Όψεις της Εφαρμοσμένης Επιστήμης και Τεχνολογίας

Διερευνώντας το αξιακό τοπίο της Τεχνοεπιστήμης

Επιστημονική επιμέλεια
Κώστας Θεολόγου – Ευγενία Τζαννίνη

ΕΛΛΗΝΟ
ΕΚΔΟΤΙΚΗ

ΕΛΛΗΝΟΕΚΔΟΤΙΚΗ

Κεντρική διάθεση:
Δ.Β. ΕΛΛΗΝΟΕΚΔΟΤΙΚΗ Α.Ε.Ε.Ε. ΑΝΩΝΥΜΗ ΕΚΔΟΤΙΚΗ ΚΑΙ ΕΜΠΟΡΙΚΗ ΕΤΑΙΡΕΙΑ
Κεντρικά γραφεία: Ιπποκράτους 82, 106 80, Αθήνα, Ελλάδα
Τηλ.: 210 36 46965 – 210 36 35229 • e-mail: info@ellinoekdotiki.gr
Βιβλιοπωλείο: Ιπποκράτους 10-12, 106 79, Αθήνα, Ελλάδα
Τηλ.: 210 36 13676 – 210 36 40632 – 210 36 35207 • Fax: 210 36 35207
e-mail: info@ellinoekdotiki.gr – diakinisi@ellinoekdotiki.gr
www.ellinoekdotiki.gr • FB: www.facebook.com/ellinoekdotiki

Μελέτες – Συλλογικοί τόμοι

Όψεις της Εφαρμοσμένης Επιστήμης και Τεχνολογίας

Διερευνώντας το αξιακό τοπίο της Τεχνοεπιστήμης

ISBN: 978-960-563-549-7

Επιστημονική επιμέλεια:
Κώστας Θεολόγου – Ευγενία Τζαννίνη

Γραφιστική επιμέλεια και σελιδοποίηση:
DTP Ελληνοεκδοτικής

Σχεδίαση εξωφύλλου:
DTP Ελληνοεκδοτικής

Πρώτη έκδοση: Δεκέμβριος 2022
Παρούσα έκδοση: Δεκέμβριος 2022, Κ.Ε.ΕΛ: 124/22

Περιεχόμενα

Εισαγωγικό σημείωμα των επιμελητών 5

Γιώργος Αραμπατζής, *Το Βιολογικό και το Ηθικό στη Βυζαντινή Σκέψη* 19

Ηλίας Βαβούρας, Ευτυχία Φιριπή, *Ο πολιτισμός πηγή δυστυχίας; Η διαλεκτική αναβίωση της διερώτησης από τον Πλάτωνα στον Freud* 35

Φίλιππος Βασιλόγιαννης, *Ζητήματα ηθικής στις σχέσεις εμπιστοσύνης: μια φιλοσοφική σύνοψη της δεοντολογίας των λειτουργημάτων* 48

Άλκης Γούναρης - Γιώργος Κωστελέτος (Εργαστήριο Εφαρμοσμένης Φιλοσοφίας, ΕΚΠΑ) *Όπλα Τεχνητής Νοημοσύνης: Προβλήματα Απόδοσης Ηθικού Καθεστώτος στις Αυτόνομες Μηχανές* 73

Μαριάννα Καραμάνου, Σπύρος Ν. Μιχαλέας (Εργαστήριο Ιστορίας της Ιατρικής και Ιατρικής Ηθικής, Ιατρική Σχολή ΕΚΠΑ) *Περί Καλλιπαιδίας στο έργο του Γάλλου ιατρού, κληρικού και ποιητή Claude Quillet (1602-1661)* 124

Νικόλαος Γ. Κόιος, *Προς μία θεολογική βιοηθική ως βάση για την ηθική θεώρηση της εφαρμοσμένης γενετικής* 130

Ελένη Μαΐστρου, Το τοπίο ως πολιτιστική κληρονομιά 150

Αντωνία Μοροπούλου, *Το Ιερό Κουβούκλιο του Παναγίου Τάφου στον Ναό της Αναστάσεως στα Ιεροσόλυμα: το μνημείο, η μελέτη και το έργο αποκατάστασης από το ΕΜΠ* 167

Κώστας Μποτόπουλος, *Τα εκ της πανδημίας ηθικά* 188

Προκόπης Παυλόπουλος, *Από τη Βιομηχανική Επανάσταση στην Τεχνολογική Επανάσταση: στον αστερισμό ενός αβέβαιου μέλλοντος* 199

Γλυκερία Σιούτη, *Η προστασία από την κλιματική αλλαγή και η αρχή της αλληλεγγύης* 247

Ελένη Σπυράκου, Βάνα Σταυρίδη, Παναγιώτης Κάβουρας, Κωνσταντίνος Χαριτίδης (Εργαστήριο Προηγμένων και Σύνθετων Υλικών, Νανοϋλικών και Νανοτεχνολογίας – R-NanoLab, Σχολή Χημικών Μηχανικών ΕΜΠ), *Ηθικά ζητήματα στην έρευνα στα οργανοειδή και σε παρόμοιες τεχνολογίες* 257

Σπυρίδων Στέλιος - Κώστας Θεολόγου, *Ο τεχνητός γρίφος της ηθικής* 298

Ευγενία Τζαννίνη, *Η προστασία της Πολιτιστικής Κληρονομιάς και η Τεχνητή Νοημοσύνη: το μέλλον του ιστορικού μας παρελθόντος* 314

Νίκος Ψαρρός, *Είναι ο άνθρωπος φύσει πολιτικό ή έλλογο ον;* 349

Όπλα Τεχνητής Νοημοσύνης: Προβλήματα Απόδοσης Ηθικού Καθεστώτος στις Αυτόνομες Μηχανές

Άλκης Γούναρης - Γιώργος Κωστελέτος

Περίληψη

Η συζήτηση περί απόδοσης ηθικού καθεστώτος σε συστήματα Τεχνητής Νοημοσύνης (Τ.Ν.) παρουσιάζει ολοένα και μεγαλύτερο ενδιαφέρον, τόσο σε ακαδημαϊκό επίπεδο όσο και σε επίπεδο θεσμών, καθώς οι εκθετικά αυξανόμενες τεχνολογικές δυνατότητες μετασχηματίζουν τη μέχρι πρότινος θεωρητική άσκηση σε ρεαλιστική προοπτική. Η υπολογιστική ισχύς των μηχανών και η αυτονόμηση της λειτουργίας τους εγείρουν πιεστικά ερωτήματα σχετικά με τη δυνατότητα απόδοσης ηθικού καθεστώτος και δη «ευθύνης» στα ίδια αυτά δρώντα συστήματα.

Στο παρόν κείμενο εξετάζουμε κριτικά τους συνήθεις τρόπους προσέγγισης των εν λόγω ερωτημάτων, λαμβάνοντας ως παράδειγμα εργασίας την περίπτωση των αυτόνομων οπλικών συστημάτων (Lethal Autonomous Weapons Systems: LAWS), καθώς η δράση τους συνυφαίνεται με αποφάσεις ζωής και θανάτου, καθιστώντας τα ίσως την πλέον αιχμηρή έκφανση της Τ.Ν. Εντούτοις, η από μέρους μας επιχειρούμενη κριτική ανάλυση γίνεται με τέτοιο τρόπο, ώστε να έχει και μια γενικότερη ισχύ, πέραν της περίπτωσης των αυτόνομων όπλων Τ.Ν. Εκκινώντας από κλασικές θέσεις υπέρ της απόδοσης ηθικού καθεστώτος στα συστήματα αυτά, αναδεικνύουμε τα λογικά σφάλματα και τα επιστημολογικά κενά των συγκεκριμένων επιχειρημάτων, υιοθετώντας εν τέλει

μια σκεπτικιστική στάση ως προς τη δυνατότητα ικανοποιητικής διευθέτησης του ζητήματος απόδοσης ηθικής ευθύνης στα συστήματα Τ.Ν.

Από τον HAL 9000 στα Αυτόνομα Συστήματα Τ.Ν.

Τον Οκτώβριο του 2017 το ανθρωποειδές Sophia έγινε η πρώτη οντότητα τεχνητής νοημοσύνης που έλαβε την ιδιότητα του πολίτη της Σαουδικής Αραβίας (Weller, 2017). Δυο χρόνια πριν η Επιτροπή Νομικών Υποθέσεων της Ευρωπαϊκής Ένωσης είχε εισηγηθεί την ανάγκη θέσπισης ενός νομικού πλαισίου αναγνώρισης πολιτικών δικαιωμάτων και υποχρεώσεων σε ευφυή «ηλεκτρονικά πρόσωπα» τα οποία λαμβάνουν αυτόνομες αποφάσεις (Delvaux, 2016, σ. 12)[1]. Το πλαίσιο αυτό προς το παρόν δεν έχει οριοθετηθεί, καθώς διατυπώνονται αντικρουόμενες απόψεις επί του θέματος και κυρίως σε ό,τι αφορά το ζήτημα απόδοσης «ευθυνών» ως αποτέλεσμα αυτόνομων πράξεων των ευφυών συστημάτων. Το ζήτημα της απόδοσης ευθυνών στα συστήματα Τ.Ν. απασχολεί σήμερα, όπως θα δούμε στη συνέχεια, τόσο τη φιλοσοφική όσο και την ερευνητική κοινότητα και συνδέεται άρρηκτα με την έννοια του προσώπου. Κατά πόσον, όμως, μπορούμε να αποδίδουμε κυριολεκτικά τον όρο «πρόσωπο»[2] σε τεχνητά ευφυή συστήματα τα οποία λαμβάνουν αποφάσεις ή και συμπεριφέρονται όπως οι πραγματικοί άνθρωποι;

---

[1] «… Η διερεύνηση των επιπτώσεων όλων των πιθανών νομικών λύσεων για την … δημιουργία ενός συγκεκριμένου νομικού καθεστώτος για αυτόματες μηχανές, έτσι ώστε, τουλάχιστον τα πιο εξελιγμένα αυτόνομα ρομπότ, να μπορούν να καθιερωθούν ως ηλεκτρονικά πρόσωπα με συγκεκριμένα δικαιώματα και υποχρεώσεις, συμπεριλαμβανομένης της αποκατάστασης οποιασδήποτε ζημίας που μπορεί να προκαλέσουν και της εφαρμογής της ηλεκτρονικής προσωπικότητας σε υποθέσεις όπου τα ρομπότ λαμβάνουν έξυπνες αυτόνομες αποφάσεις ή αλληλεπιδρούν αυτόνομα με τρίτα μέρη».

[2] Να σημειωθεί ότι η αναγνώριση νομικής προσωπικότητας σε αντικείμενα, ζώα, φυτά ή συστήματα τεχνητής νοημοσύνης αποτελεί ένα ζήτημα που έχει απασχολήσει νομικούς και φιλοσόφους επί μακρόν (Solum, L., 1992). Συγκεκριμένα, συζητείται η απόδοση της ιδιότητας του προσώπου κατ' αναλογία της απόδοσης της ιδιότητας του νομικού προσώπου σε μη φυσικές οντότητες όπως εταιρείες, θεσμικά όργανα, δήμους, κυβερνητικούς οργανισμούς κ.λπ. που προβαίνουν σε πράξεις, συμβάλλονται, έχουν δικαιώματα και υποχρεώσεις, ευθύνες και απαιτήσεις. Ωστόσο, στο παρόν κείμενο θα μας απασχολήσει μόνο η απόδοση της ιδιότητας του «ηθικού» προσώπου στα συστήματα Τ.Ν.

Σε θεωρητικό επίπεδο, η συζήτηση σχετικά με την αυτόνομη λήψη αποφάσεων και την ηθική ευθύνη των συστημάτων τεχνητής νοημοσύνης έχει ξεκινήσει πριν υπάρξει καν η δυνατότητα κατασκευής ευφυών μηχανών με βιβλιογραφικό ορόσημο τα επιχειρήματα που ανέπτυξε ο Daniel Dennett (1997) στο πλαίσιο της φιλοσοφικής υπεράσπισης του μυθιστορηματικού χαρακτήρα HAL 9000. Τα επιχειρήματά του συνοψίζονται στη θέση ότι, εφόσον τα συστήματα αυτά προβαίνουν σε καθ' όλα νοήμονες και αυτόνομες αποφάσεις και αποτελεσματικές πράξεις, μπορούν να αξιολογηθούν ηθικά όπως ακριβώς θα μπορούσαν να αξιολογηθούν ηθικά και αντίστοιχες ανθρώπινες πράξεις. Αν δηλαδή ένα σύστημα σκέφτεται, δρα και συμπεριφέρεται όπως ο άνθρωπος (ή και καλύτερα από τον άνθρωπο), θα μπορεί να τύχει αναγνώρισης ηθικού καθεστώτος και να έχει τελικώς την ηθική ευθύνη των πράξεών του.

Στο παρόν κείμενο θα εξετάσουμε αυτό ακριβώς το πρόβλημα της απόδοσης ηθικού καθεστώτος και, συνεπώς, το πρόβλημα της ηθικής ευθύνης των μηχανών, το οποίο οδηγεί σε μια σειρά από ζητήματα σχετικά με την έννοια του ηθικού προσώπου. Εκκινούμε από την παραδοχή ότι αν κάποιος ή κάτι μπορεί να χαρακτηριστεί ως ηθική οντότητα, τότε μπορεί κάλλιστα να θεωρηθεί ότι καλύπτει τα χαρακτηριστικά του προσώπου γενικότερα, ενώ το αντίστροφο δεν είναι απαραίτητο να συμβαίνει.

Επιλέξαμε να επικεντρώσουμε την συζήτησή μας στις αυτόνομες στρατιωτικές μηχανές και δη στις μηχανές που αποφασίζουν αυτόνομα για ζητήματα ζωής και θανάτου, λόγω της μεγάλης οξύτητας των ηθικών ζητημάτων που προκαλούν ο σχεδιασμός, η κατασκευή και η χρήση τους, ωστόσο θεωρούμε ότι η επιχειρηματολογία μας μπορεί κάλλιστα να έχει μια γενικότερη ισχύ προς κάθε άλλο σύστημα Τ.Ν. Εστιάζοντας τη διερεύνησή μας στις ευφυείς αυτόνομες στρατιωτικές μηχανές ερχόμαστε αντιμέτωποι με ερωτήματα όπως, για παράδειγμα, το ερώτημα αν και υπό ποιες προϋποθέσεις θα πρέπει τα νοήμονα συστήματα να λαμβάνουν «αυτόνομες» αποφάσεις ζωής και θανάτου ή αν η νοημοσύνη, η αυτονομία και η αποτελεσματικότητα αποτελούν αναγκαίες και επαρκείς προϋποθέσεις για να αποδοθεί ηθικό καθεστώς σε μια δρώσα

οντότητα[3]. Κι αν ναι, τότε μήπως οι συγκεκριμένες πολεμικές μηχανές, εκτός από την ευθύνη των πράξεών τους, θα πρέπει να λαμβάνουν στρατιωτικά αξιώματα και να εντάσσονται στη στρατιωτική ιεραρχία όχι ως όπλα αλλά ως στρατιώτες; Θα σήμαινε κάτι τέτοιο, ενδεχομένως, ότι θα πρέπει να απολαμβάνουν των ευεργεσιών των Συμβάσεων της Γενεύης για τους αιχμαλώτους πολέμου, σε περίπτωση που συλληφθούν, ή και να λογοδοτούν σε στρατιωτικά δικαστήρια για τις πράξεις και τις παραλείψεις τους ή τη μη εκτέλεση διαταγών;

Σήμερα, η εμπλοκή των συστημάτων Τ.Ν. σε κυβερνητικές, στρατιωτικές, διαστημικές και άλλες επιχειρήσεις δεν αποτελεί πλέον μυθοπλαστικό περιεχόμενο ταινιών όπως το *2001: Η Οδύσσεια του Διαστήματος*. Ήδη αυτόνομες πολεμικές μηχανές, αλλά και άλλα συστήματα, επιχειρούν για αμυντικούς ή επιθετικούς σκοπούς και δοκιμάζονται σε πραγματικές εμπόλεμες καταστάσεις. Τα εξοπλιστικά προγράμματα των κρατών έχουν ήδη ξεκινήσει έναν αγώνα δρόμου για την απόκτηση του ανταγωνιστικού πλεονεκτήματος και η αμυντική βιομηχανία ανοίγει τον δρόμο προς αυτήν την ερευνητική κατεύθυνση, την ίδια στιγμή που θεσμικά ανεξέλεγκτες και ηθικά αμφιλεγόμενες χρήσεις ποικίλων μηχανών επιχειρούν στα πεδία μαχών ή σε μυστικές αποστολές για τη θανάτωση προσώπων που αναγνωρίζονται ως στόχοι υψηλής στρατηγικής σημασίας. Χαρακτηριστικό πρόσφατο παράδειγμα αυτής της χρήσης, η εξόντωση ηγετικού στελέχους της Αλ Κάιντα (Lee, 2022). Ενώ δηλαδή το ζήτημα

[3] Στη βιβλιογραφία ο όρος «agent» αποδίδεται συνήθως σε ένα δρων υποκείμενο και ειδικότερα σε κάποιον που προβαίνει σε μια εσκεμμένη ή και εμπρόθετη πράξη. Κατ' επέκταση, σχετικώς με την ηθική πράξη ενός υποκειμένου συναντάμε τον όρο «moral agent». Στα ελληνικά ο όρος «agent» αποδίδεται συχνά με τον όρο «πράκτορας» που πάντως χρησιμοποιείται ευρέως και σε άλλα γλωσσικά πλαίσια πέραν του φιλοσοφικού και λαμβάνει πλήθος διαφορετικών εννοιοδοτήσεων. Προς τούτο, στο πλαίσιο μιας φιλοσοφικής εκφοράς λόγου που καταπιάνεται με ζητήματα όπως αυτά της προθετικότητας, της συνείδησης και της ιδιότητας του προσώπου, θεωρούμε ως ορθότερη τη χρήση του όρου «πράττουσα οντότητα». Στο παρόν κείμενο, ωστόσο, η ισοδυναμία της δράσης ενός αυτόνομου συστήματος Τ.Ν. και ενός δρώντος προσώπου τελεί υπό διερεύνηση, συνεπώς υπό διερεύνηση τελεί και η δυνατότητα χρήσης του όρου «πράττουσα οντότητα» αναφορικά προς τα αυτόνομα αυτά συστήματα. Ως εκ τούτου, κατά την ακόλουθη ανάλυσή μας και έως ότου επιτευχθεί μια ικανοποιητική απάντηση περί του οντολογικού καθεστώτος των αυτόνομων συστημάτων Τ.Ν. και περί της πιθανής διαφοροποίησης ή εξίσωσής τους με τους ανθρώπους, επιλέγουμε τη χρήση του όρου «δρώσα οντότητα» που είναι πιο ουδέτερος ως προς τα οντολογικά του συμφραζόμενα.

απόδοσης ηθικού καθεστώτος στα συστήματα Τ.Ν. δεν έχει τελεσιδικήσει τύποις, υπάρχουν περιπτώσεις που εν τοις πράγμασι τα συστήματα αυτά αξιολογούν και δρουν «αυτόνομα» με σαφέστατες ηθικές και νομικές συνέπειες. Πρόκειται ουσιαστικά για μια ασυμμετρία που θα μπορούσε να αποδειχθεί θανάσιμη για το ίδιο το ανθρώπινο είδος. Αυτός είναι και ένας βασικός λόγος για τον οποίο εστιάζουμε στα πολεμικά συστήματα της Τ.Ν. και εξετάζουμε αν η έννοια του ηθικού προσώπου μπορεί να αποδοθεί σε συστήματα ευφυών μηχανών, καθώς η σφοδρότητα των συνεπειών που προκύπτουν από τη δράση τους είναι ανάλογη της οξύτητας των ηθικής φύσεως ερωτημάτων που η τελευταία αυτή εγείρει.

Χαρακτηριστικό αυτού του πιεστικού πλαισίου εντός του οποίου φιλόσοφοι και ερευνητές της Τ.Ν. καλούνται να αντιμετωπίσουν αυτά τα καινοφανή προβλήματα αποτελεί και η πρωτοβουλία των Max Tegmark και Stuart Russell (2015) οι οποίοι δημοσίευσαν μια ανοιχτή επιστολή με σκοπό τη θεσμική θωράκιση και την απαγόρευση κατασκευής αυτόνομων οπλικών συστημάτων και φονικών ρομπότ (Russell, 2020). Μεταξύ άλλων, αναφέρουν ότι τα αυτόνομα οπλικά συστήματα αποτελούν σήμερα την τρίτη επανάσταση στις πολεμικές επιχειρήσεις (μετά την πυρίτιδα και τα πυρηνικά) και λόγω του χαμηλού σχετικά κόστους και της ευκολίας κατασκευής τους, αναμένεται να διαδοθούν ευρέως και να παραχθούν μαζικά, με κίνδυνο να χρησιμοποιηθούν για τρομοκρατικές ενέργειες, εθνοκαθάρσεις, δολοφονίες προσωπικοτήτων, αποσταθεροποίηση εθνών, υποδούλωση πληθυσμών και επιλεκτική εξόντωση εθνικών ή κοινωνικών ομάδων. Για τον λόγο αυτόν, καλούν τους ερευνητές της Τ.Ν. να αρνηθούν να συμμετάσχουν στην έρευνα και την κατασκευή τέτοιων οπλικών συστημάτων, όπως αντίστοιχα οι βιολόγοι, οι χημικοί και οι φυσικοί υποστηρίζουν ευρέως ανάλογες διεθνείς συμφωνίες για την απαγόρευση χημικών και βιολογικών όπλων ή και όπλων εφοδιασμένων με laser.

