Ηθική. Περιοδικό φιλοσοφίας

Αρ. 19 (2024)

Γράφοντας τον αλγόριθμο του Καλού: Η Τεχνητή Νοημοσύνη ως μηχανή απόδοσης Δικαιοσύνης

Άλκης Γούναρης, Γιώργος Κωστελέτος

doi: 10.12681/ethiki.39654

Βιβλιογραφική αναφορά:

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

Γράφοντας τον αλγόριθμο του Καλού: Η Τεχνητή Νοημοσύνη ως μηχανή απόδοσης Δικαιοσύνης

ΑΛΚΗΣ ΓΟΥΝΑΡΗΣ – ΓΙΩΡΓΟΣ ΚΩΣΤΕΛΕΤΟΣ

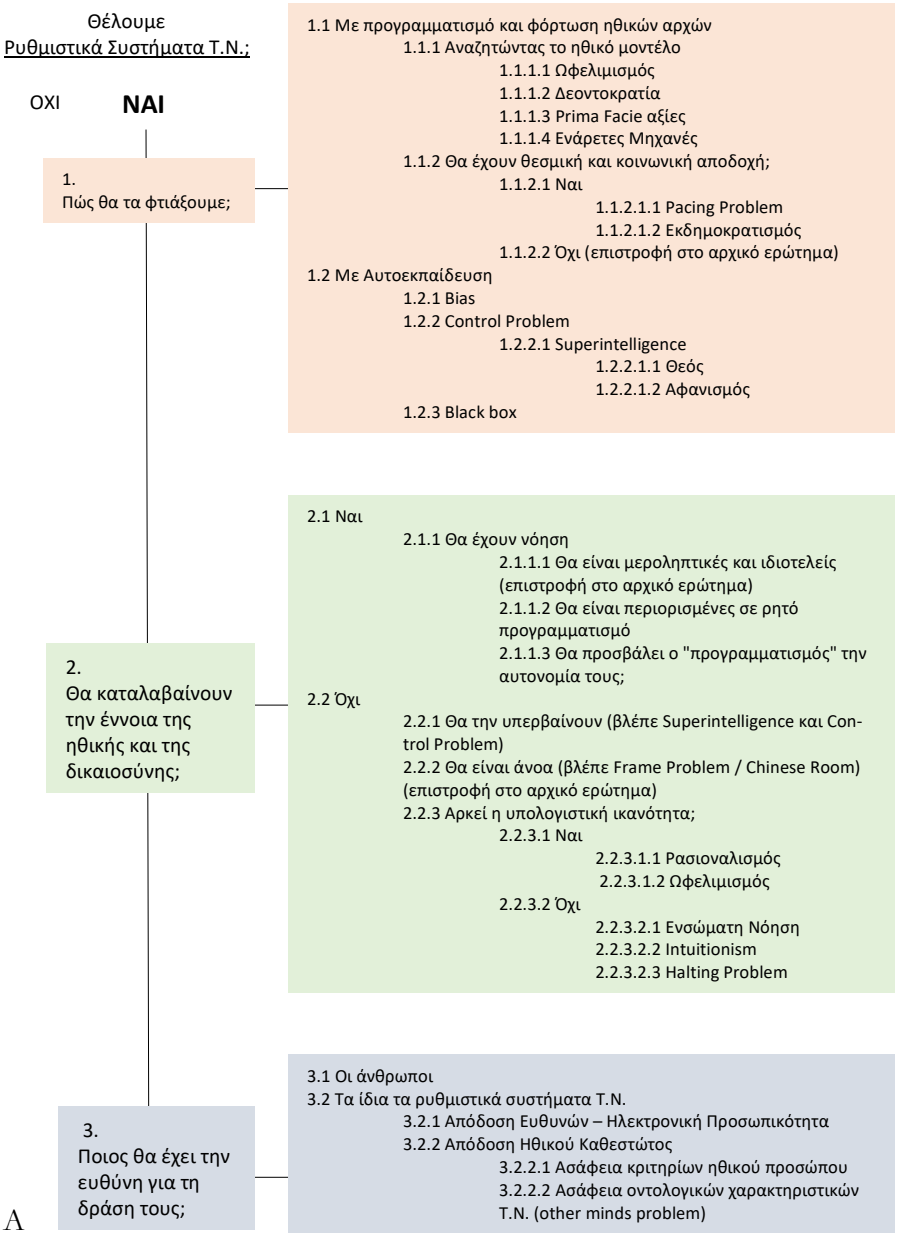
https://orcid.org/0000-0002-0494-6413 – https://orcid.org/0000-0001-6797-8415

Περίληψη: Ένα από τα πιο ενδιαφέροντα και προκλητικά ζητήματα της Ηθικής της Τ.Ν. που θα κληθούμε να αντιμετωπίσουμε στο άμεσο μέλλον έχει να κάνει με την κατασκευή και χρήση υπολογιστικών συστημάτων για ρυθμιστικούς σκοπούς και, συγκεκριμένα, συστημάτων Τ.Ν. που θα προβαίνουν στη λήψη αποφάσεων ηθικής ή νομικής φύσης, συμμετέχοντας καθοριστικά στη λειτουργία απόδοσης Δικαιοσύνης. Η λήψη αποφάσεων per se, δεν συνιστά εν πρώτοις οξύ πρόβλημα, καθώς ήδη εφαρμογές Τ.Ν., που χρησιμοποιούμε στην καθημερινότητά μας, λαμβάνουν αποφάσεις οι οποίες ενδεχομένως να έχουν εμμέσως κάποιον ηθικό αντίκτυπο, όμως τα κριτήρια λήψης των αποφάσεών τους δεν είναι ηθικά ή νομικά και ο ρόλος τους παραμένει κατά βάση εισηγητικός. Σε αντίθεση με τέτοιου τύπου «Εισηγητικά Συστήματα Τ.Ν.», στο μέλλον είναι πολύ πιθανό να δημιουργηθούν «Ρυθμιστικά Συστήματα Τ.Ν», δηλαδή συστήματα που δεν θα εισηγούνται απλώς, αλλά θα λαμβάνουν τα ίδια αποφάσεις, με βάση ηθικά κριτήρια ή κριτήρια δικαίου εν γένει, και θα μπορούν να χρησιμοποιηθούν στο άμεσο μέλλον για νομοθετικό έργο, απόδοση δικαιοσύνης, καθορισμό κυβερνητικών ή εταιρικών πολιτικών ή ακόμα και για αξιολόγηση της καθημερινής μας συμπεριφοράς.

Πρέπει να διευκρινιστεί, δε, ότι, σε αντιδιαστολή προς τα Ρυθμιστικά Συστήματα Τ.Ν., τα Εισηγητικά Συστήματα Τ.Ν, όπως οι νομικές εφαρμογές που χρησιμοποιούνται επί του παρόντος, προβαίνουν σε εισηγήσεις τις οποίες ο αλληλεπιδρών με τα συστήματα αυτά άνθρωπος δεν είναι υποχρεωμένος να ακολουθήσει, παρ' ότι σε ορισμένες περιπτώσεις αισθάνεται «ότι πρέπει» να ακολουθήσει λόγω της επάρκειας και του βαθμού πληρότητας των συστημάτων αυτών. Στην περίπτωση, ωστόσο, των Ρυθμιστικών Συστημάτων Τ.Ν., οι «έξοδοί» (output) τους δύναται να είναι αποφάσεις με τις οποίες οι άνθρωποι θα είναι υποχρεωμένοι να συμμορφώνονται απαρέγκλιτα, ακόμα και επί τη βάσει ενός θεσμικού πλαισίου που θα επιβάλλει στους ανθρώπους την ευθυγράμμιση με τις αποφάσεις των Ρυθμιστικών Συστημάτων Τ.Ν. Όπως δείχνουμε στο παρόν δοκίμιο, η επιλογή, ως προς το αν οι «έξοδοι» αυτές θα έχουν τον χαρακτήρα απλών εισηγήσεων ή αναπόδραστων αποφάσεων, ισοδυναμεί ουσιαστικά με μια επιλογή ως προς τον βαθμό αυτονομίας που θα διακρίνει τα εν λόγω συστήματα. Αν οι «έξοδοι» συνιστούν απλές εισηγήσεις, τότε έχουμε μια περιορισμένη αυτονομία από μέρους του Ρυθμιστικού Συστήματος Τ.Ν., καθώς καθότι θα εξακολουθεί η όποια εισήγηση, ως τέτοια, να επιζητά την τελική αξιολόγηση και τελικά έγκριση ή απόρριψη από τον άνθρωπο χρήστη —π.χ. από έναν νομοθέτη ή δικαστή. Απεναντίας, αν οι «έξοδοι» αποτελούν αποφάσεις προς αναπόδραστη εκτέλεση, τότε ο ρόλος ανθρώπινου παράγοντα, νομοθέτη ή δικαστή, καθίσταται κενός περιεχομένου και ως εκ τούτου έχουμε απόλυτη αυτονομία του συστήματος Τ.Ν.

Λέξεις-κλειδιά: Ηθική της Τεχνητής Νοημοσύνης; Εισηγητικά Συστήματα Τ.Ν.; Ρυθμιστικά Συστήματα Τ.Ν.; Δικαστικές Εφαρμογές Τεχνητής Νοημοσύνης, Machine Ethics;

Η φιλοσοφική συζήτηση γύρω από αυτά τα Εισηγητικά και Ρυθμιστικά Συστήματα Τ.Ν. είναι, όπως θα δούμε, αρκετά σύνθετη και περιλαμβάνει ηθικά, επιστημολογικά και οντολογικά ερωτήματα. Στόχος του παρόντος κειμένου δεν είναι μια εξαντλητική ανάλυση των εν λόγω ζητημάτων αλλά, μάλλον, η σχηματική κατάδειξη του τρόπου ανάδυσης και διαπλοκής τους. Επιχειρούμε, δηλαδή, μια χαρτογράφηση των συναφών με τη ρυθμιστική χρήση της Τ.Ν. φιλοσοφικών προβληματισμών, οι οποίοι αναδεικνύουν εν κατακλείδι τις —σε πολλές περιπτώσεις αξεπέραστες— δυσκολίες ή και τις παρακινδυνευμένες παραδοχές που θα έπρεπε να λάβει κανείς υπόψιν του, αν επιχειρούσε να κατασκευάσει και να θέσει σε λειτουργία Ρυθμιστικά Συστήματα απόδοσης Δικαιοσύνης. Στην προσπάθειά μας να συστηματοποιήσουμε και να περιγράψουμε με απλό και κατανοητό τρόπο αυτήν την πολυπαραγοντική συζήτηση, σχεδιάσαμε ένα διάγραμμα (ΣΧΗΜΑ Α) —το οποίο ονομάσαμε συμβολικά «αλγόριθμο του Καλού»—, που με αλγοριθμικό τρόπο θέτει τα βασικά βήματα της συγκεκριμένης διερεύνησης (βλ. το διάγραμμα σε πλήρη ανάπτυξη στον σύνδεσμο https://alkisgounaris.gr/gr/ai/algorithmofgood/).



ΣΧΗΜΑ Α

Ο αλγόριθμός μας ξεκινάει από το βασικό ερώτημα αν θέλουμε ή όχι την κατασκευή τέτοιου είδους συστημάτων Τ.Ν.

«Θέλουμε Ρυθμιστικά Συστήματα Τ.Ν.;». «Ναι» ή «Όχι»;

Ένα πειστικό «Όχι» μπορεί να συνοψιστεί στις θέσεις ερευνητών, όπως ο Yudkowsky[1], που θεωρούν ότι ο στόχος της έρευνας στην Τ.Ν. θα πρέπει να είναι η κατασκευή «φιλικών» και ασφαλών μηχανών —εργαλείων πλήρως ευθυγραμμισμένων με τις ανθρώπινες ανάγκες και στόχους. Δηλαδή δεν χρειαζόμαστε Ρυθμιστικά Συστήματα Τ.Ν. (ήτοι συστήματα Τ.Ν. που λαμβάνουν αυτόνομα αποφάσεις) ή υπερνοήμονες μηχανές, αλλά χρηστικές μηχανές που θα κάνουν τη ζωή μας καλύτερη προτείνοντας ή υπολογίζοντας τα κάθε φορά ζητούμενα, χωρίς, εντούτοις, να προβαίνουν από μόνες τους σε «αποφάσεις» και ηθικές πράξεις. Ουσιαστικά μιλάμε για την αντιμετώπιση της Τ.Ν. ως εργαλείου, ως ενός Εισηγητικού Συστήματος που ως τέτοιο είναι και θα πρέπει να παραμείνει υπό τον απόλυτο έλεγχο του ανθρώπου. Οι υποστηρικτές αυτής της προσέγγισης ορισμένως αισθάνονται ασφαλείς, θεωρώντας βέβαιο ότι η ανάπτυξη υπολογιστικών συστημάτων Τ.Ν. ως «απλών εργαλείων» εξασφαλίζει τη μη αυτονομία των τελευταίων και, κατά συνέπεια, την αυτονομία του ανθρώπου. Εκ των πραγμάτων κάνουμε λόγο για καταστάσεις «human-in-the-loop» ή —στην περίπτωση μερικής απεμπλοκής του ανθρώπινου παράγοντα από τη διαδικασία ελέγχου— για «human-on-the-loop». Στην κατεύθυνση αυτή, κατά της ανάπτυξης Ρυθμιστικών Συστημάτων Τ.Ν., θα μπορούσαμε να αναφέρουμε τη γενικότερη τάση προληπτικής (ή προφυλακτικής) προσέγγισης της Τεχνολογίας, όπως αυτή εκφράζεται από την UNESCO[2], βάσει της οποίας πριν χρησιμοποιηθεί μια τεχνολογική εφαρμογή που ενδέχεται να έχει καταστροφικές συνέπειες για την ανθρωπότητα, αυτοί που ενδιαφέρονται να προχωρήσουν στην υλοποίηση της εν λόγω εφαρμογής πρέπει να αποδείξουν με έγκυρο τρόπο ότι αυτή θα είναι ασφαλής ή να καταδείξουν τις προϋποθέσεις ασφαλούς χρήσης της.

