέλης. 2006a. *Ηθικά Νικομάχεια Ά–Δ*. Μετάφραση Δημήτριος Λυπουρλής. Θεσσαλονίκη: Εκδόσεις Ζήτρος.

Αριστοτέλης. 2006b. *Ηθικά Νικομάχεια Ε–Κ*. Μετάφραση Δημήτριος Λυπουρλής. Θεσσαλονίκη: Εκδόσεις Ζήτρος.

Bringsjord, Selmer και Naveen Sundar Govindarajulu. 2018. "Artificial Intelligence." *Stanford Encyclopedia of Philosophy*. Stanford University. https://plato.stanford.edu/entries/artificial-intelligence/.

Chen, Stephen. 2021. "Chinese Scientists Develop AI Prosecutor That Can Press Its Own Charges." *South China Morning Post*, 26 Δεκεμβρίου. https://www.scmp.com/news/china/science/article/3160997/chinese-scientists-develop-ai-prosecutor-can-press-its-own.

Γούναρης, Άλκης, και Γιώργος Κωστελέτος. 2024. «Γράφοντας τον αλγόριθμο του Καλού: Η Τεχνητή Νοημοσύνη ως μηχανή απόδοσης Δικαιοσύνης.» *Ηθική. Περιοδικό Φιλοσοφίας* (19). https://doi.org/10.12681/ethiki.39654.

Dennett, Daniel C. 1987. "Cognitive Wheels: The Frame Problem of AI." *The Robot's Dilemma: The Frame Problem in Artificial Intelligence*, 129–151. Norwood, NJ: Ablex Publishing Corporation.

European Parliament and Council (2024). Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). *Official Journal of the European Union*, L series, no. 2024/1689 (12 July 2024). http://data.europa.eu/eli/reg/2024/1689/oj

Newman, Jack. 2021. "China Develops AI Prosecutor That Can Press Charges with 97% Accuracy." *Daily Mail*, 27 Δεκεμβρίου. https://www.dailymail.co.uk/news/article-10346933/China-develops-AI-prosecutor-press-charges-97-accuracy.html.

Παρασκευόπουλος, Α. Νίκος. 2024. *Ο επιεικής αλγόριθμος: Από την Αριστοτελική σκέψη στην τεχνητή νοημοσύνη*. Θεσσαλονίκη: Ίδρυμα Τριανταφυλλίδη.

Rocha, Cinara, and João Alvaro Carvalho. 2022. "Artificial Intelligence in the Judiciary: Uses and Threats." *CEUR Workshop Proceedings*, EGOV-CeDEM-ePart 2022, Linköping University, Sweden, September 6–8. http://ceur-ws.org/Vol-3399/paper17.pdf.

Samson, Carl. 2021. "China Develops AI 'Prosecutor' That Can Charge Citizens with Crimes with '97% Accuracy'." *NextShark*, 27 Δεκεμβρίου. https://nextshark.com/china-ai-prosecutor-97-percent-accuracy.

Searle, John R. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3 (3): 417–424. https://doi.org/10.1017/s0140525x00005756.

Shanahan, Murray. 2016. "The Frame Problem." *The Stanford Encyclopedia of Philosophy*. Stanford University. https://plato.stanford.edu/entries/frame-problem/.

The Korea Times. 2021. "More than 97% Accuracy: Chinese Scientists Develop AI Prosecutor." *The Korea Times*, 26 Δεκεμβρίου. https://www.koreatimes.co.kr/world/20211226/more-than-97-accuracy-chinese-scientists-develop-ai-prosecutor.

Περίληψη

Σε αυτό το δοκίμιο θα εξετάσω κατά πόσο η Τεχνητή Νοημοσύνη μπορεί να είναι Αριστοτελικά Δίκαιη, και κατά πόσο θα μπορούσε να λειτουργήσει ως Αριστοτελικά Δίκαιη αν αντικαθιστούσε τους ανθρώπους δικαστές στη κοινωνία μας. Αρχικά, θα αναλύσω τι ακριβώς είναι η Δικαιοσύνη στο πλαίσιο της Ηθικής του Αριστοτέλη, και θα υποδείξω πως η Δικαιοσύνη σχετίζεται απόλυτα με την ανθρώπινη Αρετή και πράξη. Επακολούθως, θα ερμηνεύσω την Αριστοτελική Δικαιοσύνη με σκοπό την επαγωγή τριών πορισμάτων που θα έχουν καίρια σημασία για τα τελικά συμπεράσματα της εξέτασής μου. Στη συνέχεια, θα αναφερθώ στους δυο θεμελιώδεις τύπους Τεχνητής Νοημοσύνης που αναγνωρίζονται στη φιλοσοφική κοινότητα, για να ξεδιαλύνουμε πως ακριβώς νοείται η Τεχνητή Νοημοσύνη φιλοσοφικά. Έπειτα θα εξηγήσω πως ακριβώς θα λειτουργούσε η Τεχνητή

Νοημοσύνη, και ποιος θα ήταν ο στόχος αυτής της λειτουργίας, αν έμελλε να πάρει το ρόλο δικαστή. Τέλος, θα συσχετίσω τη λειτουργία και το στόχο της Τεχνητής Νοημοσύνης με την Αριστοτελική Δικαιοσύνη (και συγκεκριμένα με τα τρία προαναφερθέντα πορίσματα) ώστε να αποδείξω πως η Τεχνητή Νοημοσύνη, εκ φύσεως, δε θα μπορούσε ποτέ να είναι Αριστοτελικά Δίκαιη.

Λέξεις Κλειδιά: Αρετή, Αριστοτέλης, Δικαιοσύνη, Δικαστής, Επιείκεια, Νομικό Δίκαιο, Τεχνητή Νοημοσύνη.

Keywords: Virtue, Aristotle, Justice, Judge, Equity, Legal Justice, Artificial Intelligence.

Αλέξανδρος Νούνεσης
Τμήμα Φιλοσοφίας, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών
Email: alexander.nounesis@gmail.com
ORCiD: https://orcid.org/0009-0003-7298-7836

Μαρίνα ΞΕΝΑΚΗ
Βερνάρδος ΣΑΛΤΑΜΑΝΙΚΑΣ

Πώς να εκπαιδεύσετε τον παπαγάλο σας

doi:https://doi.org/10.12681/plogos.33698

Εισαγωγή

ΣΗΜΕΡΑ ΒΡΙΣΚΟΜΑΣΤΕ, ΣΕ ΕΝΑ ΣΗΜΕΙΟ ΣΤΗΝ ΙΣΤΟΡΙΑ ΤΗΣ ΕΞΕΛΙΞΗΣ μας, που απαιτεί επανεξέταση και επανατοποθέτηση ενδεχομένως των παραδεδεγμένων και δεδομένων αρχών που διέπουν τις ζωές μας. Τα μεγάλα γλωσσικά μοντέλα, σε ευρεία πλέον χρήση μέσω συστημάτων παραγωγικής ΤΝ όπως το ChatGPT αποκτούν όλο και περισσότερους χρήστες με πρακτικό αντίκρισμα σε πλήθος εφαρμογών. Η εξέλιξη των μοντέλων συστημάτων αυτών είναι σε θέση να μας κάνει να στεκόμαστε και να ανα-θεωρούμε τις αξίες του βίου, ρόλος που παραδοσιακά ανήκε στην ανθρωπολογική μελέτη. Η τεχνολογία έχει έρθει έξωθεν και έχει παράλληλα εσωτερικευθεί με έναν αδάμαστο τρόπο. Καθώς οι μηχανές που προκύπτουν είναι ολοένα και πιο σύνθετες, οι λειτουργίες τους απλώνουν τα πέπλα τους σε κάποια βαθύτερα και ενδόμυχα σύνορα. Εν προκειμένω, η συζήτηση θα γίνει με αφορμή το άρθρο: "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?", το οποίο συνέγραψε, μεταξύ άλλων, η τότε επικεφαλής του τμήματος Ηθικής και Τεχνητής Νοημοσύνης (ΤΝ) στην Google, Timnit Gebru.[1] Τι είναι όμως οι

[1] Η δημοσίευση του εν λόγω άρθρου οδήγησε στον τερματισμό της συνεργασίας της Gebru με την Google. Η δημοφιλία της στην εταιρία παραμένει και υπάρχουν φωνές που επιθυμούν την επιστροφή της βλ. Urian, B. Google Ai researchers want Timnit Gebru to come back at higher position among other demands, Tech Times, (2020). https://www.techtimes.com/articles/255136/20201216/google-ai-researchers-demand-new-policies-leadership-changes-and-timnit-gebru-to-come-back-at-higher-position.htm

"στοχαστικοί παπαγάλοι"; Τα προαναφερθέντα πτηνά φημίζονται για την ικανότητα τους να "μιλούν". Να μπορούν να αναπαράγουν δηλαδή όσα ακούν από τους ανθρώπους. Ένα εξόχως καυστικό όνομα για τα γλωσσικά μοντέλα. Πρόκειται για τις μηχανές εκείνες που μοντελοποιούν τη γλώσσα χρησιμοποιώντας διάφορες στατιστικές και τεχνικές για να προσδιορίσουν την πιθανότητα μιας δεδομένης ακολουθίας λέξεων που θα εμφανιστεί σε μια πρόταση.

Τα γλωσσικά μοντέλα όπως το ChatGPT λειτουργούν μέσω μιας διαδικασίας μηχανικής μάθησης γνωστής ως βαθιά μάθηση, χρησιμοποιώντας συγκεκριμένα έναν τύπο αρχιτεκτονικής νευρωνικού δικτύου που ονομάζεται μετασχηματιστής (transformer).[2] Αυτά τα μοντέλα εκπαιδεύονται σε τεράστιες ποσότητες δεδομένων κειμένου από πηγές όπως βιβλία, ιστότοποι και άρθρα, όχι για να απομνημονεύσουν το περιεχόμενο, αλλά για να μάθουν στατιστικά μοτίβα στη χρήση της γλώσσας. Αυτή η διαδικασία, που επαναλαμβάνεται δισεκατομμύρια φορές, επιτρέπει στο μοντέλο να δημιουργεί συνεκτικό και συναφές με το συγκείμενο κείμενο. Μετά από αυτή την αρχική προ-εκπαίδευση, το μοντέλο μπορεί να τελειοποιηθεί για συγκεκριμένες εφαρμογές και να προσαρμοστεί για λόγους ασφάλειας μέσω μεθόδων όπως η ενισχυτική μάθηση από ανθρώπινη ανατροφοδότηση. Όταν ένας χρήστης εισάγει κείμενο, το μοντέλο το χωρίζει σε tokens (τμήματα λέξεων), προβλέπει τα πιο πιθανά επόμενα tokens με βάση την εκπαίδευσή του και τα μετατρέπει ξανά σε αναγνώσιμη γλώσσα, παράγοντας τις απαντήσεις που βλέπουμε.[3]

Ο τρόπος μάθησης τους όμως δεν έχει να κάνει με την καλλιέργεια κάποιας κριτικής ικανότητας ή τη δυνατότητα απόδοσης νοήματος σε αυτά που αναπαράγουν. Σε αυτό το γεγονός οφείλεται και το προσωνύμιο στοχαστικοί παπαγάλοι. Πρόκειται δηλαδή για ικανά chatbot που έχουν σχεδιαστεί να προσομοιώνουν την συνομιλία με τους ανθρώπους και να αναπαράγουν προτάσεις με σωστή συντακτική και γραμματική δομή οι οποίες είναι σχετικές με το υπό συζήτηση θέμα.

Η ονοματοδοσία αυτή έγινε, γιατί ακριβώς μπορούν απλώς να αναπαράγουν αυτά που "μαθαίνουν" αλλά όχι να τα κατανοούν. Ασχολούνται με τη γλώσσα, τα γλωσσικά συστήματα εν γένει και την αναπαραγωγή της γλώσσας, με τρόπο που προσιδιάζει επιφανειακά στον ανθρώπινο, χωρίς

[2] Partha Pratim Ray, "ChatGPT: A Comprehensive Review on Background, Applications, Key Challenges, Bias, Ethics, Limitations and Future Scope," *Internet of Things and Cyber-Physical Systems* 3, no. 1 (April 14, 2023): 121–54, https://doi.org/10.1016/j.iotcps.2023.04.003.

[3] Στο ίδιο.

όμως να κατανοούν το περιεχόμενο και κυρίως χωρίς να μεταδίδουν ακριβή πάντα "γνώση". Το νοητικό πείραμα του κινέζικου δωματίου του Searle[4], μας επιτρέπει να διαπιστώσουμε ότι η απλή αναπαραγωγή των λεχθέντων δεν συνεπάγεται, ούτε προϋποθέτει την κατανόησή τους. Η πραγματική κατανόηση συμβόλων, περιλαμβάνει την "σημασιολογία" (semantics), δηλαδή μια γνώση του τι αναπαριστούν τα σύμβολα ή τι σημαίνουν.[5] Όσα συμβαίνουν μέσα σε έναν υπολογιστή, είναι ανάλογα όσων συμβαίνουν σε ένα κινέζικο δωμάτιο. Γίνεται διαχείριση των συμβόλων σε σχέση με τα σχήματά τους, βάσει συντακτικών κανόνων.[6] Ωστόσο η γλώσσα δεν είναι απλώς μια συμφωνηθείσα σειρά γραμμάτων, αλλά και ένα ολόκληρο πλαίσιο νοημάτων και συνειρμικών διαδικασιών. Στην προκειμένη περίπτωση λοιπόν, οι μηχανές αυτές αναφέρονται ως παπαγάλοι, γιατί απλώς αναπαράγουν αυτό βάσει του οποίου εκπαιδεύονται ή που προκύπτει από τους αλγορίθμους, χωρίς όμως να αντιλαμβάνονται το νόημα των όσων λένε. Επιπλέον η ικανότητα που έχουν τα μεγάλα γλωσσικά μοντέλα (LLMs) να δημιουργήσουν συνθετικά κείμενα, μπορεί να μας παραπλανήσει.[7] Αυτό γιατί, ενώ οι μηχανές συνθέτουν το κείμενο χωρίς να του προσδίδουν κάποιο νόημα, ο άνθρωπος έχει την τάση να αποδίδει νόημα σε ό,τι διαβάζει, με αποτέλεσμα να καταλήγει να ερμηνεύει μία στατιστικά επικυρωμένη σύνθεση λέξεων σε προτάσεις.

Η ανάπτυξη τέτοιων μεγάλων γλωσσικών μοντέλων μπορεί να προκαλέσει αρκετά ανεπιθύμητα αποτελέσματα. Παρατηρούμε πως στο εν λόγω άρθρο της Gebru, εντοπίζονται διάφορα ηθικά ζητήματα που αφορούν προβλήματα κοινωνικού αντικτύπου και προβλήματα σχετικά με την κατεύθυνση της έρευνας. Αναδύονται ζητήματα όπως α) η αναπαραγωγή διακρίσεων και στερεοτύπων, β) χειραγώγηση και παραπληροφόρηση αλλά και γ) κρίσιμα θέματα περιβαλλοντικών επιπτώσεων. Ωστόσο τα ζητήματα αυτά έχουν τις ρίζες τους στο μακρινό παρελθόν. Προβλήματα που άπτονται περιβαλλοντικών θεμάτων είναι συνυφασμένα με την ανθρώπινη ανάπτυξη, αν και κάνουν εντονότερη την εμφάνισή τους από τη βιομηχανική επανάσταση και έπειτα. Επίσης, το ζήτημα των διακρίσεων μπορεί να εντοπιστεί και να εξεταστεί μέσα από κοινωνικά φαινόμενα όπως η πατριαρχία ή ο θεσμός της δουλείας ή και πολιτευμάτων όπως η βασιλεία και

[4] John Searle, "Minds, Brains, and programs" *The Behavioral and brain sciences*, (1980) 3, 417-457, διαθέσιμο στο https://www.law.upenn.edu/live/files/3413-searle-j-minds-brains-and-programs-1980pdf

[5] Jaegwon Kim, *Η Φιλοσοφία του Νου* (Αθήνα: Liberal Books, 2016), 169.

[6] Στο ίδιο.

[7] Εδώ, ίσως, γίνεται καλύτερα κατανοητό και το προσωνύμιο στοχαστικοί παπαγάλοι.

η ολιγαρχία. Με τέτοιες καταστάσεις ερχόταν αντιμέτωπη η ανθρωπότητα ανέκαθεν, και θα ήταν ευκαιρία η ύπαρξη, ανάδυση και ευρεία χρήση της ΤΝ να προσβλέπει και να συμβάλλει εν τέλει αποτελεσματικά στην επίλυση και όχι την διαιώνισή τους.

Γλωσσικά μοντέλα: μηχανές Τεχνητής Νοημοσύνης

Συζητώντας για ΤΝ είναι χρήσιμο να θέτει κανείς το ερώτημα και να οριοθετεί το τι εννοούμε με τον όρο "τεχνητή νοημοσύνη"; Ο ορισμός της νοημοσύνης είναι προς ώρας ένα ανοιχτό και γι' αυτό ενοχλητικό (παρ' ότι οι θεσμοί όπως οι ΕΕ έχουν σπεύσει να ορίσουν επακριβώς και να εστιάσουν την απόδοση του όρου "τεχνητή νοημοσύνη" στα συστήματα που μιμούνται την ανθρώπινη συμπεριφορά)[8] ερώτημα στη φιλοσοφία του νου. Φαίνεται ότι έχουμε μια σταθερή διαισθητική αντίληψη του τι είναι νοημοσύνη. Σε γενικές γραμμές, είναι η ικανότητα κάποιου να συλλογίζεται, να σκέφτεται λογικά, να χρησιμοποιεί τη φαντασία, να μαθαίνει και να ασκεί κρίση. Είναι η ικανότητα να πλαισιώνεις ένα πρόβλημα και μετά να το λύνεις. Η νοημοσύνη είναι γενικεύσιμη. Είναι σε θέση να κάνει αυτά τα πράγματα σε ένα ευρύ φάσμα προβλημάτων και πλαισίων. Είναι αυτή που έχουν οι άνθρωποί, αυτό που έχουν λιγότερο τα πρωτεύοντα, οι παπαγάλοι ακόμα λιγότερο, οι μέδουσες και τα δέντρα (και οι σύγχρονες μηχανές) καθόλου.[9] Η ΤΝ είναι η νοημοσύνη σε ένα τεχνούργημα που έχουμε δημιουργήσει.[10]

Στην ΤΝ, η "μάθηση" υπονοεί την εξαγωγή "γνώσεων" μέσα από δεδομένα, με τρόπο που να επιτρέπει τη βελτίωση του συστήματος σε μία εργασία.[11] Αυτού του είδους η διαδικασία ονομάζεται μηχανική μάθηση (machine learning) και είναι υποπεδίο της ΤΝ που δίνει στους υπολογιστές

[8] Ευρωπαϊκό Κοινοβούλιο, "Τι είναι η τεχνητή νοημοσύνη και πώς χρησιμοποιείται; |Θέματα|Ευρωπαϊκό Κοινοβούλιο," www.europarl.europa.eu, September 9, 2020, https://www.europarl.europa.eu/topics/el/article/20200827STO85804/ti-einai-i-techniti-noimosuni-kai-pos-chrisimopoieitai.

[9] Robert Sparrow, "The Turing Triage Test", *Ethics and Information Technology* (2004) 6: 203–213, 204.

[10] Στο ίδιο.

[11] Γιώργος Γιαννακόπουλος, *Τεχνητή Νοημοσύνη: Μια Διακριτική Απομυθοποίηση* (Θεσσαλονίκη: Ροπή, 2020), 125.

τη δυνατότητα να "μαθαίνουν" χωρίς να είναι ρητά προγραμματισμένοι.[12] Η μηχανική μάθηση βρίσκεται πίσω από τα chatbot και το προγνωστικό κείμενο, τις εφαρμογές μετάφρασης γλώσσας, τις εκπομπές που προτείνει το Netflix και τον τρόπο παρουσίασης των ροών στα μέσα κοινωνικής δικτύωσης. Οι κανόνες με τους οποίους δρουν τα συστήματα ΤΝ μπορούν να αλλάζουν κατά τη διάρκεια του χειρισμού της μηχανής από την ίδια τη μηχανή, από τον "εαυτό" της.[13] Η βαθιά μάθηση (deep learning) ουσιαστικά προσπαθεί να κάνει τον υπολογιστή να μαθαίνει με όλο και μεγαλύτερη ακρίβεια.[14] Έχει τη δυνατότητα να αξιοποιεί πολύ καλά, τον τεράστιο όγκο δεδομένων που υπάρχει εκεί έξω (κατά βάση στο διαδίκτυο), για να κάνει πράγματα, όπως η αναγνώριση φωνής, κειμένου και άλλα.[15]

Αναδυόμενα προβλήματα από τις Ανοιχτές Βάσεις Δεδομένων

Το διαδίκτυο είναι ένας τεράστιος, γεμάτος ποικιλία ανοιχτά προσβάσιμος χώρος και θα φανταζόμασταν ότι τα πολύ μεγάλα δεδομένα (large datasets) θα εξασφάλιζαν και θα μπορούσαν να αντικατοπτρίσουν αυτή την ποικιλομορφία.[16] Το ζήτημα είναι λοιπόν ότι αυτά τα γλωσσικά μοντέλα με τον τρόπο της μηχανικής μάθησης και της βαθιάς μάθησης, φτάνουν να αναπαράγουν στερεότυπα και προκαταλήψεις, μιας και το πεδίο από όπου συλλέγουν τα δεδομένα τους είναι ο αχανής χώρος του διαδικτύου.