Σε αντίθετη κατεύθυνση, οι υπέρμαχοι των συστημάτων αυτών υποστηρίζουν ότι οι πολεμικές μηχανές θα είναι απολύτως στοχοπροσηλωμένες, έννομες, ακριβείς και θα ακολουθούν απαρέγκλιτα όσα προβλέπουν οι Διεθνείς

Συμβάσεις για τους τραυματίες, τους αμάχους, τους αιχμαλώτους κ.λπ. ενώ, σε αντίθεση με τους ανθρώπους στρατιώτες, δεν θα υφίστανται ψυχολογική πίεση, δεν θα κάνουν λάθη κόπωσης και δεν θα προβαίνουν σε εκδικητικές θηριωδίες (όπως συμβαίνει συχνά με τους ανθρώπους-στρατιώτες οι οποίοι μπορεί να αποδειχθούν ψυχικά ασταθείς και συναισθηματικά ευάλωτοι κτλ). Συνεπώς, οι ευφυείς μηχανές δύναται να αποτελούν στο μέλλον το πρότυπο του ηθικού στρατιώτη, καθώς θα σέβονται τους αντιπάλους, τους πολίτες, τις υποδομές κ.λπ. (Pagallo, 2011; Wallace & Allen, 2008).

Όπως γίνεται κατανοητό, η συζήτηση γύρω από τα ηθικά προβλήματα που εγείρονται από τον σχεδιασμό, την παραγωγή και τη χρήση αυτόνομων οπλικών συστημάτων σχετίζεται:

1. Με το αν θα πρέπει ή όχι να υπάρχουν τέτοια συστήματα, πρόβλημα που συνδέεται με: 1.1 τη σκοπιμότητα και τη χρήση τους – ενδεχομένως την κακόβουλη χρήση τους και 1.2. με το οντολογικό καθεστώς τους, καθώς η αυτονομία, η νοημοσύνη και η αποτελεσματικότητά τους καθιστούν: 1.2.1 τη δράση τους ηθικώς αξιολογήσιμη και 1.2.2 τα ίδια τα συστήματα ενδεχομένως ηθικά υπεύθυνα.
2. Με το ερώτημα: με ποια κριτήρια εξακριβώνει κάποιος αν το σύστημα έδρασε τελικά αυτόνομα και ως ηθικό πρόσωπο;

Στην παρούσα διερεύνηση επικεντρωνόμαστε στο ανωτέρω επιστημολογικό – γνωσιολογικό ερώτημα (2) η απάντηση του οποίου, ωστόσο, σχετίζεται άρρηκτα με το (1.2), με τη θεώρηση δηλαδή του οντολογικού καθεστώτος και των κριτηρίων που απαιτούνται για να θεωρηθεί κάποιος ή κάτι ως ηθικό πρόσωπο.

Λόγω της εννοιολογικής ασάφειας αλλά και των διαφορών χρήσης των ίδιων λεκτικών όρων μεταξύ του φιλοσοφικού και του τεχνικού λεξιλογίου, θεωρούμε σκόπιμο να γίνουν ορισμένες πρόσθετες εισαγωγικές διευκρινίσεις:

Μιλώντας για ευφυείς αυτόνομες στρατιωτικές μηχανές αναφερόμαστε κυρίως στα Αυτόνομα Οπλικά Συστήματα (AWS και LAWS) (Ministry of Defense, 2011). Αυτά τα συστήματα, σύμφωνα με τον ορισμό του Υπουργείου Αμύνης του

Ηνωμένου Βασιλείου, είναι ικανά να «κατανοήσουν» οδηγίες, προθέσεις, περιβάλλοντα κ.ά. και αφού εξετάσουν τις εναλλακτικές επιλογές, αποφασίζουν αυτόνομα και προβαίνουν σε πράξεις που δεν είναι δυνατόν να προβλεφθούν εκ των προτέρων.

Εν προκειμένω, αυτό που υποστηρίζεται ότι κάνει τις πολεμικές μηχανές να «αντιλαμβάνονται», να «κατανοούν», να αποφασίζουν και να δρουν μόνες, αξιοποιώντας και αξιολογώντας ένα σύνολο σύνθετων πληροφοριών προκειμένου να επιτύχουν μια συγκεκριμένη αποστολή είναι η Τεχνητή Νοημοσύνη (Singer, 2009, σ. 145)[4]. Παρότι οι φιλόσοφοι διαφωνούν ως προς τον ακριβή ορισμό της νοημοσύνης, θα μπορούσαμε να δεχθούμε ότι με τον όρο αυτόν εννοούμε την ικανότητα μιας οντότητας για επίτευξη πολύπλοκων στόχων (Tegmark, 2017, σ. 84). Πρόκειται δηλαδή για μια υπολογιστική διαδικασία κατά την οποία επιτυγχάνεται μετασχηματισμός της πληροφορίας μέσω συναρτήσεων (Tegmark, 2017, σ. 100). Σύμφωνα με τον Haugland, ωστόσο (Haugland, 1985, σσ. 11, 349), οι ερευνητές και οι προγραμματιστές της Τεχνητής Νοημοσύνης στοχεύουν στη δημιουργία μιας γνήσιας νοημοσύνης και όχι σε μια απομίμηση νοημοσύνης που απλώς προσομοιάζει στην ανθρώπινη. Με την έννοια αυτήν, οι ερευνητές επιχειρούν να φτιάξουν μια μη βιολογική νοημοσύνη η οποία θα διαθέτει τα χαρακτηριστικά των νοημόνων όντων. Στην πραγματικότητα επιχειρούν να κατασκευάσουν μηχανές με «νόηση» οι οποίες θα είναι ικανές για «νοημοσύνη» (Gounaris, 2013)[5]. Η θέση αυτή ξεκινά από

[4] Σύμφωνα με τον Singer (2009), για να θεωρηθεί μια μηχανή πλήρως αυτόνομη, θα πρέπει να διαθέτει νοημοσύνη, δηλαδή θα πρέπει να αντιλαμβάνεται και να είναι σε θέση να χρησιμοποιεί σύνθετες πληροφορίες προκειμένου να επιτύχει συγκεκριμένους στόχους η επίτευξη των οποίων απαιτεί λήψη αποφάσεων.

[5] Η πλειονότητα των ερευνητών της Τ.Ν. εξισώνουν τις έννοιες Νόηση (Cognition) και Νοημοσύνη (Intelligence). Σύμφωνα με τη θέση μας, η μεν νοημοσύνη μπορεί να οριστεί ως η ικανότητα επίτευξης πολύπλοκων στόχων και είναι άρρηκτα συνδεδεμένη με την υπολογιστική ικανότητα, η δε νόηση ορίζεται ως η ικανότητα του νοήμονος όντος να μαθαίνει, να αντιλαμβάνεται και να κατανοεί, να προβαίνει σε αξιολογικές κρίσεις και να λαμβάνει αποφάσεις, να νοηματοδοτεί τον κόσμο του κ.λπ., διαδικασίες δηλαδή που δεν συνδέονται απαραίτητα με την υπολογιστική ικανότητα.

την παραδοχή ότι και ο ανθρώπινος εγκέφαλος δεν είναι τίποτε άλλο παρά μια βιολογική υπολογιστική μηχανή η οποία παράγει την ανθρώπινη νόηση και έχει την ικανότητα να επιτυγχάνει πολύπλοκους στόχους, δηλαδή να διαθέτει νοημοσύνη.

Η ανθρωπομορφική θεώρηση της τεχνητής νοημοσύνης καθώς και η μηχανιστική θεώρηση της ανθρώπινης νόησης οριοθετεί την έρευνα και τη συζήτηση σε καθορισμένα γλωσσικά όρια (ψυχολογικό και μηχανιστικό λεξιλόγιο) τα οποία γίνονται αντιληπτά και στον τρόπο με τον οποίο ορίζουμε τις ικανότητες και τις λειτουργίες των αυτόνομων συστημάτων. Για παράδειγμα, λέμε ότι το σύστημα τεχνητής νοημοσύνης σκέφτεται, κατανοεί κ.λπ. ή ότι ο εγκέφαλος πραγματοποιεί αλγοριθμικούς υπολογισμούς. Σε αυτές τις περιπτώσεις κάνουμε μεταφορική χρήση της γλώσσας δανειζόμενοι όρους από διαφορετικά επιστημονικά λεξιλόγια, με αποτέλεσμα το προσωρινό δάνειο από το ένα γλωσσικό παιχνίδι να καθιερώνεται με άλλο νόημα εντός ενός διαφορετικού γλωσσικού παιχνιδιού. Καθώς, λοιπόν, οι έννοιες νόηση, νοημοσύνη, συνείδηση κ.ά. παραμένουν νεφελώδεις και ακαθόριστες ενώ φαίνεται να χρησιμοποιούνται με πολλούς διαφορετικούς τρόπους τόσο από τους φιλοσόφους όσο και τους τεχνικούς της Τ.Ν., η οντολογική τους διαλεύκανση καθίσταται ιδιαίτερα επίπονη (De Quincey, 2006; Levy, 2009; Sloman, 1996)[6], με αποτέλεσμα, όπως θα δούμε πιο κάτω, οι περισσότεροι στοχαστές να στρέφονται στη διατύπωση συμπεριφορικού τύπου ενδείξεων και τελικά συμπεριφορικών

[6] Μάλιστα, όπως τονίζουν οι Hoffmann και Hahn, αυτή η ασάφεια στον ορισμό της νοημοσύνης οδηγεί αντίστοιχα σε μια ασάφεια ως προς τον χαρακτηρισμό μιας μηχανής ως συστήματος Τ.Ν. (Hoffmann & Hahn, 2019). Πράγματι, μοιάζει πρακτικά αδύνατον να γνωρίζεις αν πρέπει να χαρακτηρίσεις μια μηχανή ως «σύστημα Τεχνητής Νοημοσύνης» δίχως πρώτα να έχεις έναν σαφή ορισμό του όρου «νοημοσύνη». Υπό αυτή την έννοια, η εννοιολογική ασάφεια του όρου «νοημοσύνη» οδηγεί και σε μιαν ασάφεια ως προς τον καθορισμό των ορίων του συνόλου οντοτήτων στα οποία αποδίδουμε τον χαρακτηρισμό «Τεχνητή Νοημοσύνη» και πρέπει να τονιστεί ότι αυτό έχει ήδη ως μια πρώτη σοβαρή συνέπεια να μην μπορούμε να ορίσουμε επακριβώς το σύνολο των τεχνολογικών εφαρμογών που εμπίπτουν στο πεδίο ανάλυσης της Ηθικής της Τ.Ν. Είναι, άλλωστε, χαρακτηριστική της σοβαρότητας της όλης κατάστασης η διατύπωση του *Turing-Red-Flag-Law* (Walsh, 2017), η οποία ουσιαστικά εκφράζει το αίτημα όλα τα συστήματα Τ.Ν. να είναι πράγματι αναγνωρίσιμα ως τέτοια.

κριτηρίων νοημοσύνης[7]. Αυτή τη στροφή προς συμπεριφορικού τύπου κριτήρια φαίνεται να κάνει, έστω εν μέρει, όπως θα δούμε[8] και ο Daniel Dennett, υπερασπιζόμενος την «ανθρώπινη» συμπεριφορά του HAL 9000.

Σε αντίστοιχες με τη νοημοσύνη γλωσσικές περιπέτειες, όπως θα δούμε αναλυτικότερα στη συνέχεια, έχει περιέλθει και η έννοια της αυτονομίας, καθώς χρησιμοποιείται με διαφορετικό τρόπο από τους ηθικούς φιλοσόφους και με διαφορετικό τρόπο από τους σχεδιαστές και μηχανικούς της Τεχνητής Νοημοσύνης[9]. Για τους καντιανούς ηθικούς φιλοσόφους η αυτονομία αποτελεί τη βάση για την ηθική ευθύνη και την ιδιότητα του προσώπου (Christman, 2018) και είναι συνδεδεμένη με την ελεύθερη βούληση και τον αυτοπεριορισμό, δηλαδή

[7] Η πρώτη στροφή στην εξεύρεση συμπεριφορικών κριτηρίων έγινε από τον Descartes με την από μέρους του εισήγηση του κριτηρίου της Γλώσσας αλλά και του κριτηρίου της επιτυχούς δράσης-μέσα-στον-Κόσμο (Erion, 2001; Gunderson, 1964; Savova & Peshkin, 2007). Κατά τον 20ό αιώνα αυτή η στροφή προς συμπεριφορικού τύπου κριτήρια σημαδεύτηκε από την από μέρους του Turing εισήγηση του «Παιχνιδιού της Μίμησης» –γνωστού πλέον ως Turing Test ή «Δοκιμασία Turing» (Turing, 1950)–, με τις προθέσεις του Turing όμως να είναι εκ διαμέτρου αντίθετες από αυτές του Descartes, καθώς ο πρώτος στρέφεται στη συμπεριφορά για να υποστηρίξει μια οντολογική εξίσωση ανθρώπων-μηχανών, ενώ ο δεύτερος για να υποστηρίξει την οντολογική τους διάκριση.

[8] Δείτε σχετικώς την ενότητα περί του κριτηρίου της υπέρμετρης αποτελεσματικότητας.

[9] Ο «τεχνικός» τρόπος χρήσης του όρου «αυτονομία» αναφέρεται συνήθως σε μεγάλο χρονικό διάστημα μεταξύ δύο διαδοχικών ενεργειακών εφοδιασμών, ενώ στη περίπτωση των οπλικών συστημάτων σημαίνει ότι το όπλο έχει την «ικανότητα fire and forget», δηλαδή τη δυνατότητα να διατηρεί την εστίαση και στόχευσή του στο επιλεγμένο από τον άνθρωπο-χειριστή στόχο δίχως ο χειριστής να πρέπει να παρεμβαίνει διαρκώς. Αντιθέτως, η φιλοσοφική εκφορά του όρου «αυτονομία» είναι άρρηκτα συνδεδεμένη με την ηθική ευθύνη και ταυτόχρονα είναι φορτισμένη με ένα πλήθος πλούσιων οντολογικών συμφραζομένων που, όπως θα δούμε πιο κάτω, φτάνουν ως και την έννοια της νόησης. Συχνά συμβαίνει οι ερευνητές της Τ.Ν. να εκκινούν την αναφορά τους στην «αυτονομία» των μηχανών διά του «τεχνικού» τρόπου, αλλά στην πορεία να το λησμονούν και να διεκδικούν για τις μηχανές αυτές ό,τι θα επέτασσε μια φιλοσοφική εκφορά του εν λόγω όρου. Έτσι, λόγω μιας κατά τον Wittgenstein «παραπλανητικής αναλογίας», μιας ομοιότητας στην επιφανειακή γραμματική των δύο αυτών τρόπων εκφοράς του όρου «αυτονομία», φτάνουν να υποστηρίζουν και μια ομοιότητα στην γραμματική βάθους, δηλαδή στο νόημα. Πρέπει, βέβαια, να πούμε προκαταβολικά ότι ο Dennett, το επιχείρημα του οποίου θα εξετάσουμε, δεν υποπίπτει σε ένα τέτοιο σφάλμα και χρησιμοποιεί τον όρο «αυτονομία» με τον φιλοσοφικό τρόπο χρήσης. Ωστόσο, αν το επιχείρημά του αποδειχθεί ανεπαρκές, ο μόνος τρόπος κατά τον οποίο θα μπορεί ίσως να σταθεί η χρήση του εν λόγω όρου ως προς το συστήματα της Τ.Ν. θα είναι εν τέλει ο «τεχνικός».

την ικανότητα και τη δυνατότητα του προσώπου να οριοθετεί το ίδιο τις πράξεις του. Για να έχει ηθική ευθύνη ένα πρόσωπο, θα πρέπει να είναι αυτόνομο ή εν πάση περιπτώσει να μην τελεί υπό το κράτος εξαναγκασμού. Αυτό σημαίνει ότι θα πρέπει να είναι ελεύθερο από εξωγενείς παράγοντες που του επιβάλλουν να πράξει με ορισμένο τρόπο (φερ' ειπείν, να μην απειλείται με το πιστόλι στον κρόταφο) ή να μην περιορίζεται από μη ελεγχόμενους εσωτερικούς παράγοντες που καθορίζουν την απόφασή του (για παράδειγμα, να μη βρίσκεται υπό την επήρεια κάποιου φαρμάκου ή σε κάποια μη ελεγχόμενη από το ίδιο νοητική κατάσταση). Η απόφαση, δηλαδή, που οδηγεί ένα πρόσωπο σε μια συγκεκριμένη πράξη θα πρέπει να καθορίζεται από το ίδιο το πρόσωπο με έλλογο τρόπο (Buss & Westlund, 2018). Η αυτονομία, ορισμένως, αποτελεί προαπαιτούμενο της ηθικής ευθύνης με τρόπο τέτοιο, που η ηθική ευθύνη συνεπάγεται αυτονομία. Όμως, όπως παρατηρεί ο Müller (2020), η σχέση αυτή δεν έχει και αντίστροφη συνεπαγωγή. Τα τεχνικώς νοούμενα αυτόνομα συστήματα δεν σημαίνει απαραίτητα ότι έχουν ηθική ευθύνη για τις πράξεις τους. Σύμφωνα με αυτήν την τεχνική και λιγότερο «δεσμευτική» έννοια της αυτονομίας, ένα μηχανικό σύστημα (ευφυές ή μη) θεωρείται αυτόνομο σε σχέση με τον βαθμό ελέγχου του από τον ανθρώπινο παράγοντα (Müller, 2012)[10]. Η λιγότερο δεσμευτική αυτή έννοια της αυτονομίας αφήνει ανοιχτό το ερώτημα ως προς το ποιος έχει τελικά τον έλεγχο του συστήματος και ποιος έχει την ηθική ευθύνη. Πρόκειται για το πρόβλημα που στην ηθική ονομάζουμε «Responsibility Gap» και το οποίο συναντάμε σε σύνθετες καταστάσεις (π.χ. στην οικονομία και τις επιχειρήσεις, στον πόλεμο, στις διεθνείς σχέσεις κ.α.), όπου η κρινόμενη πράξη,

[10] Ο βαθμός αυτονομίας των πολεμικών μηχανών καθορίζεται από τον βαθμό εμπλοκής του ανθρώπινου παράγοντα στην επιχειρησιακή δράση τους. Έτσι, διακρίνουμε τις περιπτώσεις που ο άνθρωπος: α. «ελέγχει» τη λειτουργία των μηχανών τεχνητής νοημοσύνης (βρίσκεται δηλαδή «in the loop») ή β. παρακολουθεί τη λειτουργία τους και ενδεχομένως αποφασίζει για την τελική δράση τους (συνθήκες «on the loop») ή γ. δεν παρεμβαίνει στην επιχειρησιακή δράση, με αποτέλεσμα οι μηχανές να λειτουργούν εν πλήρει αυτονομία (συνθήκες «human off the loop»). Η διευκρίνιση του βαθμού αυτονομίας των πολεμικών μηχανών είναι σημαντικός παράγοντας για την ηθική αξιολόγηση της δράσης τους (Welsh, 2017), εδώ ωστόσο εξετάζουμε την περίπτωση πλήρους αυτονομίας, χωρίς τη δυνατότητα ανθρώπινης παρέμβασης (συνθήκες off the loop).

ενώ προϋποθέτει τη συμμετοχή πολλών ανθρώπων ή φορέων σε προγενέστερο από την πράξη στάδιο, εν τέλει δεν μπορεί να προβλεφθεί ή να ελεγχθεί με ακρίβεια στα προηγούμενα στάδια (Matthias, 2004). Στην αυτόνομη T.N., για παράδειγμα, εγείρονται ερωτήματα ως προς το μερίδιο ευθύνης που αντιστοιχεί –αν αντιστοιχεί– στους προγραμματιστές, στους υπεύθυνους ανάπτυξης λογισμικού, στους σχεδιαστές των εφαρμογών, στους χρηματοδότες της έρευνας, στην εταιρεία που κατασκεύασε το σύστημα T.N. κ.λπ., ακόμα και στους τελικούς χρήστες.

Για τον Dennett ωστόσο, όπως θα δούμε στη συνέχεια, ένα σύστημα T.N. που δρα αυτόνομα και αποτελεσματικά, εφόσον επιδεικνύει νοήμονα συμπεριφορά αντίστοιχη με την ανθρώπινη συμπεριφορά (στους συγκεκριμένους στόχους), μπορεί να αξιολογηθεί ηθικά όπως κάθε άλλο ηθικό πρόσωπο. Ο Dennett στο κλασικό πλέον άρθρο του με τίτλο "When Hal Kills, Who's to Blame? Computer Ethics", το οποίο, σύμφωνα με τον Sparrow (2007), αποτελεί την πιο σοβαρή σύγχρονη φιλοσοφική υπεράσπιση της θέσης ότι οι μηχανές θα μπορούσαν να θεωρηθούν υπεύθυνες για τις πράξεις τους, χτίζει το επιχείρημά του επικαλούμενος αρχικά τη μνημειώδη πρώτη νίκη στο σκάκι του υπολογιστή της IBM, Deep Blue, επί του παγκοσμίου πρωταθλητή Gary Kasparov το 1996. Συγκεκριμένα, υποστηρίζει ότι αναγνωρίζουμε και θαυμάζουμε την ικανότητα του υπολογιστή να κερδίζει στο σκάκι και συγχαίρουμε τους προγραμματιστές του για το επίτευγμα, όμως η συγκεκριμένη νίκη ανήκει στον υπολογιστή και όχι στους προγραμματιστές. Οι προγραμματιστές, αν αντιμετώπιζαν τον παγκόσμιο πρωταθλητή, προφανώς θα έχαναν από αυτόν μέσα σε λίγα λεπτά. Η ευθύνη που αναλογεί στους προγραμματιστές για τη νίκη του Deep Blue είναι αντίστοιχη με την ευθύνη που αναλογεί στον προπονητή ή στον δάσκαλο του Kasparov, όμως τελικά την «ευθύνη» για το αποτέλεσμα της αναμέτρησης έχουν οι ίδιοι οι παίκτες, και συγκεκριμένα ο Kasparov και ο Deep Blue.