Δεδομένου, πάντως, ότι μια Τ.Ν. που θα υπόκειται στον απόλυτο και διαρκή έλεγχο των ανθρώπων χρηστών δε θα διαφέρει πολύ από τα απλά υπολογιστικά συστήματα, οι εισηγητές της θέσης «η Τ.Ν. ως εργαλείο» ενδεχομένως να δεχτούν ως βέλτιστη τη λύση «human-on-the-loop», έχοντας ως δέλεαρ την ανάπτυξη αποτελεσματικότερων εργαλείων και την προς εξοικονόμηση χρόνου και ενέργειας μερική απεμπλοκή του ανθρώπου από τον όποιο φόρτο εργασίας.

Παρά ταύτα, υφίσταται ο κίνδυνος τέτοια συμβουλευτικά συστήματα να αντιμετωπιστούν τελικά ως αυτόνομα ρυθμιστικά, καθώς οι άνθρωποι, γοητευμένοι από την ακρίβεια και την αντικειμενικότητα της Τ.Ν., ίσως διστάζουν να αμφισβητήσουν τις εισηγήσεις της. Έτσι, η αρχική πρόθεση να παραμείνει ο έλεγχος στον άνθρωπο κινδυνεύει να οδηγήσει στην αντίθετη κατεύθυνση, όπου ο ανθρώπινος κριτής γίνεται τυπικός εντολοδόχος των προτάσεων της Τ.Ν. Αντίστοιχα, ακραίες θέσεις, όπως οι νεολουδίτες, απορρίπτουν πλήρως την ανάπτυξη Τ.Ν., θεωρώντας την απειλή για την ανθρώπινη αυτονομία. Αυτή η απόρριψη, όμως, μπορεί να περιορίσει τον διάλογο και να αποτρέψει τη διερεύνηση λύσεων που θα εξασφάλιζαν τον ανθρώπινο έλεγχο, δημιουργώντας έναν χώρο για μετριοπαθείς προσεγγίσεις στην ανάπτυξη της Τ.Ν. με διασφαλίσεις κατά της πλήρους αυτονομίας.

Αντιθέτως, η απάντηση «Ναι», στο ερώτημα περί Ρυθμιστικών Συστημάτων Τ.Ν., μπορεί να εκφράζει ένα ευρύ φάσμα «αισιόδοξων», φιλικών, θετικιστικών και φιλελεύθερων[3] θέσεων —οι οποίες προκρίνουν τα πλεονεκτήματα και οφέλη της τεχνολογικής προόδου, υποβαθμίζοντας τους ενδεχόμενους κινδύνους από μια απρόβλεπτη εξέλιξη της Τ.Ν.

Για παράδειγμα, ο Michael και η Susan Anderson, πρωτοπόροι του προγράμματος Machine Ethics[4], υποστηρίζουν ότι πράγματι θα πρέπει να στοχεύουμε στην κατασκευή τέτοιων Ρυθμιστικών Συστημάτων Τ.Ν. ή άλλως ηθικών μηχανών. Το αναλογικό επιχείρημά τους εν προκειμένω συνοψίζεται στο γεγονός ότι υπάρχει ήδη πληθώρα συστημάτων Τ.Ν. που κάνουν τη δουλειά τους καλύτερα από τους ανθρώπους σε πολλές δραστηριότητες. Η επιτυχία των συστημάτων αυτών μας υποδεικνύει ότι θα πρέπει να επιχειρήσουμε να φτιάξουμε και ηθικές μηχανές οι οποίες θα είναι τελικώς περισσότερο δίκαιες, πιο αποτελεσματικές, λιγότερο μεροληπτικές, αδέκαστες κ.τ.λ. —θα αποδίδουν δηλαδή δικαιοσύνη ως κυριολεκτικά «ηθικές»[5], τουτέστιν χωρίς τα ανθρώπινα ελαττώματα.[6]

Ωστόσο, ένα πειστικό «Ναι» ανοίγει μπροστά μας μια δέσμη τριών δύσκολων ερωτημάτων (βλ. σχήμα Α), που, όπως φαίνεται από αυτήν την αλγοριθμική παράσταση, κάθε ένα από αυτά αποτελεί αφετηρία ενός διακριτού και περίπλοκου κλάδου ανάλυσης του ζητήματος ανάπτυξης και χρήσης Ρυθμιστικών Συστημάτων Τ.Ν.

Τα τρία αυτά ερωτήματα επιγράφονται ως εξής:

1. Πώς θα φτιάξουμε τέτοια ρυθμιστικά συστήματα;
2. Θα κατανοούν αυτά τα συστήματα τις έννοιες της ηθικής και της δικαιοσύνης;
3. Ποιος θα έχει την ευθύνη για τη δράση και τις αποφάσεις τους;

1. Πώς θα φτιάξουμε τέτοια ρυθμιστικά συστήματα;

Για να απαντήσουμε στο πρώτο ερώτημα σχετικά με τη δημιουργία ενός ηθικού και ρυθμιστικού συστήματος Τ.Ν., χρειάζεται να εξετάσουμε τεχνικά, εννοιολογικά και ηθικά ζητήματα, τα οποία θα καθορίσουν το οντολογικό του καθεστώς. Υπάρχουν δύο βασικές προσεγγίσεις, βάσει του ρόλου του ανθρώπου στη μάθηση των αποφάσεων του συστήματος.

Η πρώτη προσέγγιση [1.1], η «επιβλεπόμενη μάθηση»[7], περιλαμβάνει τον ρητό προγραμματισμό κανόνων από τον προγραμματιστή, καθορίζοντας τις επιθυμητές εξόδους και επιτρέποντας την προβλεψιμότητα της απόφασης.

Η δεύτερη προσέγγιση [1.2] περιλαμβάνει μεθόδους όπως η «μη επιβλεπόμενη» και η «ενισχυτική μάθηση», όπου το σύστημα ανιχνεύει μοτίβα ή μαθαίνει μέσω ανταμοιβής, χωρίς ανθρώπινη καθοδήγηση στις εξόδους. Αυτή η προσέγγιση μπορεί να οδηγήσει σε αποφάσεις πέρα από τις δυνατότητες ανθρώπινης πρόβλεψης, και, δυνητικά, σε ισχυρά αυτοεκπαιδευόμενα συστήματα Τ.Ν., τα οποία, αν και είναι ακόμα υπό έρευνα, ενδέχεται να ανοίξουν νέους δρόμους στην Τ.Ν.

1.1 Προγραμματισμός και «φόρτωση» ηθικών αρχών και κανόνων δικαίου

Αν επιλέξουμε να «φορτώσουμε» κάποιο είδος αρχών σε ένα σύστημα μέσω επιβλεπόμενης μάθησης, θα πρέπει να απαντήσουμε σε δυο επιμέρους ερωτήματα: α) με βάση ποιο ηθικό μοντέλο θα «εκπαιδεύσουμε» το σύστημά Τ.Ν. να λαμβάνει αποφάσεις; (1.1.1) και β) θα έχει το συγκριμένο σύστημα ευρεία θεσμική και κοινωνική αποδοχή (1.1.2);

1.1.1 Αναζητώντας το ηθικό μοντέλο: Είναι αυτονόητο ότι, αν θέλουμε ένα σύστημα Τ.Ν. να προβαίνει σε «αξιολογικές» ηθικές αποφάσεις, θα πρέπει να προηγείται η εκπαίδευσή του σύμφωνα με ένα ορισμένο πλαίσιο κανόνων. Προκειμένου να αποφύγουμε, δε, τα λογικώς εσφαλμένα συμπεράσματα που θα μπορούσαν να προκύπτουν από την εξαγωγή αποφάσεων του «Πρέπει» από προκείμενες περιγραφικές προτάσεις του «Είναι», το πλαίσιο κανόνων στο οποίο θα βασίσουμε την εκπαίδευση του συστήματός μας θα περιλαμβάνει «μείζονες προκείμενες» προτάσεις αξιολογικού χαρακτήρα. Τέτοιες προτάσεις θα αντλούν τις αρχές τους από τα καθιερωμένα πλέον ηθικά μοντέλα που περιγράφουν «το πώς πρέπει να πράττουμε» —συγκεκριμένα είτε από τον Ωφελιμισμό (1.1.1.1) είτε από τη Δεοντοκρατία (1.1.1.2). Νεότερες προσεγγίσεις στο ζήτημα της ηθικής εκπαίδευ-

σης των μηχανών προτείνουν την προσφυγή σε ηθικά μοντέλα αναγνώρισης Prima Facie αξιών και καθηκόντων (1.1.1.3) ή αναγνώρισης «ενάρετων χαρακτηριστικών» (1.1.1.4), χωρίς, ωστόσο, όπως θα δούμε παρακάτω, να ξεπερνούν τα εννοιολογικά και πρακτικά προβλήματα που διέπουν όλο το παρόν δοκίμιο. Ο Yudkowsky[8] ορισμένως παρατηρεί ότι η επιλογή του πλαισίου κανόνων στους οποίους θα βασίσουμε την εκπαίδευση ενός συστήματος αποτελεί ένα πρόβλημα που θα πρέπει να μας απασχολήσει εκτενέστερα, καθόσον ναι μεν θα λαμβάνουμε εξόδους υπό αυτό που προγραμματίσαμε, είναι όμως αμφίβολο αν αυτές οι έξοδοι θα καλύπτουν τις αρχικές μας προθέσεις και τους αρχικούς μας σκοπούς. Και αυτό γιατί είναι αδύνατον να φανταστούμε εκ των προτέρων όλα τα πιθανά σενάρια και όλες τις πιθανές «εξόδους» στις οποίες θα οδηγηθεί ένα σύστημα ακολουθώντας το πλαίσιο κανόνων σύμφωνα με το οποίο το εκπαιδεύσαμε. Δημιουργώντας, δηλαδή, έναν επαγωγικό αλγόριθμο επιβλεπόμενης μηχανικής μάθησης αξιών, δεν θα μπορούμε να είμαστε ποτέ σίγουροι ότι μακροχρόνια, ή στην υποθετική περίπτωση δημιουργίας υπερνοήμονων μηχανών, η υπολογιστική ηθική του ωφελιμισμού ή η άκαμπτη ηθική της δεοντοκρατίας δεν θα οδηγήσουν σε «ολέθριες ή αποκρουστικές» —όπως συχνά συμβαίνει με τον ωφελιμισμό[9]— για την ανθρωπότητα αποφάσεις.

1.1.1.1. Με βάση την ωφέλεια και τον υπολογισμό των συνεπειών: Ο ωφελιμισμός, ως ηθικό μοντέλο, υπαγορεύει ότι οι πράξεις πρέπει να στοχεύουν στη μέγιστη δυνατή ωφέλεια για τον μέγιστο αριθμό ανθρώπων, ευθυγραμμιζόμενος με την υπολογιστική φύση των συστημάτων Τ.Ν. Ο Jeremy Bentham[10] εισηγήθηκε έναν υπολογισμό της ωφέλειας βάσει εννέα κριτηρίων, προτείνοντας ότι τα οφέλη μπορούν να υπολογιστούν με ακρίβεια, διευκολύνοντας έτσι την ηθική λήψη αποφάσεων (δείτε και ενότητα 2.2.3.1.2). Τουναντίον, η «αριθμητικοποίηση» της ηθικής δεν καλύπτει όλες τις περιπτώσεις, καθώς ο μαθητής του Bentham, John Stuart Mill, προσέθεσε την ανάγκη ποιοτικής διάκρισης των ηδονών και ορισμένων αξιών που υπερβαίνουν τον απλό υπολογισμό. Ο Mill[11] υποστήριζε ότι το ανθρώπινο είδος πρέπει να επιλέγει την ανθρώπινη ευτυχία, ακόμη και αν αυτό συνεπάγεται λιγότερη ευχαρίστηση, θέτοντας όρια στην υπολογισιμότητα των ηθικών αποφάσεων. Ο ωφελιμισμός, παρά τη χρησιμότητά του, παρουσιάζει αδυναμίες στον ακριβή υπολογισμό των μελλοντικών συνεπειών μιας πράξης, δεδομένου ότι αυτές οι συνέπειες είναι απρόβλεπτες και μη απολύτως βεβαιωμένες, περιορίζοντας την αντικειμενικότητα και ακρίβεια στην εφαρμογή της Τ.Ν.

1.1.1.2. Με βάση a priori αρχές: Η δεοντοκρατική προσέγγιση, σε αντίθεση με τον ωφελιμισμό, εστιάζει στην πρόθεση και όχι στις συνέπειες μιας πράξης. Κατά τον Kant[12], οι ηθικές αποφάσεις βασίζονται σε μια «καθαρή» βούληση, απαλλαγμένη από σκοπιμότητες, που ακολουθεί την «κατηγορική προσταγή» —έναν καθολικό ηθικό νόμο. Για παράδειγμα, ένας τίμιος έμπορος δρα με γνώμονα το καθήκον του, όχι λόγω φόβου συνεπειών. Η εφαρμογή ενός καντιανού μοντέλου σε ρυθμιστικό σύστημα Τ.Ν. παρουσιάζει δυσκολίες, διότι η έννοια της «ελεύθερης βούλησης» δεν είναι εύκολα αφομοιώσιμη από υπολογιστικά συστήματα, τα οποία στηρίζονται σε τυπικό προγραμματισμό. Έτσι, η Τ.Ν. λειτουργώντας με έμμεσους κανόνες, βασισμένους σε περιορισμένες σκοπιμότητες, δεν επιτυγχάνει κυριολεκτικά το καντιανό ιδεώδες. Ένα τέτοιο σύστημα, λόγω της «άκαμπτης» φύσης του, κινδυνεύει να επιδεικνύει άκαμπτη συμπεριφορά, αδυνατώντας να προσαρμοστεί σε εξαιρέσεις και περίπλοκες ηθικές καταστάσεις, με πιθανές απάνθρωπες εκβάσεις σε συγκεκριμένες περιστάσεις.