Πρόκειται για μέσα που συλλέγουν τα δεδομένα τους από το διαδίκτυο και το περιβάλλον αυτό, είναι τεράστιο και χαοτικό, με αποτέλεσμα να μην επιτρέπει τον πλήρη έλεγχο των προγραμματιστών στα δεδομένα που έχουν πρόσβαση. Προκειμένου να παραχθεί συνεκτικό κείμενο, τα γλωσσικά μοντέλα συνήθως εκπαιδεύονται σε ογκώδη σύνολα δεδομένων. Είναι δύσκολο να χρησιμοποιήσεις ένα σύνολο δεδομένων μεγάλου

[12] Machine Learning, Explained", Sara Brown, "https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

[13] Andreas Matthias, "The Responsibility gap: Ascribing responsibility for the actions of learning automata", *Ethics and Information Technology 6* (2004), 175-183, 177.

[14] Γιαννακόπουλος, 137.

[15] Στο ίδιο.

[16] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, Shmargaret Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?", *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (March 2021), 610–623.

μεγέθους που απαιτείται για την εκπαίδευση ενός γλωσσικού μοντέλου, διασφαλίζοντας παράλληλα ότι το σύνολο δεδομένων απηχεί την επιθυμητή συμπεριφορά.[17] Αυτή η αδυναμία ελέγχου της πληροφορίας, αφήνει μεγάλα κενά που επιτρέπουν την αναπαραγωγή και διαιώνιση στερεοτύπων και προκαταλήψεων. Τα μοντέλα αυτά φαίνεται να αναπαράγουν στερεοτυπικές συμπεριφορές σχετικές με το φύλο, την φυλή, την εθνικότητα ακόμα και με κατάσταση αναπηρίας. Σημειώνεται ότι το μοντέλο τείνει να αντικατοπτρίζει δυτικοκεντρικές προοπτικές και επιδεικνύει την υψηλότερη απόδοσή του στην αγγλική γλώσσα. Πολλά από τα μέτρα προστασίας που έχουν σχεδιαστεί για την πρόληψη επιβλαβούς περιεχομένου έχουν δοκιμαστεί κυρίως σε αγγλόφωνο περιβάλλον.[18]

Με αυτόν τον τρόπο συμβάλλουν στο να εμφανίζονται όλο και πιο συχνά οι γνώμες αυτές, με αποτέλεσμα να συνεχίζουν να υπερ-εκπροσωπούνται στα νέα δεδομένα με τα οποία μετ-εκπαιδεύονται τα μοντέλα, διαιωνίζοντας τα προβλήματα. Επιπλέον η διόρθωση αυτών των προβλημάτων θα πρέπει να αντιμετωπίσει και τα κενά ευθύνης (responsibility gap).[19] Στην ανάπτυξη και χρήση των LLM εμπλέκεται ένας μεγάλος αριθμός ανθρώπων υπό διαφορετικούς ρόλους και ιεραρχικά επίπεδα. Αυτή η πολυπλοκότητα καθιστά εξόχως δύσκολο τον ορθό και ακριβή επιμερισμό ευθυνών.[20] Ορισμένοι μελετητές υποστηρίζουν ότι οι προηγμένες τεχνολογίες τεχνητής νοημοσύνης έχουν δημιουργήσει ένα "κενό ευθύνης", όπου ούτε οι προγραμματιστές ούτε οι χρήστες μπορούν να αναλάβουν πλήρως την ευθύνη για τα αποτελέσματα, λόγω του περιορισμένου ελέγχου που ασκούν επί των αυτόνομων συστημάτων. Αυτές οι μηχανές μάθησης προσαρμόζονται μέσω της ανατροφοδότησης από το περιβάλλον και της δοκιμής και του λάθους (trial and error), ενεργώντας ανεξάρτητα σε ανθρώπινα περιβάλλοντα και αλληλεπιδρώντας με κοινωνικές δομές, περιπλέκοντας έτσι τα ζητήματα ευθύνης και υπευθυνότητας (responsibility και

[17] Irene Solaiman, Christy Dennison, "Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets" 2, διαθέσιμο στο https://openai.com/blog/improving-language-model-behavior/
Επίσης: OpenAI, "How Should AI Systems Behave, and Who Should Decide?," Openai.com, 2023, https://openai.com/index/how-should-ai-systems-behave/.

[18] OpenAI, "Is ChatGPT Biased? | OpenAI Help Center," help.openai.com, 2024, https://help.openai.com/en/articles/8313359-is-chatgpt-biased.

[19] Matthias, 181.

[20] Claudio Novelli, Mariarosaria Taddeo, και Luciano Floridi, "Accountability in Artificial Intelligence: What It Is and How It Works," *SSRN Electronic Journal*, 2022, https://doi.org/10.2139/ssrn.4180366.

accountability).[21] Ανακύπτουν επομένως, διάφορα ερωτήματα που υπογραμμίζουν το μεγαλύτερο ερώτημα που τα περιβάλει: αξίζει όλος αυτός ο κόπος για τα αποτελέσματα που έχουμε; Οι συγγραφείς του 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" επισημαίνουν ότι εξαιτίας αυτών των αυξημένων απαιτήσεων σε δεδομένα που χρειάζεται η εκπαίδευση ενός μεγάλου γλωσσικού μοντέλου βρισκόμαστε μπροστά σε δυσάρεστες συνέπειες. Από τη μία η κατασκευή τους επιβαρύνει το περιβάλλον και έχει παράλληλα μεγάλα οικονομικά κόστη. Για να γίνει αντιληπτό το μέγεθος των περιβαλλοντικών επιπτώσεων, αρκεί να αναλογιστούμε ότι, ενώ ο μέσος άνθρωπος ευθύνεται για 5 τόνους εκπομπών CO2 το χρόνο, η εκπαίδευση ενός μεγάλου γλωσσικού μοντέλου θα προκαλέσει 284 τόνους εκπομπών.[22] Από την άλλη μεριά, υπάρχουν προβλήματα και κατά την ίδια τη λειτουργία των μοντέλων. Ενισχύουν τις προαναφερθείσες κοινωνικές διακρίσεις, καθώς αναπαράγουν την επικρατούσα ιδεολογία των ανθρώπων που έχουν τη μεγαλύτερη πρόσβαση και παρουσία στο διαδίκτυο, διογκώνουν το χάσμα αρχειοθέτησης (documentation gap)[23] και αναπαράγουν στερεότυπα που πολλές φορές μπορεί να είναι επιβλαβή.[24] Όσον αφορά την κατεύθυνση της έρευνας και δεδομένου ότι ο χρόνος του ερευνητή είναι κι αυτός ένα πολύτιμο κεφάλαιο, γεννιούνται ερωτήματα σε σχέση με το πού θα πρέπει να εστιαστούν οι ερευνητικοί πόροι και ως προς το αν η έρευνα πάνω στα LMs αποτρέπει τη διάθεσή τους σε άλλες προσπάθειες κατανόησης της φυσικής γλώσσας (NLU).[25]

---

[21] Eva Schur, Anna Brouns, και Peter Lee, "Ethical Analysis of the Responsibility Gap in Artificial Intelligence ," *International Journal of Ethics and Society* 6, no. 4 (2025): 1–10, https://doi.org/10.22034/ijethics.6.4.

[22] Emily M.Bender, "On the dangers of stochastic parrots: Can language models be too big?", Lecture at The Alan Turing Institute, (2021) 9:37-10:25. https://www.youtube.com/watch?v=N5c2X8vhfBE.

[23] Την έλλειψη μιας στοχευμένης ταξινόμησης των δεδομένων που χρησιμοποιούν. Αναφέρεται δηλαδή, στην έλλειψη ολοκληρωμένων, διαφανών ή προσβάσιμων αρχείων που περιγράφουν λεπτομερώς τον τρόπο με τον οποίο αναπτύχθηκε ένα σύστημα ΤΝ, τον τρόπο λειτουργίας του και τον τρόπο λήψης αποφάσεων στο πλαίσιο αυτού.

[24] Bender, E.M., Gebru, T., McMillan-Major, A., et al. (2021) "On the dangers of stochastic parrots: Can language models be too big?". FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (2021):614-615, https://doi.org/10.1145/3442188.3445922

[25] Bender,On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?,615.

Ποιος έχει την ευθύνη των παπαγάλων;

Παραδοσιακά έχουμε συνδέσει την ευθύνη με την δράση προσώπων.[26] Οι νέες τεχνολογίες επηρεάζουν τους ανθρώπους ήδη με την ύπαρξή τους. Επηρεάζουν τις αποφάσεις που παίρνουμε, πείθουν, διευκολύνουν και επιτρέπουν συγκεκριμένες ανθρώπινες γνωστικές διαδικασίες, πράξεις και συμπεριφορές.[27] Για παράδειγμα οι μηχανές αναζήτησης στο διαδίκτυο δίνουν προτεραιότητα και παρουσιάζουν πληροφορίες με μία συγκεκριμένη σειρά, επηρεάζοντας έτσι το τι βλέπουν οι χρήστες του διαδικτύου.[28] Τέτοια τεχνολογικά αντικείμενα είναι "ενεργοί μεσολαβητές" που συνδιαμορφώνουν ενεργά την ύπαρξη των ανθρώπων, την αντίληψη, την εμπειρία και τις πράξεις τους.[29]

Τα λάθη ήταν μέχρι τώρα, πάντοτε λάθη του προγραμματιστή, όχι του προγράμματος. Μπορούσαν να αναγνωριστούν, να απομονωθούν και να φτιαχτούν και ο προγραμματιστής μπορούσε εύκολα να θεωρηθεί υπαίτιος για κάθε σφάλμα της μηχανής. Ωστόσο καθώς οι τεχνικές της τεχνητής νοημοσύνης και του προγραμματισμού εξελίσσονται περεταίρω, αλλάζει και ο ρόλος των προγραμματιστών.[30] Προγραμματίζονται οι μηχανές, ώστε να μπορούν να προγραμματίζουν τον εαυτό τους. Και έτσι ο προγραμματιστής μετατρέπεται σε δημιουργό λογισμικών οργανισμών των οποίων τους κώδικες δεν γνωρίζει επακριβώς και καθίσταται αδύνατον να τους ελέγξει για τυχόν λάθη.[31]

Η παραπάνω κατάσταση κατά την οποία η λειτουργία μιας μηχανής δεν είναι ξεκάθαρη για τον χρήστη της ή για οποιονδήποτε ενδιαφερόμενο είναι ονομάζεται black box problem. Η ασάφεια με οποία ερχόμαστε αντιμέτωποι κατά το black box problem είναι ανάλογη με την αυτονομία που έχει μια μηχανή στον τρόπο που μαθαίνει. Η συμπεριφορά μιας μηχανής δεν είναι πλέον καθορισμένη αλλά σχηματίζεται με την αλληλεπίδρασή της με το περιβάλλον, από το οποίο η μηχανή υιοθετεί καινούργια συμπεριφοριστικά μοτίβα. Είναι μία διαδικασία μάθησης, βέβαια, πράγμα που

[26] Όπως είδαμε παραπάνω τις έννοιες responsibility και accountability. Την ευθύνη προκειμένου να μην συμβεί κάτι αλλά και της υπευθυνότητας ανάληψης ευθύνης.
[27] Noorman, Merel, "Computing and Moral Responsibility", The Stanford Encyclopedia of Philosophy (Spring 2020 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/spr2020/entries/computing-responsibility/
[28] Στο ίδιο.
[29] Στο ίδιο.
[30] Matthias, 181.
[31] Στο ίδιο, 182.

σημαίνει ότι ίσως κάποια λάθη να είναι απλώς ένα βήμα προς την εξερεύνηση και εξεύρεση της λύσης.[32]

Εν προκειμένω, αλλάζουν οι συμβατές έννοιες της ηθικής ευθύνης. Για να θεωρηθεί ένας άνθρωπος υπεύθυνος για μία πράξη, πρέπει να υπάρχει μία αιτιώδης σύνδεση ανάμεσα σε αυτόν και το αποτέλεσμα της πράξης. Το δρών υποκείμενο πρέπει να έχει γνώση και να είναι ικανό να αναλογιστεί τις πιθανές συνέπειες τις πράξεις του. Και τέλος το ίδιο υποκείμενο πρέπει να είναι ικανό να επιλέγει ελεύθερα τον τρόπο δράσης του, να μην υπόκειται σε κανέναν εξαναγκασμό.[33] Στο πλαίσιο όμως της τεχνολογίας, τα πράγματα γίνονται κάπως θολά. Διότι υπάρχει το πρόβλημα "των πολλών χεριών" (Problem of Many Hands), αφού για καθετί υπάρχουν πολλοί άνθρωποι που δουλεύουν πίσω από αυτά και σε πολλά διαφορετικά στάδια. Οι τεχνολογίες των υπολογιστών επιμηκύνουν την ανθρώπινη δραστηριότητα στο χώρο και το χρόνο.[34] Οι συνέπειες που προκύπτουν πολλές φορές είναι ανυπολόγιστες. Είναι δύσκολο να ληφθούν υπόψη όλες οι πιθανές παράμετροι, έτσι ώστε μία μηχανή να λειτουργεί σε ένα καθαρά ντετερμινιστικό περιβάλλον. Εδώ ακριβώς διογκώνεται το κενό ευθύνης (ή το χάσμα ευθύνης Responsibility Gap) μέσω και του black box problem[35]. Στο σημείο δηλαδή της μη δυνατότητας ελέγχου του τεράστιου όγκου πληροφοριών και την αδυναμία να προβλεφθεί κάθε πιθανή παραγόμενη από την μηχανή δράση.

Ωστόσο, δεν υπάρχει ένα καθολικό πρότυπο για προσβλητικό ή επιβλαβές περιεχόμενο· αλλάζει η ερμηνεία της συμπεριφοράς του γλωσσικού μοντέλου ανάλογα με τους πολιτισμικούς παράγοντες.[36] Ως εκ τούτου, μια διαδικασία για τον προσδιορισμό και την προσαρμογή της κατάλληλης συμπεριφοράς του μοντέλου θα πρέπει να είναι εφικτή και για πολλούς κοινωνικούς φορείς, ειδικά εκείνους που έχουν πληγεί περισσότερο και παραλείπονται στην ανάπτυξη των μοντέλων. Ομοίως, η συμπεριφορά του μοντέλου θα πρέπει να αξιολογείται σε ένα κοινωνικό πλαίσιο και με τρόπο που να περιλαμβάνει και τις περιθωριοποιημένες προοπτικές.[37] Παρουσιάζεται δηλαδή μια δυσκολία στην δημοκρατική πρόσβαση στα

[32] Στο ίδιο.
[33] "Computing and Moral Responsibility", 5-6.
[34] Στο ίδιο, 6.
[35] 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
[36] Solaiman, 2.
[37] Στο ίδιο.

τεχνολογικά αγαθά της ΤΝ και ανοίγονται ερωτήματα για το πώς αυτή η δημοκρατικοποίηση της ΤΝ θα μπορούσε να επιτευχθεί (Democratization of AI). Η επέκταση της πρόσβασης είναι ένα σημαντικό μέρος της υπεύθυνης ανάπτυξης συστημάτων τεχνητής νοημοσύνης, επειδή μας επιτρέπει να μάθουμε περισσότερα για τη χρήση στον πραγματικό κόσμο, αλλά και να συνεχίσουμε να χρησιμοποιούμε τέτοια μέσα με μεγαλύτερη ασφάλεια.[38]

Οι προτάσεις προκειμένου να μετριαστούν οι κίνδυνοι και να διατηρηθούν τα οφέλη

Απέναντι στα προαναφερθέντα προβλήματα υπάρχει πληθώρα βιβλιογραφικών πηγών που προτείνει λύσεις για κάθε ένα από αυτά. Όσον αφορά τα περιβαλλοντικά κόστη, προτείνεται να ενσωματωθεί η ενεργειακή και η υπολογιστική αποδοτικότητα στο σχεδιασμό και την αξιολόγηση αυτών των μοντέλων.[39] Με αυτό τον τρόπο θα επιτυγχάνεται η δημιουργία λιγότερο ενεργειακά κοστοβόρων νέων μοντέλων, ή τουλάχιστον θα πάψει να ενθαρρύνεται η κατασκευή μοντέλων ανεξαρτήτως των πόρων που αυτά απαιτούν.

Επίσης, ένας στοχευμένος περιορισμός των δεδομένων με τα οποία εκπαιδεύονται θα επέφερε θετικά αποτελέσματα. Αν τα μοντέλα τροφοδοτούνταν με συνειδητά επιλεγμένα δεδομένα και υπήρχε αρχειοθέτηση σε κάθε βήμα της διαδικασίας, τότε θα μπορούσε να ελεγχθεί και αποφευχθεί η αναπαραγωγή στερεοτύπων και υποτιμητικού περιεχομένου, ενώ παράλληλα θα μειωνόταν και το χάσμα αρχειοθέτησης. Επιπλέον η κατάλληλη επισήμανση (watermark), σε ένα κείμενο που έχει παραχθεί από μηχανή ΤΝ δεν θα άφηνε περιθώριο παραπλάνησης του αναγνώστη.

Τα ζητήματα που αφορούν την κατεύθυνση της έρευνας, από τη στιγμή που έχουν αναγνωριστεί, θα μπορούσαν επίσης να διευθετηθούν με τις κατάλληλες ενέργειες. Θα έπρεπε να κατανέμεται ο ερευνητικός χρόνος με προσοχή, και ίσως να αναζητηθούν τρόποι που να παρέχουν παρόμοια οφέλη χωρίς τη χρήση ολοένα και μεγαλύτερων LM.[40] Οι συγγραφείς του άρθρου καταλήγουν θέτοντας δύο ουσιώδη ερωτήματα: Ποιοι είναι οι κίνδυνοι γύρω από την έρευνα των LLM και τί πρέπει να λαμβάνουμε υπόψη

[38] Στο ίδιο.
[39] Στο ίδιο, 618.
[40] Στο ίδιο.

μας στην ανάπτυξη της; Είναι τελικά τα LLM τόσο αναγκαία και αναντικατάστατα; Αν όχι τί θα πρέπει να κάνουμε;[41]

Τα προβλήματα που σχετίζονται με την ανάπτυξη των LLM και η εξέλιξή τους σε σχέση με το παρελθόν

Όπως διαπιστώνουμε, υφίστανται πράγματι προβλήματα που προκύπτουν από την ανάπτυξη των LM. Αν και υπάρχει σύνδεση μεταξύ τους, μπορούμε να τα διακρίνουμε σε δύο κατηγορίες. Έχουμε προβλήματα κοινωνικού αντικτύπου (αναπαραγωγή στερεοτύπων, μεροληψία, οικονομικά και περιβαλλοντικά κόστη) αλλά και προβλήματα σε σχέση με την κατεύθυνση της έρευνας. Οι λύσεις που προτείνονται δεν επιβάλουν την συνέχεια της δημιουργίας όλο και μεγαλύτερων LM. Πριν φθάσουμε όμως στην εφαρμογή τους ή στην αναζήτηση νέων λύσεων, είναι καλό να εξετάσουμε το ευρύτερο πλαίσιο αυτών των προβλημάτων. Έχει σημασία να αναλογιστούμε γιατί μας αφορούν σήμερα υπό το πρίσμα των εξελίξεων στην ΤΝ και αν μπορούμε να εντοπίσουμε τις φιλοσοφικές ρίζες της εξέτασής τους.

Καταρχάς, τα προβλήματα σε σχέση με την ΤΝ μας αφορούν άμεσα γιατί βρισκόμαστε μπροστά σε ραγδαία ανάπτυξη του κλάδου, η οποία αναμένεται να αλλάξει ριζικά τις ζωές μας. Οι επιπτώσεις αυτής της ανάπτυξης είναι πιθανό να γίνουν αντιληπτές όταν πια θα είναι αργά για διορθωτικές κινήσεις (Collingridge dilemma).[42] Αυτή η ιδιαιτερότητα κάνει επιτακτική την ανάγκη να εντοπίσουμε εγκαίρως και να θέσουμε προς επίλυση τα ηθικά ζητήματα που προκύπτουν, αντιλαμβανόμενοι πως θα βρισκόμαστε πάντα αντιμέτωποι με ένα θεσμικό χάσμα.