Το επιχείρημα του Dennett έχει εξαιρετικά επίκαιρη σημασία, αν αναλογιστεί κανείς δυο σημαντικά επιτεύγματα της T.N. που σηματοδοτούν ουσιαστικά ένα μέλλον προς την κατεύθυνση που μας απασχολεί. Το πρώτο έχει να

κάνει με τις συνεχόμενες νίκες του συστήματος T.N. με την ονομασία AlphaGo, που κατασκεύασε η εταιρεία DeepMind της Google, το 2016 επί του Lee Sedol, παγκόσμιου πρωταθλητή κι ενός από τους σημαντικότερους παίκτες όλων των εποχών στο παιχνίδι GO. Ο Sedol εγκαταλείποντας την ενεργό δράση μετά τις ήττες του παραδέχτηκε ότι η TN είναι πλέον ανίκητη (Vincent, 2019). Η ιδιαιτερότητα του GO, σε αντίθεση με το σκάκι, είναι ότι δεν βασίζεται μόνο στην υπολογιστική ικανότητα των παιχτών αλλά και σε πιο σύνθετες γνωσιακές δεξιότητες, κάνοντας πολλούς να μιλούν ότι πρόκειται στην πραγματικότητα για ένα είδος τέχνης[11]. Το δεύτερο επίτευγμα αφορά στην ολική επικράτηση του συστήματος T.N. της DeepMind σε εικονικές αερομαχίες, το 2020, με κορυφαίους πιλότους της αεροπορίας των Ηνωμένων Πολιτειών με μαχητικά F16 Viper (Lang, 2020). Η ιδιαιτερότητα αυτής της νίκης έγκειται στο γεγονός ότι, εκτός από υπολογιστικές ικανότητες, απαιτούνται αντίληψη του τρισδιάστατου χώρου, σωματικές δεξιότητες και παραπλανητικές κινήσεις.

Ο Dennett επεκτείνει τον συλλογισμό για την απόδοση ευθυνών περνώντας από τον Deep Blue αναλογικά στον HAL 9000, έναν Ευρετικά Προγραμματισμένο Αλγοριθμικό Υπολογιστή[12] ο οποίος είναι ο βασικός πρωταγωνιστής της βασισμένης σε νουβέλα του Arthur Clarke (1968) ταινίας του Stanley Kubrick *2001, Η Οδύσσεια του Διαστήματος*. Ο HAL διαθέτει απείρως μεγαλύτερες υπολογιστικές δυνατότητες από τον Deep Blue, λειτουργεί «αυτόνομα» και προβαίνει σε πράξεις ζωής και θανάτου αφού, προκειμένου να διασφαλίσει την επιτυχία της αποστολής που έχει αναλάβει, όταν καταλαβαίνει ότι αυτή κινδυνεύει, αποφασίζει να σκοτώσει το πλήρωμα του διαστημόπλοιου μέσα στο οποίο ήταν εγκατεστημένος και να αποκτήσει τον πλήρη έλεγχο. Ο Dennett

[11] Όπως σημειώνει ο Tegmark (2017, σ. 139) υπάρχουν πολύ περισσότερες δυνατές θέσεις στο GO απ' όσα τα άτομα στο σύμπαν, γεγονός που σημαίνει ότι κανένα υπολογιστικό σύστημα δεν μπορεί να αναλύσει όλες τις ενδιαφέρουσες αλληλουχίες μελλοντικών κινήσεων.

[12] Ως «Ευρετικοί Αλγόριθμοι» (Heuristic Algorithms) ορίζονται οι ευρετικές (ή ευριστικές) υπολογιστικές τεχνικές που για οικονομία χρόνου αξιολογούν και προκρίνουν ενδιάμεσες λύσεις απορρίπτοντας τις υπόλοιπες. Στην T.N., παρότι οι τεχνικές αυτές κωδικοποιούνται αλγοριθμικά, δεν θεωρούνται «ακριβώς» αλγόριθμοι, καθώς οι αλγόριθμοι οδηγούν πάντα σε ακριβή αποτελέσματα, ενώ οι μηχανισμοί αυτοί προσομοιάζουν περισσότερο στην ανθρώπινη «διαισθητική» σκέψη και τη «σταθμισμένη εικασία».

αποδίδει στον HAL χαρακτηριστικά ηθικού προσώπου, επειδή αυτή η αυτόνομη ευφυής μηχανή επιδεικνύει ανθρώπινη συμπεριφορά ανεξάρτητα με το αν μετανιώνει, αισθάνεται τύψεις, συναισθάνεται ή κατανοεί τι σημαίνει να είναι κανείς ηθικό πρόσωπο.

Κατά την άποψή μας, όμως, τα επιχειρήματα που προβάλλει ο Dennett δεν αποδεικνύουν επαρκώς τη θέση ότι μπορεί να αποδοθεί στον HAL καθεστώς ηθικού προσώπου που φέρει την ευθύνη των πράξεών του.

Το Επιχείρημα της Ισοδυναμίας

Αρχικώς, η εικαζόμενη ισοδυναμία της σχέσης του Kasparov με τον προπονητή του και του Deep Blue με τους προγραμματιστές του δεν προκύπτει λογικά. Συγκεκριμένα, η ισοδυναμία αυτή δύναται να υποστηριχθεί με δύο τρόπους: α. ο υπολογιστής είναι οντολογικά ισοδύναμος με τον άνθρωπο-αθλητή ή β. ο υπολογιστής δεν είναι απαραίτητα οντολογικά ισοδύναμος με τον άνθρωπο-αθλητή, αλλά η σχέση «προγραμματιστή-υπολογιστή» είναι λειτουργικά ισοδύναμη με τη σχέση «προπονητή-αθλητή», δηλαδή αμφότερες οι δύο αυτές σχέσεις δύναται να περιγραφούν με κοινούς λειτουργικούς όρους. Υπό άλλη διατύπωση, η μελέτη αμφότερων των εν λόγω σχέσεων σε επίπεδο λειτουργιών δύναται να οδηγήσει σε μια ταυτόσημη περιγραφή: οι δύο σχέσεις ανάγονται στο ίδιο σύνολο επιτελούμενων λειτουργιών.

Στην περίπτωση του α., δηλαδή στην περίπτωση όπου υποστηρίζει κανείς ότι ο υπολογιστής είναι οντολογικά ισοδύναμος με έναν άνθρωπο-αθλητή, διαπράττεται ουσιαστικά ένα σφάλμα διαλληλίας, καθώς εν τέλει φτάνουμε να θεωρούμε ως δεδομένο αυτό που επιχειρούμε να αποδείξουμε. Σχετικώς, το να λέμε «Η ευθύνη που αναλογεί στους προγραμματιστές για τη νίκη του Deep Blue είναι ισοδύναμη με αυτή που αντιστοιχεί στον προπονητή ή στον δάσκαλο του Kasparov», βασιζόμενοι στην παραδοχή ότι ο υπολογιστής είναι οντολογικά ισοδύναμος με τον Kasparov, λαμβάνει ως δεδομένο το ζητούμενο, που δεν είναι άλλο παρά η απόδειξη αυτής της ισοδυναμίας. Θα περίμενε κανείς να δώσουμε επαρκείς λόγους που θα καταδεικνύουν αυτή την οντολογική

ισοδυναμία ανθρώπου-υπολογιστή, αντί να προβούμε απλώς σε μια καταφατική δήλωση που καταλήγει να είναι ουσιαστικά μια ταυτολογία, επομένως μια πρόταση δίχως πραγματικό «επιστημολογικό φορτίο»[13]. Αναλυτικότερα, το να λέμε ότι «Η νίκη ανήκει στον υπολογιστή, επειδή η σχέση 'προγραμματιστή-υπολογιστή' είναι ίδια με τη σχέση 'προπονητή-αθλητή'» και ότι «Η σχέση 'προγραμματιστή-υπολογιστή' είναι ίδια με τη σχέση 'προπονητή-αθλητή', επειδή ο υπολογιστής και ο αθλητής είναι οντολογικά ισοδύναμοι» είναι σαν να λέμε ότι «Ο υπολογιστής και ο αθλητής είναι οντολογικά ισοδύναμοι, επειδή είναι οντολογικά ισοδύναμοι». Ο μόνος τρόπος να ξεφύγει κανείς από αυτή την ταυτολογία είναι: α1. Να αντιμετωπίσει τελικά το πρόβλημα κατά μέτωπο επιχειρώντας να απαντήσει στο ερώτημα: Υπό ποια κριτήρια μπορούμε να εδραιώσουμε μια οντολογική ισοδυναμία ή διάκριση ανθρώπου-μηχανής; Πρόκειται για το πλέον κεντρικό, διαχρονικό και επίπονο φιλοσοφικό ερώτημα της Τ.Ν. α2. Να επιχειρήσει να απεμπλέξει τη συζήτηση απόδοσης ευθύνης και, τελικά, ηθικού καθεστώτος από το ζήτημα της οντολογικής ισοδυναμίας ή διάκρισης ανθρώπου-μηχανής. Αλλά πόσο εύκολο είναι στη σκέψη μας να διαχωρίσουμε αυτά τα δύο; Τι άλλο θα μπορούσε να αποτελέσει επαρκές κριτήριο απόδοσης ηθικού καθεστώτος σε μιαν οντότητα, πλην της οντολογίας της τελευταίας αυτής; Υπάρχουν παραδείγματα αποδεκτής ανθρώπινης σκέψης κατά την οποία η απόδοση ηθικού καθεστώτος και η οντολογία δεν συσχετίστηκαν με τον έναν ή τον άλλο τρόπο; Όλα τα ερωτήματα απόδοσης ηθικού καθεστώτος καταλήγουν σύντομα σε ερωτήματα απόδοσης οντολογικού καθεστώτος.

Στην περίπτωση του β., δηλαδή στην περίπτωση κατά την οποία θα επιχειρούσαμε να αποδώσουμε το ίδιο ηθικό καθεστώς τόσο στον Kasparov όσο και στον Deep Blue επί τη βάσει μιας λειτουργικής ισοδυναμίας των σχέσεων «προπονητή-αθλητή» και «προγραμματιστή-υπολογιστή», καλούμαστε να αποδείξουμε

[13] Επιπλέον, θα πρέπει κανείς να εξηγήσει υπό ποιους όρους θεωρούνται δύο οντότητες οντολογικά ισοδύναμες και να αιτιολογήσουμε επαρκώς τους όρους αυτούς. Επί παραδείγματι, θα μπορούσαμε να προτείνουμε λειτουργιστικούς όρους, αλλά τότε θα έπρεπε, για να αιτιολογήσουμε την επιλογή μας, να προβούμε σε μια λειτουργιστική περιγραφή. Επιπλέον, η λειτουργιστική περιγραφή θα μας φέρει αντιμέτωπους με τα προβλήματα που αναλύονται πιο κάτω σε σχέση με το β.

αυτήν ακριβώς τη λειτουργική ισοδυναμία των δύο σχέσεων είτε β1. μέσω μιας λειτουργικής ισοδυναμίας «προπονητή-προγραμματιστή» και «αθλητή-υπολογιστή» (σε αυτή την περίπτωση, η σχέση ισοδυναμίας των ζευγών «προπονητής-αθλητής» και «προγραμματιστής-υπολογιστής» εδραιώνεται μέσα από την κατάδειξη των σχέσεων ισοδυναμίας των αντίστοιχων μελών που απαρτίζουν αυτά τα ζεύγη)[14] είτε επειδή β2. τα αντίστοιχα μέλη των ζευγών δεν είναι λειτουργικώς ανάλογα, αλλά τα ζεύγη που τα εν λόγω μέλη απαρτίζουν τυχαίνει να είναι (σε αυτή την περίπτωση η αναλογία δεν έγκειται στα μέλη αλλά στις σχέσεις που αυτά συνάπτουν μεταξύ τους)[15].

Εξάλλου, μπροστά στην προοπτική των αυτο-προγραμματιζόμενων αλλά και αυτο-αναπαραγόμενων μηχανών η επιχειρηματολογία που θα βασίζεται στον παραλληλισμό προγραμματιστών-προπονητών και μηχανών-αθλητών ακυρώνεται εκ των πραγμάτων, καθότι ο ρόλος του ανθρώπου-προγραμματιστή καθίσταται περιττός[16].

Ας δούμε, ωστόσο, αναλυτικότερα τα προβλήματα που προκύπτουν από την προσπάθεια απόδειξης μιας λειτουργικής ισοδυναμίας. Σχετικά με το β1. πρέπει να τονίσουμε ότι η απόδειξη μιας λειτουργικής ισοδυναμίας αθλητή-υπολογιστή συνεπάγεται την απόδειξη μιας λειτουργικής ισοδυναμίας ανθρώπου-μηχανής, που με τη σειρά της δεν έχει μέχρι στιγμής καταστεί δυνατή. Η πλέον γνωστή και οργανωμένη προσπάθεια εδραίωσης μιας οντολογικής ισοδυναμίας ανθρώπου-μηχανής, η θεωρία του Λειτουργισμού και δη του Λειτουργισμού Μηχανής, έχει παρουσιάσει σημαντικά προβλήματα, κάποια εκ των οποίων εντοπίζονται ήδη στη θεμελιακή παραδοχή της εν λόγω θεωρίας, δηλαδή στη θέση ότι σκέψη = υπολογισμός. Πρόκειται για μια θέση της οποίας

[14] Παραδείγματος χάριν: α – β = γ – δ, επειδή α = γ και β = δ.

[15] Παραδείγματος χάριν: α ≠ γ και β ≠ δ, αλλά α – β = γ – δ.

[16] Για μια ενδιαφέρουσα ανάλυση των φιλοσοφικών επιπτώσεων που προκύπτουν από την πιθανή ανάπτυξη αυτο-προγραμματιζόμενων και αυτο-αναπαραγόμενων μηχανών, ανατρέξτε στους von Neumann (1966) και Rucker (1982, ειδικότερα στο Chapter 4 *Robots and Souls*). Η ιδέα των αυτο-αναπαραγόμενων μηχανών δεν είναι καινούργια. Μια από τις πρώτες τεχνικές αναλύσεις της δυνατότητας αυτο-αναπαραγόμενων μηχανών πραγματοποίησε ο Moore (1956, σσ. 118-126).

η απόδειξη δεν έχει ως τώρα επιτευχθεί, καθώς, πέραν της νεφελώδους εικόνας που έχουμε για την οντολογία του Νου, υφίστανται σημαντικά και εξακριβωμένα εμπόδια στη φύση του υπολογισμού στην οποία ενέχεται το άπειρο. Αυτό ακριβώς το πρόβλημα της πιθανότητας ενός ατέρμονου υπολογισμού ανέδειξε ο ίδιος ο Turing (στη θεωρητική μηχανή του οποίου βασίζεται εν πολλοίς ο Λειτουργισμός) με την απόδειξη του Θεωρήματος Τερματισμού (Turing, 1937, 1938)[17]. Πέραν όμως της μη υπολογισιμότητας, ο Λειτουργισμός υποπίπτει αναπόφευκτα σε Σφάλμα Διαλληλίας, καθώς αδυνατεί να ορίσει τις όποιες λειτουργίες δίχως να αναφερθεί σε νοητικούς όρους, συνεπώς δίχως να προϋποθέσει τη νοητικότητα. Έτσι, καταλήγει να προσπαθεί να εδραιώσει τη δυνατότητα νόησης στις μηχανές ουσιαστικά προϋποθέτοντάς τη (Kim, 2005, σσ. 151-153).

Εκτός από τα συγκεκριμένα λογικά σφάλματα, μια λειτουργιστική προσπάθεια απόδειξης της ως άνω ισοδυναμίας βρίσκεται αντιμέτωπη και με δύο σοβαρά οντολογικά προβλήματα που καλούνται να επιλύσουν εν γένει οι λειτουργιστές. Το πρώτο είναι ότι η λειτουργιστική περιγραφή αγνοεί –ή και αδυνατεί να περιγράψει– την *ποιοτική* και *φαινόμενη* πτυχή των νοητικών συμβάντων που ονομάζουμε qualia. Εστιάζοντας αποκλειστικά στη σχέση εισόδων-εξόδων ενός συστήματος (ανθρώπου, ζώου, μηχανής κ.λπ.), ο Λειτουργισμός αφήνει ανοικτή μια αρκετά παράδοξη πιθανότητα: δύο συστήματα δύνανται να έχουν ακριβώς την ίδια αντιστοίχιση εισόδων (ερεθισμάτων) και εξόδων (συμπεριφορικών εκδηλώσεων), αλλά εντελώς διαφορετικά ή και αντεστραμμένα qualia, δηλαδή να βιώνουν εντελώς διαφορετικές ή και αντεστραμμένες «εσωτερικές

[17] Για μια εύληπτη όσο και αναλυτική παρουσίαση του ζητήματος της *μη υπολογισιμότητας* καθώς και για τις προεκτάσεις του ως προς την T.N. δείτε: Casti (2000), Dreyfus (1992, κυρίως τα κεφάλαια 5 και 10), Rucker (1982, ειδικότερα στο Chapter 4 *Robots and Souls*) και Lucas (1961). Για την αισιόδοξη και, τελικά, αντίθετη των Dreyfus και Lucas προσέγγιση, δείτε Benacerraf (1967). Να σημειωθεί ότι ένας ελιγμός των λειτουργιστών, για να διαφύγουν από το εν λόγω αδιέξοδο της μη υπολογισιμότητας, είναι η υποστήριξη της θέσης ότι η νοημοσύνη θα μπορούσε να αναπαραχθεί πλήρως από μια κατάλληλα «πολύπλοκη» Μηχανή Turing. Η θέση αυτή ωστόσο, εκτός του ότι είναι ανάλογη της Church-Turing thesis και αποτελεί μια μη αποδεδειγμένη θέση, γεννά ένα νέο πρόβλημα, καθώς όπως σημειώνει ο Jaegwon Kim (2005, κεφ. 4), οι λειτουργιστές καλούνται τώρα να προσδιορίσουν τι εστί πολυπλοκότητα και ποιο είναι το κατάλληλο όριο πολυπλοκότητας πέραν του οποίου μια Μηχανή Turing επιτυγχάνει να επιδείξει νοημοσύνη.

καταστάσεις». Είναι, δε, πιθανό τα qualia να απουσιάζουν εντελώς από το ένα εκ των δύο συστημάτων (Block, 1980a, b; Kim, 2005, σ. 190)[18]. Το παράδοξο εδώ έγκειται στο ότι, σύμφωνα με τον Λειτουργισμό Μηχανής, τα δύο αυτά συστήματα θεωρούνται λειτουργικώς ισοδύναμα παρά τη διαφοροποίηση τους στο επίπεδο των qualia[19]. Το δεύτερο οντολογικό πρόβλημα εστιάζει στο κατά πόσον και αν μια ευφυής μηχανή ή ένα ευφυές λειτουργικό σύστημα «κατανοεί» ή αντιλαμβάνεται το νόημα της υπολογιστικής διαδικασίας και του αποτελέσματος το οποίο εξάγει. Η πιο δημοφιλής περιγραφή αυτού του προβλήματος έχει διατυπωθεί από τον John Searle με το «Επιχείρημα του Κινέζικου Δωματίου». Με το επιχείρημα αυτό ο Searle κατέδειξε ότι η επιτυχής από μέρους της μηχανής σύνταξη φυσικών συμβόλων δεν απαιτεί και την από μέρους της μηχανής κατανόηση των συμβόλων αυτών. Ως εκ τούτου, οι μηχανές δεν κατανοούν

[18] Για τις δύο αντίθετες απόψεις σχετικά με τη δυνατότητα ή μη ύπαρξης ανεστραμμένων ή απόντων qualia, δείτε Shoemaker (1984) και Block (1980, σσ. 257-274).