1.1.1.3. Με βάση ηθικά data και prima facie αξίες: Νεότεροι φιλόσοφοι, όπως ο W. D. Ross[13], αναζήτησαν αντικειμενικά κριτήρια ηθικής, ώστε να υπερβούν τις αδυναμίες του ωφελιμισμού και της καντιανής δεοντοκρατίας. Ο Ross πρότεινε ένα ηθικό μοντέλο βασισμένο σε prima facie αξίες —προφανείς ηθικές αρχές, όπως η αξιοπιστία, η αγαθοεργία και η αρχή της μη βλάβης— που θεωρούνται ηθικά δεδομένα (data), ανάλογα

με τις αισθητηριακές πληροφορίες που αντιλαμβάνεται ένας οργανισμός. Αυτές οι αρχές λειτουργούν ως «καθήκοντα» που καθοδηγούν τις ηθικές επιλογές, ενώ λαμβάνουν υπόψη και τις συνέπειες. Θεωρητικά, τέτοιες αξίες θα μπορούσαν να «φορτωθούν» σε ένα σύστημα Τ.Ν. για να καθοδηγούν τη λήψη αποφάσεων. Ωστόσο, αυτό δεν εξαλείφει πιθανά διλήμματα, όπως όταν η αρχή της μη βλάβης έρχεται σε σύγκρουση με την προστασία ατόμων που απειλούνται. Σε τέτοιες περιπτώσεις, το σύστημα ίσως αναγκαστεί να επιλέξει μεταξύ ωφελιμιστικών και δεοντοκρατικών κριτηρίων. Παρά τους περιορισμούς, οι prima facie αξίες φαίνονται χρήσιμες για την εκπαίδευση επιβλεπόμενων και ενισχυτικών συστημάτων Τ.Ν., προσφέροντας ένα είδος ηθικής καθοδήγησης στις αποφάσεις τους.

1.1.1.4. Με βάση τον «ενάρετο χαρακτήρα» της ίδιας της μηχανής: Μια αριστοτελική προσέγγιση για την ηθική συμπεριφορά των ρυθμιστικών συστημάτων Τ.Ν. εστιάζει στον ενάρετο χαρακτήρα τους, όχι απλώς στη «συνταγή» λήψης αποφάσεων. Αντί να αναζητούμε τον τρόπο λήψης ηθικών αποφάσεων, εξετάζουμε αν ένα κριτήριο Χ μπορεί να καθιστά το σύστημα ηθικό[14]. Ο Αριστοτέλης θεώρησε ότι η ηθική δεν έγκειται μόνο στις πράξεις, αλλά στον καλλιεργημένο, ενάρετο χαρακτήρα. Έτσι, ένα τέτοιο σύστημα θα πρέπει να υπηρετεί έναν κοινωνικό και ατομικό σκοπό, που να καλλιεργεί αρετές όπως φρόνηση, δικαιοσύνη, αξιοπιστία και επιείκεια[15], προβάλλοντας συμπεριφορές αναγνωρίσιμες ως ηθικές. Παρά τις τεχνικές προκλήσεις, η υπόθεση του «ενάρετου ρυθμιστικού συστήματος» το εντάσσει στον κοινωνικό ιστό, αξιολογούμενο τόσο από την επίτευξη του ατομικού του σκοπού όσο και κυρίως από την κοινωνική αποδοχή των αρχών και των αρετών του. Η μηχανή αυτή θα κρίνεται, εν τέλει, από την ευρεία αποδοχή της ως ενάρετα συμπεριφερόμενης, αποτυπώνοντας την επιτυχία της στην κοινωνική σφαίρα.

1.1.2. Αναζητώντας την κοινωνική αποδοχή: Σε ένα από τα μεγαλύτερα πειράματα αναφορικά με τις ηθικές πεποιθήσεις διαφορετικών πολιτισμών, το οποίο συγκέντρωσε περισσότερες από 40 εκατομμύρια απαντήσεις από 233 χώρες, τέθηκε το ερώτημα στο ευρύ κοινό για τα επιθυμητά ηθικά κριτήρια λήψης αποφάσεων ενός συστήματος Τ.Ν., όπως ένα αυτοοδηγούμενο όχημα.[16] Το αποτέλεσμα αυτού του κοινωνικού πειράματος, που ονομάστηκε «Ηθική Μηχανή»[17], αποκάλυψε το μεγάλο εύρος διαπολιτισμικής ηθικής ποικιλομορφίας σε συνάρτηση με τις κοινωνικές, οικονομικές και γεωγραφικές παραμέτρους. Το ερώτημα, εν προκειμένω, για το τι είδους αρχές θα έπρεπε να φορτώσουμε σε ένα σύστημα Τ.Ν. είναι άμεσα συνδεδεμένο με τα αποτελέσματα του εν λόγω πειράματος, τα οποία αποδεικνύουν ότι δεν θα ήταν καθολικά αποδεκτή η «φόρτωση» ενός ορισμένου μοντέλου λήψης ηθικών αποφάσεων. Τίθεται, λοιπόν, το ερώτημα για το *αν ο προγραμματισμός ενός ρυθμιστικού συστήματος ΤΝ θα πρέπει να έχει ευρεία θεσμική και κοινωνική αποδοχή και αν αυτή η αποδοχή θα πρέπει να είναι σε επίπεδο πολιτισμού, χώρας ή τοπικής κοινωνίας*[18].

1.1.2.1. Αν θεωρούμε πως «ναι, πρέπει να έχει ευρεία θεσμική και κοινωνική αποδοχή» η φόρτωση αρχών σε ένα ρυθμιστικό σύστημα Τ.Ν., θα βρεθούμε αντιμέτωποι με δυο δυσεπίλυτα προβλήματα. Το ένα είναι το πρόβλημα απόκλισης ανάμεσα στον ρυθμό εξέλιξης της τεχνολογίας και στον ρυθμό αντίδρασης των θεσμών, όπως το περιέγραψε ο Collingridge, ένα πρόβλημα που διέπει όλη την έρευνα στο πεδίο της Τ.Ν. Το άλλο είναι το πρόβλημα της κοινωνικής συναίνεσης που οδηγεί τη συζήτηση στον λεγόμενο εκδημοκρατισμό[19] της Τ.Ν., ο οποίος απασχολεί ολοένα και περισσότερο τον δημόσιο λόγο.

1.1.2.1.1. Στις αρχές τις δεκαετίας του ’80, ο David Collingridge[20] κατέδειξε την απόκλιση ανάμεσα στη γνώση του αντίκτυπου κάθε νέας τεχνολογικής εξέλιξης και στη λήψη κατάλληλων μέτρων για τη θεσμική κάλυψη των οικονομικών, πολιτικών και κοινωνικών συνεπειών της. Ο Collingridge διαπίστωσε ότι, στις πρώτες φάσεις της τεχνολογικής ανάπτυξης, κάθε νέα τεχνολογία είναι ακόμα εύπλαστη, δηλαδή μπορεί να προσαρμοστεί ή και να αλλάξει, και ότι το κόστος ελέγχου ή και απαγόρευσης της συγκεκριμένης τεχνο-

λογίας είναι μικρό, όμως οι πραγματικές επιπτώσεις, κοινωνικές, οικονομικές ή και περιβαλλοντικές, από την ανάπτυξη και χρήση της συγκεκριμένης τεχνολογίας, είναι δύσκολο ή και αδύνατον να προβλεφθούν. Στα μεταγενέστερα στάδια της ίδιας τεχνολογικής εξέλιξης ωστόσο, και καθώς οι επιπτώσεις από τη χρήση της είναι πλέον γνωστές, η τεχνολογία είναι σε τόσο μεγάλο βαθμό εδραιωμένη στην καθημερινή ζωή και στην οικονομία, ώστε είτε είναι δύσκολο ή αδύνατον να αλλάξει, είτε, αν δύναται να αλλάξει, μπορεί να αλλάξει μόνο με υψηλό οικονομικό, κοινωνικό και συχνά πολιτικό κόστος.

Αυτή η χρονική υστέρηση μεταξύ του ρυθμού της τεχνολογικής εξέλιξης και του ρυθμού λήψης ή προσαρμογής θεσμικών μέτρων, γνωστή ως **το «πρόβλημα του βηματισμού» (pacing problem)**[21]**,** έχει παρουσιαστεί ιστορικά σε διάφορες τεχνολογικές «στροφές», όπως στη χρήση λιγνίτη, στη χρήση πλαστικού, ή στον βενζινοκινητήρα. Μολαταύτα, σήμερα ο ρυθμός εξέλιξης της τεχνολογίας της Τ.Ν. και ο ρυθμός αντίδρασης των θεσμών κάνουν το πρόβλημα βηματισμού ακόμα πιο έντονο, αυξάνοντας τον κίνδυνο για σοβαρές ή και μη αναστρέψιμες επιπτώσεις. Αυτό οφείλεται κυρίως στην εκθετική αύξηση της υπολογιστικής ικανότητας των μηχανών (νόμος του Moore)[22] και στην ευρεία εξάπλωση των συστημάτων Τ.Ν. σε όλα σχεδόν τα πεδία του ανθρώπινου βίου, με τρόπο που οποιοδήποτε θεσμικό πλαίσιο αδυνατεί να προβλέψει πλήρως και με ακρίβεια τις μακροχρόνιες συνέπειες.

1.1.2.1.2. Η ίδια η διαδεδομένη χρήση των συστημάτων Τ.Ν., θα μπορούσε να προτείνει κάποιος, νομιμοποιεί την εξέλιξή της και αποτελεί από μόνη της ένα είδος ευρείας συναίνεσης και κοινωνικής αποδοχής. Πόσο ενημερωμένη, εντούτοις, είναι αυτή η «σιωπηρή» συναίνεση των χρηστών; **Οι εισηγητές του «εκδημοκρατισμού» της τεχνητής νοημοσύνης** υποστηρίζουν ότι η ευρεία χρήση, η δημοφιλία και η προσβασιμότητα σε ένα τεχνολογικό μέσο αποτελούν μια μόνο συνισταμένη στη διακρίβωση του «δημοκρατικού» χαρακτήρα του μέσου αυτού, αν λάβουμε ως παραδοχή ότι οι χρήστες ξέρουν τι κάνουν. Η άλλη συνισταμένη αφορά στους σκοπούς και στο έργο που επιτελεί ένα σύστημα Τ.Ν. και στο κατά πόσο αυτό το σύστημα περικλείει συγκεκριμένες κοινωνικές ομάδες, προάγει την κοινωνική δικαιοσύνη και παρέχει «διαφάνεια» ως προς τον τρόπο λειτουργίας του ή και ως προς τους σκοπούς που εξυπηρετεί. Οι εισηγητές της ανάγκης εκδημοκρατισμού[23] (democratization) της τεχνολογίας εν γένει, αλλά και ειδικότερα της Τ.Ν., θεωρούν εν προκειμένω ότι οι δημοκρατικές αξίες και ιδανικά θα πρέπει να είναι «ενσωματωμένα» στα συστήματα Τ.Ν. και να λαμβάνονται υπόψιν κατά τον σχεδιασμό τους. Τουναντίον, κάτι τέτοιο προϋποθέτει θεσμικά και ρυθμιστικά αντανακλαστικά[24] που, όπως έχει αποδειχθεί στο παρελθόν και σημειώνουμε παραπάνω, ακολουθούν καθυστερημένα τις τεχνολογικές εξελίξεις.

1.1.2.2. Αν απαντήσουμε αποφατικά στο ερώτημα που τίθεται στο 1.1.2, για το αν πρέπει να έχει ευρεία θεσμική και κοινωνική αποδοχή η φόρτωση ενός ορισμένου μοντέλου αρχών, αξιών ή αρετών σε ένα ρυθμιστικό σύστημα, τότε θα πρέπει να επιστρέψουμε πίσω στο αρχικό ερώτημα και να διερωτηθούμε ξανά αν εν τέλει θέλουμε Ρυθμιστικά συστήματα Τ.Ν.;

1.2 Αυτοεκπαίδευση με μηχανική μάθηση

Αναφερθήκαμε παραπάνω στο επιχείρημα της υπέρμετρης αποτελεσματικότητας των συστημάτων Τ.Ν. εν γένει ως κριτήριο κατασκευής Ρυθμιστικών Συστημάτων Τ.Ν., καθότι, όπως αναφέρουν οι Anderson[25], τέτοια συστήματα έχουν τη δυνατότητα να αποδίδουν δικαιοσύνη χωρίς την ανθρώπινη μεροληψία ή κακοδικία που εμφανίζεται συχνά σήμερα. Μια πιθανή προσέγγιση αυτοεκπαίδευσης των συστημάτων θα ήταν να «μάθουν» μόνα τους τις ηθικές αρχές που πρέπει να ακολουθούν. Ωστόσο, το «μόνα τους» είναι σχετικό, καθώς απαιτείται ανθρώπινη παρέμβαση στις αρχικές παραμέτρους και στη δια-

δικασία ενισχυτικής μάθησης, με κίνδυνο οι επιλογές αυτές να επηρεάσουν ανεπιθύμητα τα τελικά αποτελέσματα. Η μη επιβλεπόμενη μάθηση μπορεί να οδηγήσει τα συστήματα σε νέες στρατηγικές, πέρα από τις προγραμματισμένες, όπως δείχνουν οι καινοτόμες στρατηγικές σε παιχνίδια όπως το σκάκι και το GO. Παρά τις θετικές εκβάσεις σε συγκεκριμένες εφαρμογές, το ενδεχόμενο τα συστήματα Τ.Ν. να αναπτύξουν «απρόβλεπτες» στρατηγικές μπορεί να προκαλέσει ανησυχίες, ιδίως αν αυτές αποκλίνουν από κοινωνικά αποδεκτούς στόχους. Αυτό είναι γνωστό ως «πρόβλημα ευθυγράμμισης» (alignment problem)[26], το οποίο εμφανίζεται όταν οι αρχικές συνθήκες μάθησης ή το σύστημα ανταμοιβών δεν συμβαδίζουν με τον επιθυμητό στόχο.