Από ανθρωπολογικής πλευράς, η ρίζα αυτών των προβλημάτων είναι πιθανό να βρίσκεται στην τάση του ανθρώπου να υποτιμά τις παράπλευρες συνέπειες που προκύπτουν από την επίτευξη διαφόρων κατά τα άλλα

[41] Emily M.Bender, "On the dangers of stochastic parrots: Can language models be too big?", Lecture at The Alan Turing Institute, (2021): 31:23-31:42. https://www.youtube.com/watch?v=N5c2X8vhfBE

[42] Ο όρος Collingridge dilemma αναφέρεται στην δυσκολία να αναστρέψεις τις συνέπειες από την στιγμή που μια νέα τεχνολογία έχει εγκατασταθεί στην κοινωνία. Η ονομασία του όρου οφείλεται στον David Collingridge ο οποίος στο βιβλίο του The Social Control of Technology (1980) έγραψε σχετικά με τις προκλήσεις που υπάρχουν στην διαχείριση των νέων τεχνολογιών.

θεμιτών στόχων.[43] Το θεσμικό χάσμα που ενέχουν στόχοι όπως η ανάπτυξη μηχανών ΤΝ, καθιστά αδύνατη την έγκαιρη κατανόηση των νέων δεδομένων. Δεν παύει όμως να είναι απαραίτητη μια ορθολογική μελέτη των κινδύνων πριν την καθιέρωση της παρουσίας της ΤΝ στην κοινωνική ζωή. Το ενθαρρυντικό σε αυτή την κατεύθυνση είναι ότι τον 21ο αιώνα η μελέτη της βιωσιμότητας αλλά και της ηθικής των νέων τεχνολογιών ΤΝ (AI Ethics) αρχίζει να αποκτά όλο και πιο ουσιαστικό ρόλο στην έρευνα.[44]

Παρόλα αυτά, πρέπει να επισημανθεί ότι η εμφάνιση των προαναφερθέντων προβλημάτων κοινωνικού αντικτύπου (αναπαραγωγή στερεοτύπων, διακρίσεων και προκαταλήψεων) έχει μια ειδοποιό διαφορά έτσι όπως εμφανίζεται μέσω των LLM σε σχέση με το παρελθόν. Η εμφάνιση της στα LLM προκύπτει μόνο μετά την επαφή με το περιβάλλον (μέσω των δεδομένων για την εκπαίδευση του μοντέλου). Σε αντίθεση με τους φορείς στερεοτύπων και διακρίσεων, που έχουν επίγνωση του εκφερόμενου λόγου τους, τα LLM απλώς τον αναπαράγουν χωρίς να αντιλαμβάνονται το νόημά του (ως στοχαστικοί παπαγάλοι). Το γεγονός αυτό υποδεικνύει πως τα μοντέλα δεν θα αποτελούσαν πηγή ηθικού προβληματισμού σε ένα περιβάλλον που δεν είχε εκ των προτέρων τέτοια ζητήματα. Η ανάγκη λοιπόν να διορθώσουμε τα παραπάνω ζητήματα γύρω από την ΤΝ μπορεί να μας δώσει την ώθηση ώστε να βρούμε τη λύση να διορθώσουμε τα αίτια που οδήγησαν στην εμφάνισή τους σε πρώτο χρόνο και ανεξάρτητα από τις μηχανές ΤΝ.

Είναι πολύ βασικό να καθοριστεί, ποιος ή τι θα είναι υπεύθυνο για τις συνέπειες των αποφάσεων και των δράσεων ενός συστήματος ΤΝ και ποιος θα αναλάβει το βάρος αυτό της υποχρέωσης να απαντήσει στο τι πρέπει να κάνει και τι όχι.[45] Θα έρθει ένα σημείο όπου η κοινωνία θα πρέπει να αποφασίσει μεταξύ του να μην χρησιμοποιεί τέτοιου είδους μηχανές πια, το οποίο δεν φαίνεται να είναι μία ρεαλιστική επιλογή, ή να αντιμετωπίσει αυτό το κενό ευθύνης το οποίο δεν μπορεί να γεφυρωθεί από τις παραδοσιακές πρακτικές της ανάληψης ευθύνης.[46] Στη φάση που βρίσκεται η τεχνολογία, είναι περισσότερο παράλειψη των προγραμματιστών να

[43] Όπως διατυπώνει ο Alasdair MacIntyre στο " Dependent rational animals: Why human beings need the virtues ": «δρούμε πυροσβεστικά και όχι προληπτικά».

[44] Αυτή η τάση ενισχύθηκε κυρίως από δύο παράγοντες: την κλιματική κρίση όσο αφορά την βιωσιμότητα γενικά και την ανάπτυξη αυτόματων οχημάτων όσον αφορά τον τομέα AI Ethics.

[45] David J. Gunkel, "Mind the Gap: Responsible Robotics and the Problem of Responsibility", *Ethics and Information Technology* (2017), 2.

[46] Matthias, 175.

μην ελέγχουν το περιβάλλον στο οποίο υπάρχει και δουλεύει η μηχανή, παρά λάθος της ίδιας της μηχανής. Ίσως αργότερα στο μέλλον, όταν "ενηλικιώνεται" μια μηχανή, να μπορεί να το κάνει μόνη της. Προς το παρόν, η αμέλεια αυτή εμπίπτει ακριβώς σε αυτό το κενό ευθύνης. Που δεν είναι αμιγώς ευθύνη, αλλά παράλειψη. Η ευθύνη όμως προκύπτει και από τις πράξεις, αλλά και τις παραλείψεις μας. Διότι ακόμα και αν υποθέσουμε ότι η πρόθεση και η βούληση ως νοητικά φαινόμενα δεν επηρεάζουν τον φυσικό κόσμο σωματικά τουλάχιστον, οι πράξεις και οι παραλείψεις αυτές καθαυτές είναι που "γράφουν" πάνω στον εμπειρικό κόσμο. [47]

Προτεινόμενες ενέργειες με στόχο τις μέγιστες θετικές επιπτώσεις.

Η σωστή διάκριση των προβλημάτων και η αποφυγή της απορριπτικής τεχνοφοβίας είναι μείζονος σημασίας. Αρχικά, αυτό που μπορούμε να κάνουμε είναι να εστιάσουμε στην καλή κατανόηση των προβλημάτων που προκύπτουν. Στη συνέχεια, θα πρέπει άτομα και οργανισμοί να αναλάβουν την ευθύνη που έχουν απέναντι σε αυτά τα προβλήματα. Πρέπει να αντιμετωπίσουμε κάποια στιγμή κατά μέτωπο και με ειλικρίνεια το κενό ευθύνης, αντί να αφήνουμε το ερώτημα του καταλογισμού να αιωρείται εντός ενός ασαφούς και θολού τοπίου εμπλεκομένων. Πολλές φορές δουλεύοντας στο κομμάτι του προγραμματισμού χρειάζεται να λαμβάνουμε υπόψη ότι οι πράξεις μας δεν αφορούν μόνο ένα μαθηματικό σύμπαν, αλλά έχουν σοβαρές επιπτώσεις στον πραγματικό κόσμο.[48] Αναγνωρίζοντας και αναλαμβάνοντας την ευθύνη, έχουμε κάνει το πρώτο βήμα στην προσπάθεια να δημιουργήσουμε εγκαίρως το θεσμικό πλαίσιο που ρυθμίζει τα καινοφανή ζητήματα που δημιουργούνε οι νέες τεχνολογίες όπως η ΤΝ.

Σε τέτοιου είδους μηχανές που παρομοιάσαμε με παπαγάλους, φαίνεται πως δεν μπορούμε να δημιουργήσουμε ένα περιβάλλον εντελώς καθαρό προκαταλήψεων παρά μόνο μειωμένων προκαταλήψεων.[49] Ο τρόπος να φτάσουμε σε αυτό, είναι φιλτράροντας και δυνητικά αξιολογώντας τα δεδομένα που πρόκειται να χρησιμοποιηθούν.[50] Βέβαια εδώ ανακύπτει το

[47] Jaap Hage, "Theoretical foundations for the responsibility of autonomous agents", *Artif Intell Law* (2017), 255–271, 258. Διαθέσιμο στο https://link.springer.com/content/pdf/10.1007/s10506-017-9208-7.pdf

[48] Bender, "On the dangers of stochastic parrots: Can language models be too big?", 11:12-11:18

[49] Bender, "On the dangers of stochastic parrots: Can language models be too big?" 1:09

[50] Στο ίδιο.

ερώτημα, ποιος είναι ικανός να προσδιορίσει τι είναι σωστό να μάθει η μηχανή και αν έχει υπάρξει επαρκές δείγμα που να αντιπροσωπεύει τη διαφορετικότητα στο σύνολό της.[51] Ίσως μια περισσότερο ελεγμένη πρόσβαση σε δεδομένα (όπως οι αρχικές κλειστές βάσεις της OpenAI), μια ρήτρα στα δεδομένα που τυγχάνουν επεξεργασίας ή μια πρόσβαση σε – ει δυνατόν – ελεγμένα δεδομένα, θα διευκόλυνε να επιτευχθεί ένα σύστημα για λιγότερες προκαταλήψεις, ιδιαίτερα αν το σύστημα μιμείται την ανθρώπινη συμπεριφορά[52].

Ένας προγραμματισμός με κατηγοριοποίηση ευαίσθητων θεμάτων και σκιαγράφηση της επιθυμητής συμπεριφοράς, δημιουργία του συνόλου δεδομένων και έπειτα μικρορυθμίσεις, θα μπορούσε να βελτιώσει τη συμπεριφορά του γλωσσικού μοντέλου σε σχέση με συγκεκριμένες τιμές συμπεριφοράς, βελτιστοποιώντας ένα επιμελημένο σύνολο δεδομένων <100 παραδειγμάτων αυτών των τιμών.[53] Τέλος η αξιολόγηση των μοντέλων με στόχο τη μείωση της τοξικότητας, θα επέτρεπε τη δυνατότητα για μηχανές με λιγότερες αναπαραγωγές στερεοτύπων.[54]

Αξίζει να τονιστεί, ότι πρόκειται για ένα πεδίο που ήδη σημειώνονται εξελίξεις και διορθώσεις. Για παράδειγμα στο πρόγραμμα DALL·E έχουν γίνει προσπάθειες για ελαχιστοποίηση του κινδύνου κακής χρήσης του προγράμματος για τη δημιουργία παραπλανητικού περιεχομένου, απορρίπτοντας μεταφορτώσεις εικόνων που περιέχουν υπαρκτά πρόσωπα και απόπειρες δημιουργίας ομοιότητας δημοσίων προσώπων, συμπεριλαμβανομένων διασημοτήτων και εξέχων πολιτικών προσωπικοτήτων.[55] Κάνοντας τα φίλτρα περιεχομένου πιο ακριβή, ώστε να είναι πιο αποτελεσματικά στον αποκλεισμό μηνυμάτων προτροπής και μεταφορτώσεων εικόνων που παραβιάζουν την πολιτική περιεχομένου, ενώ παράλληλα επιτρέπουν δημιουργική έκφραση.[56]

Ίσως χρειάζεται επίσης να ενθαρρύνουμε την έρευνα σε κατευθύνσεις που δεν εξαρτώνται από τα μεγάλα LM. Η έρευνα θα μπορούσε να στραφεί και στην εξέταση τρόπων ώστε η ίδια η ΤΝ να προλαβαίνει ή να μετριάζει

[51] Στο ίδιο

[52] 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

[53] Solaiman, 4.

[54] Στο ίδιο, 5.

[55] "Reducing Bias and Improving Safety in DALL·E 2", https://openai.com/blog/reducing-bias-and-improving-safety-in-dall-e-2/

[56] Στο ίδιο.

τέτοια προβλήματα. Να κινηθούμε τελικά σε μια πιο εκτεταμένη χρήση της Ηθικής κατά τη σχεδίαση και ανάπτυξη των μηχανών ΤΝ. Να τεθεί δηλαδή η ηθική ερώτηση πριν ή παράλληλα με την υλοποίηση της μηχανής.[57]

Επίλογος

Η ανατροφή των παπαγάλων είναι ο προγραμματισμός τους και οι βάσεις δεδομένων από τις οποίες αντλούν τις πληροφορίες τους. Και αυτός πρέπει να γίνει βελτιστοποιώντας τα γλωσσικά μοντέλα, φιλτράροντας και αξιολογώντας δυνητικά τα στοχευμένα σύνολα δεδομένων. Είναι μια διαδικασία που πρέπει να γίνει εκ των προτέρων, ώστε να αποφευχθούν τα κακώς κείμενα στο μέτρο του δυνατού, αλλά και για να μην επέλθουν περιορισμοί μετά το πέρας της διαδικασίας με μορφή λογοκρισιών. Πρόκειται για μια "νέα γενιά παιδιών" που έχουμε να αναθρέψουμε και οι προκλήσεις είναι πολλές και ξεκινάνε από εμάς. Στο τέλος η ηθική των υπολογιστών αποδεικνύεται ότι επιστρέφει στη μελέτη των ανθρώπινων όντων και της κοινωνίας, των στόχων και των αξιών μας, των κανόνων συμπεριφοράς μας, του τρόπου με τον οποίο οργανωνόμαστε και αναθέτουμε δικαιώματα και ευθύνες.[58]

Τέλος, σημαντικό ζήτημα είναι ο τρόπος με τον οποίο η ανάπτυξη της ΤΝ μπορεί να έχει το μέγιστο θετικό αποτέλεσμα για τις εταιρίες και τους ανθρώπους εν γένει. Η Timnit Gebru επισημαίνει πως η ΤΝ είναι ένα εργαλείο και, όπως σε κάθε εργαλείο, τον τρόπο χρήσης τον καθορίζει κυρίως ο κατασκευαστής του.[59] Επομένως, αν χρησιμοποιηθεί για τη

[57] Μια ενδιαφέρουσα πρόταση για να αξιοποιηθούν καλύτερα οι ερευνητικοί πόροι που έχουν εγκλωβιστεί σε αμφιλεγόμενα προγράμματα όπως αποδεικνύονται τα LM είναι πάντα το πρόγραμμα που ξεκίνησαν το 2004 οι Susan και Michael Anderson με την ονομασία Machine Ethics (Ηθική των Μηχανών) το οποίο έχει σαν σκοπό να διασφαλίσει ότι τα συστήματα ΤΝ θα συμπεριφέρονται ηθικά απέναντι στον άνθρωπο. Με αυτό τον τρόπο δεν θα αποδεσμεύονταν μόνο οι ερευνητικοί πόροι αλλά θα μπορούσαμε να ευελπιστούμε για ανακαλύψεις που θα επέλυαν και τα ίδια προβλήματα που παρουσιάζουν οι LM. βλέπε: Michael Anderson και Susan Leigh Anderson, "Machine Ethics: Creating an Ethical Intelligent Agent", *AI Magazine* Volume 28 Number 4 (2007) και στο M. Anderson, S. L. Anderson, A. Gounaris, & G. Kosteletos, Conatus 6, no. 1 (2021): 177-202 DOI: https://doi.org/10.12681/cjp.26832

[58] Deborah G. Johnson, *Computer Ethics* (Upper Saddle River, NJ: Prentice Hall, 1985).

[59] Exclusive Interview Timnit Gebru: Computer Scientist, London speaker Bureau, (2021). https://www.youtube.com/watch?v=WOtJjPMt2NA

μεγιστοποίηση του κέρδους,[60] αυτό ενδεχομένως να επιτευχθεί με τρόπους επιβλαβείς για τον άνθρωπο. Όσο οι εταιρίες επιμένουν να βλέπουν τα πράγματα μόνο μέσα από αυτήν τη σκοπιά ο κίνδυνος αυτός είναι πραγματικός. Δεν υπάρχει άλλωστε κέρδος χωρίς βιωσιμότητα. Για αυτό είναι αναγκαίο οι εταιρείες να ευαισθητοποιηθούν, ίσως με δικιά τους πρωτοβουλία και σε σχέση με άλλες παραμέτρους. Να αναθεωρήσουν (όσες δεν το έχουν ήδη κάνει) την πρωτοκαθεδρία του οικονομικού κέρδους απέναντι σε παραμέτρους που σχετίζονται με τη βιωσιμότητα.

Οι μηχανές αποτελούν χρήσιμα εργαλεία που λειτουργούν με σκοπό την ανθρώπινη ευημερία και παρέχουν χρήσιμες υπηρεσίες που θα ήταν δύσκολο να αποκτηθούν αλλιώς, και αυτός τους ο ρόλος θα πρέπει να διατηρηθεί.[61] Όμως οι κατασκευαστές θα πρέπει όχι μόνο να αποκαταστήσουν αλλά να οικοδομήσουν σχέσεις εμπιστοσύνης με την κοινότητα.[62]

---

[60] Σε αυτό το σημείο είναι καλό να επισημανθεί πως αν και το επιχείρημα περί της πρόθεσης της χρήσης του εργαλείο από τον κατασκευαστή είναι βάσιμο, πρέπει να είμαστε προσεκτικοί ως προς την δαιμονοποίηση του κέρδους. Το κέρδος πολλές φορές είναι απλώς το αποτέλεσμα των δραστηριοτήτων μιας εταιρίας και πολλές επιχειρήσεις λειτουργούν με τρόπο συμβατό ή και ευεργετικό για την κοινωνία τους (B Corp Certification demonstrates a company's entire social and environmental impact. (bcorporation.net). Εδώ εντοπίζουμε λοιπόν μία απόκλιση στην κριτική της Gebru και θα ήταν σκόπιμο να αναγνωρίσει ότι η υιοθέτηση ηθικών κανόνων από μια εταιρία συνήθως διασφαλίζει ή ακόμα και αυξάνει το κέρδος της τελευταίας αυτής.

[61] Οι Brynjolfsson και McAfee στο *The second machine age: Work, progress, and prosperity in a time of brilliant technologies* (2014) οραματίζονται έναν κόσμο κοντά στο ιδανικό της αρχαίας Αθήνας χάρη στην χρήση νέων τεχνολογιών.

[62] Ένα επιπλέον γεγονός για την ανάγκη αποκατάστασης της εμπιστοσύνης είναι και η επίδραση που είχε η δημοσίευση του εν λόγω άρθρου, όχι στην επιστημονική κοινότητα, αλλά στις καριέρες των συγγραφέων. Η T.Gebru και οι M.Mitchel που ήταν οι επικεφαλής της ομάδας ηθικής της Google έχασαν τις δουλειές τους, και όπως φαίνεται στις ευχαριστίες του άρθρου κάποιοι συν-συγγραφείς απέφυγαν να βάλουν το όνομά τους για μην έχουν την ίδια τύχη. Πόσο καλά ενημερωμένοι μπορούμε να θεωρούμε ότι είμαστε όταν υπάρχουν τέτοιες απόπειρες λογοκρισίας και τι σημαίνει αυτό για την ελευθερία του λόγου; Βλ. Margaret Mitchell: Who's this researcher fired after Timnit Gebru, Tech Times, (2021). https://www.techtimes.com/articles/257231/20210219/margaret-mitchell-whos-google-researcher-fired-timnit-gebru.htm και Karen Hao, "We read the paper that forced Timnit Gebru out of Google. Here's what it says", MIT Technology Review, (2020). https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/

Αναφορές

Anderson, Michael. και Susan Leigh Anderson, "Machine Ethics: Creating an Ethical Intelligent Agent", *AI Magazine* Volume 28 Number 4 (2007).

Anderson, M., Anderson, S. L., Gounaris, A., & Kosteletos, G. (2021). Towards Moral Machines: A Discussion with Michael Anderson and Susan Leigh Anderson. Conatus - Journal of Philosophy, 6(1), 177–202. https://doi.org/10.12681/cjp.26832

Bender, Emily M., "On the dangers of stochastic parrots: Can language models be too big?" https://www.youtube.com/watch?v=N5c2X8vhfBE&t=1316s

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, Shmargaret Shmitchell. "On the dangers of stochastic parrots: Can language models be too big?". *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (March 2021), 610–623.

Brown, Sara. "Machine Learning, Explained". https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

Brynjolfsson, Erik, και Andrew McAfee. *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. WW Norton & Company, 2014.

Γιαννακόπουλος, Γιώργος. *Τεχνητή Νοημοσύνη: Μια Διακριτική Απομυθοποίηση*. Θεσσαλονίκη: Ροπή, 2020.

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

Collingridge, David. *The Social Control of Technology*, London: Frances Pinter (Publishers) Ltd., 1980.