[19] Στο σημείο αυτό έχουν υπάρξει αντιρρήσεις από ορισμένους φιλοσόφους που αρνούνται την ύπαρξη ή την επιστημολογική εγκυρότητα των qualia κατά την προσπάθεια γνώσης (inspection) του Νου (δείτε, για παράδειγμα, Churchland, 1988; Frankish, 2016, 2017; Rey, 1983, 1988; Wilkes, 1988). Ωστόσο, είναι δύσκολο να φανταστούμε πώς εν τη απουσία των qualia θα μπορούσαμε να μιλήσουμε για εμπειρίες γεύσης, οσμής, χρώματος, αφής κ.λπ. ή ακόμα και για ψευδαισθητικές εμπειρίες και τελικά πώς θα μπορούσαμε να κατηγοριοποιήσουμε τα ερεθίσματά μας, αναγνωρίζοντας λόγου χάριν τη γεύση ή το άρωμα ενός φρούτου που έχουμε φάει ξανά στο παρελθόν (για την αντίθετη άποψη δείτε: Dennett, 1988; Hardcastle, 1999). Επιπλέον, όταν πρόκειται για την ίδια τη γνώση της συνείδησης, η διάκριση ψευδαίσθησης/πραγματικότητας καταρρέει και, επομένως, η όποια κριτική ως προς την επιστημολογική εγκυρότητα των qualia για τη γνώση του Νου καθίσταται το λιγότερο προβληματική: "where consciousness is concerned the existence of the appearance is the reality" (Searle, 1997, σ. 122). Σε κάθε περίπτωση βλέπουμε εδώ, με αφορμή την παρούσα αλλά και την αμέσως προηγούμενη υποσημείωση ως προς τα qualia, ότι μια λειτουργιστική προσέγγιση του ζητήματος απόδοσης της ιδιότητας του ηθικού προσώπου στα συστήματα T.N., όπως αυτή που επιχειρεί ο Dennett, ενδέχεται να ανοίξει πολύ περισσότερα ζητήματα από αυτά που έρχεται να κλείσει. Έστω και αν ο Dennett τίθεται έναντι της ύπαρξης των qualia (Dennett, 1988), το ζήτημα παραμένει ανοικτό και ένα από τα πιο συζητήσιμα στη σύγχρονη φιλοσοφία. Συνεπώς, η επίκληση μιας λειτουργικής αναλογίας μεταξύ των σχέσεων προπονητή-αθλητή και προγραμματιστή-μηχανής θα μας έφερνε αντιμέτωπους με αυτό το σοβαρό και ακόμα εκκρεμές φιλοσοφικό οντολογικό πρόβλημα, οδηγώντας σε μια ατέρμονη συζήτηση που θα απείχε παρασάγγας από τη σαφήνεια που οφείλει να έχει ένα κριτήριο απόδοσης ηθικού καθεστώτος στο πλαίσιο ενός κλάδου Εφαρμοσμένης Ηθικής όπως η Ηθική της T.N.

και, τελικά, η από μέρους τους υλοποίηση μιας επιτυχούς σύνταξης όπως αυτή λαμβάνει χώρα κατά την εκτέλεση ενός αλγορίθμου δεν αποτελεί απαραίτητα μια επίδειξη νοημοσύνης (Searle, 1980; 1984)[20].

Τα δύο ανωτέρω οντολογικά προβλήματα που πρέπει να επιλύσουν οι λειτουργιστές αποδεικνύουν ότι δεν είναι αυτονόητο ότι μια μηχανή που προσομοιάζει στην ανθρώπινη συμπεριφορά είναι νοήμων, επειδή και μόνο επιδεικνύει μια αντιστοίχιση εισόδων-εξόδων ίδια με αυτή του ανθρώπου σε κάποιο δεδομένο έργο (task). Ο Dennett, ωστόσο, απορρίπτει a priori τον χρηστικό ρόλο των qualia στη γνωσιακή επιστήμη και διαφωνεί με τον Searle, υποστηρίζοντας ότι το Κινέζικο Δωμάτιο «καταλαβαίνει» ως ένα ενιαίο σύστημα το νόημα του αποτελέσματος που εξάγει και, τελικά, υιοθετεί μια στάση λογικού συμπεριφοριστή απέναντι στην εξομοίωση Kasparov – Deep Blue (Ανθρώπου – Μηχανής), αρκούμενος στο αποτέλεσμα και τη συμπεριφορά των συγκρινόμενων οντοτήτων. Αυτή η διχογνωμία καταδεικνύει ότι η λειτουργιστική προσέγγιση χαρακτηρίζεται από οντολογικά ζητήματα που παραμένουν ως σήμερα εκκρεμή. Συνεπώς, για την ώρα, δεν δείχνει να είναι και η πλέον κατάλληλη για την εδραίωση ενός εύχρηστου και στιβαρού κριτηρίου απόδοσης ηθικού καθεστώτος στις μηχανές. Ούτως ή άλλως, η λειτουργική εξομοίωση αποτυγχάνει να τεκμηριωθεί όπως είδαμε λογικά και, επομένως, θα πρέπει μάλλον να επιστρέψουμε στην ανάγκη ευθείας αντιμετώπισης του βασικού ερωτήματος της Τ.Ν. στο οποίο αναφερθήκαμε πιο πάνω, δηλαδή του ερωτήματος οντολογικής ισοδυναμίας ή διάκρισης ανθρώπου-μηχανής και, τελικά, στο α.

[20] Ένα ακόμα διάσημο επιχείρημα έναντι του Λειτουργισμού Μηχανής είναι αυτό της Πολλαπλής Πραγμάτωσης. Αυτός ο αντίλογος έχει εξαιρετικό ενδιαφέρον, δεδομένου ότι το Επιχείρημα της Πολλαπλής Πραγμάτωσης είχε αρχικώς αρθρωθεί για να υποστηρίξει τον Λειτουργισμό. Ωστόσο, η κατανόηση του τρόπου με τον οποίο το Επιχείρημα της Πολλαπλής Πραγμάτωσης πλήττει τον Λειτουργισμό αλλά και του αντιλόγου που ορθώνεται ως προς το πλήγμα αυτό απαιτεί μια εκτεταμένη αναφορά στη δομή και τον τρόπο λειτουργίας της Μηχανής Turing όπως και μια εκτεταμένη βιβλιογραφική προσπέλαση, οι οποίες ξεφεύγουν από τον βασικό σκοπό αλλά και τις δυνατότητες χώρου του παρόντος άρθρου. Για μια επισκόπηση του τρόπου με τον οποίο η Πολλαπλή Πραγμάτωση πλήττει τον Λειτουργισμό Μηχανής δείτε Putnam (1992). Για μια σφαιρικότερη ανάλυση της σχέσης Λειτουργισμού και Πολλαπλής Πραγμάτωσης δείτε Kim (1993, κεφ. 16 "Multiple Realization and the Metaphysics of Reduction").

Τέλος, ως προς το β2., δηλαδή τη λειτουργική σύγκριση όχι των μελών που απαρτίζουν τα ζεύγη «προπονητής-αθλητής» και «προγραμματιστής-υπολογιστής» αλλά των σχέσεων που τα ζεύγη αυτά συγκροτούν, πρέπει να παρατηρήσουμε πως ήδη η σχέση «προπονητή-αθλητή» δείχνει να χαρακτηρίζεται από πολύ περισσότερους βαθμούς ελευθερίας από ό,τι η σχέση «προγραμματιστή-υπολογιστή». Η δράση του υπολογιστή δείχνει να είναι πολύ περισσότερο εξαρτώμενη από τα κελεύσματα του προγραμματιστή από όσο δεσμεύεται η δράση του αθλητή από τα κελεύσματα του προπονητή του. Μάλιστα, στον λειτουργικό καθορισμό της σχέσης «προγραμματιστή-υπολογιστή» υπάρχει ο παράγοντας του προγράμματος, ο οποίος δεν φαίνεται να έχει ένα λειτουργικό ανάλογο στην περίπτωση της σχέσης «προπονητή-αθλητή» («δασκάλου-μαθητή» κ.λπ.). Επιπλέον, θα μπορούσε να υποστηρίξει κανείς ότι ο προγραμματιστής και η μηχανή εμπλέκονται σε έναν ατέρμονο βρόχο δυναμικής αλληλεπίδρασης και σε έναν διαρκή διάλογο που καθορίζει ταυτοχρόνως τη δράση και των δύο. Σε κάθε περίπτωση, αυτή η συζήτηση περί της χαλαρότητας ή μη της σχέσης «προπονητή-αθλητή» έναντι της σχέσης «προγραμματιστή-υπολογιστή» μάς φέρνει αντιμέτωπους με τους όρους του περιβαλλοντικού προγραμματισμού (το περιβάλλον ως προγραμματιστής) και, τελικά, της ιδιότητας να δρα κανείς αυτόνομα, δηλαδή του αυτόνομου δρᾶν (autonomous agency). Πρόκειται, δε, για όρους που έχουν μια διαχρονική παρουσία στην προσπάθεια αντιμετώπισης του θεμελιώδους φιλοσοφικού ερωτήματος της Τ.Ν. ως προς την οντολογική ταύτιση ή διάκριση ανθρώπου-μηχανής[21].

[21] Στο σημείο αυτό εντοπίζεται ξανά μια τομή –ή καλύτερα μια κοινή κατάληξη– των α και β. Φαίνεται, λοιπόν, ότι ακόμα και υπό μια λειτουργιστική προσπάθεια παράκαμψης της ευθείας αντιμετώπισης του ερωτήματος οντολογικής ταύτισης ή διάκρισης ανθρώπου-μηχανής, δηλαδή ακόμα και υπό έναν ελιγμό αναγωγής ενός οντολογικού ερωτήματος σε όρους λειτουργιών, οι βασικοί όροι του ερωτήματος και τα όποια αδιέξοδά τους παραμένουν εν πλήρει ισχύ.

Το Επιχείρημα της Αυτονομίας

Το γεγονός ότι ο Dennett επικαλείται, μεταξύ των άλλων, την αυτονομία του HAL για να του αποδώσει την ευθύνη του ηθικού προσώπου δεν είναι κάτι καινοφανές στο πεδίο της Ηθικής της Τ.Ν. Άλλοι στοχαστές και ερευνητές της Τ.Ν. συνδέουν επίσης το ζήτημα απόδοσης της ιδιότητας του ηθικού προσώπου στις μηχανές με το ζήτημα της αυτονομίας (για παράδειγμα Calverley, 2005; Sparrow, 2007) ενώ πειραματικές έρευνες στην ψυχολογία της Αλληλεπίδρασης Ανθρώπου-Μηχανής (Human-Computer Interaction) καταδεικνύουν ότι η πλειονότητα των ανθρώπων θεωρούν τη δυνατότητα ενός συστήματος Τ.Ν. να προβαίνει σε επιλογές ως ένα εκ των βασικών κριτηρίων για την απόδοση ηθικής ευθύνης στο σύστημα αυτό (Monroe, Dillon & Malle, 2014). Η εν λόγω σύνδεση του ζητήματος απόδοσης ηθικού προσώπου στα συστήματα Τ.Ν. με την έννοια της αυτονομίας δείχνει σε πρώτη ανάγνωση εύλογη, και μάλιστα, υπό μια καντιανή φιλοσοφική προσέγγιση, απαραίτητη.

Στην περίπτωση του HAL, ο Dennett ξεπερνά το ζήτημα της ενδεχόμενης ετερονομίας ενός προγραμματισμένου υπολογιστή, συγκρίνοντάς τον με τον γενετικά ή εμπειρικά «προγραμματισμένο» ηθικό δράστη. Αν ο γενετικός προγραμματισμός και οι ανθρώπινες εμπειρίες απαλλάσσουν τον άνθρωπο από την ηθική του ευθύνη, τότε απαλλάσσουν και τον HAL. Εδώ έχουμε ουσιαστικά την άρθρωση του επιχειρήματος ότι το περιβάλλον λειτουργεί για τους ανθρώπους όπως οι προγραμματιστές για τις μηχανές. Συνεπώς, θα λέγαμε ότι έστω και αν θα ήταν, όπως είδαμε πιο πάνω, εξαιρετικά δύσκολο –αν όχι αδύνατον– να εδραιωθεί με λεπτομερείς λειτουργιστικούς όρους μια αναλογία προπονητή-προγραμματιστή, μπορεί τουλάχιστον να υπάρξει ένας κάποιος παραλληλισμός προγραμματιστή-περιβάλλοντος που θα μπορούσε ίσως να προλειάνει το έδαφος για την υποστήριξη μιας οντολογικής εξίσωσης ανθρώπων και συστημάτων Τ.Ν.

Σε αυτό το σημείο, θα μπορούσε κανείς να αντιτείνει ότι ο Dennett παραβλέπει μια σημαντική παράμετρο που καθιστά τη χρήση του όρου «αυτονομία» στην Τ.Ν. μεταφορική. Σχετικώς, θα μπορούσε να υποστηριχθεί ότι, σε

αντίθεση με τον άνθρωπο, κάθε αυτόνομο σύστημα Τ.Ν. εμπεριέχει έναν δεδομένο σκοπό – μια δεδομένη αποστολή. Δεν είναι, για παράδειγμα, δυνατόν, αν θέλει, να αναστείλει προσωρινά την αποστολή του και να καθίσει σε ένα καφέ να διαβάσει ένα βιβλίο φιλοσοφίας. Κάθε αποστολή ενός συστήματος Τ.Ν. είναι δεδομένη, αναπόδραστη και έξωθεν ορισμένη με τέτοιο τρόπο, ώστε κάθε έννοια αυτονομίας καταργείται. Κι αυτό συμβαίνει, όχι γιατί απλώς είναι προγραμματισμένο με κάποιον τρόπο, αλλά γιατί ο σκοπός για τον οποίο το εν λόγω σύστημα υπάρχει εμπεριέχεται στην «ουσία» του. Κάθε σύστημα Τ.Ν. είναι μια μηχανή «για να» – έχει δηλαδή κατασκευαστεί για να επιτελεί μια ορισμένη λειτουργία και για να επιτύχει κάποιους στόχους, ανεξάρτητα από την πολυπλοκότητα των στόχων αυτών. Μοιάζει πραγματική πρόκληση στην έρευνα της Τ.Ν. η σύλληψη (πόσω μάλλον η κατασκευή) μιας ευφυούς μηχανής χωρίς κανέναν ειδικό σκοπό – χωρίς αποστολή[22].

Εντούτοις, αν θέλουμε να είμαστε πραγματικά δίκαιοι με τον Dennett, οφείλουμε να αναρωτηθούμε πόσο διαφορετικοί είναι οι άνθρωποι από τις μηχανές σε αυτό το ζήτημα της ενσωμάτωσης ενός σκοπού στην ύπαρξή τους. Έρχονται, πράγματι, οι άνθρωποι στη ζωή και αναπτύσσονται ελεύθεροι από σκοπούς που δεν επιλέγονται από αυτούς τους ίδιους αλλά από το περιβάλλον τους; Συχνά οι άνθρωποι γαλουχούνται, περισσότερο ή λιγότερο ρητά, ώστε να έχουν έναν σκοπό στη ζωή τους, από τις ακραίες περιπτώσεις κλειστών κοινοβίων και θρησκευτικών ταγμάτων, στις πιο συνηθισμένες περιπτώσεις των πολιτικών νεολαιών, της Εκκλησίας και της μύησης και εκπαίδευσης ιερέων και καλογριών, της ένταξης στον στρατό και, τελικά, μέχρι τους πιο αδιόρατους τρόπους περιβαλλοντικής εκπαίδευσης, όπως ο παραδειγματισμός από το στενό οικογενειακό περιβάλλον και η ανάληψη των ευθυνών μας απέναντι στην οικογένεια ή ο διαχωρισμός των κοινωνικών ρόλων του κάθε φύλου. Αλλά πριν φτάσουμε

[22] Αξίζει να σημειώσουμε πως, παρότι η δημιουργία μηχανών γενικού σκοπού ήταν ένα από τα βασικότερα οράματα των ερευνητών ήδη από τα πρώτα χρόνια του προγράμματος της Τ.Ν. (δείτε, για παράδειγμα, την ιδέα της Καθολικής Μηχανής Turing, ερευνητικά προγράμματα όπως ο General Problem Solver των Allen Newel και Herbert Simon ή τη γνωσιακή αρχιτεκτονική SOAR), δεν έχει επιτευχθεί μέχρι σήμερα η δημιουργία συστημάτων Τ.Ν. άνευ σκοπού, ίσως εξαιτίας οντολογικών περιορισμών στον χαρακτήρα των ίδιων αυτών των συστημάτων.

σε όλα αυτά, ήδη η πράξη γέννησης ενός ανθρώπου εμπεριέχει έναν εξωγενή προς αυτόν σκοπό: την επιλογή των γονιών του να τον φέρουν στη ζωή (με σκοπό να διαιωνίσουν το γένος τους, το όνομά τους, να ικανοποιήσουν τα γονεϊκά ένστικτα, όπως το μητρικό φίλτρο, ή τα κοινωνικά πρότυπα-απαιτήσεις του οικογενειακού περιβάλλοντός τους κ.λπ.). Τίθεται, συνεπώς, εδώ το ερώτημα: Μέχρι ποιον βαθμό περιβαλλοντικής παρέμβασης θεωρείται ο άνθρωπος αυτόνομος; Υπό άλλη διατύπωση, ποιο είναι το όριο, το κατώφλι παρέμβασης πέρα από το οποίο οι περιβαλλοντικές παρεμβάσεις θεωρούνται προγραμματισμός ή ετερονόμηση της βούλησης μιας οντότητας; Δηλαδή, ποιο το όριο παρέμβασης πέρα από το οποίο η οντότητα θεωρείται ότι έχει εγκολπώσει στην ουσία της ύπαρξής της έναν εξωγενή προς αυτήν σκοπό; Εδώ απαιτείται ο καθορισμός ενός ποσοτικού κριτηρίου (π.χ. το κατώφλι εξωγενούς παρέμβασης), διότι ανάλογα ποσοτικά αντιμετωπίζεται η απόδοση ηθικού καθεστώτος (ως τώρα, το ηθικό καθεστώς αποδίδεται υπό διαφορετικές διαβαθμίσεις σε διαφορετικές ανθρώπινες ή ζωικές οντότητες)[23]. Φαίνεται, λοιπόν, πως τουλάχιστον μέχρι να μπορέσουμε να εντοπίσουμε αυτό το κατώφλι, δεν μπορούμε να απορρίψουμε εντελώς το επιχείρημα του Dennett περί του παραλληλισμού περιβάλλοντος και προγραμματιστών. Δεδομένου ότι οι άνθρωποι υφίστανται ένα είδος προγραμματισμού από το περιβάλλον τους, η απόλυτη αυτονομία ενδέχεται να μην υφίσταται ούτε καν για τον άνθρωπο. Συνεπώς, για την ώρα, δεν έχουμε το δικαίωμα να υποστηρίζουμε μια διάκριση ανθρώπου-μηχανών επί τη βάσει ενός επιχειρήματος επιβολής σκοπών στις μηχανές από τους δημιουργούς και προγραμματιστές τους.

Αν θέλουμε, λοιπόν, να εντοπίσουμε πράγματι κάποιο πρόβλημα στη χρήση του κριτηρίου της αυτονομίας, θα πρέπει να μετατοπίσουμε τη συζήτηση από τη σχέση μηχανής-προγραμματιστή και να εστιάσουμε στην ίδια την οριοθέτηση της φιλοσοφικής έννοιας «αυτονομία» και στον τρόπο με τον οποίον αυτή

[23] Ειδικότερα για τη διαβάθμιση απόδοσης ηθικής ευθύνης για πράξεις πολέμου και συγκεκριμένα για τη διάκριση ανάμεσα σε ενήλικες και παιδιά πολεμιστές αλλά και τον παραλληλισμό των τελευταίων αυτών με τα LAWS δείτε την ενδιαφέρουσα ανάλυση του Robert Sparrow (Sparrow, 2007).

σχετίζεται με την απόδοση ηθικού καθεστώτος καθώς και στην από μέρους μας δυνατότητα εξακρίβωσης της εμφάνισης της αυτονομίας σε μια οντότητα.

Είδαμε πιο πάνω ότι για να θεωρηθεί ένα πρόσωπο ως αυτόνομος δράστης δεν θα πρέπει να τελεί υπό καθεστώς εσωτερικού ή εξωτερικού εξαναγκασμού, δηλαδή να μην είναι με το πιστόλι στον κρόταφο ή να μην βρίσκεται σε μη ελεγχόμενη από αυτό το ίδιο νοητική κατάσταση.

Ωστόσο, αν θέλουμε να ορίσουμε επακριβώς την έννοια της αυτονομίας, θα πρέπει να είμαστε σε θέση να απαντήσουμε σε τέσσερα ερωτήματα:

1. Τι προϋποθέτει το αυτόνομο δρᾶν (autonomous agency); Ποια τα χαρακτηριστικά και οι ιδιότητες που πρέπει να έχει μια οντότητα, ώστε να μπορεί να δρα ως αυτόνομη δρώσα οντότητα (autonomous agent); Υπό άλλη διατύπωση, πώς οριοθετείται το αυτόνομο δρᾶν;
2. Πώς εξακριβώνει κανείς την αυτονομία; Ποιες οι ενδείξεις που πρέπει να έχουμε, ώστε να θεωρήσουμε μια οντότητα ως αυτόνομη δρώσα οντότητα[24];
3. Είναι η αυτόνομη δράση και ιδιαίτερα η δράση μιας ηθικής αυτόνομης δρώσας οντότητας (moral agent) *αναγκαστικά* συνδεδεμένη με την ιδιότητα του νοήμονος όντος; Είναι, με άλλα λόγια, η νοητική κατάσταση του δρώντος –και κατά μία έννοια η νόηση γενικότερα– αναγκαία συνθήκη για την επίδειξη αυτόνομης δράσης;
4. Είναι το ζήτημα απόδοσης της ιδιότητας του αυτόνομου δρᾶν απολύτως συμμετρικό με το ζήτημα απόδοσης ηθικού καθεστώτος; Το να χαρακτηρίζεται κάποιος ή κάτι ως αυτόνομη δρώσα οντότητα συνεπάγεται αυτομάτως ότι μπορεί να χαρακτηριστεί και ως δρώσα οντότητα με ηθικό καθεστώς;

Θα πρέπει, καταρχάς, να δούμε ότι τα ερωτήματα 1 και 2 συνδέονται, καθώς κάποια από τα χαρακτηριστικά και τις ιδιότητες που απαιτεί το αυτόνομο δρᾶν δύναται να τροφοδοτήσουν και τα κριτήρια επίδειξης του τελευταίου αυτού.