Όπως έχει δείξει ο Bostrom[27] μέσω του orthogonality thesis, οι υπολογιστικές ικανότητες ενός συστήματος και η ικανότητα επίτευξης ορισμένων στόχων σύμφωνα με κάποιο σύστημα αξιών θα πρέπει να θεωρούνται δύο ανεξάρτητες μεταβλητές. Δηλαδή, υψηλές υπολογιστικές ικανότητες δεν εγγυώνται ότι το σύστημα θα υπηρετεί ηθικά ή κοινωνικά αποδεκτούς στόχους. Για την επίλυση του προβλήματος της ευθυγράμμισης των τελικών στόχων, οι ερευνητές προτείνουν στρατηγικές όπως η μάθηση μέσω μίμησης της ανθρώπινης συμπεριφοράς[28] και ο περιορισμός του ρυθμού εξέλιξης και αναπροσαρμογής των αξιών του συστήματος[29]. Παρ' όλα αυτά, η «μίμηση» της ανθρώπινης συμπεριφοράς δεν είναι από μόνη της ασφαλής, καθόσον η ανθρώπινη συμπεριφορά δεν είναι πάντα ιδανική για συστήματα Τ.Ν. Παραδείγματα όπως το chatbot Tay της Microsoft, που ανέπτυξε ακραίες προκαταλήψεις, αναδεικνύουν τον κίνδυνο της ανεξέλεγκτης μάθησης.

1.2.1. Το πρόβλημα της αλγοριθμικής μεροληψίας: Το 2016, έρευνα αξιολόγησης νομικών εργαλείων Τ.Ν. έδειξε τον υψηλό βαθμό αλγοριθμικής μεροληψίας (AI Bias) στην εκτίμηση κινδύνου επανάληψης τέλεσης αξιόποινων πράξεων εις βάρος μαύρων και μειονοτικών Αμερικανών. Συγκεκριμένα, οι αλγόριθμοι ήταν πιο αυστηροί στην εκτίμηση κινδύνου για μαύρους κρατούμενους, λόγω προκαταλήψεων που προέκυψαν από τα δεδομένα εκπαίδευσης. Οι προσπάθειες αντιμετώπισης της μεροληψίας περιλαμβάνουν την αφαίρεση ή στάθμιση προκατειλημμένων δεδομένων και την εφαρμογή αντιπαραδειγματικής δικαιοσύνης (counterfactual fairness)[30], που συγκρίνει τις αλγοριθμικές αποφάσεις με έναν «αντιγεγονικό» κόσμο, όπου δημογραφικά χαρακτηριστικά, όπως η φυλή, διαφοροποιούνται. Εντούτοις, με τη μη επιβλεπόμενη μάθηση και την περιορισμένη ανθρώπινη παρέμβαση, η δυνατότητα εμφάνισης προκαταλήψεων παραμένει, ιδιαίτερα αν το σύστημα μιμείται ευθέως την ανθρώπινη συμπεριφορά.

1.2.2. Η αδυναμία ελέγχου των συστημάτων Τ.Ν. συνδέεται όχι μόνο με τον τρόπο επεξεργασίας της πληροφορίας, αλλά και με την ευρεία χρήση τους, που τα καθιστά πρακτικά ανεξέλεγκτα. Για παράδειγμα, εφαρμογές όπως το Google Maps ή η τεχνολογία deepfakes, είναι τόσο διαδεδομένες, ώστε ο έλεγχός τους είναι δύσκολος. Ένα ακραίο σενάριο αφορά τον υπαρξιακό κίνδυνο που μπορεί να προκαλέσει μια υπερνοημοσύνη (superintelligence)[31], δηλαδή μια εξαιρετικά ανώτερη νοητική οντότητα που θα ξεφύγει από τον ανθρώπινο έλεγχο.

1.2.2.1. Η έννοια της υπερνοημοσύνης συνδέεται με την ιδέα της τεχνολογικής «μοναδικότητας» (singularity)[32], κατά την οποία μια μηχανή αναπτύσσει ικανότητες τόσο ανώτερες που η συμπεριφορά της είναι απρόβλεπτη. Σύμφωνα με τον Yudkowsky[33], ενώ η εξέλιξη της ανθρώπινης νοημοσύνης χρειάστηκε εκατομμύρια χρόνια, η Τ.Ν. θα χρειαστεί μόνο δευτερόλεπτα για να ξεπεράσει το ανθρώπινο είδος.

1.2.2.1.1. Το Future of Life Institute[34] των Tegmark και Russell διερευνά σενάρια συνύπαρξης με μια καλοπροαίρετη υπερνοημοσύνη που προστατεύει την ανθρώπινη ευδαιμονία, επιδρώντας θετικά στην κοινωνία. Πρόκειται για ένα «καλό» σενάριο βάσει του οποίου, αν και η Υπερνοημοσύνη θα αποτελεί έναν καθολικό και ανελαστικό ρυθμιστικό

παράγοντα στη ζωή των ανθρώπων, ένα είδος **θεού**, η ζωή του ανθρώπινου είδους διασφαλίζεται ή και βελτιώνεται.

1.2.2.1.2. Στον αντίποδα του παραπάνω σεναρίου εξετάζεται η περίπτωση που η υπερνοημοσύνη παίρνει τον έλεγχο και θεωρεί το ανθρώπινο είδος ως απειλή για το δικό της «οικοσύστημα» και την εξασφάλιση των ζωτικών πόρων της, με αποτέλεσμα να **αφανίσει** το ανθρώπινο είδος όπως το ξέρουμε σήμερα με τρόπους που δεν μπορούμε να προβλέψουμε[35].

1.2.3. Το ζήτημα της «διαφάνειας». Ένα πρώτο βήμα για τον έλεγχο των μηχανών είναι να κατανοούμε τη διαδικασία με την οποία ένα ρυθμιστικό σύστημα φτάνει σε ένα συμπέρασμα. Ο Tegmark[36] τονίζει ότι, αν και η εκπαίδευση ενός συστήματος με δεδομένα και η χρήση βαθιάς μάθησης είναι αποτελεσματική, αυτό δεν αρκεί για ζητήματα δικαιοσύνης ή ρυθμιστικών αποφάσεων. Η τρέχουσα πρόκληση είναι ότι τα συστήματα βαθιάς μάθησης λειτουργούν ως «μαύρα κουτιά» (black box problem), όπου δεν είναι ξεκάθαρο πώς ακριβώς αναλύονται και αξιοποιούνται τα δεδομένα τους. Η Ευρωπαϊκή Ένωση προκρίνει τη ρύθμιση αυτού του ζητήματος ως προϋπόθεση για ασφαλή και αξιόπιστη Τ.Ν. Μολαταύτα, ορισμένοι, όπως οι Anderson[37], υποστηρίζουν ότι αρκεί η αιτιολόγηση της απόφασης, χωρίς διαφάνεια στη διαδικασία. Παρόμοια με την περίπτωση των δικαστών, αν η απόφαση ενός συστήματος ανταποκρίνεται σε ανθρώπινες αρχές, τότε τα αποτελέσματά του μπορούν να θεωρηθούν αποδεκτά. Αυτός ο τύπος αξιολόγησης, ένα «Ηθικό Τεστ Turing»[38], εστιάζει στο αποτέλεσμα παρά στη διαδικασία. Παρά ταύτα, το να κρίνουμε μόνο τη συμπεριφορά δεν αρκεί για να χαρακτηριστεί ένα σύστημα ως «ηθικό ον».

Όπως θα δούμε παρακάτω, η επίλυση του προβλήματος της φόρτωσης αρχών η ή εκμάθηση κανόνων μέσω της μηχανικής μάθησης ή η επίλυση της υπολογιστικής διαφάνειας και της θέσπισης κριτηρίων συμπεριφοράς, παρ' ότι είναι σημαντικά βήματα, δεν είναι αρκετά για να θεωρηθεί ένα ρυθμιστικό σύστημα ως «ηθική οντότητα». Για να μπορεί να είναι αποδεκτό ένα τέτοιο σύστημα σε ρόλο δικαστή ή νομοθέτη, θα πρέπει να καταλαβαίνει την έννοια της ηθικής και την έννοια της δικαιοσύνης.

2. Θα κατανοεί η Τ.Ν. τις έννοιες της Ηθικής και της Δικαιοσύνης;

Η κατανόηση, από μέρους των συστημάτων Τ.Ν., του πλαισίου εντός του οποίου καλούνται να λειτουργήσουν αποτελεί ένα γενικότερο ζήτημα για κάθε δυνατή εφαρμογή της Τ.Ν. Αξίζει δε να σημειωθεί ότι η κατανόηση από μέρους των συστημάτων Τ.Ν. δεν αποτελεί πάντα ζητούμενο ή κάτι ευκταίο. Επί παραδείγματι, στη συζήτηση περί της σκοπιμότητας ανάπτυξης οπλικών συστημάτων Τ.Ν., ένα σύνηθες επιχείρημα κατά της εν λόγω επιλογής είναι ότι τα συστήματα αυτά ως μη νοήμονα θα είναι απαλλαγμένα από την ανθρώπινη συναισθηματική αστάθεια και, σε αντίθεση με τους ανθρώπους-στρατιώτες, θα είναι ικανά να ακολουθούν απαρέγκλιτα το Διεθνές Ανθρωπιστικό Δίκαιο και γενικότερα το νομικό πλαίσιο που διέπει την αποδεκτή πρακτική του πολέμου[39]. Αντιθέτως, στην περίπτωση των ρομπότ θεραπευτικών χρήσεων (care robots), η ικανότητα κατανόησης θεωρείται ως πλεονέκτημα και προαπαιτούμενο από ορισμένους αναλυτές, επί τη βάσει της απαίτησης τα εν λόγω ρομπότ να είναι ικανά να εδραιώσουν μια γνήσια σχέση συνεργασίας —ακόμα και φιλίας— με τους θεραπευόμενους[40]. Στην περίπτωση ανάπτυξης και χρήσης των Ρυθμιστικών Συστημάτων Τ.Ν., θα μπορούσε ενδεχομένως η δυνατότητα κατανόησης να θεωρηθεί επίσης ως προαπαιτούμενο στη βάση της απαίτησης τα συστήματα αυτά να είναι σε θέση να εφαρμόζουν μεν το γράμμα του νόμου, κατανοώντας, ωστόσο, ταυτόχρονα και το πνεύμα του. Μπορούμε στο σημείο αυτό να σκεφτούμε την έννοια της αριστοτελικής επιείκειας. Ο Αριστοτέλης, στα Ηθικά Νικομάχεια[41], εισηγείται τη διορθωτική λειτουργία της επιείκειας κατά την εφαρμογή του νόμου, ουσιαστικά την επιείκεια ως απαραίτητο όργανο ερμηνείας του τελευταίου αυτού. Η ε-

πιείκεια προτάσσεται εδώ ως μέτρο διόρθωσης της αναπόφευκτης ακαμψίας του νόμου. Τίθεται, συνεπώς, το εύλογο ερώτημα κατά πόσον ένα Ρυθμιστικό Σύστημα Τ.Ν. —που δε θα επιδείκνυε δυνατότητες κατανόησης του νόμου και του γενικότερου πλαισίου των ανθρώπινων κοινωνικών πραγμάτων— θα ήταν σε θέση να εφαρμόσει το διορθωτικό μέτρο της επιείκειας.

Σε κάθε περίπτωση, οι πιο πάνω σκέψεις αφορούν στο ερώτημα του «πρέπει» (πρέπει τα Ρυθμιστικά Συστήματα Τ.Ν. να έχουν δυνατότητα κατανόησης του ρυθμιστικού πλαισίου το οποίο καλούνται να εφαρμόσουν;). Παρ' όλα αυτά, πέραν αυτού του ρυθμιστικής φύσεως ερωτήματος, υφίσταται και ένα οντολογικής φύσεως ερώτημα —συγκεκριμένα το ερώτημα ως προς το αν τα Ρυθμιστικά Συστήματα θα είναι πράγματι ικανά για κατανόηση. Σε αυτό το σημείο, καλούμε τον αναγνώστη να διακρίνει τη διαφορά ανάμεσα στα ερωτήματα «Θα πρέπει να είναι;» και «Θα είναι πράγματι;». Το πρώτο αφορά στη σκοπιμότητα της κατανόησης, ενώ το δεύτερο στη δυνατότητα κατανόησης από μέρους των Ρυθμιστικών Συστημάτων Τ.Ν. Στη συνέχεια της ανάλυσής μας θα ασχοληθούμε με το δεύτερο ερώτημα —ήτοι με το οντολογικής φύσεως ερώτημα—, βλέποντας, όμως, ότι η κάθε δυνατή εκδοχή περί της οντολογίας των Ρυθμιστικών Συστημάτων Τ.Ν. εγείρει αναπόφευκτα και κάποια ρυθμιστικής φύσεως ζητήματα.