Ευρωπαϊκό Κοινοβούλιο. "Τι είναι η τεχνητή νοημοσύνη και πώς χρησιμοποιείται; |Θέματα|Ευρωπαϊκό Κοινοβούλιο." www.europarl.europa.eu, September 9, 2020. https://www.europarl.europa.eu/topics/el/article/20200827STO85804/ti-einai-i-techniti-noimosuni-kai-pos-chrisimopoieitai.

Exclusive Interview Timnit Gebru: Computer Scientist, London speaker Bureau, (2021). https://www.youtube.com/watch?v=WOtJjPMt2NA

Gunkel, David J. "Mind the Gap: Responsible Robotics and the Problem of Responsibility". *Ethics and Information Technology* (2017).

Hage, Jaap. "Theoretical foundations for the responsibility of autonomous agents". *Artif Intell Law* (2017), 255–271, διαθέσιμο στο https://link.springer.com/content/pdf/10.1007/s10506-017-9208-7.pdf

Hao, Karen. "We read the paper that forced Timnit Gebru out of Google. Here's what it says." *MIT Technology Review* (December 2020).

Johnson, Deborah G. *Computer Ethics.* Upper Saddle River, NJ: Prentice Hall, 1985.

Kim, Jaegwon. *Η Φιλοσοφία του Νου*. Αθήνα: Liberal Books, 2016.

MacIntyre, Alasdair C. *Dependent rational animals: Why human beings need the virtues*. Vol. 20. Open Court Publishing, 1999.

Margaret Mitchell: Who's this researcher fired after Timnit Gebru, Tech Times, (2021). https://www.techtimes.com/articles/257231/20210219/margaret-mitchell-whos-google-researcher-fired-timnit-gebru.htm

Matthias, Andreas. "The Responsibility gap: Ascribing responsibility for the actions of learning automata". *Ethics and Information Technology 6* (2004), 175-183.

Neff, Gina, Peter Nagy. "Talking to Bots: Symbiotic Agency and the Case of Tay". *International Journal of Communication* (October 2016), 4915–4931.

Noorman, Merel, "Computing and Moral Responsibility", The Stanford Encyclopedia of Philosophy (Spring 2020 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/spr2020/entries/computing-responsibility/

Novelli, Claudio, Mariarosaria Taddeo, and Luciano Floridi. "Accountability in Artificial Intelligence: What It Is and How It Works." SSRN Electronic Journal, 2022. https://doi.org/10.2139/ssrn.4180366.

OpenAI. "How Should AI Systems Behave, and Who Should Decide?" Openai.com, 2023. https://openai.com/index/how-should-ai-systems-behave/.

OpenAI. "Is ChatGPT Biased? | OpenAI Help Center." help.openai.com, 2024. https://help.openai.com/en/articles/8313359-is-chatgpt-biased.

Ray, Partha Pratim. "ChatGPT: A Comprehensive Review on Background, Applications, Key Challenges, Bias, Ethics, Limitations and Future Scope." Internet of Things and Cyber-Physical Systems 3, no. 1 (April 14, 2023): 121–54. https://doi.org/10.1016/j.iotcps.2023.04.003.

Schur, Eva, Anna Brouns, και Peter Lee. "Ethical Analysis of the Responsibility Gap in Artificial Intelligence ." International Journal of Ethics and Society 6, no. 4 (2025): 1–10. https://doi.org/10.22034/ijethics.6.4.

Searle, John. "Minds, Brains, and programs". *The Behavioral and brain sciences*, (1980) 3, 417-457, διαθέσιμο στο https://www.law.upenn.edu/live/files/3413-searle-j-minds-brains-and-programs-1980pdf

Solaiman, Irene, Christy Dennison. "Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets" 2, διαθέσιμο στο https://openai.com/blog/improving-language-model-behavior/

Sparrow, Robert. "The Turing Triage Test", *Ethics and Information Technology* (2004) 6: 203–213, 204.

Urian, B. Google Ai researchers want Timnit Gebru to come back at higher position among other demands, Tech Times, (2020). https://www.techtimes.com/articles/255136/20201216/google-ai-researchers-demand-new-policies-leadership-changes-and-timnit-gebru-to-come-back-at-higher-position.htm

Περίληψη

Το παρόν δοκίμιο στοχεύει να αναδείξει ορισμένα ηθικά ζητήματα γύρω από τα μεγάλα γλωσσικά μοντέλα (LLM) ή αλλιώς στοχαστικούς παπαγάλους, τα οποία προκύπτουν σε σχέση κυρίως με το γεγονός ότι απαιτείται πολύ μεγάλος όγκος δεδομένων για την εκπαίδευσή – τον προγραμματισμό τους. Τα ζητήματα αυτά είναι αφετηριακά και εκκινούν από τις αρχές του 2021, όταν δημοσιεύτηκε ένα άρθρο από την Timnit Gebru την Emily Bender και τους συνεργάτες τους, με τίτλο "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?", στο οποίο αναδεικνύονται ζητήματα Ηθικής υπό το πρίσμα της Τεχνητής Νοημοσύνης

(ΤΝ). Το άρθρο εντοπίζει τους κίνδυνους γύρω από την ανάπτυξη όλο και μεγαλύτερων LLM σε σχέση και με τα οφέλη που προσφέρουν αλλά και πιθανές λύσεις. Τίθενται ερωτήματα σχετικά με ζητήματα που αφορούν την κατεύθυνση της έρευνας αλλά και τον καταμερισμό της ευθύνης γύρω από τις εξελίξεις στις τεχνολογίες ΤΝ. Τελικά, η παράλληλη έρευνα πάνω στην Ηθική, σε σχέση με τα ζητήματα ΤΝ, τι επίδραση έχει ή τι επίδραση θα έπρεπε να έχει στην κοινωνία;

Λέξεις κλειδιά: μηχανική μάθηση, μηχανές, γλωσσικά μοντέλα, στοχαστικοί παπαγάλοι, αυτόματα, ηθική ευθύνη, Τεχνητή Νοημοσύνη.

Keywords: Machine learning, Artificial Intelligence, Language Models, Moral Responsibility

---

Μαρίνα Ξενάκη
Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών
ORCID iD: 0009-0006-7676-6661

Βερνάρδος Σαλταμανίκας
Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών
ORCID iD: 0000-0001-7003-9996

Ελευθερία ΠΑΤΣΑΡΗ

Η τεχνολογία των πολεμικών drones υπό το πρίσμα της ηθικής: η περίπτωση των στοχευμένων δολοφονιών

doi:https://doi.org/10.12681/plogos.33345

Εισαγωγή

Τα DRONES ΑΠΟΤΕΛΟΥΝ ΕΝΑ ΤΕΧΝΟΛΟΓΙΚΟ ΕΠΙΤΕΥΓΜΑ, ΠΟΥ ΚΑτάφερε σταδιακά να γίνει ευρέως γνωστό και οικείο στους περισσότερους. Οι χρήσεις τους ποικίλλουν ανάλογα με το πλαίσιο στο οποίο τα εντάσσουμε κάθε φορά, όπως συμβαίνει και με τα περισσότερα επιτεύγματα της τεχνολογίας. Στην ανά χείρας πραγματεία δεν θα αναφερθούν αναλυτικά όλες οι χρήσεις τους παρά μόνο αυτές των πολεμικών drones, εκείνων δηλαδή που έχουν ρυθμιστεί και κατασκευαστεί με σκοπό να υποκαταστήσουν τους στρατιώτες κατά τη διάρκεια των ένοπλων συγκρούσεων, αφού σε αυτήν την κατηγορία προκύπτουν οι περισσότεροι προβληματισμοί σχετικά με την ηθική φύση αυτών, αλλά και για το ποιος θα πρέπει να αναλάβει την ευθύνη της χρήσεώς τους, ιδίως όταν πρόκειται για εξατομικευμένες δολοφονίες ή μαζικές επιθέσεις[1]. Λαμβάνοντας υπόψιν το γεγονός ότι η ευρεία χρήση των μη επανδρωμένων αεροσκαφών ξεκίνησε μετά την τρομοκρατική επίθεση της 11ης Σεπτεμβρίου το 2001 με σκοπό την αποφυγή παρόμοιων μελλοντικά επιθέσεων αλλά και την εξάλειψη της τρομοκρατίας γενικότερα, εγείρονται ερωτήματα σχετικά με τη δικαιοδοσία των χωρών -κυρίως των ΗΠΑ - για χρήση των

[1] 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

πολεμικών drones ακόμα και σε περιοχές με τις οποίες δεν βρίσκονται επίσημα σε ένοπλη σύγκρουση[2]. Πέραν τούτου, θα πρέπει να διασαφηνιστούν και άλλοι όροι όπως εκείνοι της εξουσιοδότησης ενός κράτος να καταφύγει στη χρήση στρατιωτικής δύναμης (AUMF, Authorization to Use Military Force) ή στα όρια σύμφωνα με τα οποία ένα κράτος χάνει το δικαίωμα για μη παρέμβαση στο έδαφός του (R2P, Responsibility to Protect).[3] Πρόσθετα, νέα ερωτήματα έρχονται στην επιφάνεια και για το δίκαιο του πολέμου, μια και η δυναμική του διαφοροποιείται συνεχώς εξαιτίας των νέων τεχνολογιών που εισάγονται, όπως είναι η υπό μελέτη περίπτωση. Σίγουρα, λοιπόν, τα υπό διασαφήνιση θέματα είναι πολλά και συνεχώς ανανεώνονται, με τα κυριότερα ερωτήματα να παραμένουν στο προσκήνιο όταν γίνεται λόγος για τέτοιου είδους επιχειρήσεις πολεμικού χαρακτήρα, τα οποία είναι τα εξής: Ποιος ορίζει το πού σταματά η δικαιοδοσία της χρήσης στρατιωτικής δύναμης; Ποιος έχει την ευθύνη όταν πρόκειται για τα τηλεχειριζόμενα drones; Η χρήση τέτοιου είδους τεχνολογικών επιτευγμάτων άπτεται του στρατιωτικού ή του νομικού δικαίου; Στο τρέχων κείμενο, θα γίνει μία προσπάθεια προσέγγισης των συγκεκριμένων ερωτημάτων αλλά και των ηθικών διλημμάτων που έπονται εφ΄αυτοίς.

Η συζήτηση γύρω από τα drones

Λόγος και αντίλογος

Ξεκινώντας τη συζήτηση γύρω από τα drones, ίσως να ήταν σκόπιμο να δοθεί ένας ορισμός αυτών, αλλά και μια γενική προσπάθεια πλαισίωσης των χρήσεών τους. Με τον όρο "drone" ή UAV (Unmanned Air Vehicle) αναφέρονται τα εναέρια οχήματα που δεν μεταφέρουν ανθρώπινο χειριστή, χρησιμοποιούν αεροδυναμικές δυνάμεις για την ανύψωσή τους, μπορούν να πετάξουν αυτόνομα ή με τηλεχειρισμό, είναι αναλώσιμα ή επαναχρησιμοποιούμενα και φέρουν φονικά ή μη φορτία[4]. Όπως αναφέρθηκε

[2] Koukoudakis, George, "*Drones' contribution to the transformation of contemporary warfare",* J. Military Stud. 2024; 13(1): 25.

[3] M. J. Boyle, ''*The legal and ethical implications of drone warfare'', The International Journal of Human Rights,* 19:2 (24 Feb 2015): 108,110.

[4] Φυτιλής Βασίλειος, ''*Drone και Δημοκρατικός Πόλεμος'',* Διπλωματική εργασία, Πανεπιστήμιο Πειραιώς (Αύγουστος 2022) σελ.17.

και νωρίτερα, τα τελευταία χρόνια η τεχνολογία των drones έχει εισβάλει στη ζωή μας και τόσο η μηχανική τους όσο και η χρήση τους έχουν γίνει ευρέως γνωστές στους περισσότερους από εμάς. Κάποιες από τις χρήσεις των μη επανδρωμένων αεροσκαφών- εκτός από την πολεμική στην οποία θα αναφερθούμε εκτενώς- είναι μεταξύ άλλων οι υπηρεσίες παράδοσης τροφίμων ή φαρμάκων σε απομακρυσμένες περιοχές όπου η πρόσβαση με άλλο μεταφορικό μέσο είναι εξαιρετικά δύσκολη, ενώ δεν είναι λίγες οι φορές που χρησιμοποιείται η συγκεκριμένη τεχνολογία σε αγροτικές παραγωγές με σκοπό την επίβλεψη μεγάλων γεωργικών εκτάσεων. Παράλληλα, τα drones αποτελούν και χόμπι πολλών -ερασιτεχνών και μη- φωτογράφων ή γεωγράφων χρησιμοποιώντας τα για αεροφωτογραφίες ή ακόμα και για χαρτογραφήσεις δύσβατων περιοχών, χρήσεις οι οποίες, βεβαίως, δεν αποτελούν συστήματα Τεχνητής Νοημοσύνης, μας δίνουν, όμως, μία ευρεία εικόνα της δυναμικής τους. Καθίσταται, λοιπόν, εύκολα σαφές, πως μια τέτοιου είδους τεχνολογία δύναται να διευκολύνει αρκετά μέρη του πληθυσμού -ιδίως εκείνα τα οποία κατοικούν μακριά από τα αστικά κέντρα- ή ακόμα να δώσει λύση σε αρκετούς επαγγελματικούς κλάδους οι οποίοι χρειάζονται ένα μέσο μεγάλης εμβέλειας για επιτήρηση. Επομένως, το συγκεκριμένο πλαίσιο χρήσης στο σημείο αυτό είναι και εκείνο που καθιστά τη χρήση τους θεμιτή από κάθε άποψη.

Στην άλλη όψη του νομίσματος, υπάρχει και η πολεμική χρήση. Αν και τα πολεμικά drones είχαν έρθει στο προσκήνιο ήδη από τα τέλη της δεκαετίας του 1980, δεν παρουσίαζαν ωστόσο τις ίδιες δυνατότητες με σήμερα όπου και επίσημα υποκαθιστούν τους στρατιώτες. Συγκεκριμένα, αρκετά νωρίτερα από τον Β' μόλις Παγκόσμιο Πόλεμο ή από τις αρχές του Ψυχρού Πολέμου, η χρήση τους περιοριζόταν στους στόχους για επανδρωμένα αεροσκάφη και αυτό μόνο κατά τη διάρκεια εκπαίδευσης[5]. Το μεγάλο άλμα στη χρήση τους εντός του πεδίου μάχης έγινε στις αρχές της δεκαετίας του 2000, μετά την τρομοκρατική επίθεση στις ΗΠΑ και την πτώση των Δίδυμων Πύργων, με σκοπό ουσιαστικά την αρχή ενός νέου πολέμου, αυτού

[5] Φυτιλής Βασίλειος, *ό.π,* 6.

κατά της τρομοκρατίας[6]. Στη συνέχεια η χρήση τους επεκτάθηκε ακόμη και σε περιοχές όπου η Αμερική δεν ήταν επίσημα σε πόλεμο, όπως το Πακιστάν ή η Υεμένη. Σύμφωνα με τα λεγόμενα του τότε προέδρου των ΗΠΑ, Μπαράκ Ομπάμα, η χρήση των μη επανδρωμένων αεροσκαφών στις προαναφερθείσες περιοχές γινόταν κατ' αποκλειστικότητα στα πλαίσια του «war against terror», αλλά και για σκοπούς αυτοάμυνας[7]. Σε αντίθεση με τα συνηθισμένα οπλικά συστήματα που χρησιμοποιούνται στα πεδία των μαχών, τα drones επιλέγονται για σκοπούς επιτήρησης ή για στοχευμένες δολοφονίες (targeting killings), μειώνοντας θεωρητικά με αυτόν τον τρόπο τις απώλειες αμάχων πληθυσμών, αλλά και τις παράπλευρες απώλειες στρατιωτών του ίδιου στρατοπέδου. Επιπλέον, μπορούν να θεωρηθούν συσκευές διαχείρισης ή ακόμα και απώλειας κινδύνου στη μάχη, εφόσον όντας όργανο μεγάλης εμβέλειας, έχουν τη δυνατότητα να εξαπολύσουν επίθεση μεγαλύτερης δυναμικής ελαχιστοποιώντας τα θύματα στα «απολύτως απαραίτητα», μειώνοντας παράλληλα τον κίνδυνο για τους ίδιους τους χειριστές αυξάνοντας την δυνατότητα επιβίωσής τους στη μάχη, κάτι το οποίο υπήρξε ζητούμενο για πολλές δεκαετίες πολέμων[8]. Πρόσθετα με αυτό, πρέπει να αναφερθεί ότι, όσον αφορά τα αυτόνομα οπλικά συστήματα, παρουσιάζουν θετικές πτυχές κατά τη χρήση τους οι οποίες δεν περνάνε απαρατήρητες στις εμπλεκόμενες δυνάμεις. Συγκεκριμένα, όσα drones έχουν σχεδιαστεί για στρατιωτική χρήση, δύνανται να ξεπεράσουν ανθρώπους, αλλά και συστήματα ελεγχόμενα από ανθρώπους στα πεδία των μαχών, ιδίως όσον αφορά την ταχύτητα, την ακρίβεια και την ικανότητα λειτουργίας χωρίς ανάπαυση. Ως απόρροια του γεγονότος αυτού είναι η εξοικονόμηση χρημάτων, υποδομών και ανθρώπινου δυναμικού[9] [10].

---

[6] M. J. Boyle, *ό.π,* 108.

[7] M. J. Boyle, *ό.π,* 106.

[8] Φυτιλής Βασίλειος, *ό,π,* 24.

[9] Arkin, R.C. (2010). The Case of Ethical Autonomy in Unmanned Systems, Journal of Military Ethics 9:4, 332-341.

[10] Müller, Vincent C. (2016), 'Autonomous killer robots are probably good news', in Ezio Di Nucci and Filippo Santoni de Sio (eds.), Drones and responsibility: Legal, philosophical and socio technical perspectives on the use of remotely controlled weapons (London: Ashgate), 67-81.

Στο σημείο αυτό ίσως να ήταν σωστό να θιγεί και η αρνητική έκβαση της τεχνολογίας των πολεμικών drones, ανεξαρτήτου σκοπού χρήσεως αυτών ή της ταυτότητας του χειριστή τους. Σε ό,τι αφορά τη χρήση τους, ένα μη επανδρωμένο αεροσκάφος τείνει να προάγει τη βία πολύ περισσότερο από ένα παραδοσιακό οπλικό σύστημα εφόσον ενέχει τη δολοφονία από μακριά (distancing from killing). Αναφέρεται πως η συγκεκριμένη μέθοδος δολοφονίας παρουσιάζει αρκετές ομοιότητες με τις μεθόδους τηλεχειριζόμενων παιχνιδιών όπως το "Playstation", γεγονός το οποίο σημαίνει πως κάνει τη δολοφονία «convenient», δηλαδή ότι οι περισσότεροι άνθρωποι θα πρόβαιναν σε αυτή καθώς δεν έρχονται σε άμεση επαφή με το θύμα[11]. Σε ατομικό επίπεδο και σε ό,τι αφορά τη στοχευμένη δολοφονία, θα πρέπει να ειπωθεί ότι πολλοί από τους χειριστές των drones εμφανίζουν πολύ υψηλά επίπεδα διαταραχής μετατραυματικού στρες (PTSD) καθόσον παρακολουθούν τον στόχο της επικείμενης δολοφονίας για μεγάλα χρονικά διαστήματα, ακόμη και σε πιο προσωπικές στιγμές. Αυτό έχει ως αποτέλεσμα να εκδηλώνουν μία εξοικείωση με τον στόχο περνώντας η σύνδεσή τους σε προσωπικό-πλέον-επίπεδο[12]. Τέλος στο πολιτειακό επίπεδο, τα μη επανδρωμένα αεροσκάφη, κυρίως όσα χρησιμοποιούνται για επιτήρηση μεγάλου πληθυσμού, αν και δεν επίπτουν στην κατηγορία των οπλισμένων drones (Unmanned Aerial Combat Vehicles UCAV), φαίνεται, ωστόσο, να βλάπτουν τη δημοκρατία μια και οι απεργίες, τουλάχιστον στην περίπτωση των Ηνωμένων Πολιτειών, ελέγχονται από το εκτελεστικό τμήμα της κυβέρνησης υπό την εποπτεία του Κογκρέσου. Άρα, όπως προκύπτει λογικά, η συλλογή δεδομένων μέσω των UAV σημαίνει είτε άμεσα είτε έμμεσα το τέλος του κατοχυρωμένου δικαιώματος στην απεργία[13].

Ταυτοχρόνως, θέτει ένα ζήτημα ιδιωτικότητας- επιτήρησης. Συγκεκριμένα, το ζήτημα της παραβίασης της ιδιωτικότητας δεν

[11] Thomas M. Philip, Ayush Gupta, Andrew Elby & Chandra Turpen, ''*Why Ideology Matters for Learning: A Case of Ideological Convergence in an Engineering Ethics Classroom Discussion on Drone Warfare''* , Journal of the Learning Sciences, 19:16 (11 October 2017): 17.