[24] Ο καθορισμός των χαρακτηριστικών και, επομένως, ασφαλών ενδείξεων είναι κρίσιμος, καθότι επί αυτών των ενδείξεων θα βασιστεί η οντολογική αξιολόγηση και ταξινόμηση των υπό εξέταση οντοτήτων για τις οποίες θα τεθεί το ερώτημα απόδοσης ηθικού καθεστώτος. Δείτε αμέσως πιο κάτω, στο κυρίως κείμενο.

Αν, λόγου χάριν, η ιδιότητα Ι απαιτείται ώστε μια οντότητα Ο να είναι πράγματι αυτόνομη δρώσα οντότητα, τότε ένα ασφαλές κριτήριο εξακρίβωσης αυτόνομου δρᾶν στην Ο είναι η εξακρίβωση σε αυτήν της ιδιότητας Ι ως ιδιότητας της Ο. Το ερώτημα 1 είναι ένα οντολογικού τύπου ερώτημα (Τι εστί αυτόνομο δρᾶν;), ενώ το ερώτημα 2 είναι ένα επιστημολογικού τύπου ερώτημα (Πώς μπορούμε να γνωρίσουμε το αυτόνομο δρᾶν σε μια υπό εξέταση περίπτωση;). Ωστόσο, συμβαίνει συχνά η απάντηση του επιστημολογικού ερωτήματος να καθορίζεται εν πολλοίς από την απάντηση του οντολογικού ερωτήματος[25].

Σε κάθε περίπτωση, ως προς το ερώτημα 1 παρατηρείται μια πολυαρχία ορισμών και, τελικά, προϋποθέσεων για το αυτόνομο δρᾶν[26]. Ποια από όλες αυτές τις θεωρήσεις είναι η ορθή; Επομένως, βάσει ποιας εξ αυτών θα έπρεπε να διεξαχθεί η συζήτηση απόδοσης της ιδιότητας του ηθικού προσώπου στα συστήματα της Τ.Ν.; Στον χώρο της Ηθικής της Τ.Ν. φαίνεται πως μέχρι τώρα οι προσεγγίσεις των περισσότερων ερευνητών είναι «εσωτερικιστικές» (υπό την έννοια ότι αναφέρονται στη συνείδηση και σε νοητικές καταστάσεις όπως πίστεις, πεποιθήσεις, προθέσεις, συναισθήματα κ.λπ.) και, ως εκ τούτου, προσεγγίζουν ή είναι σε πλήρη ευθυγράμμιση με ό,τι στο πεδίο ανάλυσης του

[25] Χαρακτηριστικό παράδειγμα αυτής της διασύνδεσης οντολογικού και επιστημολογικού ερωτήματος είναι η από μέρους του Thomas Reid εισήγηση του Προβλήματος των Άλλων Έμψυχων Όντων (στη βιβλιογραφία απαντά και υπό τον ενδεχομένως πιο σύγχρονο όρο «Πρόβλημα των Άλλων Νόων» – δείτε και υποσημείωση 30) εν είδει κριτικής στον τρόπο με το οποίο αντιλαμβανόταν τον Νου ο Berkeley (Avramides, 2001). Πρέπει εδώ να επισημανθεί ότι, πέραν της έννοιας της αυτονομίας, η εν λόγω σύνδεση του οντολογικού και του επιστημολογικού ερωτήματος υφίσταται και για κάθε άλλη έννοια που έχει συσχετιστεί με την απόδοση της ιδιότητας του ηθικού προσώπου. Έννοιες όπως η συνείδηση, η νόηση και η νοημοσύνη είναι λογικό να έχουν επίσης ένα οντολογικό και ένα επιστημολογικό ερώτημα, με την απάντηση του πρώτου να επηρεάζει την απάντηση του δευτέρου και την απάντηση του δευτέρου να επηρεάζει την οντολογική ταξινόμηση των υπό εξέταση οντοτήτων.

[26] Επιστρέφοντας σε όσα αναφέραμε πιο πάνω για το «περιβάλλον ως προγραμματιστή», πρέπει να αναφέρουμε ότι αυτή η πολυαρχία ορισμών και προϋποθέσεων του αυτόνομου δρᾶν επηρεάζει αρνητικά και τη δυνατότητα προσδιορισμού του ορίου περιβαλλοντικής επίδρασης πέραν του οποίου ένας άνθρωπος πρέπει να θεωρείται ότι εγκολπώνει στην ουσία της ύπαρξής του έναν εξωγενή προς αυτόν σκοπό. Για μια επισκόπηση του τρόπου με τον οποίο η οριοθέτηση του αυτόνομου δρᾶν επηρεάζεται από την οριοθέτηση των εξωγενών προς τον άνθρωπο παρεμβάσεων δείτε στο Buss & Westlund, 2018.

αυτόνομου δρᾶν αποκαλείται ως Συναφειοκρατική προσέγγιση (coherrentist view)[27]. Σύμφωνα με τη Συναφειοκρατική προσέγγιση του αυτόνομου δρᾶν, μια δρώσα οντότητα ελέγχει την πράξη της, αν και μόνο αν το κίνητρο της πράξης της είναι σε συνάφεια με μια νοητική κατάσταση που αφορά στην προοπτική υπό την οποία η δρώσα αυτή οντότητα αντιλαμβάνεται τον Κόσμο και τη θέση της μέσα σε αυτόν (Frankfurt, 1988a). Όμως διαφορετικοί απολογητές της Συναφειοκρατικής προσέγγισης προτείνουν διαφορετικές νοητικές καταστάσεις ως συναφείς με το αυτόνομο δρᾶν. Συγκεκριμένα, αυτές οι νοητικές καταστάσεις μπορεί να είναι σχετιζόμενες με κάποιους μακροπρόθεσμους στόχους, κίνητρα και σχέδια (Bratman, 1979, 2007; Watson, 1975) ή και με συναισθήματα και κυρίως συναισθήματα ενδιαφέροντος και φροντίδας (emotions of 'caring') για τους ανθρώπους (Frankfurt, 1988b, 1999; Jaworska, 2007a, 2007b, 2009; Shoemaker, 2003). Έτσι, τίθεται και πάλι ένα ζήτημα πολυαρχίας ορισμών που οδηγεί στο εύλογο ερώτημα: Ποια από όλα αυτά τα κριτήρια είναι σωστό; Βάσει ποιας εξ αυτών των προτάσεων θα πρέπει να κρίνουμε ως προς την αυτονομία ανθρώπους, ζώα και μηχανές; Το πρόβλημα της εννοιολογικής ασάφειας κάνει εδώ την εμφάνισή του.

Επιστρέφοντας, δε, στη σχέση των ερωτημάτων 1 και 2 βλέπουμε, με αφορμή τη Συναφειοκρατική προσέγγιση, τον τρόπο με τον οποίο η αδυναμία οριστικής και καθολικώς αποδεκτής απάντησης του ερωτήματος 1 οδηγεί σε αδυναμία οριστικής απάντησης του ερωτήματος 2. Σχετικώς, η υπό τη Συναφειοκρατική

[27] Για τη Συναφειοκρατική Προσέγγιση δείτε: Calverlay, 2005; De Landa, 1991; Levy, 2007; Pincker, 1997; Solum, 1992; Sparrow, 2007; Torrance, 2004). Για μια αναλυτική παρουσίαση όλων των ως τώρα προτεινόμενων θεωρήσεων για το αυτόνομο δρᾶν δείτε στο: Buss & Westlund, 2018. Δεδομένου του περιορισμού χώρου στο παρόν άρθρο, επιλέγουμε να επικεντρωθούμε μόνο στη Συναφειοκρατική θεώρηση, καθώς αυτή διακρίνει την ως τώρα συζήτηση στο πεδίο της Ηθικής της Τ.Ν. Η ανάλυση των προβλημάτων ή των λύσεων που θα μπορούσαν να δώσουν οι υπόλοιπες προσεγγίσεις θα μπορούσε να αποτελέσει έναν νέο γόνιμο τρόπο προβληματισμού που θα παρουσιαζόταν σε άλλο άρθρο. Για την ώρα και τις ανάγκες του παρόντος κειμένου, ο σχολιασμός μας ως προς τις υπόλοιπες θεωρήσεις του αυτόνομου δρᾶν θα αρκεστεί στην επισήμανση πως η ύπαρξή τους αυξάνει τον «θόρυβο» στη συζήτηση για την απόδοση ηθικού καθεστώτος στις μηχανές.

θεώρηση πολλαπλότητα υποψηφίων νοητικών καταστάσεων συναφών με το αυτόνομο δρᾶν –επομένως η πολλαπλότητα απαντήσεων του ερωτήματος 1– πλήττει τη δυνατότητα ξεκάθαρης και τελεσίδικης απάντησης του ερωτήματος 2. Πώς μπορούμε να γνωρίζουμε ποιες νοητικές καταστάσεις και λειτουργίες πρέπει να αναζητήσουμε και να παρατηρήσουμε σε μια υπό εξέταση οντότητα, ώστε αυτή να θεωρηθεί αυτόνομη και τελικά δικαιούχος της απόδοσης ηθικού καθεστώτος;

Το πρόβλημα της εξακρίβωσης νοητικών καταστάσεων σε άλλες δρώσες οντότητες, πέραν της εννοιολογικής ασάφειας, μας φέρνει ευθέως αντιμέτωπους με ένα εκ των κεντρικότερων προβλημάτων της Φιλοσοφίας του Νου και συγκεκριμένα με το Πρόβλημα των Άλλων Νόων[28]. Πώς μπορεί να εξακριβώσει κανείς την ύπαρξη νοητικών καταστάσεων στις οντότητες γύρω του; Μάλιστα, το πρόβλημα αυτό επιμερίζεται στα ακόλουθα δύο ερωτήματα:

α. πώς μπορεί να γνωρίσει κάποιος ότι τα άλλα όντα γύρω του είναι φορείς νοητικών καταστάσεων και

β. αν όντως είναι φορείς, πώς μπορεί να καταστεί γνωστό το είδος των νοητικών καταστάσεών τους (Avramides, 2001).

Επιχειρώντας, λοιπόν, να προσεγγίσουμε το ζήτημα απόδοσης της ιδιότητας του ηθικού προσώπου στα συστήματα Τ.Ν. μέσω των πιο πάνω Συναφειακρατικών προσεγγίσεων του αυτόνομου δρᾶν, βρισκόμαστε αντιμέτωποι με τις εκφάνσεις αυτού του προβλήματος και συγκεκριμένα με το ερώτημα: Πώς γνωρίζουμε αν μια μηχανή είναι φορέας νοητικών καταστάσεων, και μάλιστα νοητικών καταστάσεων που αφορούν στην προοπτική της ίδιας της μηχανής; Πώς γνωρίζουμε αν ένα σύστημα Τ.Ν. έχει κίνητρα και σχέδια και αν αυτά είναι μακροπρόθεσμα; Πώς θα μπορούσαμε να γνωρίζουμε αν ένα σύστημα Τ.Ν. έχει

[28] Στη βιβλιογραφία απαντά και με τον παλαιότερο όρο «Πρόβλημα των Άλλων Έμψυχων Όντων», που πάντως ενέχει ενδεχομένως και μια καρτεσιανή παραδοχή, καθώς περιλαμβάνει τον όρο «έμψυχων».

συναισθήματα και αν αυτά είναι συναισθήματα φροντίδας και ενδιαφέροντος[29];

Θα πρέπει σε αυτό το σημείο να τονιστεί η κρισιμότητα που έχει η απάντηση του ερωτήματος 2 όταν εργαζόμαστε σε ένα πλαίσιο εφαρμοσμένης Ηθικής. Σε ένα τέτοιο πλαίσιο φιλοσοφικής ενατένισης, απαιτούνται στιβαρά και εύχρηστα οντολογικά κριτήρια, που με τη σειρά τους θα οδηγήσουν σε στιβαρά και εύχρηστα κριτήρια απόδοσης ηθικού καθεστώτος, με τις όποιες διαβαθμίσεις του τελευταίου αυτού. Συνεπώς, θα λέγαμε ότι η έως τώρα αντιμετώπιση του ερωτήματος 1 δεν έχει οδηγήσει σε αποτελέσματα πραγματικά χρήσιμα για την αντιμετώπιση του ερωτήματος 2. Υπό άλλη διατύπωση, το ερώτημα 1 προσεγγίζεται από τους φιλοσόφους με έναν τρόπο αντιπαραγωγικό ως προς τις απαιτήσεις και ανάγκες της Εφαρμοσμένης Ηθικής και δη της Ηθικής της Τ.Ν.

Λόγω του αδιεξόδου στο οποίο οδηγεί η αντιμετώπιση αυτού του προβλήματος, μια στρατηγική θα ήταν να επιχειρήσει κάποιος να απεμπλακεί από το Πρόβλημα των Άλλων Νόων και να εξετάσει το κριτήριο του αυτόνομου δρᾶν ανεξάρτητα από το πρόβλημα αυτό. Μια τέτοια στρατηγική όμως θα τον έφερνε αντιμέτωπο με το ερώτημα 3 («Είναι η αυτόνομη δράση *αναγκαστικά* συνδεδεμένη με την ιδιότητα του νοήμονος όντος;»).

---

[29] Βέβαια, το Πρόβλημα των Άλλων Νόων συνιστά ένα εμπόδιο και για κάθε άλλη «εσωτερικιστική» προσέγγιση του ζητήματος απόδοσης της ιδιότητας του ηθικού προσώπου, ακόμα και για προσεγγίσεις που δεν χρησιμοποιούν το κριτήριο της αυτονομίας. Μπορούμε επιγραμματικά να αναφερθούμε σε μια τάση της Ηθικής της Τ.Ν. που εξετάζει την απόδοση ηθικού καθεστώτος στα συστήματα Τ.Ν. μέσω της έννοιας του ηθικώς παθόντα (moral patiency - Δείτε σχετικώς: Floridi & Sanders, 2004; Hajdin, 1994; Hoffman & Hahn, 2020; Levy, 2009; Regan, 1983; Wallach & Allen, 2009), η οποία συνήθως συνδέεται και με την έννοια της αισθητότητας (sentience). Η τελευταία αυτή χρησιμοποιήθηκε για πρώτη φορά ως κριτήριο απόδοσης ηθικού καθεστώτος σε μη ανθρώπινες οντότητες από τον Peter Singer σε σχέση με τα ζώα (Singer, 1975, δείτε και Singer, 1993) αλλά έχει πλέον εισαχθεί και στη συζήτηση για τα συστήματα της Τ.Ν. (Levy, 2009; Owen & Osley, 2007). Η βασική γραμμή σκέψης σε σχέση με την ιδιότητα του ηθικώς παθόντα προβλέπει ότι τα συστήματα της Τ.Ν ενδεχομένως να πρέπει να θεωρηθούν παθόντες, αν πράγματι δύνανται να αισθάνονται –και δη να αισθάνονται ψυχικό και σωματικό πόνο– και επομένως υποφέρουν. Ωστόσο, τίθεται το ερώτημα κατά πόσον μπορείς να γνωρίζεις ότι τα συστήματα Τ.Ν. υποφέρουν. Πράγματι, ορισμένοι ερευνητές της Ηθικής της Τ.Ν. έχουν αρχίσει να επισημαίνουν το εμπόδιο του Προβλήματος των Άλλων Νόων (επί παραδείγματι: Gunkel, 2012; Hoffman & Hahn, 2020; Levy, 2009).

Ας σκεφτούμε, χάριν παραδείγματος, ένα όχημα του οποίου έχει καταστραφεί το σύστημα πλοήγησης ή ένα συμβατικό αυτοκίνητο του οποίου έχει «μπλοκάρει» το τιμόνι ή έχουν χαλάσει τα φρένα. Μπορούμε να υποστηρίξουμε ότι ένα τέτοιο όχημα επιδεικνύει κάποιο είδος αυτονομίας, υπό την έννοια ότι η δράση του δεν ελέγχεται ή δεν καθορίζεται από τον οδηγό[30]; Είναι γεγονός ότι δεν έχουμε ποτέ μπει στον πειρασμό να θεωρήσουμε ότι η μη ελεγχόμενη δράση του εν λόγω οχήματος μοιάζει με την αυτονομία που θεωρούμε ότι επιδεικνύουν τα ανθρώπινα όντα. Κι αυτό, γιατί ό,τι αναζητούμε εδώ είναι μια *ορισμένου τύπου* αυτονομία και, τελικά, μια αυτονομία που συνδέεται με ένα *ορισμένου τύπου* δρᾶν (agency)[31]. Ποιο είναι το ουσιώδες χαρακτηριστικό αυτού του τρόπου; Γιατί δεν τίθεται καν το ερώτημα απόδοσης αυτού του αυτόνομου δρᾶν σε ένα μη ελεγχόμενο αυτοκίνητο που κινείται με «σπασμένα τα φρένα», αλλά τίθεται για έναν υπολογιστή –αλλά και για ένα «έξυπνο όχημα»– και πολύ περισσότερο για τους ανθρώπους;

Ενδεχομένως, επειδή, σε αντίθεση με το μη ελεγχόμενο συμβατικό αυτοκίνητο, στην περίπτωση του ανθρώπου έχουμε αποδεχθεί a priori την ιδιότη-

[30] Σε αυτό το παράδειγμα θα μπορούσε να επισημάνει κανείς ότι το αυτοκίνητο είναι μεν ακυβέρνητο από τον άνθρωπο-οδηγό, αλλά υπόκειται πλήρως στους νόμους της Φυσικής, οι οποίοι και διέπουν, εν τέλει, πλήρως την κίνησή του. Συνεπώς, ο μη έλεγχός του από τον άνθρωπο δεν σημαίνει απαραίτητα και ένα αυτόνομο δρᾶν. Βέβαια, υπό μια ακραιφνώς ντετερμινιστική προσέγγιση, θα μπορούσε να υποστηριχθεί ότι το ίδιο συμβαίνει και με την ανθρώπινη συμπεριφορά. Συνεπώς, μια άρνηση της αυτονομίας δράσης του αυτοκινήτου μέσω της επίκλησης στους νόμους της Φύσης θα μπορούσε ανάλογα να εφαρμοστεί και στη περίπτωση των ανθρώπων πλήττοντας την ιδέα και της δικής τους αυτονομίας. Εντούτοις, αυτό θα ήταν ένας ελιγμός μοιραίος για το όλο εγχείρημα της Ηθικής, συνεπώς ένας ελιγμός που θα τερμάτιζε την παρούσα συζήτηση και την αναζήτηση λύσεων στα προβλήματα της Εφαρμοσμένης Ηθικής και δη της Ηθικής της Τ.Ν., όχι διά της παροχής λύσεων αλλά διά της άρνησης ύπαρξης του όλου πλαισίου εντός του οποίου εμφανίζονται τα εν λόγω προβλήματα. Στην παρούσα ανάλυση, ωστόσο, υιοθετούμε μια συμβατοκρατική στάση (compatibilitism) θεωρώντας ότι οι φυσικοί νόμοι και τα κριτήρια απόδοσης ηθικών ευθυνών ή ηθικού καθεστώτος ανήκουν σε ξεχωριστά εννοιολογικά πεδία (συγκεκριμένα στο οντολογικό και το αξιολογικό).

[31] Βλέπε εδώ τη διάκριση μεταξύ της φιλοσοφικής και τεχνικής έννοιας της αυτονομίας που περιγράφουμε πιο πάνω.

τα του νοήμονος όντος[32] (cognitive being) η οποία συνεπάγεται νοημοσύνη (intelligence).

Φαίνεται, δηλαδή, ότι εν γένει αποδεχόμαστε πως το αυτόνομο δρᾶν δε μπορεί παρά να είναι ένα νοήμον δρᾶν το οποίο επιδεικνύει νοημοσύνη (intelligence) ή, πολύ περισσότερο, νόηση (cognition). Συνεπώς, σε σχέση με το ερώτημα 3 («Είναι η αυτόνομη δράση αναγκαστικά συνδεδεμένη με την ιδιότητα του νοήμονος όντος;»), θα απαντούσαμε πως βάσει των ως τώρα κυρίαρχων προσεγγίσεων στο πεδίο της Ηθικής και ειδικότερα των προσεγγίσεων στο πεδίο της Ηθικής της Τ.Ν., η αυτόνομη δράση δείχνει να είναι πράγματι *αναγκαστικά* συνδεδεμένη με τη νόηση.