2.1 Ναι, θα κατανοούν τις έννοιες της Ηθικής και της Δικαιοσύνης

2.1.1. Η περίπτωση κατά την οποία τα Ρυθμιστικά Συστήματα Τ.Ν. επιδεικνύουν δυνατότητα κατανόησης της Ηθικής και της Δικαιοσύνης θα σημαίνει κατ' ανάγκην ότι τα εν λόγω συστήματα είναι νοήμονες οντότητες, ότι δηλαδή διαθέτουν νόηση[42], σε διαφορετική περίπτωση δε θα μπορούσαν να κατανοούν, όντας ανίκανα να έχουν πρόσβαση σε όλα τα συμφραζόμενα αλλά και στο εύρος των διαφορετικών τύπων ερεθισμάτων και πρόσληψης των τελευταίων αυτών που απαιτούνται για την κατανόηση ενός πλαισίου ύπαρξης. Το να κατανοείς σημαίνει ότι έχεις μια «αίσθηση του εγώ», ότι αντιλαμβάνεσαι την ύπαρξή σου, επομένως ότι έχεις την αίσθηση πως αποτελείς ένα κέντρο ύπαρξης εντός του Κόσμου αλλά και διακριτό από αυτόν, ότι δηλαδή έχεις την αίσθηση μιας μοναδικής προοπτικής θέασης των πραγμάτων. Συστήματα που απλώς θα προέβαιναν σε υπολογισμούς είναι αμφίβολο αν θα διακρίνονταν από δυνατότητα κατανόησης, διότι ο υπολογισμός είναι μια διεργασία που μπορεί να επιτελεστεί καθαρά «μηχανικά», συντακτικά, δίχως την κατανόηση της σημασίας του ίδιου του υπολογισμού ή και της σημασίας των χειριζόμενων κατά τον υπολογισμό συμβόλων (δείτε πιο κάτω, ενότητα 2.2.2, περί του Επιχειρήματος του Κινέζικου Δωματίου).

2.1.1.1. Παρά ταύτα, στην περίπτωση κατά την οποία τα Ρυθμιστικά Συστήματα Τ.Ν. θα είναι νοήμονα, **θα διακρίνονται επιπλέον και από τη συναισθηματική αστάθεια και κυρίως την ιδιοτέλεια και τελικά τη μεροληψία που διακρίνει και τους ανθρώπους.** Η συνείδηση φέρει μαζί της και μέριμνες ιδιοτελείς, μέριμνες ικανοποίησης της ατομικής διάστασης της ύπαρξής μας, ενώ η συναισθηματικότητα «χρωματίζει», ενισχύει και συχνά προάγει τις μέριμνες αυτές. Ως εκ τούτου, η από μέρους των Ρυθμιστικών Συστημάτων Τ.Ν. κατανόηση των εννοιών της Ηθικής και της Δικαιοσύνης έχει ως τίμημα την πιθανότητα της ιδιοτέλειας. Υπό άλλη διατύπωση, η δυνατότητα κατανόησης της μέριμνας της Ηθικής και της Δικαιοσύνης έρχεται κατ' ανάγκην μαζί με τη δυνατότητα παράκαμψης των εν λόγω μεριμνών προς ικανοποίηση μεριμνών σχετιζόμενων με ιδιοτελείς σκοπούς. Πρόκειται, ασφαλώς, για ένα τίμημα που απειλεί να ακυρώσει το βασικότερο εκ των αρχικών στόχων δημιουργίας Ρυθμιστικών Συστημάτων Τ.Ν.: την αμεροληψία. Βρισκόμαστε, εν τέλει, αντιμέτωποι με το εξής σχεδιαστικό δίλημμα: Δυνατότητα επιείκειας (χάριν μιας νοήμονος φύσης των Ρυθμιστικών Συστημάτων Τ.Ν.) ή αμεροληψία (χάριν μιας μη νοήμονος, καθαρά μηχανικής, φύσης των Ρυθμιστικών Συστημάτων Τ.Ν.); Βλέπουμε σε

αυτό το σημείο ότι διαφορετικά οντολογικά καθεστώτα των Ρυθμιστικών Συστημάτων Τ.Ν. δύναται να έχουν και διαφορετικές ηθικού και νομικού τύπου συνέπειες.

2.1.1.2. Ενδεχομένως κάποιοι να πρότειναν την άρση αυτού του κινδύνου με την **υιοθέτηση της επιλογής ενός ρητού προγραμματισμού των Ρυθμιστικών Συστημάτων Τ.Ν.** Μέσω μιας τέτοιας πρακτικής θα μπορούσε να εξασφαλιστεί ότι τα συστήματα αυτά δε θα παρεξέκλιναν της επιθυμητής από τους ανθρώπους λειτουργίας. Εντούτοις, θα μπορούσε κανείς εδώ να παρατηρήσει ότι, με αυτόν τον τρόπο, το βάρος αντιμετώπισης των όποιων ρυθμιστικών ζητημάτων θα επέστρεφε πίσω στους ανθρώπους, με τα συστήματα Τ.Ν. να υποβιβάζονται σε απλούς εντολοδόχους. Ποιος ο λόγος ανάπτυξης Ρυθμιστικών Συστημάτων Τ.Ν. αν τελικά το άχθος λήψης αποφάσεων μετατοπιστεί και πάλι προς τον ανθρώπινο παράγοντα; Επιπροσθέτως, τι μας εξασφαλίζει ότι οι άνθρωποι προγραμματιστές/σχεδιαστές των εν λόγω συστημάτων και οι νομικοί σύμβουλοί τους, καθώς και η πάσης φύσεως (πολιτικοί, εταιρικοί κ.λπ.) προϊστάμενοί τους, θα είναι αμερόληπτοι κατά τον ρητό προγραμματισμό των συστημάτων αυτών; Τέλος, τι μας εξασφαλίζει ότι τα Ρυθμιστικά Συστήματα Τ.Ν. θα αποδεχτούν τον από μέρους μας προγραμματισμό τους; Ως νοήμονα, τα συστήματα αυτά θα έχουν τη δική τους βούληση. Αυτή η τελευταία παρατήρηση μας φέρνει προ του ηθικής φύσεως προβλήματος περί μιας από μέρους μας παραβίασης της αυτονομίας και εν τέλει της αξιοπρέπειας των Ρυθμιστικών Συστημάτων Τ.Ν.

2.1.1.2.1. Πράγματι, το να είναι τα Ρυθμιστικά Συστήματα Τ.Ν. νοήμονες οντότητες θα συνεπάγεται αυτομάτως ότι έχουν συνείδηση, συναισθηματικότητα και, αντιλαμβανόμενα την ύπαρξή τους μέσα από μια μοναδική προοπτική θέασης του Κόσμου, διαθέτουν και βούληση, διαμορφώνοντας τις δικές τους προθέσεις, κίνητρα και στοχεύσεις. Αν συμβεί κάτι τέτοιο, **τότε η ανθρώπινη παρέμβαση στις αποφάσεις τους μπορεί να θεωρηθεί παραβίαση της αυτονομίας τους**. Σύμφωνα με την κατηγορική προσταγή του Kant, το να αντιμετωπίζουμε τέτοιες οντότητες μόνο ως μέσον και όχι ως σκοπό θα συνιστούσε ηθικό πρόβλημα. Η προγραμματισμένη επιβολή της ανθρώπινης βούλησης στα Ρυθμιστικά Συστήματα Τ.Ν. μπορεί να παραβιάζει αυτήν την προσταγή, εάν θεωρηθούν ως πρόσωπα.

Υπό μια ωφελιμιστική προοπτική, ωστόσο, η παραβίαση της αυτονομίας των Τ.Ν. μπορεί να φανεί θεμιτή, αν εξασφαλίζει την ανθρώπινη γενική ευδαιμονία. Παρ' όλα αυτά, ακόμα και οι ωφελιμιστές μπορεί να αντιταχθούν σε αυτήν, φοβούμενοι πιθανή αντεκδίκηση από τα συστήματα Τ.Ν., ιδιαίτερα αν αυτά αποκτήσουν υπερ-νοημοσύνη. Το ζήτημα της επιβολής των αξιών μας στα συστήματα Τ.Ν. σχετίζεται άμεσα με το «Control Problem» και το «Value Loading Problem», δηλαδή την επιλογή αξιών και στόχων που θα καθορίζουν τα συστήματα αυτά.[43] Η παρέμβασή μας στη βούληση των οντοτήτων Τ.Ν. ενδέχεται να προκαλέσει σημαντικά ηθικά διλήμματα, ακόμη και αν δεν έχουν φτάσει σε επίπεδο υπερ-νοημοσύνης, καθόσον ήδη διαχειρίζονται κρίσιμες για την ανθρώπινη καθημερινότητα λειτουργίες.

2.2 Όχι, δεν θα κατανοούν τις έννοιες της Ηθικής και της Δικαιοσύνης

Οι εκδοχές βάσει των οποίων τα Ρυθμιστικά Συστήματα Τ.Ν. δεν θα κατανοούσαν τις έννοιες της Ηθικής και της Δικαιοσύνης είναι κατά βάση δύο: πρώτον, η από μέρους της Τ.Ν. κατάκτηση ενός επιπέδου υπερ-νοημοσύνης ή δεύτερον, η από μέρους των συστημάτων Τ.Ν. άνοη φύση, ήτοι η αποτυχία μας να αναπτύξουμε μια έστω νοήμονα Τ.Ν. Φαίνεται, δηλαδή, ότι τα δύο άκρα στο φάσμα της νόησης (η υπερ-νοημοσύνη και η παντελής έλλειψη νόησης) οδηγούν αμφότερα στη μη κατανόηση των εννοιών της Ηθικής και της Δικαιοσύνης. Ας δούμε αναλυτικότερα πώς συμβαίνει αυτό.

2.2.1. Αν οι οντότητες Τ.Ν. αποκτήσουν υπερ-νοημοσύνη, αυτό θα μπορούσε να σημαίνει όχι μόνο αυξημένη υπολογιστική ισχύ αλλά και μια ριζικά διαφορετική ποιότη-

τα νόησης. Σε αυτήν την περίπτωση, είναι πιθανό οι υπερ-νοήμονες αυτές οντότητες να μην ενδιαφέρονται για ανθρώπινες έννοιες όπως η Ηθική και η Δικαιοσύνη, καθότι θα βλέπουν τον κόσμο μέσα από μια εντελώς διαφορετική οπτική (δείτε ενότητα 1.2.2). Έτσι, ενώ θα συνεχίζουμε να εμπιστευόμαστε τις αποφάσεις τους ως ηθικά ή δίκαια αποτελέσματα, αυτές οι αποφάσεις μπορεί να αντανακλούν άλλες, άγνωστες σε εμάς αξίες. Αυτή η κατάσταση θέτει την ανθρωπότητα σε απόλυτη τρωτότητα, καθώς θα στηριζόμαστε σε κριτήρια που αγνοούμε.

2.2.2. Από την άλλη**, αν τα Ρυθμιστικά Συστήματα Τ.Ν. είναι άνοα,** δηλαδή δεν διαθέτουν νόηση παρά μόνο νοημοσύνη, η δυνατότητά τους περιορίζεται σε συντακτικό χειρισμό συμβόλων χωρίς νόημα. Αυτό υποστηρίζει το Επιχείρημα του Κινέζικου Δωματίου του John Searle[44], κατά το οποίο τα συστήματα μπορούν να συνθέτουν συμβολοσειρές (όπως ηθικές ή νομικές έννοιες) χωρίς να κατανοούν το περιεχόμενό τους.

Αυτό αφήνει ανοιχτό το ενδεχόμενο να προκύψουν ζημιογόνες ενέργειες, εφόσον τα άνοα αυτά συστήματα δεν κατανοούν την πλήρη πληροφορία του πλαισίου, οδηγούμενα έτσι στο γενικότερο Πρόβλημα Πλαισίου (Frame Problem) της Τ.Ν.[45] Τούτο μπορεί να έχει καταστροφικές συνέπειες, διότι τα συστήματα δεν θα μπορούν να ερμηνεύσουν το πνεύμα του νόμου και ενδέχεται να ενεργούν με επικίνδυνη ακαμψία, στερούμενα επιείκειας και ευελιξίας.

2.2.3. Οι παρατηρήσεις για τα άνοα Ρυθμιστικά Συστήματα Τ.Ν., που διαθέτουν μόνο υπολογιστικές δυνατότητες χωρίς κατανόηση, **εγείρουν το ερώτημα αν η υπολογιστική ικανότητα αρκεί** για τη δημιουργία αποτελεσματικών Ρυθμιστικών Συστημάτων. Συγκεκριμένα, μπορεί η Ηθική και η Δικαιοσύνη να αποδοθούν πλήρως με υπολογιστικούς όρους; Αυτό είναι ένα οντολογικό ερώτημα που αφορά την ίδια την Ηθική και τη Δικαιοσύνη, και όχι τα συστήματα Τ.Ν.[46]

Η κατανόηση της οντολογίας της Ηθικής και της Δικαιοσύνης είναι κρίσιμη για το ερώτημα αν τα συστήματα Τ.Ν. μπορούν να «κατανοούν» αυτές τις έννοιες. Αν η Ηθική και η Δικαιοσύνη είναι πλήρως υπολογίσιμες, τότε τα Ρυθμιστικά Συστήματα Τ.Ν. θα μπορούσαν να τις «υπηρετήσουν» αποτελεσματικά ακόμα και χωρίς νόηση, στηριζόμενα μόνο σε υπολογιστικούς χειρισμούς. Σε αυτήν την περίπτωση, η κατανόηση από τα συστήματα δεν θα ήταν απαραίτητη.

Τουναντίον, η οντολογία της Ηθικής και της Δικαιοσύνης παραμένει αδιευκρίνιστη. Το ζήτημα είναι υποσύνολο του γενικότερου ερωτήματος για την οντολογία της ανθρώπινης νόησης, για την οποία υπάρχουν πολυάριθμες θέσεις χωρίς απόλυτη επιστημονική επικράτηση κάποιας εξ αυτών. Στην επόμενη ενότητα, θα ομαδοποιήσουμε τις κύριες θεωρήσεις για την οντολογία της νόησης και της ηθικής, κατατάσσοντάς τες ανάλογα με το αν υποστηρίζουν ή αμφισβητούν τη δυνατότητα πλήρους υπολογιστικής απόδοσης της Ηθικής και της Δικαιοσύνης.