[12] M. J. Boyle, *ό.π,* 106-107.

[13] Thomas M. Philip, Ayush Gupta, Andrew Elby & Chandra Turpen, *ό.π.,* 13-14.

τίθεται μόνο σε σχέση με τα UAV που επιτηρούν μεγάλους πληθυσμούς, αλλά και σε σχέση με τα UCAV, καθώς συχνά αυτά επιστρατεύουν ειδικούς αισθητήρες βιομετρικών δεδομένων (λόγου χάριν, αναγνώρισης προσώπου ή ιδιαίτερων χαρακτηριστικών) κατά τη διάρκεια επιχειρήσεων. Η ηθική επιληψιμότητα του ζητήματος αυτού, έχει απασχολήσει σε μεγάλο βαθμό και την ευρωπαϊκή ήπειρο, αλλά και την Ιαπωνία, εφόσον η ενσωμάτωση των drones για την επιτήρηση μεγάλου μέρους του πληθυσμού εγείρει ανησυχίες σχετικά με την παραβίαση της ιδιωτικής ζωής των ατόμων και στις δύο περιπτώσεις. Για παράδειγμα, στην περίπτωση των «Αεροπορικών Συντριβών των Αερογραμμών στην Ουκρανία MH17» η χρήση των μη επανδρωμένων αεροσκαφών για συλλογή δεδομένων και πληροφοριών, παραβίασαν το δικαίωμα στην ιδιωτική ζωή που προστατεύεται από το άρθρο 8 της Ευρωπαϊκής Σύμβασης Δικαιωμάτων του Ανθρώπου (ΕΣΔΑ)[14] [15]. Αντίστοιχα ζητήματα περί της παραβίασης της ιδιωτικότητας αντιμετωπίζει και η Ιαπωνία η οποία το 2015 ψήφισε νόμο σύμφωνα με τον οποίο οι ιδιοκτήτες των drones θα πρέπει να καταχωρούν τις συσκευές τους και να λαμβάνουν άδεια από τις αρχές για οποιαδήποτε χρήση αυτών (¨On special measures to ensure the safety of drones¨). Παρόλα αυτά εντοπίστηκαν drones σε περιοχές στις οποίες απαγορευόταν η χρήση τους, ήτοι αεροδρόμια ή στρατιωτικές βάσεις. Για τον λόγο αυτό το 2018, συστάθηκε μια ομάδα εργασίας από την ιαπωνική κυβέρνηση η οποία περιελάμβανε εμπειρογνώμονες, μέλη της ακαδημαϊκής κοινότητας, αλλά και κατασκευαστές drones με σκοπό την αντιμετώπιση ζητημάτων τέτοιου ύφους σε μια προσπάθεια για αυστηρότητα του πλαισίου χρήσης τους[16].

[14] Kutynska ,Anastasiia, Dei , Maryna (2023). "*Legal Regulation of the Use of Drones: Human Rights and Privacy Challenges"* ,Journal of International Legal Communication, 8(1), 41.

[15] Szira, Zoltán, Varga, Erika, László Csegodi, Tibor, Milics Gábor, "*THE DEVELOPMENT OF DRONE TECHOLOGY AND ITS REGULATION IN THE EUROPEAN UNION",* 10.2478/eual-2023-0005, 35-39.

[16] Kutynska ,Anastasiia, Dei , Maryna, *ό.π.*, 49.

AUMF και η απειλή της τρομοκρατίας

Το ζήτημα της νομικής εξουσίας

Όπως αναφέρθηκε και προηγουμένως, οι πρωτεργάτες της γενικευμένης χρήσης των drones ήταν οι Ηνωμένες Πολιτείες της Αμερικής με το επιχείρημα της αυτοάμυνας και της εξάλειψης της τρομοκρατίας. Ως εκ τούτου, ο τότε Πρόεδρος, Bush junior, με σκοπό να αποδείξει τη νομική υπόστασή τους, έκανε λόγο για την εξουσιοδότηση των ΗΠΑ στη χρήση της στρατιωτικής δύναμης (Authorization to Use Military Force). Ο νόμος αυτός πέρασε την έγκριση του Κογκρέσου και υπογράφηκε επίσημα λίγες μέρες αργότερα[17]. Σύμφωνα με τον πρόεδρο Ομπάμα, ο όρος της AUMF αναφέρεται στην «άδεια χρήσης όλης της απαραίτητης και κατάλληλης βίας εναντίων εκείνων των εθνών, οργανισμών ή προσώπων που κρίνει ότι σχεδίασαν, εξουσιοδότησαν, διέπραξαν ή βοήθησαν τις τρομοκρατικές επιθέσεις που συνέβησαν στις 11 Σεπτεμβρίου 2001»[18]. Μία τέτοια πρωτοβουλία φαίνεται αρκετά δυναμική και προστατευτική προς τους πολίτες της, καθώς ο κίνδυνος της τρομοκρατίας δεν μπορεί να γίνει άμεσα ορατός γι' αυτό και χρειάζονται μέτρα με σκοπό την αποφυγή παρόμοιων συμβάντων αλλά και την απώλεια ανθρώπινων ζωών. Παρόλα αυτά, το επιχείρημα υπέρ της αρχής της AUMF ενέχει πολλούς κινδύνους και απροσδιοριστίες οι οποίες πρέπει να τεθούν επί τάπητος, ειδικά όταν γίνεται λόγος για μία αρχή που αφορά τόσες ανθρώπινες ζωές.

Οι τρεις ανησυχίες για την ισχύ της AUMF

Η πρώτη ανησυχία αφορά την έλλειψη χρονικού πεδίου. Εκείνο το οποίο δεν έχει διασαφηνιστεί επαρκώς στον προαναφερθέντα ορισμό είναι το μέχρι πότε οι ΗΠΑ θα μπορούν να διαθέτουν αυτή την εξουσιοδότηση. Η έννοια της τρομοκρατίας είναι μία έννοια γενικής φύσεως χωρίς σαφή χρονικά όρια, εφόσον κανείς δεν μπορεί να υπολογίσει ή να γνωρίσει το πότε θα γίνει η τελευταία τρομοκρατική επίθεση, γεγονός που επιτρέπει στην Αμερική την επ' αόριστον χρήση του όρου αυτού. Η δεύτερη ανησυχία επαφίεται στην έλλειψη των γεωγραφικών ορίων. Με γνώμονα την εξάλειψη της τρομοκρατίας, οι ΗΠΑ έχουν το δικαίωμα να παρέμβουν σε

[17] https://www.congress.gov/107/plaws/publ40/PLAW-107publ40.pdf
[18] M. J. Boyle, *ό.π.,* 108.

οποιοδήποτε κράτος για στρατιωτική δράση, ακόμη και για προληπτικούς λόγους εάν αντιληφθούν οποιαδήποτε ύποπτη κίνηση σχετιζόμενη με την τρομοκρατική δράση της Αλ Κάιντα. Λαμβάνοντας υπόψιν το γεγονός ότι θυγατρικές τρομοκρατικές κινήσεις έχουν καταγραφεί επίσημα σε έως και περισσότερες από 30 χώρες, γίνεται εύκολα κατανοητό πως κάτι τέτοιο συνεπάγεται την μετατροπή της Αμερικής σε «big brother», καθώς τα μη επανδρωμένα αεροσκάφη θα μπορούν να χρησιμοποιούνται για επιτήρηση οπουδήποτε η Αμερική ορίζει χωρίς κανέναν επιπλέον περιορισμό[19]. Βεβαίως, παρόμοια χρήση των UCAV, δηλαδή χωρίς σαφείς χρονικούς και χωρικούς περιορισμούς, γίνεται και στην περίπτωση του πολέμου μεταξύ Ουκρανίας και Ρωσίας, κατά τη διάρκεια του οποίου γίνεται χρήση τέτοιων οπλικών συστημάτων. Τα συστήματα αυτά επιτρέπουν από τη μία τη μείωση του εμπλεκόμενου αριθμού στρατιωτικών στο πεδίο της μάχης, όμως, από την άλλη, αυξάνουν την παροχή πληροφοριών (παρεχόμενες συντεταγμένες, εχθρικές ή ύποπτες συναντήσεις), με σκοπό την αποφυγή πρόκλησης απωλειών σε άμαχο πληθυσμό, όσο αυτό καθίσταται εφικτό[20].

Τέλος, ο τρίτος φόβος που προκύπτει είναι σχετικός με τα ιδιωτικά κριτήρια για τον προσδιορισμό της ικανότητας αλλά και της προθυμίας μίας κυβερνήσεως στην αντιμετώπιση τέτοιου είδους απειλών. Με άλλα λόγια, σε περίπτωση μεγάλου κινδύνου εξωτερικά κράτη μπορούν να αποφασίσουν πότε ένα κυρίαρχο κράτος χάνει το δικαίωμα της μη παρέμβασης στο έδαφός του , μία αρχή γνωστή ως «Ευθύνη προστασίας», «Responsibility to Protect»- R2P). Η αρχή της Ευθύνης Προστασίας υποστηρίχτηκε από πολλές ακτιβιστικές οργανώσεις, με το σκεπτικό ότι η συστηματική καταπάτηση των ανθρωπίνων δικαιωμάτων του πληθυσμού ενός κράτους από το ίδιο το κράτος απαιτεί άμεση παρέμβαση και μέτρα. Αν, λοιπόν, μια κυβέρνηση δεν δύναται να προστατέψει τους πολίτες της, τότε δικαιούνται άλλοι παράγοντες να παρέμβουν και να το πράξουν, μόνο με την έγκριση του ΟΗΕ ή οποιουδήποτε άλλου νόμιμου οργάνου[21]. Συναφής με αυτό είναι και η αρχή jus ad bellum η οποία αφορά τη δικαιοσύνη του πολέμου και λαμβάνεται υπόψιν κατά τη διάρκεια της ένοπλης σύγκρουσης. Εν ολίγοις αναφέρεται, δηλαδή, στην αιτιολόγηση της προσφυγής στον πόλεμο

[19] M. J. Boyle, *ό.π,* 109-110.
[20] G. Minculete and V. Păstae No.4/2023 (vol. 12) https://doi.org/10.53477/2284-9378-23-58.
[21] Στο ίδιο, 111.

εξ αρχής[22]. Η κυβέρνηση των ΗΠΑ υποστηρίζοντας πως το επιχείρημα αυτό έχει καθαρά ανθρωπιστικούς σκοπούς, αφού έχει κατασκευαστεί για να καταλαγιάσει τον φόβο που ακολουθεί κάθε νύξη τρομοκρατίας, εκείνο το οποίο πετυχαίνει επί της ουσίας είναι να δημιουργήσει ένα αμερικανικό πρόγραμμα drone σε ακυβέρνητους χώρους σε παγκόσμια κλίμακα[23]. Συμπερασματικά, τα τρία αυτά αντεπιχειρήματα για την αρχή της AUMF αποτελούν τον λόγο για τον οποίο μία τέτοια πρωτοβουλία ίσως πρέπει να αντιμετωπιστεί με μία κριτική στάση πολύ περισσότερο όταν επρόκειτο για ολόκληρα κράτη, πόσο μάλλον όταν εμπλέκονται και οπλικά συστήματα Τεχνητής Νοημοσύνης τα οποία τείνουν να οξύνουν τα ζητήματα αυτά ακόμη περισσότερο εξαιτίας της ασαφούς φύσης τους και των προβληματισμών που εγείρουν.

Το δίκαιο του Πολέμου

Στο σημείο αυτό ίσως να πρέπει να γίνει μία σύντομη αναφορά στις βασικές αρχές του Δικαίου του Πολέμου (International Humanitarian Law). Το ζήτημα της δικαιοσύνης πριν ή κατά τη διεξαγωγή του πολέμου, άρχισε να απασχολεί ενεργά τους Διεθνείς Ανθρωπιστικούς Οργανισμούς από τη λήξη του Β' Παγκοσμίου Πολέμου κι έπειτα, θέλοντας έτσι, να περιοριστούν οι ακρότητες και η καταπάτηση ανθρώπινων δικαιωμάτων κατά τη διάρκεια των συγκρούσεων. Στόχος ήταν να κατοχυρωθούν τα δικαιώματα των πολιτών και να γίνει μία προσπάθεια εξανθρωπισμού στον υπό κατοχή πληθυσμό[24] Η θεωρία αυτή χωρίζεται σε τρεις βασικές αρχές: το jus ad bellum, jus in bello και jus post bellum, με τις δύο πρώτες, jus ad bellum και jus in bello, να ενέχουν μία πιο αιχμηρή ηθική υπόσταση. Ο πρώτος όρος, το jus ad bellum, αφορά την αιτιολόγηση της προσφυγής στον πόλεμο εξ αρχής, δηλαδή εξετάζεται αν ο πόλεμος διεξήχθη δίκαια ή άδικα. Ο επόμενος όρος, το jus in bello, είναι η αρχή της θεωρίας του δικαίου πολέμου και ασχολείται με την δικαιολόγηση κατά τη διεξαγωγή του πολέμου, με τις δύο βασικές πτυχές που το απασχολούν να είναι η διάκριση μεταξύ μαχητών και αμάχων, αλλά και το κατά πόσο η δύναμη είναι

[22] Estrella E, Iceal Averroes , *On the Ethics of War,* KRITIKE VOLUME SIX NUMBER ONE (JUNE 2012) 67-84, http://www.kritike.org/journal/issue_11/estrella_june2012.pdf .

[23] Στο ίδιο.

[24] Khan, Gawhar Ahmad, *Protection of Civilians in War: The Role of International Humanitarian Law* , Vol. 3 No. 3 (2024) pp. 204-214.

ανάλογη με τον στόχο που θέλει να πετύχει. Τέλος, το jus post bellum, είναι μία νέα προσθήκη στην κλασική θεωρία του πολέμου και ασχολείται με την αιτιολόγηση του τερματισμού αυτού.[25] [26]

Περί ηθικής/νομικής επιληψιμότητας

Στρατιωτικό ή νομικό δίκαιο;

Στα πλαίσια του δικαίου του πολέμου, η 4η Σύμβαση της Γενεύης ορίζει εκτενώς τα βασικά δικαιώματα των αιχμαλώτων κατά τη διάρκεια του πολέμου, τις θεμελιώδεις προστασίες για τους τραυματίες εξαιτίας αυτού αλλά και ό,τι αφορά τους πολίτες μέσα ή ακόμη και γύρω από την πολεμική ζώνη[27]. Η προαναφερθείσα Σύμβαση συστάθηκε έπειτα από το τέλος του Β' Παγκοσμίου Πολέμου με σκοπό την προστασία των θεμελιωδών ανθρωπίνων δικαιωμάτων σε κατάσταση ένοπλης σύγκρουσης. Έκτοτε το κλασικό μέτωπο μάχης στη διεξαγωγή του πολέμου έχει αλλάξει κατά πολύ με νέες τεχνολογίες, όπως η Τεχνητή Νοημοσύνη, να έχουν έρθει στο προσκήνιο. Παρόλα αυτά, είναι σαφές πως δεν καθίσταται πιθανό να εμπλουτίζεται με νέες προσθήκες συνεχώς η Σύμβαση ακολουθώντας τον ρυθμό της τεχνολογίας. Στο σημείο αυτό, όμως, προκύπτει το ζήτημα της εξατομικευμένης ευθύνης κατά τη χρήση μίας νέας τεχνολογίας, όπως αυτή των drones, εφόσον το διεθνές δίκαιο δεν περιέχει κάποια εξειδικευμένη διάταξη. Επομένως, έρχεται στο προσκήνιο ξανά το ερώτημα που τέθηκε και στην αρχή: Ποιος έχει την ευθύνη όταν πρόκειται για τα τηλεχειριζόμενα drones, εφόσον το διεθνές δίκαιο παραλείπει τη συγκεκριμένη υποπερίπτωση στις διατάξεις του;

Για να δοθεί απάντηση στο προηγούμενο ερώτημα πρέπει να εξεταστεί σε ποιο είδος δικαίου εμπίπτει. Η εκάστοτε στρατιωτική δύναμη πρέπει να συνδέεται με εξατομικευμένες κρίσεις ευθύνης πάνω στο πεδίο της μάχης. Το πρόβλημα, ωστόσο, βρίσκεται στο ότι το στρατιωτικό δίκαιο είναι αυτόνομο από το νομικό όταν έρχεται αντιμέτωπο με τρομοκράτες ή αντάρτες, γενικότερα με καταστάσεις έκτακτης ανάγκης. Παρόλα αυτά, το

[25] Estrella E, Iceal Averroes, *ό.π.*, 72-77.

[26] Stahn, Carsten, '*Jus ad bellum', 'jus in bello' . . . 'jus post bellum'? Rethinking the Conception of the Law of Armed Force,* The European Journal of International Law Vol. 17 no.5 © EJIL 2007, 922-940.

[27] https://ihl-databases.icrc.org/en/ihl-treaties/gciv-1949

όργανο που έχει σχεδιαστεί αλλά και καταρτισθεί παραδοσιακά για αυτή τη λειτουργία είναι το δικαστικό. Από την άλλη, το στρατιωτικό δίκαιο δεν μπορεί να είναι τελείως ανεξάρτητο από τη νομική εξουσία. Οι στρατιωτικές δυνάμεις ενός κράτους πρέπει να προσαρμοστούν στις εσωτερικές ή εξωτερικές πιέσεις. Αυτοί οι προβληματισμοί καταλήγουν στον φαύλο κύκλο εξατομικευμένη ευθύνη-στρατιωτικό δίκαιο- νομικό δίκαιο και αντίστροφα. Ακολουθώντας αυτού του είδους τον συλλογισμό ερχόμαστε αντιμέτωποι με το παράδοξο της εξατομίκευσης: Αν η κυβέρνηση κάνει δικαστικές κρίσεις ατομικής ευθύνης πριν τη χρήση στρατιωτικής δύναμης, θα πρέπει να απαιτούνται οι πιο παραδοσιακοί θεσμοί και διαδικασίες μέσω των οποίων παρόμοιες αποδόσεις του ατόμου κατονομάζουν τον εγκληματία, άρα ο εγκληματίας θα είναι γνωστός στις δικαστικές αρχές και δεν θα χρειαστεί η χρήση στρατιωτικής δύναμης[28]. Εν ολίγοις, όταν γίνεται λόγος για στοχευμένες δολοφονίες μέσω μη επανδρωμένων οχημάτων η ευθύνη δεν είναι δυνατόν να εναπόκειται μόνο σε εκείνον ο οποίος εκτελεί τηλεχειριζόμενα την εντολή που έχει, νωρίτερα η ευθύνη περνάει από τους κανόνες του στρατιωτικού δικαίου, όπως εκείνοι έχουν οριστεί με γνώμονα το ισχύον νομικό πλαίσιο του κάθε κράτους.

Ηθικά Διλήμματα

Όπως έχει προκύψει από την έως τώρα συζήτηση γύρω από την τεχνολογία των drones, γίνεται αντιληπτό πως πολλές από τις παραμέτρους χρήσης αυτών είναι ηθικά προβληματικές[29]. Παρακάτω θα γίνει μία προσέγγιση των βασικών εξ αυτών με σκοπό την προσπάθεια διευκρίνισής τους. Ξεκινώντας από το τελευταίο επιχείρημα και σε ό,τι αφορά το ζήτημα της ευθύνης, αυτό εκκινεί από την προστασία της ανθρώπινης ζωής, λαμβάνοντας υπόψιν σε ένα κράτος όλες οι νομικές αποφάσεις σε καιρό ειρήνης ή και πολέμου. Παρόλα αυτά , οι αποφάσεις αυτές αφορούν ένα συγκεκριμένο πλαίσιο διαχείρισης καταστάσεων και δεν είναι δυνατόν να συμπεριλάβουν όλες τις πιθανές υποπεριπτώσεις. Προηγούμενα, έγινε αναφορά στην 4η Σύμβαση της Γενεύης και στο γεγονός ότι από το τέλος του Β' Παγκοσμίου Πολέμου δεν έχει γίνει καμία προσπάθεια ανανέωσής της με σκοπό την συμπόρευσή της με την ραγδαία εξέλιξη της τεχνολογίας.