Ωστόσο, η νόηση δεν αντιμετωπίζεται μονοσήμαντα, δηλαδή ως ένα χαρακτηριστικό του «όλα ή τίποτα», αλλά μάλλον ως ένα χαρακτηριστικό που παρουσιάζει τόσο ποσοτικές όσο και ποιοτικές διαφορές. Ως εκ τούτου, υπάρχουν ακόμα και περιπτώσεις ανθρώπινων όντων στα οποία αρνούμαστε την απόδοση της ιδιότητας του αυτόνομου δρᾶν και, τελικά, την απόδοση πλήρους ή έστω μερικού ηθικού καθεστώτος. Για παράδειγμα, τα βρέφη ή ορισμένες κατηγορίες ψυχικά ασθενών ή άνθρωποι σε κωματώδη κατάσταση ή σε κατάσταση «φυτού» είναι μερικές μόνο περιπτώσεις ανθρώπινων όντων στα οποία δυσκολευόμαστε να φτάσουμε σε καθολικώς αποδεκτές και τελεσίδικες αποφάνσεις ως προς την απόδοση ή μη της ιδιότητας του νοήμονος δρᾶν και, τελικά, ηθικού

[32] Παρά το γεγονός ότι στην περίπτωση του ανθρώπου έχουμε αποδεχθεί την ιδιότητα του νοήμονος όντος η οποία συνεπάγεται νοημοσύνη, δεν ισχύει το ίδιο για την περίπτωση των έξυπνων μηχανών. Για τις έξυπνες μηχανές το ερώτημα του νοήμονος όντος (cognitive being) παραμένει ανοιχτό, παρότι απαντάμε καταφατικά ως προς την ικανότητά τους να επιδεικνύουν νοημοσύνη (intelligence) – έστω και σε διαβαθμίσεις. Αντίθετα, στην περίπτωση μιας ετεροκαθοριζόμενης μηχανής, όπως το αυτοκίνητο, απαντάμε εξαρχής αρνητικά τόσο για την απόδοση της ιδιότητας του νοήμονος όντος όσο και για την ικανότητα νοημοσύνης. Παρατηρούμε, λοιπόν, ότι έχουμε τρεις διαβαθμίσεις απόδοσης της ιδιότητας του νοήμονος όντος. Τα ευφυή συστήματα ΤΝ βρίσκονται σε ένα μέσο έδαφος, ανάμεσα στην πλήρη αναγνώριση της ιδιότητας του νοήμονος όντος και στην πλήρη απόρριψή της· και αυτό, διότι συμπεριφέρονται με έναν «intelligent» (νοήμονα) τρόπο ο οποίος αφήνει ανοιχτή την πιθανότητα να είναι τελικά «cognitive» (νοητικός) τρόπος (Ανατρέξτε στην υποσημείωση 7 για τη διάκριση ανάμεσα στους όρους «intelligence» και «cognition»).

καθεστώτος. Συνεπώς, αν και στη σκέψη μας το αυτόνομο δρᾶν είναι συνήθως άρρηκτα συνδεδεμένο με τη νόηση, η τελευταία αυτή δείχνει να χαρακτηρίζεται από πολλές διαφορετικές διαβαθμίσεις και εκφάνσεις που οδηγούν εν τέλει σε προβληματισμούς περί της ανάγκης υιοθέτησης ανάλογων διαβαθμίσεων κατά την απόδοση ηθικού καθεστώτος μέσω του κριτηρίου της αυτονομίας. Ως εκ τούτου, θα ολοκληρώναμε την απάντησή μας σε σχέση με το ερώτημα 3 («Είναι η αυτόνομη δράση *αναγκαστικά* συνδεδεμένη με την ιδιότητα του νοήμονος όντος;») ως εξής: Βάσει των ως τώρα κυρίαρχων προσεγγίσεων στο πεδίο της Ηθικής της Τ.Ν., η αυτόνομη δράση είναι άρρηκτα συνδεδεμένη με την ικανότητα νόησης, αλλά δεδομένων των διαβαθμίσεων και ποιοτικών διαφοροποιήσεων που αναγνωρίζουμε στην τελευταία, αυτή η άρρηκτη σύνδεση οδηγεί σε μια *μη μονοσήμαντη* συσχέτιση της αυτόνομης δράσης με το ζήτημα απόδοσης ηθικού καθεστώτος. Τελικά, λαμβάνοντας υπόψη μας και τα όσα αναλύσαμε πιο πάνω σε σχέση με το Πρόβλημα των Άλλων Νόων, η άρρηκτη σχέση αυτόνομου δρᾶν και νόησης κληροδοτεί στο πρώτο όλα τα εννοιολογικά, οντολογικά και επιστημολογικά προβλήματα της δεύτερης, με αποτέλεσμα η αυτόνομη δράση να καθίσταται ένα δύσχρηστο κριτήριο απόδοσης ηθικού καθεστώτος.

Αξίζει πάντως εδώ να σημειωθεί ότι, τουλάχιστον ως προς τη Συναφειοκρατική του προσέγγιση, το αυτόνομο δρᾶν καθίσταται δύσχρηστο ως κριτήριο λόγω της σύνδεσής του και με άλλες έννοιες. Σχετικώς, η Συναφειοκρατική θεώρηση συνιστά ένα σημείο τομής της συζήτησης για το αυτόνομο δρᾶν με τους παραδοσιακούς όσο και επίπονους προβληματισμούς σχετικά με την ιδιότητα του προσώπου, και αυτό γίνεται με τρεις τρόπους: i. Η απαίτηση για την ύπαρξη στοχεύσεων και νοητικών καταστάσεων υπό την προοπτική μιας δρώσας οντότητας ισοδυναμεί με την απαίτηση για την οριοθέτηση μιας προσωπικής προοπτικής. ii. Η ύπαρξη μακροπρόθεσμων στοχεύσεων ως χαρακτηριστικών του αυτόνομου δρᾶν νοητικών καταστάσεων προϋποθέτει τη «διαχρονική ενότητα» αυτής της προσωπικής προοπτικής. Συνεπώς, προϋποθέτει την επιβίωση της

δρώσας οντότητας μέσα στον χρόνο, άρα την επιβίωση του ίδιου προσώπου[33]. iii. Η υπό τη Συναφειοκρατική θεώρηση των μακροπρόθεσμων ή βραχυπρόθεσμων αποβλεπτικών[34] νοητικών καταστάσεων ως συναφών με το αυτόνομο δρᾶν μπορεί να εγείρει ένα ερώτημα οριοθέτησης της βούλησης της δρώσας οντότητας. Ποιες βουλητικές εκδηλώσεις θεωρούνται εξωγενώς προερχόμενες και ποιες ενδογενώς; Υπάρχουν καθαρά ενδογενείς βουλητικές εκδηλώσεις; Ποιο το όριο διαχωρισμού των μεν από τις δε; Με άλλα λόγια, ποιο το όριο διαχωρισμού του προσώπου από τον υπόλοιπο κόσμο που το περιβάλλει; Επιπλέον, δύναται οι παρορμήσεις μας και οι βραχύβιες έντονες επιθυμίες μας να θεωρηθούν προϊόντα της βούλησής μας; Τέλος, τα βασικά χαρακτηριστικά της προσωπικότητάς μας (personality traits) αποτελούν ενδογενείς ή εξωγενείς παράγοντες της βούλησής μας; Τα εν λόγω χαρακτηριστικά έχει αποδειχθεί μέσω ερευνών στο πεδίο της Πειραματικής Ψυχολογίας ότι είναι, σε μεγάλο βαθμό, ρυθμιστές των πράξεών μας. Αυτή η σύζευξη του ζητήματος του αυτόνομου δρᾶν με το πρόβλημα του προσώπου συνιστά ένα παράδειγμα του τρόπου με τον οποίο η φιλοσοφική ανάλυση των διαφόρων εννοιών και η μεταξύ τους σύνδεση εν τέλει οδηγεί όχι σε μια ελάττωση αλλά σε μια αύξηση των προβλημά-

[33] Εδώ γίνεται εμφανές ότι η εκδοχή της Συναφειοκρατικής προσέγγισης, η οποία προτάσσει ως συναφείς με το αυτόνομο δρᾶν τις μακροπρόθεσμες αποβλεπτικές νοητικές καταστάσεις, απαιτεί μια «ψυχολογική (ή νοητική) συνέχεια» της δρώσας οντότητας, που είναι ταυτόσημη με την «ψυχολογική (ή νοητική) συνέπεια» που απαιτείται για τη διατήρηση της «αίσθησης του εγώ» και τελικά της ταυτότητας του προσώπου. Σε αυτό το σημείο, η συζήτηση οριοθέτησης της αυτόνομης δράσης δείχνει να εφάπτεται με τα προβλήματα αντιμετώπισης διατήρησης στον χρόνο της ιδιότητας του προσώπου αλλά και μιας ορισμένης ταυτότητας (identity) αυτού. Υπό άλλη διατύπωση, η Συναφειοκρατική προσέγγιση του αυτόνομου δρᾶν μας φέρνει αντιμέτωπους με το ερώτημα της διατήρησης (persistence question) και του χαρακτηρισμού (characterization question) σχετικά με την ιδιότητα του προσώπου. Για μια αναλυτική περιγραφή των εν λόγω προβλημάτων δείτε στο Olson, 2019.

[34] Οι όροι «αποβλεπτικές», «προθετικές», ή «κατευθυντικές» καταστάσεις αναφέρονται γενικώς στις νοητικές καταστάσεις (intentional states) κατά τις οποίες ένα υποκείμενο (S) αναφέρεται σε κάτι ή σχετίζεται με κάτι (p) έξω από αυτό. Για παράδειγμα ένα νοήμων ον S μπορεί να σκέφτεται, να πιστεύει ή να λαμβάνει μια «προθετική στάση» (επιθυμία, φόβο, συναίσθημα κ.λπ.) για το p, να γνωρίζει p ή να νοεί p συγκροτώντας ή έχοντας ένα ορισμένο νοητικό περιεχόμενο, αναπαριστώντας τελικά p (Gounaris, 2014).

των, καθώς κάθε νέα έννοια (π.χ. πρόσωπο) που εισάγεται στη συζήτηση προς διασάφηση της προηγούμενης έννοιας (π.χ. αυτόνομο δρᾶν) φέρει μαζί της τα δικά της προβλήματα οριοθέτησης.

Το ζήτημα οριοθέτησης της βούλησης και συμπερίληψης ή μη των παρορμήσεων και των βραχύβιων έντονων επιθυμιών μας εντός των ορίων αυτής φέρνει στο προσκήνιο και το ερώτημα 4 («Είναι το ζήτημα απόδοσης της ιδιότητας του αυτόνομου δρᾶν απολύτως συμμετρικό με το ζήτημα απόδοσης ηθικού καθεστώτος; Το να χαρακτηρίζεται κάποιος ή κάτι ως αυτόνομη δρώσα οντότητα συνεπάγεται αυτομάτως ότι μπορεί να χαρακτηριστεί και ως δρώσα οντότητα με ηθικό καθεστώς;»).

Όπως συναντάμε στη βιβλιογραφία αλλά και στην καθημερινή πρακτική, διαφορετικές τοποθετήσεις περί των ορίων της βούλησης οδηγούν σε διαφορετικές απαντήσεις του εν λόγω ερωτήματος. Επί παραδείγματι, συνήθως δεν αποδίδουμε πλήρη αυτονομία σε άτομα εθισμένα σε ουσίες. Αυτό έχει ως συνέπεια να μην τους αποδίδουμε και πλήρες ηθικό καθεστώς (είναι γνωστό ότι υφίσταται μια μεγάλη και επίπονη συζήτηση ως προς τα όρια της ηθικής ευθύνης τους). Συνεπώς, στην περίπτωση των ατόμων που εθίζονται σε ουσίες η σχέση αυτονομίας-ηθικού καθεστώτος δείχνει να είναι συμμετρική, δηλαδή η έλλειψη αυτονομίας αντιστοιχεί σε έλλειψη ηθικού καθεστώτος. Απεναντίας, σε άλλες περιπτώσεις, σε άτομα που έχουν υποστεί «πλύση εγκεφάλου» ή κατήχηση και στα οποία συνήθως, σύμφωνα με τη βιβλιογραφία, δεν αποδίδεται αυτονομία, συμβαίνει να αποδίδεται ηθική ευθύνη. Δηλαδή, ενώ υπό τις περισσότερες φιλοσοφικές προσεγγίσεις της έννοιας του αυτόνομου δρᾶν τα άτομα αυτά δεν θεωρούνται ως αυτόνομες δρώσες οντότητες, προσλαμβάνουν ωστόσο ηθικό καθεστώς. Βλέπουμε, συνεπώς, ότι η συμμετρικότητα της σχέσης αυτονομίας-ηθικού καθεστώτος μεταβάλλεται κατά περίσταση, γεγονός που κάνει τον ορισμό της αυτονομίας περισσότερο προβληματικό.

Εν κατακλείδει, θα λέγαμε πως η από μέρους του Dennett επίκληση στην αυτονομία δεν τροφοδοτεί το επιχείρημά του με στιβαρότητα και σαφήνεια. Η αυτονομία είναι ένα κριτήριο που για την ώρα χαρακτηρίζεται από εννοιολογική

ασάφεια αλλά και από πρόσθετες επιστημολογικού τύπου δυσκολίες, λόγω της συσχέτισής της (τουλάχιστον υπό την πιο δημοφιλή στο πεδίο της Ηθικής της Τ.Ν. Συναφειοκρατική της προσέγγιση) με τις έννοιες της νόησης και του προσώπου.

Το Επιχείρημα της Υπέρμετρης Αποτελεσματικότητας

Στην προσπάθειά του να τεκμηριώσει πειστικότερα το επιχείρημά του περί απόδοσης ευθυνών στα συστήματα Τ.Ν., ο Dennett υποστηρίζει ότι αναγνωρίζουμε και θαυμάζουμε την ικανότητα του υπολογιστή να κερδίζει στο σκάκι και συγχαίρουμε τους προγραμματιστές του για το επίτευγμα, όμως η νίκη ανήκει στον υπολογιστή και όχι στους προγραμματιστές. *Οι τελευταίοι, εάν αντιμετώπιζαν τον παγκόσμιο πρωταθλητή, προφανώς θα έχαναν μέσα σε λίγα λεπτά*. Εδώ φαίνεται ότι ο Dennett διατυπώνει ένα επιχείρημα επί τη βάσει της υπέρμετρης αποτελεσματικότητας του Deep Blue. Ο εν λόγω υπολογιστής έχει πράγματι αποδειχθεί εξαιρετικά αποτελεσματικός στο σκάκι και συγκεκριμένα έχει αποδειχθεί πολύ πιο αποτελεσματικός από τους ανθρώπους-προγραμματιστές του. Σύμφωνα με τον Dennett, αυτή η υπεροχή αποτελεσματικότητας τού υπολογιστή έναντι των ανθρώπων-προγραμματιστών του είναι επαρκής λόγος απόδοσης της νίκης στον υπολογιστή και όχι στους ανθρώπους-προγραμματιστές του. Θα μπορούσε η συγκεκριμένη επιχειρηματολογία του Dennett να ανοίξει τον δρόμο για μια επιτυχή απάντηση του ερωτήματος περί του ενδεχομένου να αποδώσουμε ηθικά δικαιώματα αλλά και ευθύνες στα συστήματα Τ.Ν.; Θα μπορούσε δηλαδή η υπέρμετρη αποτελεσματικότητα να αποτελέσει ένα στιβαρό, επαρκές, καθολικό κριτήριο απόδοσης ηθικού καθεστώτος στις οντότητες της Τ.Ν. αλλά και γενικότερα στις δρώσες οντότητες γύρω μας (ανθρώπους, ζώα, μηχανές κ.λπ.); Αυτό το ενδεχόμενο μας καλεί να εξετάσουμε το ακόλουθο ερώτημα: Έχει μέχρι σήμερα υπάρξει μια επιτυχής ανάλογη εφαρμογή του κριτηρίου της υπέρμετρης αποτελεσματικότητας στους ανθρώπους, στα ζώα ή στις μηχανές;

Όπως είδαμε στην αρχή του δοκιμίου μας, οι Max Tegmark (MIT) και Stuart Russell (Berkeley) επικαλούνται επίσης το κριτήριο της υπέρμετρης αποτελεσματικότητας για τον περιορισμό ή την απαγόρευση των αυτόνομων οπλικών συστημάτων. Εδώ γεννιέται το ερώτημα: Γιατί δεν αποδίδουμε ευθύνη και στα πυρηνικά ή τα χημικά όπλα επί τη βάσει της υπέρμετρης αποτελεσματικότητάς τους όπως ακριβώς μας καλεί να κάνουμε για τον Deep Blue ο Dennett; Τόσο ο υπερ-υπολογιστής αυτός όσο και τα πυρηνικά ή χημικά οπλικά συστήματα είναι μηχανές που επιδεικνύουν υπέρμετρη αποτελεσματικότητα. Περιορίζοντας τη συζήτηση μόνο στο επίπεδο της αποτελεσματικότητας, βλέπουμε ότι αν ο Deep Blue είναι πολύ πιο αποτελεσματικός από τους ανθρώπους-δημιουργούς του στο να κερδίζει μια παρτίδα σκάκι, εντελώς ανάλογα τα πυρηνικά και χημικά όπλα είναι πολύ πιο αποτελεσματικά από τους ανθρώπους-δημιουργούς τους στο να σκοτώνουν. Γιατί δεν έχει ως σήμερα αρθρωθεί κάποιο επιχείρημα απόδοσης ηθικής ευθύνης στα ίδια αυτά όπλα επί τη βάσει της αποτελεσματικότητάς τους όπως αρθρώνεται για τον Deep Blue; Η πιο πάνω έκκληση των Tegmark και Russell εξισώνει τα όπλα μαζικής καταστροφής με τα συστήματα της T.N. στη βάση μιας ανάλογης επικινδυνότητας, που με τη σειρά της θα λέγαμε ότι υπονοεί μια ανάλογη υπέρμετρη αποτελεσματικότητα. Αν η επικινδυνότητα και τελικά η υπεροχή αποτελεσματικότητας έναντι των ανθρώπων-δημιουργών είναι ανάλογη στις περιπτώσεις του Deep Blue και των όπλων μαζικής καταστροφής, γιατί δεν είμαστε πρόθυμοι να ανοίξουμε μια ανάλογη συζήτηση απόδοσης ηθικής ευθύνης στα ίδια τα όπλα μαζικής καταστροφής, όπως το επιχειρεί εδώ ο Dennett στην περίπτωση του Deep Blue; Φαίνεται ότι η χρήση του κριτηρίου της υπέρμετρης αποτελεσματικότητας δεν χρησιμοποιείται με τρόπο συνεπή ως προς τις μηχανές.

Θα μπορούσε κανείς στο σημείο αυτό να απαντήσει ότι, σε αντίθεση με τον Deep Blue και τα περισσότερα συστήματα T.N., τα όπλα μαζικής καταστροφής δεν δρουν με νοήμονα ή έστω με φαινομενικά νοήμονα τρόπο και, κατά

συνέπεια, δεν τίθεται θέμα απόδοσης ηθικής ευθύνης στα δεύτερα[35]. Ωστόσο, εδώ θα επισημαίναμε ότι με ένα τέτοιο επιχείρημα πρώτον καλείται κανείς να ορίσει τι εννοεί με τον όρο «νοήμονα» (ή «φαινομενικά νοήμονα»)[36] και, επομένως, να βρεθεί και πάλι ευθέως αντιμέτωπος με το πρόβλημα οριοθέτησης των όρων «νοημοσύνη» (intelligence) και «νόηση (cognition)[37]. Έχουμε, επομένως, ένα συμπεριφορικό κριτήριο απόδοσης ηθικού καθεστώτος, που όχι μόνο αποτυγχάνει να μας απαλλάξει από την ανάγκη ευθείας αντιμετώπισης του προβλήματος οριοθέτησης της νόησης, αλλά απεναντίας μας ρίχνει ακριβώς πάνω σε αυτό το επίπονο πρόβλημα[38]. Δεύτερον, εκτρέπει τη συζήτηση από

[35] Σε αυτό το σημείο πρέπει πάντως να επισημάνουμε ότι τα περισσότερα όπλα μαζικής καταστροφής κατευθύνονται και ελέγχονται πλέον από συστήματα Τ.Ν. Επομένως, η Τ.Ν. αποτελεί πια αναπόσπαστο τμήμα των όπλων μαζικής καταστροφής, στον βαθμό που τα τελευταία να μπορούν να ταξινομηθούν ως οπλικά συστήματα Τ.Ν. Ως εκ τούτου, η διάκριση ανάμεσα σε συστήματα της Τ.Ν. και σε όπλα μαζικής καταστροφής δεν στέκει στην πράξη. Ωστόσο, χάριν της πιο πάνω συζήτησης, ας υποθέσουμε ότι τα όπλα μαζικής καταστροφής δεν διακρίνονται από δυνατότητες Τ.Ν. και ανήκουν σε άλλης τάξεως κατηγορία μηχανών. Η ίδια η αναφορά των Tegmark και Russell τα αντιμετωπίζει έτσι, ώστε να πετύχει την επιθυμητή από αυτούς αντιδιαστολή και τελικά συσχέτιση με τα οπλικά συστήματα της Τ.Ν. όχι επί τη βάσει του γεγονότος ότι τα όπλα μαζικής καταστροφής επιστρατεύουν Τ.Ν. αλλά επί τη βάσει της ανάλογης επικινδυνότητάς τους. Μπορούμε, δε, θέλοντας να περιορίσουμε την ανάλυση στο στοιχείο της επικινδυνότητας, να αναφερθούμε μόνο στα παλαιότερα όπλα μαζικής καταστροφής, για παράδειγμα στις πρώτες ατομικές βόμβες, που ήταν απλές βόμβες άφεσης από αεροσκάφος και δεν διέθεταν ούτε καν το απλούστερο αυτόματο σύστημα πλοήγησης.

[36] Ήδη η διάκριση ανάμεσα σε «νοήμονα» και «φαινομενικά νοήμονα» τρόπο φέρνει ξανά στο προσκήνιο το Επιχείρημα του Κινέζικου Δωματίου και την πιθανότητα απλής μίμησης της νοήμονος συμπεριφοράς, μας θέτει δηλαδή αντιμέτωπους με παραδοσιακά ερωτήματα της Φιλοσοφίας του Νου που προσπαθούσαμε να αποφύγουμε εισάγοντας το κριτήριο της υπέρμετρης αποτελεσματικότητας.

[37] Για τις διαφορές ανάμεσα στους όρους «cognition» και «intelligence» δείτε και σημείωση 7.