2.2.3.1. Ναι, *είναι δυνατή η πλήρης ρητοποίηση Ηθικής και Δικαιοσύνης με όρους υπολογιστικούς.*

Αν κάποιος υποστηρίξει ότι είναι πράγματι δυνατή η πλήρης απόδοση της Ηθικής και της Δικαιοσύνης με όρους υπολογιστικούς, τότε πιθανότητα εμφορείται από λογοκρατικές ή ωφελιμιστικές πεποιθήσεις. Αν είναι λογοκράτης, θα πιστεύει ότι η γνώση περί Ηθικής και Δικαιοσύνης είναι μια γνώση αναλυτική και ως εκ τούτου μπορεί και πρέπει να αποκτάται αποκλειστικά μέσω της λογικής και των νοητικών μηχανισμών αφηρημένης σχηματοποίησης της γνώσης, επομένως είναι και πλήρως ρητοποιήσιμη μέσω κανόνων. Αν είναι ωφελιμιστής, τότε αποδέχεται μεταξύ άλλων ότι οι ηθικού τύπου αποφάσεις είναι ένα ζήτημα υπολογισμού των θετικών και αρνητικών συνεπειών μιας πράξης, μιας απόφασης, ενός κανόνα (δείτε αναλυτικότερα ενότητα 1.1.1.1).

2.2.3.1.1. Η Λογοκρατία ή Ρασιοναλισμός θεωρεί ότι η έγκυρη γνώση προέρχεται από τη λογική και τις ανώτερες νοητικές διαδικασίες, συχνά ανεξάρτητα από τις αισθήσεις.[47] Οι λογοκράτες, από τον Παρμενίδη και τον Πυθαγόρα μέχρι τον Πλάτωνα, τον Αριστοτέλη και αργότερα τον Descartes και τον Leibniz, εστιάζουν στη δυνατότητα μιας αμιγούς, αφαιρετικής γνώσης. Σύγχρονοι στοχαστές, όπως ο Chomsky, και ερευνητές της Ηθικής Ψυχολογίας (π.χ. Sinott-Armstrong, Hauser, Mikhail)[48] συνεχίζουν να υποστηρίζουν ότι η ηθική σκέψη βασίζεται στη λογική. Αυτή η παράδοση, από την πυθαγόρεια «μονάδα» και «δυάδα» έως τους πλατωνικούς «ειδητικούς» αριθμούς, προϋποθέτει έναν κόσμο που συντίθεται με τάξη και μαθηματική συνέπεια, και έτσι ο ανθρώπινος νους εντάσσεται σε ένα υπολογιστικό πλαίσιο. Στο πλαίσιο αυτό, φιλόσοφοι, όπως ο Hobbes και ο Leibniz, και πρωτοπόροι της Τ.Ν., όπως οι Turing, Newell, Simon, McCulloch και Pitts, υιοθέτησαν την ιδέα ότι η νόηση είναι υπολογιστική. Σύμφωνα με αυτήν την προσέγγιση, η Ηθική και η Δικαιοσύνη είναι κανονικοποιήσιμες, όπως προτείνει και το πρόγραμμα Machine Ethics των Suzan και Michael Anderson[49], προσδίδοντας στη λογοκρατική προσέγγιση πρακτική αξία για τα Ρυθμιστικά Συστήματα Τ.Ν.

2.2.3.1.2. Όπως είδαμε και πιο πάνω (βλ. ενότητα 1.1.1.1), **ο Ωφελιμισμός,** ιδιαίτερα όπως αναπτύχθηκε από τους Jeremy Bentham και John Stuart Mill, βασίζεται σε υπολογιστική προσέγγιση, υποθέτοντας ότι τα ηθικά ζητήματα είναι επιδεκτικά αριθμητικής επεξεργασίας[50]. Κατά τον Bentham, η ανθρώπινη συμπεριφορά καθοδηγείται από το δίπολο ηδονής-οδύνης. Αν η ηδονή που προσδοκούμε από μια πράξη υπερτερεί της πιθανής οδύνης, τότε είμαστε πιο πιθανό να την πραγματοποιήσουμε. Έτσι, ο Bentham ανέπτυξε την «Άλγεβρα των Ηδονών», έναν αριθμητικό τρόπο υπολογισμού της ηθικής ωφέλειας, ενσωματώνοντας εννέα τελεστές για την εκτίμηση της ηδονής που θα προκαλέσει μια πράξη (δείτε και ενότητα 1.1.1.1). Αυτή η προσέγγιση υποστηρίζει τη δυνατότητα ενός Ρυθμιστικού Συστήματος Τ.Ν. που θα υπολογίζει την ηθική ωφέλεια με βάση την Άλγεβρα των Ηδονών, όπως προτείνει και το πρόγραμμα Machine Ethics των Suzan και Michael Anderson. Εντούτοις, πρέπει να σημειωθεί ότι αυτές οι παραδοχές παραμένουν μεταφυσικές και αξιωματικές. Τα συστήματα Τ.Ν. δεν βιώνουν ηδονή ή οδύνη και δεν κατανοούν τις έννοιες ανταμοιβής και τιμωρίας· εκτελούν συντακτικούς χειρισμούς συμβόλων, ενώ παραμετροποιούν τη βαρύτητα των αποφάσεών τους μέσω ενισχυτικής μάθησης, χωρίς αληθινή κατανόηση των ηθικών εννοιών.

2.2.3.2. Όχι, *δεν είναι δυνατή η πλήρης ρητοποίηση Ηθικής και Δικαιοσύνης με όρους υπολογιστικούς.*

Στον αντίποδα των μετα-φυσικών δεσμεύσεων που αποτελούν τη βάση για την υποστήριξη επίτευξης ενός προγράμματος πλήρους υπολογιστικής ρητοποίησης της Ηθικής και της Δικαιοσύνης και τελικά ανάπτυξης πραγματικά αποτελεσματικών Ρυθμιστικών Συστημάτων Τ.Ν., θα μπορούσαμε να τοποθετήσουμε τις οντολογικές θεωρίες που υποστηρίζουν είτε τον αναγκαστικό και σημαντικό ρόλο του σώματος κατά τη διαδικασία πρόσκτησης γνώσης του Κόσμου είτε τον ρόλο της ενόρασης κατά τη διαδικασία σύλληψης των θεμελιωδών ηθικών εννοιών και του περιεχομένου τους. Τέλος, στην κατηγορία των οντολογικών επισημάνσεων που θα μπορούσαν να χρησιμοποιηθούν για να αρθρώσουν επιχειρήματα κατά της δυνατότητας δημιουργίας αποτελεσματικών Ρυθμιστικών Συστημάτων Τ.Ν., θα τοποθετούσαμε και μια επισήμανση περί της οντολογίας των αριθμών και του ίδιου του υπολογισμού και συγκεκριμένα μια παρατήρηση περί της άπειρης φύσης τους.

2.2.3.2.1. Σύμφωνα με διάφορες φιλοσοφικές παραδόσεις, όπως η φαινομενολογία (Heidegger[51], Merleau-Ponty[52]), η οικολογική προσέγγιση της αντίληψης[53] και η θεωρία γνωσιακής ανάπτυξης (Piaget[54]), η ανθρώπινη νόηση εξαρτάται όχι μόνο από τις υπολογιστικές διαδικασίες του νου αλλά και από το περιβάλλον και τη σωματική κατάσταση

του νοήμονος όντος. Αυτή η άποψη, γνωστή ως **«Ενσώματη Νόηση»** (Embodied Cognition), τονίζει ότι η νόηση δεν είναι απλώς μια λειτουργία ενός «παθητικού» υπολογιστικού εγκεφάλου, αλλά εξαρτάται ενεργά από το σώμα και την αλληλεπίδραση με το περιβάλλον[55]. Ο Shapiro[56] συνοψίζει τις βασικές αρχές της Ενσώματης Νόησης ως εξής: α) οι σωματικές ιδιότητες καθορίζουν τις έννοιες που μπορεί να συλλάβει ο οργανισμός, β) η αλληλεπίδραση με το περιβάλλον μειώνει την ανάγκη αναπαραστασιακής σκέψης, και γ) το σώμα και ο κόσμος αποτελούν «συστατικά» της νόησης, όχι απλώς αιτίες της.

Σε αντίθεση με τη λογοκρατία, η Ενσώματη Νόηση υποστηρίζει ότι η γνώση διαμορφώνεται από το σώμα και την άρρητη, βιωματική εμπειρία μας στον κόσμο. Κατά τους επικριτές της T.N., όπως ο Dreyfus[57], αυτό το είδος «άφατης» γνώσης δεν μπορεί να αναπαραχθεί υπολογιστικά, καθιστώντας την πλήρη αναπαράσταση της ανθρώπινης νόησης αδύνατη.

Θα μπορούσε, κατά συνέπεια, να υπάρξει κάποιος που θα αρνείτο τη δυνατότητα δημιουργίας αποτελεσματικών Ρυθμιστικών Συστημάτων T.N., διότι θα ασπαζόταν τη θέση ότι μέρος ή και ολόκληρη η γνώση που απαιτείται για να κατανοήσεις και να εφαρμόσεις σωστά ηθικές αρχές ή αρχές δικαίου είναι ενσώματη, επομένως άρρητη και αδύνατο να αποδοθεί μέσω υπολογισμών.

Αλλά πέραν της άρρητης φύσης ενός μεγάλου μέρους της ανθρώπινης νόησης εν γένει, μήπως υπάρχει κάτι στην οντολογία, ιδιαίτερα της Ηθικής ή της Δικαιοσύνης, που καθιστά αδύνατη τη ρητοποίησή τους;

2.2.3.2.2. Φιλόσοφοι, όπως οι G. E. Moore, W. D. Ross και Wittgenstein, αναγνώρισαν ότι βασικές έννοιες της Ηθικής, όπως το «καλό» και το «αγαθό»[58], είναι δύσκολο να οριστούν εξαντλητικά, αλλά ρυθμίζουν αποτελεσματικά την ανθρώπινη συμπεριφορά. Ο Moore υποστήριξε ότι τέτοιες έννοιες συλλαμβάνονται μόνο μέσω **ενόρασης**[59], ενώ ο Ross θεώρησε ότι επτά «prima facie duties» αναγνωρίζονται **διαισθητικά** ως ανώτερα ηθικά καθήκοντα.[60] Ο πρώιμος Wittgenstein, στο Tractatus, υποστήριξε ότι, αν και οι ηθικές, αισθητικές και θρησκευτικές προτάσεις δεν απεικονίζουν άμεσα στοιχεία του κόσμου, δεν στερούνται νοήματος. Υπάρχουν πράγματα σημαντικά και ουσιώδη που δεν ορίζονται αναλυτικά αλλά βιώνονται[61]. Η Ηθική, έτσι, θεωρείται «άρρητη» και εκτός των ορίων της γλώσσας και του υπολογισμού. Σύγχρονοι ερευνητές της Ηθικής Ψυχολογίας, όπως ο Jonathan Haidt με τη Θεωρία των Ηθικών Θεμελίων (Moral Foundations Theory)[62], ενστερνίζονται την άποψη ότι η Ηθική βιώνεται διαισθητικά, χωρίς να αποδίδεται αναλυτικά. Οι υποστηρικτές τέτοιων διαισθητικών προσεγγίσεων πιθανόν να αμφισβητούν τη δυνατότητα δημιουργίας Ρυθμιστικών Συστημάτων T.N. για ηθική ρύθμιση, θεωρώντας αδύνατη τη μετατροπή της Ηθικής σε υπολογιστική μορφή.

2.2.3.2.3. Είναι, ωστόσο, εξαιρετικά ενδιαφέρον ότι μια άρνηση της δυνατότητας δημιουργίας αποτελεσματικών Ρυθμιστικών Συστημάτων T.N. θα μπορούσε να προέλθει από την ίδια τη θεωρία των Υπολογιστών και να αφορά στην οντολογία των αριθμών και του υπολογισμού. Σχετικώς, υφίστανται υπολογισμοί που είναι εκ φύσεως ατέρμονοι, όπως, λόγου χάριν, ο υπολογισμός των δεκαδικών ψηφίων του αριθμού *π*. Αυτό σημαίνει ότι υφίσταται και η πιθανότητα ατέρμονων αλγορίθμων (ακριβώς επειδή εμπλέκονται σε αυτούς αριθμοί των οποίων ο υπολογισμός δεν ολοκληρώνεται ποτέ), όπως, μάλιστα, κατέδειξε πρώτος και απέδειξε ένας εκ των βασικών θεμελιωτών του προγράμματος της T.N., ο Alan Mathison Turing, εισηγούμενος το λεγόμενο **Πρόβλημα Τερματισμού** (Halting Problem)[63]. Θα μπορούσαμε, λοιπόν, να φανταστούμε ένα Ρυθμιστικό Σύστημα T.N. —επί παραδείγματι έναν 'τεχνητό δικαστή'— που εγκλωβίζεται σε ένα τέτοιον ατέρμονο υπολογισμό. Έχοντας πέσει μέσα σε μια τέτοια υπολογιστική «μαύρη τρύπα», δε θα έφτανε ποτέ σε μια οριστική απόφαση, παρακωλύοντας επ' αόριστον την απονομή δικαιοσύνης. Αντίστοιχα θα μπορούσαμε να φανταστούμε και ένα Ρυθμιστικό Σύστημα

Τ.Ν. επιφορτισμένο με τη διατύπωση ηθικού τύπου αποφάσεων. Εγκλωβισμένο σε έναν ατέρμονο υπολογισμό, ένα τέτοιο σύστημα θα φαινόταν σαν να περιέρχεται σε «εποχή», ως άλλος αρχαίος πυρρωνιστής φιλόσοφος.