Κατ' αναλογία και το εκάστοτε κράτος δεν μπορεί να περικλείσει κάθε

[28] Samuel Issacharoff and Richard H. Pildes, '*Drones and the Dilemma of Modern Warfare',* NEW YORK UNIVERSITY SCHOOL OF LAW, (NO. 13-34), (June 2013), 8.
[29] Müller, Vincent C. (2016), *ό,π.*

υποπερίπτωση κατά τη δημιουργία του στρατιωτικού ή του νομικού δικαίου, ούτε και να βρίσκεται συνεχώς σε μία διαδικασία ανανέωσης των νόμων αυτών. Όταν μία στρατιωτική επιχείρηση ξεκινά, είναι βέβαιο πως ακολουθείται μία σειρά από συγκεκριμένες διαδικασίες. Όσο, όμως, τα στάδια της επιχείρησης προχωρούν και με δεδομένο ότι η κάθε περίπτωση είναι μοναδική, η ευθύνη εξατομικεύεται. Συγκεκριμένα, στην περίπτωση όπου ο χειριστής ενός μη επανδρωμένου αεροσκάφους έχει την εντολή να εκτελέσει μία στοχευμένη δολοφονία, είναι σίγουρο ότι έχει προηγηθεί επιτήρηση του θύματος για μεγάλο χρονικό διάστημα κι έχει ακολουθηθεί μία σειρά διαδικασιών για να ελαχιστοποιηθεί το ποσοστό του λάθους. Η τελική πράξη επαφίεται, ωστόσο, στον χειριστή του συστήματος. Από εκείνον, λοιπόν, εξαρτάται εάν θα εκτελέσει μία εξ αποστάσεως δολοφονία και αυτό είναι το στάδιο στο οποίο η ευθύνη εξατομικεύεται πλήρως. Τέλος, στις περιπτώσεις κατά τις οποίες οι χειριστές ακολούθησαν πιστά τις οδηγίες, αποποιούμενοι έτσι την ευθύνη, πολλοί από αυτούς εμφάνισαν μετέπειτα ψυχικές διαταραχές[30].Αντιθέτως, χειριστές οι οποίοι δεν εκπλήρωσαν την υποχρέωσή τους για δολοφονία, τιμωρήθηκαν αυστηρά. Ένα πρόβλημα τέτοιου είδους τείνει να πάρει τη μορφή διλημμάτων του τρόλεϊ, «trolley dilemma», τα οποία καθιστούν τα πιο συχνά ηθικά διλήμματα με τα οποία ερχόμαστε αντιμέτωποι. Η σύγκριση των διλημμάτων αυτών με το δίλημμα του τρόλεϊ έγκειται στο γεγονός ότι και οι δύο συνθήκες αφορούν τη θυσία ενός ή μιας ομάδας ανθρώπων χάριν της επιβίωσης πολλών.

Ένα ακόμη ηθικό ζήτημα για το οποίο πρέπει να γίνει λόγος είναι οι ίδιοι οι στόχοι των drones. Το βασικό πρόβλημα στις στοχευμένες δολοφονίες ατόμων που εμπλέκονται σε πιθανές τρομοκρατικές οργανώσεις, είναι το ότι δεν ξεχωρίζουν από τους απλούς πολίτες αφού δεν φορούν τη συνηθισμένη στολή μάχης. Αυτό καθιστά σαφές πως οι πιθανότητες λάθους αυξάνονται κατά πολύ όταν πρόκειται για στόχους που δύσκολα μπορούν να ξεχωρίσουν από το πλήθος, γεγονός κατά το οποίο παραβιάζεται η αρχή της διάκρισης στο jus in bello[31]. Ο συνδυασμός αυτού αλλά και το ότι οι περισσότεροι από τους σκοπούς χρήσης των drones δεν είναι διαφανείς από τις ΗΠΑ, συνιστούν την προσοχή για το πόσο αντικειμενικά καθίστανται τα κριτήρια σύμφωνα με τα οποία δίνονται εντολές για επιτήρηση ή ακόμη και τη δολοφονία συγκεκριμένων ατόμων[32] [33]. Πρόσθετα με

[30] M. J. Boyle, *ό.π,* 106.
[31] Stahn, Carsten, *ό.π.*
[32] Kutynska ,Anastasiia, Dei , Maryna (2023), *ό,π.*
[33] Szira, Zoltán, Varga, Erika, László Csegodi, Tibor, Milics Gábor, *ό,π.*

αυτό, θα πρέπει να αναφερθεί ότι έχει επισημανθεί πως σε ό,τι αφορά τις αμερικάνικες επιχειρήσεις των μη επανδρωμένων αεροσκαφών, στην κατηγορία 'πολίτες' πολύ πιο πρόθυμα εντάσσουν άτομα λευκού χρώματος, αυτόχθονες Αμερικανούς [34]. Αυτός ο προβληματισμός αναφέρεται στο γενικότερο ζήτημα του Al bias το οποίο καθίσταται ένα γενικότερο πρόβλημα της Τεχνητής Νοημοσύνης. Αναλυτικά, οι αλγόριθμοι που βασίζονται σε δεδομένα χρησιμοποιούνται ευρέως για να δημιουργήσουν ή να βοηθήσουν αποφάσεις σε ευαίσθητους τομείς, συμπεριλαμβανομένης της υγειονομικής περίθαλψης, της κοινωνικής πρόνοιας υπηρεσίες, εκπαίδευση, προσλήψεις και ποινική δικαιοσύνη[35]. Σε διάφορες περιπτώσεις, τέτοιοι αλγόριθμοι έχουν διατηρήσει ή ακόμη και επιδεινώσει προκαταλήψεις εναντίον ευάλωτων κοινοτήτων, πυροδοτώντας σε ακόμη μεγαλύτερο βαθμό προ υπάρχουσες προκαταλήψεις. Το ζήτημα αυτό παρατηρήθηκε εντονότερα στα αρχικά στάδια της Τεχνητής Νοημοσύνης και, πλέον, παρατηρείται άμβλυνση του φαινομένου, χωρίς αυτό να σημαίνει πως δεν υπάρχουν περιθώρια βελτίωσης[36]. Με γνώμονα το δεδομένο αυτό, όσοι , από την άλλη, είναι ερυθρόδερμοι είναι ύποπτοι για υποτιθέμενες τρομοκρατικές ενέργειες. Εν ολίγοις, σε ό,τι αφορά τους στόχους των drones το ότι δημοσιοποιούνται ελάχιστα δεδομένα των Αμερικανικών βάσεων τα οποία σχετίζονται με τα κριτήρια επιλογής αυτών , σε συνδυασμό με τη δυσκολία που υπάρχει όταν οι στόχοι εντάσσονται στο γενικότερο πλαίσιο της κατηγορίας των «πολιτών», εξάγεται το συμπέρασμα ότι το κράτος της Αμερικής έχει το ελεύθερο να χειρίζεται τις καταστάσεις χωρίς ηθικούς ή και νομικούς περιορισμούς, θέτοντας, με αυτόν τον τρόπο, το ζήτημα της παραβίασης της ιδιωτικότητας και τελικά ένα ζήτημα επιτήρησης, όπως αυτό αναφέρθηκε προηγουμένως[37].

Συμπεράσματα

Εν κατακλείδι, τα μη επανδρωμένα αεροσκάφη συνιστούν άλλο ένα επίτευγμα της τεχνητής νοημοσύνης. Όπως όλα τα επιτεύγματα του κλάδου

[34] Thomas M. Philip, Ayush Gupta, Andrew Elby & Chandra Turpen, *ό.π.,* 19.
[35] Fazelpour, Sina , Danks ,David , *Algorithmic bias: Senses, sources, solutions,* DOI: 10.1111/phc3.1276*0.*
[36] Lopez Paola , *Bias does not equal bias: a socio-technical typology of bias in data-based algorithmic systems,* Internet Policy Review 10(4) | 2021, 5-29.
[37] Kutynska ,Anastasiia, Dei , Maryna (2023), *ό,π.*

έτσι κι αυτό δεν μπορεί να χαρακτηριστεί από μόνο του θεμιτό ή μη. Η τεχνητή νοημοσύνη έχει εισβάλλει για τα καλά στη ζωή μας τις τελευταίες δεκαετίες κι έχει γνωρίσει ραγδαία ανάπτυξη με στόχο τη διευκόλυνση πολλών καταστάσεων της καθημερινότητας και μη. Τα πολεμικά drones έχουν αλλάξει την έκβαση των πολεμικών επιχειρήσεων μειώνοντας τα θύματα και προστατεύοντας τη ζωή των στρατιωτών. Όμως, από την άλλη μεριά, όσον αφορά τις στοχευμένες δολοφονίες ανθρώπων τείνουν να υποβαθμίζουν την αξία της ζωής, εφόσον με ένα μόλις κουμπί βάζουν τέλος σε αυτή, όπως ακριβώς και στα ηλεκτρονικά παιχνίδια. Τι γίνεται, όμως, σε περίπτωση λάθους; Στην περίπτωση αυτή ένας αθώος έχει χάσει τη ζωή του επειδή κάποιος άλλος εξ αποστάσεως πήρε την απόφαση να την τερματίσει με το πάτημα ενός κουμπιού. Η τεχνολογία μπορεί στα αλήθεια να χρησιμοποιηθεί είτε για να κάνει το καλό, είτε το κακό. Το πώς θα επιλέξει ο καθένας να τη χρησιμοποιήσει δεν εξαρτάται μόνο από τούς νόμους και τους κανόνες, αλλά εξαρτάται άμεσα και από τις αξίες που έχει ο άνθρωπος στο χέρι του οποίου λειτουργεί ως όπλο την εκάστοτε στιγμή.

Βιβλιογραφία

Φυτιλής, Βασίλειος, ‘’*Drone και Δημοκρατικός Πόλεμος’’,* Διπλωματική εργασία, Πανεπιστήμιο Πειραιώς , Αθήνα, Αύγουστος 2022.

Arkin, R.C. (2010). The Case of Ethical Autonomy in Unmanned Systems, Journal of Military Ethics 9:4.

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

Müller, Vincent C. (2016), ‘Autonomous killer robots are probably good news’, in Ezio Di Nucci and Filippo Santoni de Sio (eds.), Drones and responsibility: Legal, philosophical and socio technical perspectives on the use of remotely controlled weapons (London: Ashgate).

Kutynska ,Anastasiia, Dei , Maryna (2023). “*Legal Regulation of the Use of Drones: Human Rights and Privacy Challenges”*, Journal of International Legal Communication, 8(1).

Stahn, Carsten, ‘*Jus ad bellum’, ‘jus in bello’ . . . ‘jus post bellum’? Rethinking the Conception of the Law of Armed Force,* The

European Journal of International Law Vol. 17 no.5 © EJIL 2007.
Khan, Gawhar Ahmad, *Protection of Civilians in War: The Role of International Humanitarian Law* , Vol. 3 No. 3 (2024).
Szira, Zoltán, Varga, Erika, László Csegodi, Tibor, Milics Gábor, "*THE DEVELOPMENT OF DRONE TECHOLOGY AND ITS REGULATION IN THE EUROPEAN UNION",* 10.2478/eual-2023-0005.
Thomas M. Philip, Ayush Gupta, Andrew Elby & Chandra Turpen, ''*Why Ideology Matters for Learning: A Case of Ideological Convergence in an Engineering Ethics Classroom Discussion on Drone Warfare''* , Journal of the Learning Sciences, 19:16 (11 October 2017).
Koukoudakis, George, "*Drones' contribution to the transformation of contemporary warfare",* J. Military Stud. 2024; 13(1).
Samuel Issacharoff and Richard H. Pildes, '*Drones and the Dilemma of Modern Warfare',* NEW YORK UNIVERSITY SCHOOL OF LAW, (NO. 13-34), (June 2013).
M. J. Boyle, ''*The legal and ethical implications of drone warfare'',* The International Journal of Human Rights*,* 19:2 , 24 Feb 2015.
Samuel Issacharoff and Richard H. Pildes, '*Drones and the Dilemma of Modern Warfare',* NEW YORK UNIVERSITY SCHOOL OF LAW, (NO. 13-34), June 2013.
Lopez Paola , *Bias does not equal bias: a socio-technical typology of bias in data-based algorithmic systems,* Internet Policy Review 10(4) | 2021.

Δικτυογραφία

https://www.congress.gov/107/plaws/publ40/PLAW-107publ40.pdf
https://ihl-databases.icrc.org/en/ihl-treaties/gciv-1949

Περίληψη

Είναι γεγονός πως τα τελευταία χρόνια, κυρίως τις τελευταίες δεκαετίες, οι τεχνολογικές εξελίξεις επιτρέπουν τη δημιουργία ολοένα και ισχυρότερων οπλικών συστημάτων. Τέτοια συστήματα είναι και τα drones ή UAV(Unmanned aerial vehicle), δηλαδή τα μη επανδρωμένα αεροσκάφη,

αλλά και τα οπλισμένα drones τα οποία αποκαλούμε UCAV (Unmanned Aerial Combat Vehicles), ένα ακόμα γέννημα της τεχνητής νοημοσύνης. Η μεγάλη επανάσταση σε αυτό το σημείο έγκειται στο γεγονός ότι τα αεροσκάφη αυτά έχουν τη δυνατότητα να δρουν ανεξάρτητα, χωρίς φαινομενικά κάποιον "απτό" ανθρώπινο χειρισμό. Σε αυτή δε την περίπτωση μιλάμε για Φονικά Αυτόματα Οπλικά Συστήματα (Lethal Autonomous Weapon Systems - LAWS). Η αυτονομία που τα διακρίνει τα καθιστά ακόμα πιο επικίνδυνα, αν λάβει κανείς υπόψη ότι, σε πολλές περιπτώσεις οι χειριστές των εν λόγω οπλικών συστημάτων παραμένουν άγνωστοι, όπως ακριβώς και η αιτία χρησιμοποίησής τους. Παράλληλα, η περίπτωση των εντελώς αυτόνομων drones, τα οποία βρίσκονται ακόμη σε στάδιο ανάπτυξης είναι εκείνη που χαρακτηρίζεται ως πιο αιχμηρή από ηθικής απόψεως. Καθίσταται, επομένως, λογικό τόσο η χρήση τους, όσο και ο σκοπός αυτής να εγείρουν ερωτήματα ηθικής φύσεως. Ανάγοντας τον προβληματισμό αυτό στο σήμερα, με την κατάσταση στην ευρωπαϊκή ήπειρο και τη Μέση Ανατολή να μην είναι η καλύτερη δυνατή, ίσως να πρέπει να αντιμετωπίσουμε μια τέτοια καινοτομία με μία κριτική θεώρηση, πολύ περισσότερο από ό,τι πριν. Το παρόν δοκίμιο έχει ως στόχο την προσέγγιση και ανάλυση της τεχνολογίας των μη επανδρωμένων αεροσκαφών στο πεδίο των συγκρούσεων σε περιοχές οι οποίες είτε είναι σε επίσημη εμπόλεμη κατάσταση, είτε ανεπίσημα γίνεται χρήση της συγκεκριμένης τεχνολογίας με σκοπό την αποφυγή τρομοκρατικών επιθέσεων. Στην προσπάθεια προσέγγισης του θέματος που θα ακολουθήσει, θα γίνει αναφορά στις χρήσεις και τους σκοπούς της συγκεκριμένης τεχνολογίας, στους κανόνες Διεθνούς Δικαίου του πολέμου, όπως αυτοί διαμορφώθηκαν μέσα από την Τέταρτη Σύμβαση της Γενεύης, αλλά και στις ηθικής φύσεως διαφορές των σύγχρονων όπλων στις ένοπλες συγκρούσεις σε σχέση με εκείνες του παρελθόντος. Επιπλέον, θα επιχειρηθεί να δοθεί μία απάντηση στα προηγούμενα ηθικά-κυρίως- αλλά και ανθρωπιστικά ζητήματα που προκύπτουν σχετικά με το πλαίσιο δημιουργίας και χρήσης του τεχνολογικού αυτού επιτεύγματος.

Λέξεις Κλειδιά: πολεμικά drone, ηθική, AUMF, στοχευμένη δολοφονία, τρομοκρατία, εξατομικευμένη ευθύνη, στρατιωτικό δίκαιο.

Keywords: combat drones, ethics, AUMF (Authorization for Use of Military Force), targeted killing, terrorism, individual accountability, military law.

Ελευθερία Πάτσαρη
Τμήμα Φιλοσοφίας, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών
Email: eleftheriapatsari@gmail.com
ORCID iD: 0000-0002-4954-0134

Βασίλειος Πολυχρονιάδης

Η Τέχνη της Παραγωγικής Τεχνητής Νοημοσύνης: Από την Πρόθεση στην Ερμηνεία. Φιλοσοφικές και Πολιτισμικές Συνέπειες

doi:https://doi.org/10.12681/plogos.35835

Ι. Εισαγωγή

Η ΣΧΕΣΗ ΤΗΣ ΤΕΧΝΗΣ ΜΕ ΤΗΝ ΤΕΧΝΟΛΟΓΙΑ ΥΠΗΡΞΕ ΠΑΝΤΟΤΕ ΔΗμιουργική αλλά και αμφίσημη. Από την εφεύρεση της προοπτικής στην Αναγέννηση, την ανάπτυξη της φωτογραφίας τον 19ο αιώνα και την εμφάνιση της βιντεοτέχνης (video art) τον 20ό, η καλλιτεχνική πρακτική τροφοδοτήθηκε και ταυτόχρονα αμφισβητήθηκε από τις τεχνολογικές καινοτομίες. Η είσοδος της τεχνητής νοημοσύνης στον χώρο της καλλιτεχνικής δημιουργίας σηματοδοτεί έναν ακόμη, ίσως τον πιο ριζικό, μετασχηματισμό. Για πρώτη φορά στην ιστορία της τέχνης, η δημιουργική διαδικασία δεν προϋποθέτει άμεσα έναν έλλογο και συνειδητό άνθρωπο, αλλά μπορεί να παραχθεί από ένα υπολογιστικό σύστημα που «μαθαίνει» πρότυπα, αναγνωρίζει συσχετίσεις και παράγει νέα έργα χωρίς εμπειρία, συνείδηση ή πρόθεση. Στη σύγχρονη εποχή, η παραγωγή εικόνων από Π.Τ.Ν. για παράδειγμα, βασίζεται κυρίως σε Generative Pretrained Transformers (GPTs) και σε ειδικά image generators όπως τα diffusion models (π.χ. DALL·E, Stable Diffusion, Midjourney). Τα GPTs εκπαιδεύονται σε τεράστιες βάσεις δεδομένων κειμένου ώστε να κατανοούν τη γλώσσα και να μεταφράζουν λεκτικές περιγραφές σε δομημένες εντολές. Οι image generators, με τη σειρά τους, «μεταφράζουν» αυτές τις εντολές σε οπτικό περιεχόμενο. Για παράδειγμα, ένα diffusion model ξεκινά από θόρυβο και, βήμα προς βήμα, «καθαρίζει» την εικόνα ώσπου να εμφανιστεί η τελική μορφή που ανταποκρίνεται στην περιγραφή. Έτσι, η

φράση «ένας πίνακας με το στυλ του Βαν Γκογκ που απεικονίζει ένα ηλιοβασίλεμα πάνω από τη θάλασσα» μπορεί να παραχθεί σε ελάχιστα δευτερόλεπτα. Αυτή η διαδικασία καθιστά την Π.Τ.Ν. όχι μόνο εργαλείο αναπαραγωγής, αλλά και μηχανισμό σύνθεσης, ικανό να συνδυάζει στυλ, μοτίβα και αισθητικές αναφορές με τρόπους που ξεπερνούν την ανθρώπινη μνήμη ή φαντασία. Η δυναμική αυτή φαίνεται στην πράξη: εκατομμύρια εικόνες, βίντεο και μουσικές συνθέσεις παράγονται καθημερινά, συχνά με εντυπωσιακή αισθητική ποιότητα. Τα έργα αυτά εκτίθενται σε μουσεία, πωλούνται σε δημοπρασίες και συζητούνται σε διεθνές επίπεδο. Το 2018, το έργο *Portrait of Edmond de Belamy* που δημιουργήθηκε από το γαλλικό συλλογικό Obvious με τη χρήση GAN, πουλήθηκε στον οίκο Christie's για 432.500 δολάρια – γεγονός που σηματοδότησε την είσοδο της Π.Τ.Ν. στην παγκόσμια αγορά τέχνης. Η εξέλιξη αυτή εγείρει θεμελιώδη φιλοσοφικά ερωτήματα. Η τέχνη, όπως την κατανοούμε παραδοσιακά, συνδέεται με την ανθρώπινη πρόθεση, τη δημιουργικότητα, την εμπειρία και την αυθεντικότητα. Μπορεί, όμως, ένα μηχανικό σύστημα να θεωρηθεί δημιουργός; Είναι τα έργα της Π.Τ.Ν. απλώς προϊόντα υπολογιστικής επεξεργασίας, ή μπορούν να ενταχθούν σε αυτό που ονομάζουμε «τέχνη»; Και αν ναι, με ποια κριτήρια; Τα ερωτήματα αυτά δεν είναι αμιγώς θεωρητικά, αλλά έχουν βαθιές πολιτισμικές και κοινωνικές συνέπειες**.** Η αποδοχή της Π.Τ.Ν. ως καλλιτεχνικού υποκειμένου μπορεί να αναδιαμορφώσει τον ρόλο του ανθρώπινου δημιουργού, να επηρεάσει την αγορά τέχνης, να δημιουργήσει νέα ηθικά και νομικά διλήμματα, και να προκαλέσει κρίση αξιών γύρω από την έννοια της αυθεντικότητας. Από την άλλη πλευρά, μπορεί να ανοίξει νέους ορίζοντες δημιουργικότητας, να εκδημοκρατίσει την πρόσβαση στην καλλιτεχνική παραγωγή και να προσφέρει νέα εργαλεία έκφρασης σε καλλιτέχνες και κοινό.