[38] Αυτό μπορεί εύκολα να φανεί από το γεγονός ότι η πιο πάνω παράθεση επιχειρημάτων-αντεπιχειρημάτων οδηγεί τον απολογητή της θέσης του Dennett σε ένα σφάλμα διαλληλίας και, τελικά, σε μια ταυτολογία. Συγκεκριμένα, το αρχικό επιχείρημα του Dennett μπορεί να διατυπωθεί αφαιρετικά ως εξής: «Πρέπει να αποδίδουμε την ηθική ευθύνη μιας πράξης Π σε μιαν οντότητα Ο, όταν η Ο επιτελεί την Π με υπέρμετρη αποτελεσματικότητα». Για να αντιμετωπιστεί το αντεπιχείρημα ότι και οντότητες στις οποίες συνήθως δεν αποδίδουμε ηθική ευθύνη επιδεικνύουν ανάλογη υπέρμετρη αποτελεσματικότητα, το αρχικό επιχείρημα διασκευάστηκε ως εξής: «Πρέπει να αποδίδουμε την ηθική ευθύνη μιας πράξης Π σε μιαν οντότητα Ο, όταν η Ο επιτελεί την Π με έναν νοήμονα (ή φαινομενικά νοήμονα) τρόπο [και με μιαν υπέρμετρη αποτελεσματικότητα]». Όμως, ούτως ή άλλως, ακόμα και σε επίπεδο της καθημερινής ψυχολογικής γλώσσας (naive psychology), η απόδοση ηθικής ευθύνης σε μιαν οντότητα υπονοεί

το κριτήριο της υπέρμετρης αποτελεσματικότητας, εισάγοντας έναν επιπλέον όρο –αυτόν του νοήμονα (ή φαινομενικά νοήμονα) τρόπου– ο οποίος μάλιστα δείχνει πλέον να είναι και πιο επαρκής ή καθοριστικός από αυτόν που αρχικά επιχειρούσαμε να στηρίξουμε (δηλαδή τον όρο της υπέρμετρης αποτελεσματικότητας). Αν, εν τέλει, ό,τι διακρίνει τα συστήματα της Τ.Ν. από τα όπλα μαζικής καταστροφής είναι ο νοήμων (ή φαινομενικά νοήμων) τρόπος, ποιος ο λόγος να αναφερόμαστε στο κριτήριο της υπέρμετρης αποτελεσματικότητας; Η εστίαση της ανάλυσής μας έχει πλέον μετατοπιστεί οριστικά προς ένα άλλο κριτήριο. Οποιαδήποτε αναφορά στην υπέρμετρη αποτελεσματικότητα δείχνει πια να πλεονάζει. Αν, δε, εξετάσουμε με προσοχή τον τρόπο με τον οποίο αντιπαρατέθηκαν τα πιο πάνω επιχειρήματα, η υπέρμετρη αποτελεσματικότητα φαίνεται τελικά να είναι στοιχείο ομοιότητας και όχι ειδοποιού διαφοράς ανάμεσα στα συμβατικά-μη σχετιζόμενα με Τ.Ν. όπλα μαζικής καταστροφής και τα συστήματα της Τ.Ν. Η δικαιολογημένη επισήμανση της παρόμοιας αποτελεσματικότητας ήταν αυτή που επιστράτευσε το κριτήριο του νοήμονα (ή φαινομενικά νοήμονα) τρόπου δράσης.

Συνεχίζοντας την ανάλυσή μας ως προς τη χρήση του κριτηρίου της αποτελεσματικότητας στις μηχανές, ας παρακάμψουμε, χάριν της συζήτησης, το πρόβλημα ορισμού της έννοιας «νοήμων τρόπος δράσης». Ο ασυνεπής τρόπος με τον οποίο χρησιμοποιείται το κριτήριο της αποτελεσματικότητας δύναται να φανερωθεί ακόμα και αν η ανάλυσή μας περιοριστεί εντός του συνόλου των συστημάτων της Τ.Ν. Συγκεκριμένα, αν και προτείνεται η απόδοση ηθικών δικαιωμάτων στον HAL 9000, δεν συμβαίνει ωστόσο κάτι ανάλογο στα οπλικά

ότι η οντότητα αυτή είναι νοήμων (π.χ. χαρακτηρίζεται από αποβλεπτικού τύπου νοητικές καταστάσεις). Συνεπώς, τι περισσότερο ήρθε να προσθέσει έναντι της καθημερινής γλωσσικής προσέγγισης το κριτήριο της υπέρμετρης αποτελεσματικότητας; Η τελευταία εκδοχή του επιχειρήματος γίνεται πλέον η εξής: «Μια οντότητα Ο είναι νοήμων, αν δρα με έναν νοήμονα (ή φαινομενικά νοήμονα) τρόπο». Εδώ πρέπει να προσέξουμε ότι, αν επιλέξουμε την εκδοχή με τον όρο «νοήμονα», έχουμε μια ταυτολογία (ακόμα και αν διακρίνουμε ανάμεσα στους όρους «νοημοσύνη» και «νόηση», η γνωσιολογική αξία της πιο πάνω πρότασης δεν γίνεται ανώτερη μιας ταυτολογίας, καθώς η νοημοσύνη είναι υποσύνολο της νόησης). Αν πάλι επιλέξουμε την εκδοχή με τον όρο «φαινομενικά νοήμονα», «γλιτώνουμε» την ταυτολογία, αλλά βρισκόμαστε αντιμέτωποι με το Επιχείρημα του Κινέζικου Δωματίου.

συστήματα της Τ.Ν. Η υπέρμετρη αποτελεσματικότητα χαρακτηρίζει τόσο τον πρώτο όσο και τα δεύτερα. Γιατί, λοιπόν, ανοίγουμε τη συζήτηση απόδοσης ηθικών δικαιωμάτων μόνο στον HAL; Ποια είναι η ειδοποιός διαφορά μεταξύ τους; Μήπως ότι ο HAL μετέχει σε μια κατεξοχήν ανθρώπινη δράση ως μέλος μιας διαστημικής αποστολής, σε αντίθεση με τα όπλα μαζικής καταστροφής (καθώς το να σκοτώνεις δεν είναι ίδιον μόνο των ανθρώπων). Ωστόσο, θα απαντούσαμε εδώ ότι έτσι:

1. υποκαθιστούμε και πάλι το κριτήριο της υπέρμετρης αποτελεσματικότητας με ένα άλλο κριτήριο, συγκεκριμένα με το κριτήριο του πεδίου των ανθρώπινων δράσεων, οπότε και βρισκόμαστε αντιμέτωποι με το πρόβλημα σαφούς οριοθέτησης του πεδίου αυτού και
2. αποδεχόμαστε μια οριοθέτηση του όρου «νόηση» που ταυτίζεται αποκλειστικά με την οριοθέτηση του όρου «ανθρώπινη δράση» και, επομένως, αποδεχόμαστε σιωπηρά μια άρνηση απόδοσης της ιδιότητας του νοήμονος όντος στα ζώα , ζήτημα που είναι ακόμα υπό συζήτηση και για το οποίο πολλοί από αυτούς που θα ήθελαν να αρνηθούν την απόδοση ηθικών δικαιωμάτων στα οπλικά συστήματα της Τ.Ν. απαντούν θετικά, τασσόμενοι δε υπέρ των δικαιωμάτων των ζώων.

Βλέπουμε, εν τέλει, ότι ακόμα και αν περιορίσουμε τη συζήτηση εντός του πεδίου των συστημάτων της Τ.Ν., η χρήση του κριτηρίου της υπέρμετρης αποτελεσματικότητας παραμένει ασυνεπής από μέρους μας.

Ως προς τη χρήση του εν λόγω κριτηρίου στους ανθρώπους, τα πράγματα δεν είναι καλύτερα. Είναι κοινώς αποδεκτό και εμπράκτως επιβεβαιωμένο ότι το ανθρώπινο είδος παρουσιάζει μια αξιοσημείωτη ποικιλομορφία δεξιοτήτων, που πάντως δεν διανέμονται ομοιόμορφα εντός του. Οι άνθρωποι ποικίλλουν ως προς τις ιδιαίτερες ικανότητές τους, τα «ταλέντα» τους αλλά και ως προς τις αδυναμίες τους. Ωστόσο, προσπαθούμε να μην είναι ανάλογα ποικιλόμορφη η απόδοση ηθικών ευθυνών ή δικαιωμάτων σε αυτούς, δίχως πάντως αυτό να επιτυγχάνεται απολύτως. Συχνά οι άνθρωποι θεωρούνται ηθικά υπεύθυνοι για πράξεις τους σε πεδία δράσης στα οποία δεν επιδεικνύουν υπέρμετρη αποτελε-

σματικότητα, ενώ υπάρχουν περιπτώσεις στις οποίες επιχειρείται η άμβλυνση της απόδοσης ηθικών ευθυνών για πράξεις στις οποίες οι άνθρωποι επιδεικνύουν υπέρμετρη αποτελεσματικότητα. Χαρακτηριστική η περίπτωση ατόμων που έχουν υποστεί βλάβη σε συγκεκριμένα εγκεφαλικά κέντρα επιφορτισμένα με την πρόκληση και τον έλεγχο των λεγόμενων φιλο-κοινωνικών συναισθημάτων. Συχνά τέτοια άτομα καταλήγουν να γίνονται κατά συρροή δολοφόνοι, καθώς συνδυάζουν μια υψηλή ικανότητα και τελικά αποτελεσματικότητα στο να σχεδιάσουν έναν φόνο με μια έλλειψη ηθικών αναστολών (Allely, et al., 2014). Τα άτομα αυτά συχνά επιχειρείται να αντιμετωπιστούν ως νοητικά νοσούντα και, επομένως, ως άτομα που έχουν μειωμένη αυτονομία, δηλαδή ως άτομα των οποίων η βούληση είναι ετερονομημένη από την ασθένειά τους. Βλέπουμε, εν τέλει, ότι όχι μόνο η απόδοση ηθικού καθεστώτος δεν είναι συμμετρική με την επίδειξη υπέρμετρης αποτελεσματικότητας –ότι δηλαδή η οριοθέτηση του ηθικού καθεστώτος δεν σχετίζεται με τη οριοθέτηση της όποιας υπέρμετρης αποτελεσματικότητας–, αλλά μάλλον ότι βασίζεται σε εντελώς άλλα κριτήρια όπως, για παράδειγμα, αυτό της αυτονομίας, γεγονός που μας επαναφέρει στην προηγηθείσα συζήτηση για το αυτόνομο δρᾶν. Αν, λοιπόν, στην περίπτωση των ανθρώπων αποφεύγουμε να συνδέσουμε την αποτελεσματικότητα με την απόδοση ηθικής ευθύνης, γιατί να το πράξουμε στην περίπτωση των συστημάτων Τ.Ν.;

Μάλιστα, ως προς την περίπτωση των μηχανών αλλά και των ζώων, η υπέρμετρη αποτελεσματικότητα έχει χρησιμοποιηθεί άλλοτε ως ένδειξη μιας άνοης, «αυτόματης» –και, επομένως, υποδεέστερης έναντι της ανθρώπινης– φύσης και άλλοτε ως απόδειξη της ηθικής ανωτερότητας των οντοτήτων αυτών έναντι των ανθρώπων. Σχετικώς, πρώτος ο Descartes υποστήριξε ότι η από μέρους μιας οντότητας επίδειξη μιας υπέρμετρης αποτελεσματικότητας σε συγκεκριμένες συμπεριφορικές πτυχές είναι μια ασφαλής ένδειξη –επομένως ένα στιβαρό συμπεριφορικό κριτήριο– της αυτόματης φύσης της οντότητας αυτής (Descartes, 1646, *Letter to the Marquess of Newcastle*. Δείτε στο: Cottingham, Stoothoff & Murdoch, 2006, σ. 304). Κατά τον Descartes, τα «αυτόματα» (ζώα και μηχανές) λειτουργούν όχι βάσει του Λόγου, αλλά αποκλειστικά βάσει των

ιδιαιτεροτήτων του σώματός τους και, ως εκ τούτου, παρουσιάζουν μια υπέρμετρη αποτελεσματικότητα σε συγκεκριμένους τομείς όπως τούτη εξασφαλίζεται από αυτές τις υλισμικού τύπου ιδιαιτερότητες (Descartes, 1637 *Discourse of the Method,* part 5. Δείτε στο: Cottingham, Stoothoff & Murdoch, 2006, σ. 141). Πρόκειται για μια θέση που έχει υιοθετηθεί και από ορισμένους σύγχρονους ερευνητές της Τ.Ν., ώστε να υπάρξει ένα συμπεριφορικό κριτήριο επί του οποίου θα μπορούσε να εδραιωθεί μια ασφαλής περί νοημοσύνης κρίση στα πλαίσια της Δοκιμασίας Turing (Michie, 2002). Εδώ, τη θέση των υλισμικών ιδιαιτεροτήτων έχουν πάρει ό,τι θα αποκαλούσαμε ως «λογισμικές ιδιαιτερότητες», δηλαδή η εξειδίκευση του προγράμματός τους. Μια εκ διαμέτρου αντίθετη χρήση του κριτηρίου της υπέρμετρης αποτελεσματικότητας γίνεται από έναν σχεδόν σύγχρονο του Descartes φιλόσοφο, τον Michel de Montaigne, κατά την από μέρους του τελευταίου εισήγηση του δόγματος της Θηριοφιλίας. Σύμφωνα με τον Montaigne, το γεγονός ότι τα ζώα επιδεικνύουν μια θαυμαστή και πολύ υψηλότερη έναντι των ανθρώπων αποτελεσματικότητα σε συγκεκριμένες δράσεις στοιχειοθετεί επαρκή λόγο θεώρησης μιας ανωτερότητας των ζώων έναντι των ανθρώπων και, τελικά, του δικαιώματός τους να απολαμβάνουν ένα ηθικό καθεστώς πλήρως σεβαστό από τον άνθρωπο (Montaigne, 1580, *An Apology for Raymond Sebond,*. Δείτε στο: Screech, 2003, *Montaigne Essays*, Book II, chapter 12).

Φαίνεται, συνεπώς, ότι η φιλοσοφική ανάλυση δεν έχει κατασταλάξει ως προς τη σχέση της υπέρμετρης αποτελεσματικότητας με την απόδοση νοητικών δυνατοτήτων και, τελικά, ενός οντολογικού καθεστώτος που θα μπορούσε να σχετίζεται με την απόδοση ηθικού καθεστώτος. Απεναντίας, η ως τώρα συζήτηση χαρακτηρίζεται από εκ διαμέτρου αντίθετους τρόπους χρήσης του κριτηρίου της υπέρμετρης αποτελεσματικότητας. Στον βαθμό που η ηθική ευθύνη συναρτάται με τη νόηση, μπορούμε να θεωρήσουμε ότι η πιο πάνω θέση του Dennett, βάσει της οποίας η υπέρμετρη αποτελεσματικότητα στοιχειοθετεί επαρκή λόγο απόδοσης της νίκης στον Deep Blue και της ηθικής ευθύνης στον HAL, είναι εκ διαμέτρου αντίθετη με αυτήν του Descartes και όσων σύγχρονων

ερευνητών της Τ.Ν. επιχειρούν να εδραιώσουν τη Δοκιμασία Turing συσχετίζοντας την υπέρμετρη αποτελεσματικότητα με τη μηχανική, άνοη φύση. Κατά τον Descartes και όσους υιοθετούν την υπέρμετρη αποτελεσματικότητα ως ένδειξη «αυτόματης» φύσης, ο Deep Blue δεν θα δικαιούνταν την απόδοση της νίκης ακριβώς επί τη βάσει της υπέρμετρης αποτελεσματικότητας που επέδειξε. Αντιθέτως, κατά τον Dennett, η υπέρμετρη αποτελεσματικότητα είναι επαρκής λόγος απόδοσης της ευθύνης της νίκης στον Deep Blue και της ηθικής ευθύνης στον HAL. Θα μπορούσε κανείς να υποστηρίξει ότι η θέση του Dennett δείχνει πιο συμβατή με αυτήν του Montaigne. Ωστόσο, σε αντίθεση με ό,τι υποστηρίζει ο Montaigne για τα ζώα, ο Dennett δεν χρησιμοποιεί την υπέρμετρη αποτελεσματικότητα για να υποστηρίξει μια ανωτερότητα του Deep Blue και κατ' αναλογία του HAL έναντι των ανθρώπων. Μάλλον διεκδικεί μια εξίσωση των εν λόγω υπερ-υπολογιστών με τους ανθρώπους. Υπό μια αδρά περιγραφή, φαίνεται πως τελικά έχουμε ως τώρα τρεις διαφορετικούς τρόπους χρήσης της υπέρμετρης αποτελεσματικότητας από τους φιλοσόφους:

1. την υπό τον Descartes χρήση της υπέρμετρης αποτελεσματικότητας ως τεκμηρίου κατωτερότητας άλλων οντοτήτων έναντι των ανθρώπων,
2. την υπό τον Montaigne χρήση της ως τεκμηρίου ανωτερότητας άλλων οντοτήτων έναντι των ανθρώπων και
3. την υπό το Dennett χρήση της ως τεκμηρίου εξίσωσης των άλλων οντοτήτων με τους ανθρώπους.

Ποια εξ αυτών είναι ορθή; Το μόνο που μπορούμε, για την ώρα, να αποφανθούμε με σιγουριά είναι πως το κριτήριο της υπέρμετρης αποτελεσματικότητας δεν χρησιμοποιείται με σταθερό και μονοσήμαντο τρόπο κατά την οντολογική σύγκριση των ανθρώπων με άλλες οντότητες. Αυτό οδηγεί λογικά και σε έναν ανάλογα μη μονοσήμαντο τρόπο χρήσης της υπέρμετρης αποτελεσματικότητας κατά την απόδοση ηθικού καθεστώτος στις οντότητες αυτές.

Συμπεράσματα

Στο παρόν άρθρο εξετάσαμε το ζήτημα απόδοσης ηθικού καθεστώτος στα συστήματα της Τ.Ν. Η ανάλυσή μας αναφέρεται στα πολεμικά συστήματα της Τ.Ν. λόγω της σφοδρότητας των συνεπειών που προκύπτουν από τη δράση τους, η οποία είναι και ανάλογη της οξύτητας των ηθικής φύσεως ερωτημάτων που η τελευταία αυτή εγείρει. Αναφερθήκαμε, δηλαδή, ειδικά στην περίπτωση των πολεμικών συστημάτων της Τ.Ν., διότι συνιστά το οξύτερο και πλέον πιεστικό πλαίσιο εντός του οποίου οι φιλόσοφοι και οι ερευνητές της Τ.Ν. καλούνται να αντιμετωπίσουν το ερώτημα απόδοσης ηθικού καθεστώτος σε οντότητες Τ.Ν. Ωστόσο, θεωρούμε ότι τα επιχειρήματα και συμπεράσματα που αρθρώθηκαν έχουν γενική ισχύ για κάθε σύστημα που χαρακτηρίζεται ως σύστημα Τ.Ν., καθώς δεν βασίστηκαν σε πτυχές που προσιδιάζουν μόνο στα αυτόνομα οπλικά συστήματα αλλά, αντιθέτως, θα μπορούσαν να εφαρμοστούν σε κάθε άλλο σύστημα Τ.Ν. Επιπλέον, επιλέξαμε να βασίσουμε την ανάλυσή μας σε έναν σκεπτικιστικό αντίλογο στα επιχειρήματα του Dennett όπως αυτά εκφράζονται στο κείμενό του *When HAL Kills, Who's to Blame? Computer Ethics*, το οποίο θεωρείται ως ένα ορόσημο της σύγχρονης φιλοσοφικής ανάλυσης υπέρ της απόδοσης ηθικού καθεστώτος στις μηχανές. Η δε αναφορά στον HAL και τον φόνο που αυτό το σύστημα διαπράττει στη διάσημη πλέον ταινία *A Space Odyssey* καθιστά την ανάλυση του Dennett συναφή με τα ζητήματα της Ηθικής της Τ.Ν. για τα αυτόνομα οπλικά συστήματα.

Συγκεκριμένα, θεωρούμε ότι η επιχειρηματολογία του Dennett τεκμηριώνεται, μεταξύ των άλλων, με τρία βασικά επιχειρήματα: την αναλογία της σχέσης προγραμματιστή-μηχανής με τη σχέση προπονητή-αθλητή (ένα επιχείρημα που μας οδηγεί στην εξέταση της οντολογικής και λειτουργικής ισοδυναμίας τόσο των ρόλων όσο και της νοημοσύνης των συστημάτων), το επιχείρημα περί της μηχανικής αυτονομίας (machine autonomy) και το επιχείρημα της υπέρμετρης αποτελεσματικότητας (με τα δύο τελευταία να χρησιμοποιούνται από αρκετούς και ως επαρκής βάση για τη σύσταση κριτηρίων απόδοσης της ιδιότητας του ηθικού προσώπου σε ένα σύστημα Τ.Ν.).

Ως προς το πρώτο επιχείρημα, καταδείξαμε ότι ήδη η υποστήριξη μιας αναλογίας μεταξύ των σχέσεων προγραμματιστή-μηχανής και προπονητή-αθλητή για

την απόδοση ηθικού καθεστώτος στη μηχανή είναι λογικά εσφαλμένη. Πρώτον, διότι συνιστά σφάλμα διαλληλίας, καθώς προϋποθέτει το ζητούμενο, δηλαδή την οντολογική ισοδυναμία ανθρώπου-μηχανής, και, δεύτερον, σε περίπτωση που θεωρηθεί ότι η ισοδυναμία αυτή είναι εν τέλει λειτουργική, έρχεται αντιμέτωπη με τα προβλήματα λογικής θεμελίωσης του Λειτουργισμού Μηχανής. Επιπροσθέτως, η περίπτωση της λειτουργικής ισοδυναμίας έχει να αντιμετωπίσει και τα οντολογικά ζητήματα που προκύπτουν από τον Λειτουργισμό Μηχανής.