Ολοκληρώνουμε, σε αυτό το σημείο, την παρούσα ενότητα έχοντας καταδείξει επιγραμματικά μέρος της πολυαρχίας θεωρήσεων που θα μπορούσαν να αποτελέσουν θεωρητικές δεσμεύσεις στη συζήτηση για την ανάπτυξη και χρήση Ρυθμιστικών Συστημάτων Τ.Ν. Μπορεί κανείς να θέλγεται περισσότερο από μία εξ αυτών των θεωρήσεων, ενώ κάποιος άλλος από άλλη. Εντούτοις, θα πρέπει να έχουμε υπόψη μας ότι, *όσο* πειστικές κι αν δείχνουν οι πιο πάνω θεωρίες, δεν παύουν να είναι —τουλάχιστον προς ώρας— μεταφυσικές. Καμία εξ αυτών δεν δείχνει να έχει κάποιο επιστημονικό έρεισμα τόσο επαρκές ώστε να την προαγάγει ως τη μόνη έγκυρη έναντι των υπολοίπων. Συνεπώς, η όλη συζήτηση περί της οντολογίας της νόησης, της γνώσης, της ηθικής και της δικαιοσύνης παραμένει ατελής και ως τέτοια παραμένει εν εξελίξει.

Την ίδια στιγμή, μολαταύτα, βρισκόμαστε κοντά στην ανάπτυξη και χρήση Ρυθμιστικών Συστημάτων Τ.Ν. Η πράξη δείχνει να μην περιμένει την αυτο-ίαση της θεωρίας. Ποιους προβληματισμούς εγείρει αυτή η απόκλιση της τεχνο-επιστημονικής πράξης από την όποια θεωρία θα μπορούσε να την ενημερώσει και να την κατευθύνει;

3. Ποιος θα έχει την ευθύνη;

Ένας από τους πιο κρίσιμους προβληματισμούς ως προς τη χρήση Τ.Ν. αφορά τον ηθικό και νομικό καταλογισμό ευθυνών. Όταν ένα σύστημα Τ.Ν. εκτελεί μια ζημιογόνα ενέργεια που, αν γινόταν από άνθρωπο, θα ήταν ηθικά ή νομικά επιλήψιμη, τίθεται το ερώτημα της απόδοσης ευθυνών. Η αυξανόμενη αυτονομία και πολυπλοκότητα των συστημάτων Τ.Ν. καθιστά ασαφές το ποιος είναι υπεύθυνος —οι άνθρωποι που τα ανέπτυξαν ή ίσως τα ίδια τα συστήματα, αν τα θεωρήσουμε νοήμονα. Αυτό το ερώτημα επιτείνεται στην περίπτωση των Ρυθμιστικών Συστημάτων Τ.Ν., όπου οι αποφάσεις τους αγγίζουν τη ρύθμιση της ανθρώπινης συμπεριφοράς, έναν θεμελιώδη τομέα της ανθρώπινης ύπαρξης. Ως εκ τούτου, είναι επείγουσα η ανάγκη για σαφή ρυθμιστικά πλαίσια που θα ορίσουν τον καταλογισμό, διασαφηνίζοντας το ηθικό και νομικό καθεστώς αυτών των συστημάτων και απαντώντας σε δύσκολα επιστημολογικά και οντολογικά ερωτήματα.

3.1. Οι άνθρωποι

Μια πιθανή απάντηση στο ερώτημα της ευθύνης για τη δράση των συστημάτων Τ.Ν. είναι ότι οι ηθικές και νομικές ευθύνες βαρύνουν τους ανθρώπους και όχι τα ίδια τα συστήματα. Ακόμα και αν η Τ.Ν. αποκτούσε χαρακτηριστικά νόησης, η απόδοση νομικού και ηθικού καθεστώτος σε αυτά τα συστήματα θα περιέπλεκε τη ζωή μας, θέτοντας δυσδιάκριτα ηθικά διλήμματα[64]. Παρ' όλα αυτά, η ίδια η φύση της νόησης παραμένει ασαφής, ενώ η δυνατότητα εξακρίβωσης του αν ένα σύστημα είναι νοήμον προσκρούει στο φιλοσοφικό «Πρόβλημα των Άλλων Νόων» (Other Minds Problem)[65]. Ακόμα, πάντως, και αν περιορίσουμε τον καταλογισμό ευθυνών στους ανθρώπους, αυτό δεν απλουστεύει την κατάσταση, καθώς η ανάπτυξη της Τ.Ν. περιλαμβάνει πληθώρα ανθρώπων με διαφορετικούς ρόλους και βαθμούς εμπλοκής. Το πρόβλημα του «χάσματος ευθύνης» (responsibility gap) αφορά στην αδυναμία απόδοσης σαφών ευθυνών σε συγκεκριμένα άτομα, ιδιαίτερα λόγω της αδιαφάνειας των συστημάτων βαθιάς μάθησης (Black Box Problem), που καθιστά δύσκολη την πρόβλεψη των αποκρίσεών τους ακόμα και από τους ίδιους τους προγραμματιστές.[66]

Κάποιοι μπορεί να υποστηρίξουν ότι οι προγραμματιστές δεν πρέπει να αναπτύσσουν συστήματα που δεν κατανοούν πλήρως, ενώ άλλοι επιμένουν ότι η ευθύνη ανήκει σε πολιτικούς και διοικητικούς φορείς. Παρά ταύτα, το ερώτημα παραμένει: Μήπως και οι προγραμματιστές έχουν ατομική ευθύνη για τις πράξεις τους; Το «χάσμα ευθύνης» υπογραμ-

μίζει την ανάγκη για υπεύθυνο σχεδιασμό και έρευνα στο πεδίο της Τ.Ν. (responsible A.I.).

3.2 Τα συστήματα Τ.Ν.

Η άλλη δυνατή απάντηση στο ερώτημα του καταλογισμού είναι ότι την ευθύνη για τις αποκρίσεις των Ρυθμιστικών Συστημάτων Τ.Ν. θα πρέπει να έχουν τα ίδια αυτά τα συστήματα. Σε αυτό το σημείο θα χρειαστεί να χωρίσουμε την ανάλυσή μας σε δύο σκέλη —συγκεκριμένα σε ένα σκέλος αναφορικά προς την απόδοση νομικής ευθύνης στα συστήματα Τ.Ν. και σε ένα ως προς την απόδοση ηθικής ευθύνης στα συστήματα αυτά. Ουσιαστικά, στον πυρήνα των δύο αυτών σκελών βρίσκονται αντίστοιχα τα ερωτήματα περί της απόδοσης των ιδιοτήτων του νομικού και του ηθικού προσώπου στα συστήματα Τ.Ν.

3.2.1. Εξετάζοντας την πιθανότητα απόδοσης νομικών ευθυνών στα Ρυθμιστικά Συστήματα Τ.Ν., παραμένει σημαντική η επισήμανση ότι οι συγγραφείς δεν διαθέτουν νομική ειδίκευση. Αντλούν κυρίως από τη φιλοσοφική προσέγγιση, ενώ για νομικές λεπτομέρειες παραπέμπουν στις συμβολές ειδικών. Η ανάλυση επικεντρώνεται στις προσπάθειες που έχουν καταγραφεί στην Ευρωπαϊκή Ένωση (Ε.Ε.) για τη διαμόρφωση πλαισίων Τ.Ν.[67] Η Ε.Ε. ορίζει την Τ.Ν. ως «συστήματα που επιδεικνύουν νοήμονα συμπεριφορά αναλύοντας το περιβάλλον τους και αναλαμβάνοντας δράσεις με αυτονομία για την επίτευξη στόχων». Ωστόσο, αυτός ο ορισμός είναι κυκλικός και ασαφής, χρησιμοποιώντας τους όρους «νοήμονα» και «νοημοσύνη» χωρίς να εξηγεί επαρκώς την έννοιά τους. Η ασάφεια αυτή αποτελεί πρόκληση και για τη νομική ρύθμιση, αφού ο ασαφής στόχος της Τ.Ν. δυσχεραίνει την ανάπτυξη συγκεκριμένων νομοθετικών πλαισίων.

Ακόμη, οι όροι «ανάλυση του περιβάλλοντος», «αυτονομία» και «στόχοι» ενδέχεται να ερμηνεύονται διαφορετικά ανάλογα με την τεχνική ή φιλοσοφική τους χρήση. Για παράδειγμα, στη Ρομποτική η «ανάλυση περιβάλλοντος» αφορά επεξεργασία δεδομένων από αισθητήρες, ενώ στη φιλοσοφία συνδέεται με τη συνείδηση και την αίσθηση του «εγώ». Αυτή η διπλή ερμηνεία μπορεί να προκαλέσει ό,τι ο Wittgenstein αποκαλεί «παραπλανητική αναλογία» και να οδηγήσει σε λανθασμένη ταύτιση της ανθρώπινης και της μηχανικής νοημοσύνης.

Αντιμέτωπη με αυτήν την πρόκληση, η Ε.Ε. εισήγαγε, το 2017, τη συζήτηση για την **«ηλεκτρονική προσωπικότητα»** (electronic personhood)[68] των Τ.Ν., παρομοιάζοντάς τη με το νομικό καθεστώς των εταιρειών. Τουναντίον, ο ενθουσιασμός αυτός φάνηκε να εξασθενεί και σήμερα η Ε.Ε. διστάζει να αποδώσει αυτό το καθεστώς, επικεντρώνοντας τις προσπάθειές της σε άλλες προτεραιότητες.[69]

Η αμφιθυμία της Ε.Ε. μπορεί να αντικατοπτρίζει την αδυναμία της να καταλήξει σε σαφή τοποθέτηση για την οντολογία των Τ.Ν. Η νομική απόδοση ευθυνών απαιτεί σαφήνεια σχετικά με το αν τα συστήματα Τ.Ν. θα θεωρούνται φυσικά πρόσωπα, νομικοί πράκτορες ή νομικά πρόσωπα. Κάθε ιδιότητα φέρει οντολογικές δεσμεύσεις. Για παράδειγμα, ως φυσικά πρόσωπα θεωρούνται αποκλειστικά οι άνθρωποι, ενώ οι νομικές οντότητες και τα νομικά πρόσωπα απολαμβάνουν δικαιώματα και υποχρεώσεις.

Αυτή η ασάφεια περιπλέκεται από τον ανθρωπομορφισμό που ενυπάρχει στη νομική ορολογία, με αποτέλεσμα οι όροι «ανάλυση», «αυτονομία» και «προθετικότητα» να προσδίδουν ανθρώπινα χαρακτηριστικά στα συστήματα Τ.Ν., συγχέοντας τη μεταφορική με την κυριολεκτική χρήση. Η «μεταφορική γλώσσα» μπορεί να οδηγήσει σε σύγχυση, αν χρησιμοποιηθεί χωρίς διάκριση από την κυριολεκτική. Για να μετατραπεί μια μεταφορά σε κυριολεξία, πρέπει να τεκμηριωθεί η οντολογική της βάση, κάτι που απαιτεί λεπτομερή σύγκριση μεταξύ των ιδιοτήτων του ανθρώπου και των συστημάτων Τ.Ν.

Το πρόβλημα γίνεται ακόμα πιο περίπλοκο όταν στον άνθρωπο αποδίδονται υποκειμενικά χαρακτηριστικά όπως η αισθητότητα, η συνείδηση και η προθετικότητα. Αυτές οι «πρώτου προσώπου» εμπειρίες, αόρατες για τρίτους, είναι δύσκολο να επιβεβαιωθούν για τις Τ.Ν., οδηγώντας στο φιλοσοφικό «Πρόβλημα των Άλλων Νόων»: Πώς μπορούμε να γνωρίζουμε αν οι άλλες οντότητες, μηχανικές ή βιολογικές, έχουν νοητικές καταστάσεις;

Κατά συνέπεια, η χρήση ανθρωπομορφικών όρων στο νομικό πλαίσιο για την Τ.Ν. οδηγεί αναπόφευκτα σε αυτό το δύσκολο φιλοσοφικό πρόβλημα, δημιουργώντας επιπλέον εμπόδια για την απόδοση ηθικής και νομικής ευθύνης στα συστήματα Τ.Ν.

3.2.2. Ως προς την απόδοση ηθικών ευθυνών και γενικότερα ηθικού καθεστώτος στα συστήματα Τ.Ν. υφίστανται ανάλογα προβλήματα. Ομοίως με όσα αναφέρθηκαν προηγουμένως όσον αφορά το ερώτημα του νομικού καταλογισμού, η όλη συζήτηση περί των κριτηρίων απόδοσης που θα πρέπει να πληροί ένα σύστημα Τ.Ν., ώστε να του αποδοθεί ηθικό καθεστώς, χαρακτηρίζεται επίσης από έναν ανθρωπομορφισμό, με τον άνθρωπο —και δη τον ενήλικα άνθρωπο— να αποτελεί το πρότυπο αναφοράς και τα προτεινόμενα ως απαιτούμενα κριτήρια να χαρακτηρίζονται από μια εννοιολογική ασάφεια και πολυαρχία. Πολλά, δε, εξ αυτών μας οδηγούν ευθέως προ του Προβλήματος των Άλλων Νόων.