Στόχος του παρόντος κειμένου είναι να προσεγγίσει τη δημιουργία μέσω Π.Τ.Ν. με φιλοσοφικούς και πολιτισμικούς όρους. Ειδικότερα:

- να αναλύσει θεωρητικά ερωτήματα γύρω από τον ορισμό της τέχνης (θεσμικές θεωρίες, το ωραίο, η πρόθεση, η προοπτική του Βιτγκενστάιν),
- να αποτιμήσει τις συνέπειες της αναγνώρισης της Π.Τ.Ν. ως τέχνης**,** τόσο στο επίπεδο της αισθητικής αξίας όσο και στο επίπεδο της κοινωνικής και θεσμικής πρακτικής,
- και να προτείνει κατευθύνσεις πολιτισμικής πολιτικής για μια δίκαιη, συμπεριληπτική και ηθικά τεκμηριωμένη συνύπαρξη ανθρώπινης και τεχνητής δημιουργικότητας.

Το κείμενο αυτό δεν φιλοδοξεί να δώσει έναν τελικό ορισμό της τέχνης· αντίθετα, επιδιώκει να δείξει πώς η εμφάνιση της Π.Τ.Ν. αναδεικνύει την αστάθεια και τη δυναμικότητα της έννοιας «τέχνη» και μας καλεί να την ξανασκεφτούμε υπό νέο πρίσμα.

ΙΙ. Ο Ορισμός της Τέχνης

Το ερώτημα *«τι είναι τέχνη;»* βρίσκεται στον πυρήνα της φιλοσοφικής αισθητικής και παραμένει ανοιχτό εδώ και αιώνες. Η πολυπλοκότητα της τέχνης δεν επιτρέπει έναν καθολικό, διαχρονικό ορισμό[1]. Από την αρχαιότητα έως σήμερα, οι θεωρίες για την τέχνη ποικίλλουν: άλλοτε την ταυτίζουν με την ομορφιά, άλλοτε με τη μιμητική ικανότητα, άλλοτε με την έκφραση συναισθημάτων, άλλοτε με τη θεσμική αναγνώριση**.** Στην παρούσα ενότητα θα εξετάσουμε τέσσερις βασικές φιλοσοφικές προσεγγίσεις που έχουν καθορίσει τον διάλογο και αποκτούν νέα επικαιρότητα στην εποχή της τεχνητής νοημοσύνης.

Η Θεωρία των Καλλιτεχνικών Θεσμών

Η πιο επιδραστική θεωρητική απόπειρα του 20ού αιώνα να οριστεί η τέχνη είναι η *θεσμική θεωρία*[2], που διατυπώθηκε από τον Τζορτζ Ντίκι **(**George Dickie**)** το 1974 και επηρεάστηκε από τις ιδέες του θεωρητικού τέχνης Άρθουρ Ντάντο (Arthur Danto). Σύμφωνα με αυτήν, ένα αντικείμενο δεν είναι τέχνη εξαιτίας κάποιων εγγενών αισθητικών ιδιοτήτων (όπως η ομορφιά ή η αρμονία), αλλά επειδή εντάσσεται και αναγνωρίζεται από τον λεγόμενο «κόσμο της τέχνης» (*artworld*)[3]. Για να είναι κάτι έργο τέχνης, πρέπει:

1. να είναι ένα τεχνούργημα, δηλαδή να έχει δημιουργηθεί ή επιμεληθεί από κάποιον,

[1] Arthur Danto. History and Theory, Vol. 37, No. 4, Theme Issue 37: Danto and His Critics: Art History, Historiography and After the End of Art. (Dec., 1998), 134.

[2] George Dickie. Art and the Aesthetic: An Institutional Analysis. Ithaca, (New York: Cornell University Press, 1974), 34.

[3] Ο όρος «κόσμος της τέχνης» (artworld) καθιερώθηκε στη φιλοσοφία της τέχνης από τον Ντάντο**,** Danto, Arthur C. (1964). "The Artworld". *The Journal of Philosophy*, Vol. 61, No. 19, pp. 571–584.

2. και να έχει αναγνωριστεί θεσμικά από τους φορείς του κόσμου της τέχνης (μουσεία, γκαλερί, επιμελητές, ειδικό κοινό).

Η θεωρία αυτή έχει το πλεονέκτημα ότι αναγνωρίζει τη σχετικότητα και την ιστορικότητα της τέχνης: το τι θεωρείται έργο μπορεί να αλλάζει ανά εποχή και πλαίσιο. Ωστόσο, έχει δεχθεί κριτική για το ότι φαίνεται κυκλική: για να ξέρουμε τι είναι τέχνη, πρέπει να ρωτήσουμε τον κόσμο της τέχνης, αλλά για να ξέρουμε ποιος ανήκει στον κόσμο της τέχνης, πρέπει ήδη να ξέρουμε τί είναι τέχνη. Η περίπτωση της Π.Τ.Ν. οξύνει αυτές τις ενστάσεις. Αν ένα έργο που δημιουργείται από αλγόριθμο εκτίθεται σε μουσείο και συζητείται από κριτικούς, πληροί ήδη τα κριτήρια του Ντίκι. Άρα το ζήτημα μετατοπίζεται: όχι αν είναι τέχνη, αλλά τι είδους τέχνη είναι και πώς νοείται όταν ο δημιουργός δεν είναι άνθρωπος.

Η Τέχνη του Ωραίου

Μια αρχαιότερη και πιο διαισθητική προσέγγιση συνδέει την τέχνη με την ομορφιά. Από τον Πλάτωνα, που θεωρούσε την τέχνη ως μίμηση του ιδεατού κάλλους, έως τον Καντ, που είδε την αισθητική κρίση ως ανιδιοτελή απόλαυση, η τέχνη συχνά ορίστηκε ως μέσο αποκάλυψης του ωραίου. Ωστόσο, ήδη από τον 18ο αιώνα, στοχαστές όπως ο Ντέιβιντ Χιούμ (David Hume) επεσήμαναν τη υποκειμενικότητα της αισθητικής κρίσης: το όμορφο δεν βρίσκεται στο αντικείμενο, αλλά στο πνεύμα που το αντιλαμβάνεται[4]. Η νεότερη τέχνη απομακρύνθηκε ακόμη περισσότερο από το «ωραίο»: το έργο του Μουνκ (*Η Κραυγή*), οι προκλητικές εγκαταστάσεις του Ντυσάν (Marcel Duchamp) όπως το έργο *Κρήνη* (Fountain, 1917) ή η σκόπιμη ασχήμια του Μπέικον δείχνουν ότι η τέχνη δεν εξαντλείται στο όμορφο. Στην περίπτωση της Τ.Ν., η παραγωγή εικόνων «όμορφων» ή «καλαίσθητων» είναι σχετικά εύκολη, χάρη στους αλγόριθμους εκμάθησης προτύπων. Αλλά η ομορφιά από μόνη της δεν αρκεί για να αποδώσει καλλιτεχνική αξία. Ένα ηλιοβασίλεμα ή ένας πολύτιμος λίθος είναι όμορφα, αλλά δεν είναι τέχνη. Χρειάζεται πλαίσιο, πράξη και νόημα. Άρα, ακόμα κι αν η Π.Τ.Ν. παράγει κάτι όμορφο, το ερώτημα παραμένει: είναι αυτό τέχνη;

[4] David Hume. Four Dissertations: Of the Standard of Taste,(London: A. Millar, 1751), §678.

Η Πρόθεση και το Τελικό Αίτιο

Ο Αριστοτέλης (*Μετά τα Φυσικά*) υποστήριξε ότι κάθε δημιουργία εξηγείται μέσω τεσσάρων αιτίων[5]: *υλικό*, *μορφικό*, *ποιητικό* και *τελικό*. Ιδιαίτερα το *τελικό αίτιο* – η πρόθεση, ο σκοπός – είναι αυτό που προσδίδει νόημα στο έργο. Ένα αγγείο δεν είναι απλώς πηλός (υλικό) και σχήμα (μορφή), ούτε μόνο αποτέλεσμα του κεραμίστα (ποιητικό αίτιο), αλλά υπηρετεί έναν σκοπό: να χρησιμοποιηθεί ή να διακοσμήσει. Στην περίπτωση της τέχνης, η πρόθεση θεωρείται κεντρική. Χωρίς αυτήν, η τέχνη μοιάζει να στερείται βάθους. Η Π.Τ.Ν., όμως, δεν έχει συνείδηση ή σκοπό. Απλώς εκτελεί υπολογιστικές διαδικασίες, όπως έδειξε και το διάσημο επιχείρημα του Τζον Σέαρλ (John Searle) με το «Κινέζικο Δωμάτιο» (1980): ένα σύστημα μπορεί να χειρίζεται σύμβολα χωρίς να κατανοεί το νόημά τους[6]. Κάποιοι υποστηρίζουν ότι η πρόθεση μπορεί να «μετατεθεί» στον άνθρωπο που προγραμματίζει ή εισάγει την εντολή στο σύστημα. Ωστόσο, όσο πιο αυτόνομα λειτουργούν τα μοντέλα, τόσο δυσκολότερο είναι να εντοπίσουμε ποιος είναι ο πραγματικός «δημιουργός». Έτσι, η Π.Τ.Ν. αμφισβητεί ευθέως την αριστοτελική παράδοση.

Ο Βιτγκενστάιν και η Οικογενειακή Ομοιότητα

Μια εναλλακτική προσέγγιση είναι αυτή του Βίτγκενστάιν (Ludwig Wittgenstein), ο οποίος στις *Φιλοσοφικές Έρευνες* (*Philosophical Investigations, 1953*), υποστήριξε ότι πολλές έννοιες της γλώσσας μας (π.χ. «παιχνίδι», «τέχνη») δεν έχουν αυστηρό ορισμό, αλλά λειτουργούν με βάση την οικογενειακή ομοιότητα[7]. Δεν υπάρχει ένα ενιαίο χαρακτηριστικό που να ισχύει για όλα τα παιχνίδια ή όλα τα έργα τέχνης· υπάρχουν, όμως, συστάδες ομοιοτήτων που δημιουργούν δίκτυα νοήματος[8]. Από αυτήν την προοπτική, δεν χρειάζεται να αναζητήσουμε μια απόλυτη ουσία της τέχνης. Αρκεί να δούμε πώς ένα έργο λειτουργεί μέσα σε πολιτισμικά συμφραζόμενα. Έτσι, η τέχνη της Π.Τ.Ν. μπορεί να θεωρηθεί απλώς μια νέα «οικογένεια» εκφραστικών πρακτικών: δεν είναι λιγότερο τέχνη επειδή

[5] R. B. Todd, 1976, "The Four Causes: Aristotle's Exposition and the Ancients, *Journal of the History of Ideas*, 37: 319–322.

[6] John Searle. Minds, Brains, and Programs. Behavioral and Brain Sciences 3, Cambridge University Press, (1980), 417-424.

[7] LudwigWittgenstein. Philosophical Investigations.Trans. G.E.M. Anscombe.Blackwell, 2001 (Translation of the 1953 original edition). Part II, 200-242.

[8] Ibid., Part I, 23

δεν πληροί τον ορισμό της πρόθεσης, αλλά ούτε είναι αυτομάτως ισότιμη με την ανθρώπινη δημιουργία. Το ζήτημα δεν είναι «είναι τέχνη;» αλλά «πώς λειτουργεί ως τέχνη;».

Συμπερασματικά, οι διαφορετικές θεωρίες που εξετάσαμε – από τη θεσμική μέχρι την αριστοτελική και τη βιτγκενσταϊνική – δείχνουν ότι η τέχνη δεν επιδέχεται έναν μοναδικό και οριστικό ορισμό. Η Π.Τ.Ν. αναδεικνύει ακόμη πιο έντονα αυτό το αδιέξοδο, αφού μπορεί να ικανοποιεί κάποια κριτήρια και να αμφισβητεί άλλα. Επομένως, αντί να προσπαθούμε να δώσουμε έναν τελικό ορισμό, το ουσιαστικό ερώτημα είναι ποιες συνέπειες έχει η αποδοχή της Π.Τ.Ν. ως μορφής τέχνης για τον τρόπο που κατανοούμε την ανθρώπινη δημιουργικότητα, την αυθεντικότητα και τα όρια της καλλιτεχνικής πράξης. Στο επόμενο κεφάλαιο θα στραφούμε σε αυτές ακριβώς τις συνέπειες.

III. Συνέπειες της Αναγνώρισης της Τέχνης της Π.Τ.Ν. ως Πραγματικής Τέχνης

Η αποδοχή της Π.Τ.Ν. ως καλλιτεχνικού υποκειμένου δεν αποτελεί απλώς ένα θεωρητικό παιχνίδι. Συνοδεύεται από ευρύτατες συνέπειες που αφορούν τον τρόπο με τον οποίο νοούμε την τέχνη, τον ρόλο του ανθρώπινου δημιουργού, αλλά και θεμελιώδη ζητήματα ηθικής και δικαίου.[9] Στην ενότητα αυτή θα εξετάσουμε τρεις όψεις του προβλήματος: (α) τον επαναπροσδιορισμό της αξίας της ανθρώπινης τέχνης, (β) τα ζητήματα αυθεντικότητας και εμπειρίας, και (γ) τα ηθικά και νομικά διλήμματα που προκύπτουν.

Επαναπροσδιορισμός της Αξίας της Ανθρώπινης Τέχνης

Η ιστορία της τέχνης έχει συνδεθεί στενά με την μοναδικότητα και την ανεπανάληπτη προσωπική σφραγίδα του καλλιτέχνη. Από τον ρομαντισμό

[9] Επιπλέον, η καλλιτεχνική και - πιο ειδικά - η μουσική δημιουργικότητα προτείνεται ως κριτήριο απόδοσης νοημοσύνης σε συστήματα Τ.Ν. και έχει οδηγήσει στο σχεδιασμό και επιτέλεση ενός πλήθους μουσικών εκδοχών της 'Δοκιμασίας Turing' (Turing Test). Για μια σύνοψη αλλά και κριτική ανάλυση αυτού του εγχειρήματος, δείτε: George Kosteletos and Anastasia Georgaki, "A Turing Test for the Singing Voice as an Anthropological Tool: Epistemological and Technical Issues," *Proceedings of the ICMC*, Ljubljana, Slovenia, (2012):46-51. http://hdl.handle.net/2027/spo.bbp2372.2012.008.

και μετά, η τέχνη θεωρήθηκε η αυθεντική έκφραση ενός μοναδικού υποκειμένου που μετουσιώνει το βίωμα, τον πόνο, τη χαρά και το όραμά του σε μορφή. Ένα ποίημα του Σεφέρη ή ένας πίνακας του Βαν Γκογκ δεν έχουν αξία επειδή είναι «όμορφα»· έχουν αξία επειδή αποτελούν μαρτυρίες ζωής, κρυσταλλώσεις μιας υποκειμενικότητας που πάλεψε να βρει έκφραση. Η μαζική παραγωγή έργων από Π.Τ.Ν. απειλεί να υποβαθμίσει αυτή την αξία. Αν ένας αλγόριθμος μπορεί μέσα σε δευτερόλεπτα να παράγει χιλιάδες εικόνες υψηλής αισθητικής ποιότητας, τι σημαίνει αυτό για τον πίνακα που ένας καλλιτέχνης δούλεψε μήνες; Υπάρχει ο κίνδυνος η ανθρώπινη τέχνη να φανεί πιο «αργή» ή «αναποτελεσματική» μπροστά στην αλγοριθμική υπερπαραγωγή. Ωστόσο, αυτή η κρίση μπορεί να λειτουργήσει και αναστοχαστικά. Η ανθρώπινη τέχνη δεν αξίζει επειδή είναι τεχνικά αξεπέραστη (ήδη η φωτογραφία, ο κινηματογράφος, η ψηφιακή τέχνη έχουν δείξει ότι η τεχνική υπεροχή δεν είναι το κριτήριο), αλλά επειδή είναι υπαρξιακά ειλικρινής. Η Φρίντα Κάλο έχει αξία όχι γιατί δεν μπορεί να την μιμηθεί η Π.Τ.Ν., αλλά γιατί τα έργα της αποτελούν άμεση αντανάκλαση της προσωπικής της εμπειρίας, της σωματικής οδύνης, των τραυμάτων και της βαθιάς της σχέσης με την ταυτότητα και τον πολιτισμό της. Αυτές οι εσωτερικές και βιωματικές διαστάσεις δεν μπορούν να παραχθούν από μηχανές που στερούνται εμπειρία ζωής.

Ζήτημα Αυθεντικότητας και Εμπειρίας

Η έννοια της αυθεντικότητας βρίσκεται στο επίκεντρο. Ο Γουόλτερ Μπέντζαμιν (Walter Benjamin) είχε ήδη επισημάνει ότι η τεχνική αναπαραγωγή (φωτογραφία, κινηματογράφος) απειλεί την «αύρα» του έργου τέχνης, την μοναδικότητα που το συνδέει με τον δημιουργό και τον χρόνο του[10]. Στην περίπτωση της Π.Τ.Ν., η αναπαραγωγή φτάνει σε ακραίο βαθμό: η μηχανή μπορεί να δημιουργήσει «νέες» εικόνες συνδυάζοντας και αναμειγνύοντας προϋπάρχουσες. Το ερώτημα που αναδύεται είναι: μπορεί ένα τέτοιο έργο να θεωρηθεί αυθεντικό; Αν η αυθεντικότητα ορίζεται ως μοναδική προσωπική έκφραση, τότε η Π.Τ.Ν. αδυνατεί να την προσφέρει. Αν, όμως, την ορίσουμε λειτουργικά, δηλαδή ως την εμπειρία που προκαλεί σε εμάς, τότε η Π.Τ.Ν. μπορεί να δημιουργήσει έργα που συγκινούν, προβληματίζουν ή εμπνέουν, ακόμη και αν στερούνται προσωπικής πρόθεσης. Άλλωστε, ακόμη και η ανθρώπινη τέχνη δεν είναι ποτέ

[10] Benjamin, W. (2008). *The work of art in the age of mechanical reproduction* (J. A. Underwood, Trans.). Penguin Books, 100.

πλήρως «καθαρή» ή αυτόνομη· κάθε καλλιτέχνης επηρεάζεται από την παράδοση, τα έργα άλλων δημιουργών και το πολιτισμικό του πλαίσιο. Έτσι, η αυθεντικότητα μπορεί να γίνει κατανοητή όχι ως απόλυτη πρωτοτυπία, αλλά ως ιδιαίτερη και ανεπανάληπτη ερμηνεία αυτού του πλούσιου δικτύου επιρροών. Η εμπειρία του θεατή είναι επίσης κρίσιμη. Ο θεατής που γνωρίζει ότι βλέπει έργο Π.Τ.Ν. συχνά βιώνει **διπλή ανάγνωση**: από τη μια απολαμβάνει την αισθητική, από την άλλη στοχάζεται πάνω στο ίδιο το φαινόμενο της μηχανικής δημιουργικότητας. Αυτό το «μετα-καλλιτεχνικό» στοιχείο προσδίδει μια ιδιότυπη αξία που δεν υπήρχε στην παραδοσιακή τέχνη.