Ως προς το επιχείρημα της αυτονομίας, παρατηρήσαμε ότι ο από μέρους του Dennett παραλληλισμός του περιβάλλοντος με τους προγραμματιστές και, επομένως, η αμφισβήτηση της άνευ όρων απεριόριστης αυτονομίας του ανθρώπου, έχει αρχικώς κάποια βάση. Ωστόσο, καταδείξαμε ότι η επίκληση του Dennett στο κριτήριο της αυτονομίας αντιμετωπίζει το πρόβλημα της εννοιολογικής ασάφειας, όπως αυτή προκύπτει μέσα από μια πολυαρχία ορισμών της αυτονομίας. Επιπλέον, υπό την πλέον δημοφιλή στο πεδίο της Ηθικής της Τ.Ν. προσέγγιση του αυτόνομου δρᾶν, δηλαδή υπό την «εσωτερικιστική» –και δη τη Συναφειοκρατική– προσέγγιση, βρίσκεται κανείς αναπόφευκτα ευθέως αντιμέτωπος με το Πρόβλημα των Άλλων Νόων, αλλά και με παραδοσιακούς ως προς την ιδιότητα του προσώπου προβληματισμούς, όπως, επί παραδείγματι, το ερώτημα της διατήρησης (persistence question) και του χαρακτηρισμού (characterization question). Επίσης, καταδείξαμε ότι η επίκληση στο κριτήριο της αυτονομίας μάς φέρει αντιμέτωπους με την ασάφεια κατά την οριοθέτηση της βούλησης που οδηγεί σε μια μη συμμετρικότητα στη σχέση αυτονομίας-απόδοσης ηθικού καθεστώτος, δηλαδή σε μιαν ασυνεπή χρήση του κριτηρίου της αυτονομίας.

Ασυνεπής αποδεικνύεται ως τώρα και η χρήση του κριτηρίου της υπέρμετρης αποτελεσματικότητας, όχι μόνο ως προς τη διάκριση μεταξύ ανθρώπων-μηχανών αλλά και ως προς τη διάκριση ανθρώπων-ζώων. Η δε εφαρμογή τού εν λόγω κριτηρίου δείχνει να γίνεται με έναν αυθαίρετα επιλεκτικό τρόπο μόνο προς τα συστήματα της Τ.Ν και όχι και ως προς άλλες μηχανές, όπως τα όπλα μαζικής καταστροφής για παράδειγμα.

Το σύνολο των ως άνω αντεπιχειρημάτων θεωρούμε ότι απαντούν σε ένα μεγάλο μέρος της σημερινής συζήτησης για την κυριολεκτική απόδοση ηθικού καθεστώτος στις ευφυείς αυτόνομες μηχανές –και δη στα αυτόνομα οπλικά συστήματα–, συνεπώς και στην απόδοση σε αυτές τις μηχανές των χαρακτηριστικών του ηθικού προσώπου που έχει την ευθύνη των πράξεών του. Θεωρούμε ότι οποιαδήποτε συζήτηση προς αυτή την κατεύθυνση μπορεί προς το παρόν και μέχρι να επιλυθούν ικανοποιητικά τα οντολογικά και επιστημονικά προβλήματα που συνδέονται με την ανθρώπινη νόηση και την τεχνητή νοημοσύνη να γίνεται με μεταφορική χρήση των εννοιών της προσωπικότητας, της αυτονομίας και της ηθικής ευθύνης. Δεν πρέπει να παραβλέπουμε το γεγονός ότι είναι διαφορετικό να επιδεικνύουν οι μηχανές συμπεριφορά ή να προβαίνουν σε πράξεις οι οποίες μπορούν να αξιολογηθούν ηθικά και διαφορετικό να έχουν οι ίδιες οι μηχανές ηθική ευθύνη για τις πράξεις αυτές (Anderson, & Anderson, 2007, σσ. 15-26). Άλλωστε, όπως έχουν τονίσει οι πρωτοπόροι του προγράμματος Machine Ethics, ενδεχομένως να αποτελεί ερευνητική ματαιοδοξία η αναπαραγωγή της ανθρώπινης νόησης, της ανθρώπινης νοημοσύνης και των ανθρώπινων, ενδεχομένως «ατελών», ηθικών αξιών στα συστήματα Τ.Ν. Οι ίδιοι περιόρισαν σκοπίμως το ερευνητικό τους ενδιαφέρον σε χρήσιμα και πρακτικά προβλήματα τα οποία απέχουν από την «κυριολεκτικά» αυτόνομη ηθική δράση των μηχανών (Anderson, Anderson, Gounaris, & Kosteletos, 2021, σ. 180).

Αντίθετα, η τοποθέτηση του Dennett τάσσεται υπέρ της «θεωρητικής» προς το παρόν απόδοσης ηθικού καθεστώτος στα συστήματα της Τ.Ν. Η παρούσα ανάλυσή μας δεν στόχευσε στην υποστήριξη μιας αντίθετης θέσης, δηλαδή μιας θέσης κατά της απόδοσης ηθικού καθεστώτος στα συστήματα αυτά, αλλά στην κατάδειξη του γεγονότος ότι βάσει της ως τώρα συνήθους επιχειρηματολογίας επί του θέματος, αυτό δεν μπορεί παρά να παραμείνει *αναποκρίσιμο* (undecidable). Βρισκόμαστε, έτσι, για άλλη μια φορά αντιμέτωποι με μια συγκρουσιακή κατάσταση, γνωστή σε όσους εργάζονται στο πεδίο της Εφαρμοσμένης Φιλοσοφίας, και συγκεκριμένα με το δίλημμα ανάμεσα στην απαίτηση για σαφή και έγκυρα κριτήρια λήψης αποφάσεων και την ατέρμονη φύση μιας ανάλυσης που θέλει να είναι συνεπής.

Βιβλιογραφικές αναφορές

Allely, C., Minnis, H., Thompson, L., Wilson, P. and Gilberg, C. (2014). "Neurodevelopmental and psychosocial risk factors in serial killers and mass murderers". In *Aggression and Violent Behavior* 19, pp. 288-301.

Anderson, M., Anderson, S. (2007). "Machine Ethics: Creating an Ethical Intelligent Agent". In *AI Magazine*, 28 (4), pp. 15-26.

Anderson, M., Anderson, S., Gounaris, A. & Kosteletos, G. (2021). "Towards Moral Machines: A Discussion with Michael Anderson and Susan Leigh Anderson". In *Conatus - Journal of Philosophy*, 6 (1), pp. 177-202.

Avramides, A. (2001). *Other Minds*. London, New York: Routledge.

Benacerraf, P. (1967). "God, the Devil and Gödel". In *The Monist* 51, pp. 9-32.

Block, N. (1980a). "Are Absent Qualia Impossible?". In *Philosphical Review*, 89, pp. 257-274.

_____ (1980b). "Troubles with Functionalism". In *Readings in Philosophy of Psychology*, vol. 1, edited by Ned Block, pp. 268-305. Cambridge: Harvard University Press.

Bratman, M. (1979). "Practical Reasoning and Weakness of the Will". In *Noûs,* 13 (2), pp. 153-171.

_____ (2007). *Structures of Agency: Essays*. Oxford: Oxford University Press.

Buss, S., Westlund, A. (2018). "Personal Autonomy". In *The Stanford Encyclopedia of Philosophy* (Spring 2018 Edition). Edward N. Zalta (ed.). URL = <https://plato.stanford.edu/archives/spr2018/entries/personal-autonomy/>.

Calverley, D. (2005). "Toward a method for determining the legal status of a conscious machine". In *Proceedings of the AISB 2005 symposium on next generation approaches to machine consciousness: imagination, development, intersubjectivity, and embodiment*, edited by R. Chrisley, R. Clowes, and S. Torrance, pp. 75-84. Hatfield: University of Hertfordshire.

Casti, J. L., & De Pauli, W. (2000). *Gödel, a Life of Logic*. Cambridge: Basic Books.

Christman, J. (2018). "Autonomy in Moral and Political Philosophy". In *The Stanford Encyclopedia of Philosophy* (Spring 2018 Edition). Edward N. Zalta (ed.). URL = <https://plato.stanford.edu/archives/spr2018/entries/autonomy-moral/>.

Churchland, P. M. (1988). *Matter and Consciousness* (revised edition). Cambridge, Massachusetts: MIT Press.

Clarke, A. C. (1968). *2001: A Space Odyssey*. New York: New American Library.

De Landa, M. (1991). *War in the Age of Intelligent Machines*. New York: Swerve Editions.

Delvaux, M. (2016). Draft Report. Committee on Legal Affairs. European Parliament https://www.europarl.europa.eu/doceo/document/JURI-PR-582443_EN.pdf

Dennett, D. (1988). "Quining Qualia". In *Consciousness in Contemporary Science*, edited by Antony J. Marcel and E. Bisiach, pp. 42-77. New York: Oxford University Press.

_____ (1997). "When Hal Kills, Who's to Blame? Computer Ethics". In *Hal's Legacy: 2001's Computer as Dream and Reality*, edited by David G. Stork, pp. 351-365. Cambridge, MA: MIT Press.

De Quincey, C. (2006). "Switched-on consciousness: clarifying what it means". In *Journal of Consciousness Studies*, 13 (4), pp. 6-10.

Descartes, R. (1637/2006). "Discourse of the Method of Rightly Conducting One's Reason and of Seeking Truth in the Sciences". In *The Philosophical Writings of Descartes: Volume 1*, translated by John Cottingham, Robert Stoothoff, and Murdoch Dugald, pp. 111-151. New York: Cambridge University Press.

_____ (1646/2006). "Letter to the Marquess of Newcastle". In *The Philosophical Writings of Descartes: Volume 3*, translated by John Cottingham, Robert Stoothoff, and Murdoch Dugald, pp. 302-304. New York: Cambridge University Press.

Dreyfus, H. L. (1992). *What computers still can't do: α critique of artificial reason*. New York: MIT Press.

Erion, G. J. (2001). "The Cartesian test for Automatism". In *Minds and Machines*, 11 (2), pp. 29-39.

Floridi, L., Sanders, J. W. (2004). "On the morality of artificial agents". In *Minds and Machines*, 14, pp. 349-379.

Frankfurt, H. (1988a). "Freedom of the Will and the Concept of a Person". In *The Importance of What We Care About*, edited by Harry Frankfurt, pp. 11-25. Cambridge: Cambridge University Press.

_____ (1988b). "The Importance of What We Care About". In *The Importance of What We Care About*, edited by Harry Frankfurt, pp. 80-94. Cambridge: Cambridge University Press.

_____ (1999). "On Caring". In *Necessity, Volition and Love*, edited by Harry Frankfurt, pp. 155-180. Cambridge: Cambridge University Press.

Frankish, K. (2016). "Illusionism as a Theory of Consciousness". In *Journal of Consciousness Studies*, 23 (11-12), pp. 11–39. Accessed July 25, 2020. https://www.keithfrankish.com/illusionism-as-a-theory-of-consciousness/

_____ (2017). *Illusionism: As a Theory of Consciousness*. Exeter: Imprint Academic Publishing.

Gounaris, A. (2013). *Human Cognition and Artificial Intelligence: Searching for the fundamental differences of meaning in the boundaries of metaphysics*. Accessed January 14, 2019. https://alkisgounaris.gr/gr/research/human-cognition-artificial-intelligence/. DOI: 10.13140/RG.2.2.17433.67681.

_____ (2014). *Can we talk about Intentionality in Eliminative Materialism? The point of view of Embodied Cognition Theories*. https://www.researchgate.net/publication/331354816_Can_we_talk_about_Intentionality_in_Eliminative_Materialism_The_point_of_view_of_Embodied_Cognition_Theories. Accessed August 14, 2022. DOI: 10.13140/RG.2.2.18587.11041

Gunderson, K. (1964). "Descartes, La Mettrie, Language, and Machines". In *Philosophy*, 39 (149), pp. 193-222.

Gunkel, D. (2012). *The Machine Question: critical perspectives on AI, robots and ethics*. Cambridge, Massachusetts: MIT Press.

Hajdin, M. (1994). *The boundaries of moral discourse*. Chicago: Loyola University Press.

Hardcastle, V. (1999). *The Myth of Pain. Cambridge*. Massachusetts: MIT Press.

Haugeland, J. (1985). *Artificial intelligence: The very idea*. Cambridge, Massachusetts: MIT Press. Chicago.

Hoffmann, C., Hahn, B. (2020). "Decentered ethics in the machine era and guidance for AI regulation". In *AI & Society*, 35(3), pp. 635-644.

Jaworska, A. (2007a). "Caring and Full Moral Standing". *Ethics*, 117 (3), pp. 460-97.

_____ (2007b). "Caring and Internality". *Philosophy and Phenomenological Research*, 74 (3), pp. 529-568.

_____ (2009). "Caring, Minimal Autonomy, and the Limits of Liberalism". In *Naturalized Bioethics: Toward Responsible Knowing and Practice*, edited by Hilde Lindemann, Marian Verkerk, and Margaret Walker, pp. 80-105. Cambridge: Cambridge University Press.

Kim, J. (1993). *Supervenience and the Mind*: *Selected philosophical essays*. Cambridge: Cambridge University Press.

_____ (2005). *Η Φιλοσοφία του Νου*. Αθήνα: Leader Books.

Lang, F. (2020). "AI Flawlessly Beats US Air Force F-16 Pilot in Simulated Dogfight". *Interesting Engineering*. Accessed August 21, 2020. https://interestingengineering.com/ai-flawlessly-beats-us-air-force-f-16-pilot-in-simulated-dogfight

Lee, P. (2022). "Bladed 'Ninja' missile used to kill al-Qaida leader is part of a scary new generation of unregulated weapons". https://theconversation.com/bladed-ninja-missile-used-to-kill-al-qaida-leader-is-part-of-a-scary-new-generation-of-unregulated-weapons-188316

Levy, D. (2007). *Intimate relationships with artificial partners*. Ph.D. Diss., Maastricht University.

_____ (2009). "The ethical treatment of artificially conscious robots". In *International Journal of Social Robotics*, 1 (3), pp. 209-216.

Lucas, J. R. (1961). "Minds, Machines and Gödel". *Philosophy, XXXVI*, pp. 112-127.

Matthias, A. (2004). "The responsibility gap: Ascribing responsibility for the actions of learning automata". *Ethics and Information Technology*, 6(3) (2004), pp. 175-183.

Michie, D. (2002). "Turing's Test and Conscious Thought". In *Machines and Thought. The Legacy of Alan Turing*, vol.1, edited by Peter Millican, and Andy Clark, pp. 27-51. Oxford, New York: Oxford University Press.

Ministry of Defense. (2011). *Joint Doctrine Note 2/11, The UK Approach To Unmanned Aircraft Systems*. Accessed July 20, 2020. https://www.law.upenn.edu/live/files/3890-uk-ministry-of-defense-joint-doctrine-note-211-the

Monroe, A. E., Dillon, K. D., Malle, B. F. (2014). "Bringing free will down to earth: people's psychological concept of free will and its role in moral judgment". *Consciousness and Cognition*, 27, pp. 100-108.

Montaigne, Michel de. (1580/2003). "An Apology for Raymond Sebond". In *Michel de Montaigne: The complete essays*, translated by Michael A. Screech. Penguin Books.

Moore, E. F. (1956). "Artificial Living Plants". *Scientific American*, 195(4), pp. 118-126.

Müller, V. C. (2012). "Autonomous Cognitive Systems in Real-World Environments: Less Control, More Flexibility and Better Interaction". *Cognitive Computation*, 4(3), pp. 212-215.

_____ (2020). "Ethics of Artificial Intelligence and Robotics". In *The Stanford Encyclopedia of Philosophy* (Fall 2020 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2020/entries/ethics-ai/>

Neumann, J. von (1966). *Theory of Self-Reproducing Automata*. Urbana: University of Illinois Press.

Olson, E. T. (2019). "Personal Identity". In *The Stanford Encyclopedia of Philosophy* (Fall 2019 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2019/entries/identity-personal/>

Owen, J., Osley, R. (2007). "Bill of rights for abused robots: experts draw up an ethical charter to prevent humans exploiting machines". In *The Independent*. Last modified April 1, 2007. https://www.independent.co.uk/news/science/bill-of-rights-for-abused-robots-5332596.html

Pagallo, U. (2011). "Robots of just war: a legal perspective". *Philosophy & Technology*, 24(3), pp. 307-323.

Pinker, S. (1997). "Can a computer ever be conscious?". In *US News & World Report* 123, no 7 (1997). Accessed July 28, 2020. https://stevenpinker.com/files/pinker/files/computer.pdf

Putnam, H. (1992). *Representation and Reality*. Cambridge: MIT Press.

Regan, T. (1983). *The case for animal rights*. Berkeley & Los Angeles: University of California Press.

Rey, G. (1983). "A Reason for Doubting the Existence of Consciousness". In *Consciousness and Self-Regulation*, vol. 3, edited by Richard J. Davidson, Gary E. Schwartz, and David Shapiro, pp. 1-39. New York: Plenum.

_____ (1988). "A Question About Consciousness". In *Perspectives on Mind*, edited by Herbert R. Otto, and James A. Tuedio, pp. 5-24. Dordrecht: D. Reidel Publishing.

Rucker, R. (1982). *Infinity and the Mind: The Science and Philosophy of the Infinite*. Princeton, N.J.: Princeton University Press.

Russell, S., Tegmark, M., et al. (2020). *Autonomous Weapons: An Open Letter From AI & Robotics Researchers*. Future of Life Institute. Accessed July 25, 2020. https://futureoflife.org/open-letter-autonomous-weapons/

Savova, V., Peshkin, L. (2007). "Is the Turing Test Good Enough? The Fallacy

of Resource-Unbounded Intelligence". In International Joint Conferences on Artificial Intelligence Organization: *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence*, IJCAI-07: pp. 545-550. Accessed August 29, 2020. https://www.ijcai.org/Proceedings/07/Papers/086.pdf

Searle, J. (1980). "Minds, Brains, and Programs". *Behavioral and Brain Sciences*, 3(3), pp. 414-457.

_____ (1984). *Minds, Brains, and Science*. Cambridge Massachusetts: Harvard University Press.

_____ (1997). *The Mystery of Consciousness*. New York: The New York Review of Books.

Shoemaker, D. (2003). "Caring, Identification, and Agency". *Ethics*, 114(1), pp. 88-118.

Shoemaker, S. (1984). *Identity, Cause, and Mind*. Cambridge: Cambridge University Press.

Singer, P. (1975). *Animal liberation: a new ethics for our treatment of animals*. New York: New York Review of Books.

_____ (1993). *Practical Ethics* (2nd edition). Cambridge: Cambridge University Press.

_____ (2009). *Wired for War*. New York: Penguin Press.

Sloman, A. (1996). *A systematic approach to consciousness (How to avoid talking nonsense?)*. Accessed July 28, 2020. http://www.cs.bham.ac.uk/research/projects/cogaff/misc/consciousness.rsa.text

Solum, L. (1992). "Legal personhood for artificial intelligences". *North Carolina Law Review*, 70(4), pp. 1231-1287.

Sparrow, R. (2007). "Killer Robots". *Journal of Applied Philosophy*, 24(1), pp. 62-77.

Tegmark, M. (2017). *Life 3.0: being human in the age of artificial intelligence*. New York: Knopf.

Torrance, S. (2004). "Could we, should we, create conscious robots?". *Journal of Health Social and Environmental Issues*, 4 (2), pp. 43-46.

Turing, A. (1937). "On Computable Numbers With an Application to the Entscheidungsproblem". *Proceedings of the London Mathematical Society, 42* (2), pp. 230-265.

_____ (1938). "On Computable Numbers with an Application to the Entscheidungsproblem". *Proceedings of the London Mathematical Society*, 43(2), pp. 544-546.

_____ (1950). "Computing, Machinery and Intelligence". *Mind, LIX*, pp. 433-660.

Vincent, J. (2019). "Former Go champion beaten by DeepMind retires after declaring AI invincible". In *The Verge*. Accessed August 1, 2020. https://www.theverge.com/2019/11/27/20985260/ai-go-alphago-lee-se-dol-retired-deepmind-defeat

Wallach, W., Allen, C. (2009). *Moral machines: teaching robots right from wrong*. Oxford: Oxford University Press.

Walsh, T. (2017). *It's alive: Artificial Intelligence from the Logic Piano to Killer Robots*. Hamburg: Edition Körber.

Watson, G. (1975). "Free Agency". *Journal of Philosophy*, 72(8), pp. 205-220.

Weller, C. (2020). "Meet the first-ever robot citizen — a humanoid named Sophia that once said it would 'destroy humans'". In *Business Insider*. Accessed July 30, 2020. https://www.businessinsider.com/meet-the-first-robot-citizen-sophia-animatronic-humanoid-2017-10?r=UK

Welsh, S. (2017). "Clarifying the Language of Lethal Autonomy in Military Robots". In Aldinhas Ferreira, M., Silva Sequeira, J., Tokhi, M. E., Kadar, E., Virk, G. (eds) *A World with Robots. Intelligent Systems, Control and Automation: Science and Engineering*, vol 84. Springer, Cham.

Wilkes, K. (1988). "Yishi, Duh, Um and Consciousness". In *Consciousness in Contemporary Science*, edited by Antony Marcel and Edoardo Bisiach, pp. 16-41. Oxford: Oxford University Press.