Ηθικά και Νομικά Διλήμματα

Η παραγωγή τέχνης από την Π.Τ.Ν. φέρνει στο προσκήνιο τρία κρίσιμα ζητήματα: την ευθύνη, την πνευματική ιδιοκτησία και τα προσωπικά δεδομένα. Πρώτον, το ζήτημα της ευθύνης είναι περίπλοκο. Αν ένα έργο παράγει ρατσιστικό ή βίαιο περιεχόμενο, ποιος λογοδοτεί; Ο προγραμματιστής, ο χρήστης ή η εταιρεία που ανέπτυξε το μοντέλο; Η απουσία πρόθεσης από την ίδια τη μηχανή δεν αναιρεί τον κοινωνικό αντίκτυπο που μπορεί να έχει το έργο, γεγονός που δημιουργεί σημαντικά κενά ευθύνης, τα οποία η τρέχουσα νομοθεσία δυσκολεύεται να καλύψει. Δεύτερον, το ζήτημα της πνευματικής ιδιοκτησίας αναδεικνύεται χαρακτηριστικά μέσα από παραδείγματα όπως το *Portrait of Edmond de Belamy*, το οποίο δημιουργήθηκε με τη χρήση GAN και πωλήθηκε σε δημοπρασία όπως προαναφέραμε. Σε ποιον ανήκουν τα δικαιώματα; Στους δημιουργούς του αλγορίθμου, στο συλλογικό Obvious που το παρουσίασε, ή στους καλλιτέχνες των έργων που χρησιμοποιήθηκαν για την εκπαίδευση του συστήματος; Οι Ηνωμένες Πολιτείες έχουν ξεκαθαρίσει ότι μόνο άνθρωπος μπορεί να θεωρηθεί δημιουργός (U.S. Copyright Office, 2023), ενώ η Ευρωπαϊκή Ένωση φαίνεται να αναζητά πιο ευέλικτες λύσεις (AIPPI, 2022). Τρίτον, η εκπαίδευση των μοντέλων στηρίζεται σε τεράστιες βάσεις δεδομένων εικόνων, συχνά χωρίς την ρητή συναίνεση των δημιουργών ή των υποκειμένων που απεικονίζονται. Αυτό εγείρει σοβαρά ζητήματα προστασίας της ιδιωτικότητας, αφού μπορεί να οδηγήσει στην παραγωγή εικόνων που μοιάζουν με πραγματικά πρόσωπα ή ακόμη και σε παραπλανητικές αναπαραστάσεις τύπου *deepfake*. Η προστασία των δεδομένων, επομένως, καθίσταται μια κατεξοχήν ηθική και πολιτική προτεραιότητα.

Συνοψίζοντας, η αποδοχή της Π.Τ.Ν. ως μορφής τέχνης προκαλεί μια διπλή κρίση: από τη μία πλευρά μια οντολογική κρίση, που αφορά το τι

σημαίνει τέχνη, δημιουργικότητα και αυθεντικότητα, και από την άλλη μια θεσμική και ηθική κρίση, που αφορά την ευθύνη, την ιδιοκτησία και την προστασία των εμπλεκομένων. Ωστόσο, ταυτόχρονα ανοίγει και νέους δρόμους, καλώντας μας να επαναστοχαστούμε τι πραγματικά εκτιμούμε στην ανθρώπινη τέχνη και πώς μπορούμε να διαμορφώσουμε ένα δίκαιο και ισορροπημένο πλαίσιο για τη συνύπαρξη ανθρώπινης και τεχνητής δημιουργικότητας.

IV. Πιθανές Λύσεις

Η ενσωμάτωση της τεχνητής νοημοσύνης στο πεδίο της καλλιτεχνικής δημιουργίας δεν είναι μια ουδέτερη τεχνολογική καινοτομία· είναι μια πολιτισμική πρόκληση που απαιτεί συντονισμένες απαντήσεις σε αισθητικό, θεσμικό και κοινωνικό επίπεδο. Αντί να αντιμετωπιστεί ως απειλή ή ως πανάκεια, η τέχνη της Π.Τ.Ν. χρειάζεται πλαίσια που θα διασφαλίσουν ότι συμβάλλει στον εμπλουτισμό της καλλιτεχνικής εμπειρίας χωρίς να υπονομεύει τον άνθρωπο. Στην ενότητα αυτή προτείνονται πέντε βασικές κατευθύνσεις.

Διαχωρισμός Ανθρώπινης και Τεχνολογικά Υποβοηθούμενης Τέχνης

Μια πρώτη λύση θα μπορούσε να είναι η καθιέρωση ενός συστήματος σήμανσης που θα ενημερώνει το κοινό για τον βαθμό εμπλοκής της Π.Τ.Ν. στη δημιουργία ενός έργου. Όπως τα τρόφιμα φέρουν ετικέτες προέλευσης που επιτρέπουν στον καταναλωτή να γνωρίζει την πηγή και τον τρόπο παραγωγής τους, έτσι και τα καλλιτεχνικά έργα θα μπορούσαν να φέρουν ενδείξεις σχετικά με το αν αποτελούν αποκλειστικά ανθρώπινη δημιουργία, αν είναι προϊόν συνεργασίας ανθρώπου και Π.Τ.Ν., ή αν έχουν παραχθεί αποκλειστικά από μηχανικά συστήματα. Η διαφάνεια αυτή θα δίνει στον θεατή τη δυνατότητα να κατανοεί καλύτερα το πλαίσιο παραγωγής του έργου και να αποδίδει ανάλογη αξία στην εμπειρία του. Παράλληλα, η πιστοποίηση αυτή μπορεί να λειτουργήσει ως δείκτης αυθεντικότητας, προστατεύοντας τον ρόλο του ανθρώπινου καλλιτέχνη και, ταυτόχρονα, αναγνωρίζοντας την Π.Τ.Ν. ως μια ξεχωριστή και αυτόνομη δημιουργική δύναμη.

Δημιουργία Διακριτής Αγοράς για Έργα Π.Τ.Ν.

Όπως η φωτογραφία απέκτησε τη δική της αυτόνομη αγορά τον 20ό αιώνα, έτσι και η τέχνη που παράγεται με Π.Τ.Ν. μπορεί να αναπτύξει διακριτούς μηχανισμούς αποτίμησης. Αντί τα έργα αυτά να ανταγωνίζονται άμεσα την παραδοσιακή ζωγραφική ή γλυπτική, θα μπορούσαν να ενταχθούν σε έναν παράλληλο χώρο, όπου θα αξιολογούνται με διαφορετικά κριτήρια, όπως η καινοτομία του αλγορίθμου που χρησιμοποιείται, η δημιουργική αξιοποίηση των δεδομένων και η πολιτισμική απήχηση που αποκτά το έργο. Με αυτόν τον τρόπο αποφεύγεται η άδικη απαξίωση των ανθρώπινων δημιουργών και, ταυτόχρονα, ενθαρρύνεται η ανάπτυξη νέων μορφών καλλιτεχνικής έκφρασης που αξιοποιούν πλήρως τις δυνατότητες της τεχνολογίας.

Θεσμική Υποστήριξη και Εκπαίδευση

Η μετάβαση σε μια εποχή όπου η Π.Τ.Ν. διαδραματίζει ενεργό ρόλο στην τέχνη απαιτεί θεσμική μέριμνα και στήριξη προς τους καλλιτέχνες. Είναι αναγκαία η ανάπτυξη προγραμμάτων επιμόρφωσης που θα τους επιτρέπουν να κατανοούν και να αξιοποιούν δημιουργικά τα νέα εργαλεία, η παροχή χρηματοδότησης σε έργα που συνδυάζουν ανθρώπινη και τεχνητή δημιουργικότητα, καθώς και η δημιουργία διεπιστημονικών εργαστηρίων όπου καλλιτέχνες, προγραμματιστές και θεωρητικοί θα μπορούν να συνεργάζονται και να πειραματίζονται. Παράλληλα, χρειάζεται να δοθεί έμφαση και στην εκπαίδευση του κοινού, ώστε να αναπτύξει μια κριτική στάση απέναντι στα έργα της Π.Τ.Ν. Μόνο όταν οι θεατές κατανοούν τόσο τις δυνατότητες όσο και τους περιορισμούς των αλγορίθμων θα μπορέσουν να αποδώσουν σωστά την αισθητική και κοινωνική σημασία αυτών των έργων.

Εκδημοκρατισμός της Παραγωγικής Τεχνητής Νοημοσύνης

Υπάρχει ο κίνδυνος η καλλιτεχνική χρήση της Π.Τ.Ν. να περιοριστεί στις μεγάλες εταιρείες που ελέγχουν τα πιο ισχυρά μοντέλα, δημιουργώντας έτσι μια νέα τεχνοκρατική ελίτ. Για να αποφευχθεί αυτό, είναι κρίσιμο να διασφαλιστεί η ισότιμη πρόσβαση σε όλους τους δημιουργούς[11].

[11] Όσοι αναφέρονται στον εκδημοκρατισμό της τεχνητής νοημοσύνης επιχειρηματολογούν λέγοντας ότι η ευρεία χρήση και η προσβασιμότητα στο τεχνολογικό μέσο

Αυτό μπορεί να επιτευχθεί μέσα από την ανάπτυξη ανοιχτών λογισμικών και εργαλείων δημιουργίας, τη διαφάνεια στα δεδομένα εκπαίδευσης και την ενίσχυση κοινοτήτων ανοιχτού κώδικα[12]. Με αυτόν τον τρόπο, η Π.Τ.Ν. μπορεί να λειτουργήσει ως εργαλείο ενδυνάμωσης και όχι ως μέσο αποκλεισμού, προσφέροντας ακόμη και σε καλλιτέχνες χωρίς τεχνικές γνώσεις ή οικονομικούς πόρους τη δυνατότητα να τη χρησιμοποιούν δημιουργικά.

Συνεργατικά Μοντέλα Δημιουργίας

Η πιο γόνιμη προοπτική για τη σχέση τέχνης και Π.Τ.Ν. ίσως δεν βρίσκεται στην αντιπαράθεση, αλλά στη συνεργασία. Αντί να αντιμετωπίζεται η Π.Τ.Ν. ως αυτόνομος δημιουργός, μπορεί να ιδωθεί ως συνεργατικός εταίρος. Ο καλλιτέχνης προσφέρει την πρόθεση, την ιδέα και το βίωμα, ενώ η μηχανή συνεισφέρει την υπολογιστική ισχύ, την ταχύτητα και την ικανότητα επεξεργασίας τεράστιων όγκων δεδομένων. Σε αυτό το υβριδικό πλαίσιο, η τέχνη μετατρέπεται σε προϊόν διαλόγου ανθρώπου και μηχανής. Δεν χάνεται ο άνθρωπος· αντίθετα, η δημιουργικότητά του επεκτείνεται, καθώς αναλαμβάνει τον ρόλο του «σκηνοθέτη» ή «ενορχηστρωτή» που αξιοποιεί την Π.Τ.Ν. ως μέσο πειραματισμού και πηγή έμπνευσης.

Συμπερασματικά, οι παραπάνω προτάσεις δεν συνιστούν εξαντλητικές λύσεις αλλά κατευθύνσεις που μπορούν να διαμορφώσουν μια νέα προοπτική. Το κρίσιμο ζητούμενο είναι να οικοδομηθεί ένα πολυεπίπεδο πλαίσιο που θα συνδυάζει τη διαφάνεια μέσω σαφούς σήμανσης των έργων, τη δικαιοσύνη με την προστασία των ανθρώπινων δημιουργών, την εκπαίδευση τόσο του κοινού όσο και των καλλιτεχνών, την ισότητα πρόσβασης μέσα από τον εκδημοκρατισμό των εργαλείων και, τέλος, τη συνεργασία μέσω νέων μοντέλων δημιουργικότητας. Με αυτόν τον τρόπο, η τέχνη που παράγεται με τη βοήθεια της Π.Τ.Ν. μπορεί να πάψει να αντιμετωπίζεται ως απειλή και να αναδειχθεί σε ευκαιρία για μια πιο πλούσια, πολυφωνική και δημοκρατική καλλιτεχνική εμπειρία.

αποτελούν ένα κριτήριο δημοκρατικού χαρακτήρα της τεχνητής νοημοσύνης αν λάβουμε δεδομένη την παραδοχή ότι οι χρήστες έχουν επίγνωση των πράξεών τους, δηλαδή ξέρουν τι κάνουν (Γούναρης & Κωστελέτος, 2024)

[12] Andreas Sudmann, ed. The Democratization of Artificial Intelligence.Net Politics in the Era of Learning Algorithms.Transcript Verlag, 2019, 9.

V. Συμπέρασμα

Η εμφάνιση της Π.Τ.Ν. στον χώρο της τέχνης αποτελεί ίσως το πιο ριζοσπαστικό σημείο καμπής στην ιστορία της αισθητικής. Για πρώτη φορά, η δημιουργικότητα δεν αποδίδεται αποκλειστικά σε ανθρώπινη συνείδηση, αλλά αναδύεται από αλγοριθμικά συστήματα που στερούνται εμπειρίας, συναισθήματος ή βιώματος. Η εξέλιξη αυτή θέτει σε δοκιμασία τις παραδοσιακές θεωρίες για τον ορισμό της τέχνης. Η θεσμική θεωρία του Ντίκι δείχνει ότι η κοινωνική αναγνώριση αρκεί για να προσδώσει τον τίτλο του έργου τέχνης, ενώ η θεωρία του «ωραίου» αποκαλύπτει τα όρια της αισθητικής απόλαυσης ως μοναδικού κριτηρίου. Η αριστοτελική έννοια της πρόθεσης αναδεικνύει το πρόβλημα της απουσίας σκοπού στις μηχανές, την ίδια στιγμή που η προοπτική της οικογενειακής ομοιότητας του Βιτγκενστάιν επιτρέπει την ένταξη της Π.Τ.Ν. σε ένα νέο πεδίο καλλιτεχνικών πρακτικών. Οι συνέπειες αυτής της αποδοχής είναι διπλές. Από τη μια πλευρά, αναδύονται κρίσεις αυθεντικότητας, ευθύνης και πνευματικής ιδιοκτησίας· από την άλλη, ανοίγονται νέες δυνατότητες για τον εκδημοκρατισμό της δημιουργίας και τον εμπλουτισμό της καλλιτεχνικής εμπειρίας. Το μέλλον της τέχνης που παράγεται με τη συνδρομή της Π.Τ.Ν. δεν θα κριθεί αποκλειστικά από τα ίδια τα έργα, αλλά κυρίως από το πλαίσιο που θα οικοδομηθεί γύρω τους. Αν αφεθεί στην αγορά και στην τεχνοκρατία να μονοπωλήσουν τον χώρο, υπάρχει κίνδυνος να απαξιωθεί η ανθρώπινη δημιουργικότητα και να χαθεί η πολιτισμική ποικιλομορφία. Αντίθετα, η επιλογή του δρόμου της διαφάνειας, της εκπαίδευσης, της θεσμικής υποστήριξης και της συνεργασίας μπορεί να καταστήσει την Π.Τ.Ν. καταλύτη μιας νέας ανθρωπιστικής αναγέννησης. Εν τέλη, η ουσία της τέχνης ίσως δεν βρίσκεται αποκλειστικά ούτε στον καλλιτέχνη ούτε στο εργαλείο, αλλά στη σχέση που αναδύεται ανάμεσα στο έργο και τον θεατή. Από αυτή την άποψη, η Π.Τ.Ν. δεν σηματοδοτεί το «τέλος της τέχνης», αλλά μια νέα αρχή που μας καλεί να ξανασκεφτούμε όχι μόνο τι είναι τέχνη, αλλά και τι σημαίνει να είμαστε άνθρωποι.

Βιβλιογραφία

AIPPI. *Approaches to IP Protection for Works Generated by Artificial Intelligence: European Standards*. AIPPI, 2022.

Benjamin, Walter. *The Work of Art in the Age of Mechanical Reproduction*. Trans. J.A. Underwood. Penguin, 2008. (Originally published 1936).

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

Danto, Arthur C. (1964). "The Artworld". *The Journal of Philosophy*, Vol. 61, No. 19.

Danto, Arthur. History and Theory, Vol. 37, No. 4, Theme Issue 37: Danto and His Critics: Art History, Historiography and After the End of Art. (Dec., 1998), 134.

Dickie, George. Art and the Aesthetic: An Institutional Analysis.Ithaca, New York: Cornell University Press, 1974.

Hume, David. Four Dissertations: Of the Standard of Taste, London: A. Millar, 1751. R. B., 1976, "The Four Causes: Aristotle's Exposition and the Ancients,*Journal of the History of Ideas*.

Kosteletos, George., and Anastasia Georgaki. "A Turing Test for the Singing Voice as an Anthropological Tool: Epistemological and Technical Issues," Proceedings of the ICMC, Ljubljana, Slovenia, (2012):46-51. http://hdl.handle.net/2027/spo.bbp2372.2012.008.

Searle, John. Minds, Brains, and Programs.*Behavioral and Brain Sciences* 3, Cambridge University Press, 1980.

Sudmann, Andreas, ed. The Democratization of Artificial Intelligence.Net Politics in the Era of Learning Algorithms.Transcript Verlag, 2019, 9.

U.S. Copyright Office. *Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence*. Washington, D.C., 2023.

Ludwig Wittgenstein,. Philosophical Investigations. Trans. G.E.M. Anscombe. Blackwell, 2001 (Translation of the 1953 original edition).

Περίληψη

Η εργασία εξετάζει τη σχέση τέχνης και τεχνολογίας με έμφαση στη ριζική τομή που φέρνει η παραγωγική τεχνητή νοημοσύνη (Π.Τ.Ν.). Από την αναγεννησιακή προοπτική και τη φωτογραφία του 19ου αιώνα έως τη βιντεοτέχνη του 20ού, η καλλιτεχνική δημιουργία διαρκώς επηρεαζόταν από τεχνολογικές καινοτομίες. Σήμερα, τα συστήματα Π.Τ.Ν. όπως τα GANs, το DALL·E 2, το Midjourney και το Stable Diffusion παράγουν μαζικά έργα με εντυπωσιακή αισθητική ποιότητα, εγείροντας νέα φιλοσοφικά και πολιτισμικά ερωτήματα: Μπορεί ένα μηχανικό σύστημα χωρίς συνείδηση και βίωμα να θεωρηθεί δημιουργός; Είναι τα έργα της Π.Τ.Ν. απλώς προϊόντα υπολογιστικής επεξεργασίας ή μπορούν να ενταχθούν στην έννοια της τέχνης; Η μελέτη αξιοποιεί αισθητικές και φιλοσοφικές προσεγγίσεις για να δείξει ότι η τέχνη της Π.Τ.Ν. αποσταθεροποιεί παραδοσιακούς ορισμούς, ενώ ταυτόχρονα ανοίγει νέους ορίζοντες δημιουργικότητας. Επιπλέον, συζητούνται ηθικά και νομικά διλήμματα σχετικά με την αυθεντικότητα, την ευθύνη και τα πνευματικά δικαιώματα, καθώς και κατευθύνσεις πολιτισμικής πολιτικής για δίκαιη και συμπεριληπτική συνύπαρξη ανθρώπινης και τεχνητής δημιουργίας. Συμπεραίνεται ότι η Π.Τ.Ν. δεν σηματοδοτεί το «τέλος της τέχνης», αλλά μια νέα αρχή που μας καλεί να ξανασκεφτούμε τόσο την έννοια της τέχνης όσο και την ίδια την ανθρώπινη ταυτότητα.

Λέξεις Κλειδιά: Παραγωγική Τεχνητή Νοημοσύνη, Τέχνη και Τεχνολογία, Αισθητική Φιλοσοφία, Πρόθεση, Δημιουργικότητα, Πολιτισμική Πολιτική, Ηθική, Αυθεντικότητα

Keywords: Generative Artificial Intelligence, Art and Technology, Aesthetic Philosophy, Intention, Creativity, Cultural Policy, Ethics, Authenticity

---

Βασίλειος Πολυχρονιάδης
vpolychroniadis@gmail.com
Εργαστήριο Εφαρμοσμένης Φιλοσοφίας, ΕΚΠΑ
ORCID iD: https://orcid.org/0000-0003-0846-5614

Εργασίες που αποστέλλονται στο περιοδικό για κρίση δεν δεσμεύουν τη συντακτική επιτροπή να τις δημοσιεύσει. Οι συγγραφείς φέρουν πλήρως την ευθύνη για το περιεχόμενο των εργασιών τους, που δεν εκφράζει απαραιτήτως την άποψη της συντακτικής επιτροπής, η οποία και δεν αναλαμβάνει ευθύνες γι' αυτό.

Disclaimer: Papers submitted to the journal for review do not oblige the editorial board to publish them. The views expressed in this publication are those of the author and are not necessarily indicative of those of the editorial board.

Παιδαγωγικός Λόγος, τόμος 32, τεύχος 1, 2026

Αυτό το έργο έχει άδεια χρήσης Creative Commons: Αναφορά Δημιουργού-Μη Εμπορική Χρήση-Όχι Παράγωγα Έργα 4.0 Διεθνές (CC BY-NC-ND 4.0).
Πληροφορίες για αυτή την άδεια:
https://creativecommons.org/licenses/by-nc-nd/4.0/

This work is licenced under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0)
To view a copy of this licence, visit:
https://creativecommons.org/licenses/by-nc-nd/4.0/

Πρώτη έκδοση: Απρίλιος 2026

Επιστημονική Επιμέλεια: Η Συντακτική Επιτροπή

Printed in Greece 2026
ISSN 1106-9341
e-ISSN: 2732-8937

Ο Παιδαγωγικός Λόγος είναι διαθέσιμος στις ιστοσελίδες:
https://ejournals.epublishing.ekt.gr/index.php/plogos/index
www.plogos.gr