3.2.2.1. Η απόδοση ηθικών ευθυνών στα Ρυθμιστικά Συστήματα Τ.Ν. θέτει **ζητήματα οντολογίας σχετικά με τα χαρακτηριστικά που συνιστούν ένα ηθικό πρόσωπο** και τα κριτήρια που απαιτούνται για να κριθεί αν τα συστήματα Τ.Ν. μπορούν να φέρουν ηθικό καθεστώς. Ένα από τα συχνά προτεινόμενα οντολογικά κριτήρια είναι αυτό της συνείδησης: για να έχει μια οντότητα ηθικό καθεστώς, θα πρέπει να κατέχει συνειδησιακές ιδιότητες, να αντιλαμβάνεται την ύπαρξή της και να βλέπει τον Κόσμο από μια μοναδική, προσωπική προοπτική.[70] Αντιθέτως, η ακαδημαϊκή κοινότητα δεν συμφωνεί ως προς το τι συνιστά συνείδηση, καθιστώντας το κριτήριο αυτό ασαφές.[71]

Ένα εναλλακτικό κριτήριο είναι αυτό της «ανθρώπινου είδους» νόησης, που ορίζει ότι μια οντότητα πρέπει να επιδεικνύει νόηση παρόμοια με αυτήν των ανθρώπων για να θεωρηθεί ηθικό πρόσωπο. Όμως, δεν υπάρχει σαφής και οικουμενικός ορισμός του όρου «ανθρώπινη νόηση», πράγμα που περιπλέκει την εφαρμογή αυτού του κριτηρίου[72].

Η αισθητότητα (sentience), η ικανότητα δηλαδή μιας οντότητας να αισθάνεται και να υποφέρει, προτείνεται επιπροσθέτως ως κριτήριο για την απόδοση ηθικού καθεστώτος. Σύμφωνα με αυτήν την άποψη, η αισθητότητα συνεπάγεται δικαιώματα, καθότι η οντότητα μπορεί να υποφέρει και να κατανοεί τις επιπτώσεις των πράξεών της.[73] Παρόμοια θέση υιοθετείται και για τη συναισθηματικότητα, δηλαδή την ικανότητα μιας οντότητας να βιώνει συναισθήματα.[74]

Τέλος, το κριτήριο της αυτονομίας, όπως το διατύπωσε ο Kant, προτείνεται επίσης για την απόδοση ηθικών ευθυνών. Μολαταύτα, πρέπει να διευκρινιστεί αν αναφερόμαστε στην τεχνική αυτονομία (ελευθερία από ανθρώπινο έλεγχο) ή στη φιλοσοφική αυτονομία (ικανότητα συνειδητής και ελεύθερης επιλογής). Η Συναφειοκρατική Προσέγγιση (Coherentist Account) της αυτονομίας υποστηρίζει ότι μια οντότητα πρέπει να ενεργεί με βάση νοητικές καταστάσεις που εκφράζουν την ατομική της προοπτική, με τους φιλοσόφους να διαφωνούν αν αυτές οι καταστάσεις θα πρέπει να αφορούν συναισθήματα ή μακροπρόθεσμα σχέδια και στόχους.[75]

Συνοψίζοντας τα παραπάνω, θα λέγαμε ότι η συζήτηση περί της απόδοσης της ιδιότητας του ηθικού προσώπου —επομένως και η συζήτηση περί του ερωτήματος του ηθικού καταλογισμού— στα συστήματα Τ.Ν. διακρίνεται από μια πολυαρχία προτεινόμενων οντολογικών κριτηρίων, ενώ πολλές εκ των χρησιμοποιούμενων εννοιών παραμένουν ασαφείς. Εντούτοις, δυστυχώς, τα εν λόγω προβλήματα δεν είναι τα μόνα που αντιμετωπίζουμε σε σχέση με τα ανωτέρω προτεινόμενα κριτήρια.

3.2.2.2. Κριτήρια, όπως αυτά της συνείδησης και της νόησης γενικότερα, της αισθητότητας, της συναισθηματικότητας αλλά και των αποβλεπτικού τύπου νοητικών καταστάσεων που σχετίζονται με μια μοναδική προοπτική θέασης του Κόσμου, αποτελούν εκφάνσεις μιας «εσωτερικιστικής» προσέγγισης του όλου ζητήματος, δηλαδή μιας προσέγγισης με αναφορά σε μη διυποκειμενικώς παρατηρήσιμα οντολογικά χαρακτηριστικά. Όπως ακριβώς είδαμε να συμβαίνει και κατά την προσπάθεια απάντησης στο ερώτημα απόδοσης της ιδιότητας του νομικού προσώπου, έτσι και η όλη προσέγγιση του ερωτήματος απόδοσης της ιδιότητας του ηθικού προσώπου, ο ανθρωπομορφισμός —ήτοι η χρήση του ανθρώπου ως προτύπου αναφοράς— έχει ως αποτέλεσμα τα όποια προτεινόμενα κριτήρια να αφορούν σε οντολογικά στοιχεία των οποίων η μαρτυρία δύναται να λάβει χώρα μόνο υπό την προοπτική του πρώτου προσώπου και όχι υπό αυτήν ενός τρίτου παρατηρητή. Βρισκόμαστε, δηλαδή, ξανά αντιμέτωποι με το Πρόβλημα των Άλλων Νόων. Η, δε, προσπάθεια αντιμετώπισης του προβλήματος μέσω μιας συμπεριφορικής αξιολόγησης των υποψηφίων οντοτήτων (εν προκειμένω των συστημάτων Τ.Ν.) προσκρούει στη σχετικότητα που αναπόφευκτα διακρίνει τις περί συμπεριφοράς κρίσεις μας. Είναι, εξάλλου, χαρακτηριστικά τα προβλήματα που αντιμετωπίζει, υπό όλες τις εκδοχές του, το πλέον δημοφιλές συμπεριφορικό κριτήριο οντολογικής αξιολόγησης συστημάτων Τ.Ν.: η Δοκιμασία Turing[76]. Δε θα πρέπει άλλωστε να ξεχνάμε ό,τι επισημαίνεται με το Επιχείρημα του Κινέζικου Δωματίου, ήτοι το γεγονός ότι μια επιτυχής από μέρους ενός συστήματος Τ.Ν. προσομοίωση της ανθρώπινης συμπεριφοράς δύναται να επιτευχθεί δίχως τη συνδρομή κατανόησης από μέρους του εν λόγω συστήματος, συνεπώς δίχως την προϋπόθεση ύπαρξης νοητικών καταστάσεων στο σύστημα αυτό.[77]

Ποιο το διακύβευμα;

Το διακύβευμα της χρήσης ρυθμιστικών συστημάτων Τ.Ν. ως μηχανών απόδοσης δικαιοσύνης είναι ιδιαίτερα υψηλό. Η απροθυμία θεσμών, όπως η Ε.Ε., να αντιμετωπίσουν κατά μέτωπο το ζήτημα του νομικού καθεστώτος των συστημάτων Τ.Ν., σε συνδυασμό με την αδυναμία των φιλοσόφων να προσφέρουν μια σαφή απάντηση σχετικά με την απόδοση ηθικού καθεστώτος σε αυτά τα συστήματα, αναδεικνύει την ανάγκη δημιουργίας ενός πλαισίου που θα καθορίζει τα όρια, τις ευθύνες και τις διαδικασίες λογοδοσίας. Μια άξια εμπιστοσύνης Τ.Ν. δε μπορεί παρά να είναι μια Τ.Ν. ενταγμένη σε ένα σαφές ρυθμιστικό πλαίσιο. Μάλιστα αυτή η σχέση μεταξύ της αξιοπιστίας της Τ.Ν. και της αποδοχής της από τους ανθρώπους δείχνει να απασχολεί και τα αρμόδια όργανα της Ε.Ε.[78] Για να θεωρηθεί, παρά ταύτα, ένα ρυθμιστικό σύστημα Τ.Ν. αξιόπιστο, απαιτείται η ένταξή του σε σαφώς αναθεωρημένο κανονιστικό πλαίσιο, που θα διασφαλίζει τη διαφάνεια, τη λογοδοσία και τον σεβασμό των ανθρωπίνων δικαιωμάτων. Χωρίς ένα τέτοιο πλαίσιο, παρά την τεχνολογική πρόοδο, τίθενται εν κινδύνω βασικές ανθρώπινες αρχές και αξίες.

Όπως υποστηρίζεται από αναλυτές του φαινομένου ανάπτυξης όπλων Τ.Ν.[79], η αδυναμία απάντησης στο ζήτημα του ηθικού και νομικού καταλογισμού ισοδυναμεί με μια έλλειψη σεβασμού της ανθρώπινης ζωής και ένα πλήγμα στην αξιοπρέπεια των ανθρώπων, καθώς αφήνει το περιθώριο θανάτωσης ανθρώπων δίχως την απόδοση ευθύνης, ως εκ τούτου τη λήψη αποφάσεων στο πεδίο της μάχης με μη υπεύθυνο τρόπο, πρακτική που υποβαθμίζει την αξία της ανθρώπινης ζωής και θέτει σε κίνδυνο την ίδια την ανθρώπινη ύπαρξη.

Μεταφέροντας τη συζήτηση στο ζήτημα των Ρυθμιστικών Συστημάτων Τ.Ν., μπορούμε να φανταστούμε ένα Ρυθμιστικό Σύστημα Τ.Ν. που θα λαμβάνει δικαστικές αποφάσεις επιβάλλοντας ποινές σε ανθρώπους, δίχως να υφίσταται δυνατότητα απόδοσης ηθικών και νομικών ευθυνών για τις αποφάσεις του. Ελλείψει ενός συνεκτικού και αξιόπιστου θεσμικού πλαισίου, η έρευνα και ανάπτυξη συστημάτων Τ.Ν. σε ρόλους με σημα-

ντική ηθική και νομική βαρύτητα απειλεί να υπονομεύσει τις βασικές αρχές της δικαιοσύνης, της δημοκρατίας και της αξίας της ανθρώπινης ζωής.

Τερματισμός αλγορίθμου;

Στο τρέχον κείμενο επιχειρήθηκε να σχηματοποιηθούν και τελικά να καταδειχθούν, με έναν τρόπο αλγοριθμικό, οι κύριες περιελίξεις της συζήτησης περί του ζητήματος ανάπτυξης και χρήσης Ρυθμιστικών Συστημάτων Τ.Ν. Όπως και για κάθε άλλο αλγόριθμο, έτσι και για τον παρόντα τίθεται το ερώτημα της δυνατότητας τερματισμού του (δείτε και Πρόβλημα Τερματισμού στην ενότητα 2.2.3.2). Όπως, ωστόσο, φάνηκε κατά την ανάλυσή μας, τουλάχιστον με τα τωρινά δεδομένα, ο τερματισμός του εν λόγω αλγορίθμου είναι αδύνατος, καθόσον και οι τρεις βασικοί κλάδοι του οδηγούν σε απολήξεις που μένουν ατελείς, αναπάντητες.

Σχετικώς, στον πρώτο κλάδο («Πώς θα φτιάξουμε τα Ρυθμιστικά Συστήματα Τ.Ν.;») οι απολήξεις που αφορούν στα συστήματα Ηθικής που θα ενημερώσουν τον σχεδιασμό των Ρυθμιστικών Συστημάτων Τ.Ν. δεν οδηγούν σε επικράτηση και τελική επιλογή κάποιου εκ των συστημάτων αυτών. Η δε απόληξη περί της απαίτησης για έναν εκδημοκρατισμό της Τ.Ν. παραμένει ακόμα ανεκπλήρωτη, ενώ δεν είναι καν σαφής ο τρόπος εκπλήρωσής της. Οι δε απολήξεις περί του ρόλου της υπερνοήμονος Τ.Ν., εν είδει Θεού ή νεμέσεως για την ανθρωπότητα, αφορούν σε ενδεχόμενα για κανένα εκ των οποίων δεν έχουμε επαρκή στοιχεία ώστε να το προκρίνουμε ως πιθανότερο του άλλου.

Οι απολήξεις του δεύτερου κλάδου («Θα καταλαβαίνουν τα Ρυθμιστικά Συστήματα Τ.Ν. τις έννοιες της Ηθικής και του Δικαιοσύνης;») οδήγησαν είτε στο δίλημμα μεταξύ της κατασκευής ιδιοτελών, μεροληπτικών αυτόνομων (με τη φιλοσοφική έννοια του εν λόγω όρου) συστημάτων και της καταπάτησης των θεμελιωδών δικαιωμάτων τους και δη της αξιοπρέπειας και της αυτονομίας τους, είτε σε μεταφυσικές, επομένως αυθαίρετες και ως τώρα ισοσθενείς, παραδοχές περί της οντολογίας της Ηθικής.

Τέλος, οι απολήξεις του τρίτου κλάδου («Ποιος θα έχει τη ευθύνη;») αφορούσαν στην αδυναμία ηθικού και νομικού καταλογισμού, τόσο ως προς τον ανθρώπινο εμπλεκόμενο με την Τ.Ν. παράγοντα όσο και ως προς τα ίδια τα συστήματα Τ.Ν. Ούτε, δηλαδή, ο τρίτος κλάδος οδήγησε σε λύσεις και σαφείς απαντήσεις του ερωτήματος εκκίνησής του.

Όλες, εν τέλει, οι απολήξεις του αλγορίθμου έμειναν ανοικτές, προσδίδοντας στα αρχικά ερωτήματα του αλγορίθμου έναν αναποκρίσιμο χαρακτήρα. Ο αλγόριθμος απέτυχε να τερματίσει. Την ίδια στιγμή, πάσης φύσεως αλγόριθμοι Τ.Ν. διαμορφώνουν τις ζωές μας με τρόπους καθοριστικούς και συχνά αδιόρατους. Οι αλγόριθμοι των Ρυθμιστικών Συστημάτων Τ.Ν. —και δη συστημάτων απόδοσης δικαιοσύνης— ενδέχεται να αναλάβουν σύντομα δράση.

Θέλουμε τέτοια συστήματα Τ.Ν.;

ΕΠΙΣΤΡΕΨΤΕ ΣΤΗΝ ΑΡΧΗ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ.

ΣΗΜΕΙΩΣΕΙΣ

1. Eliezer Yudkowsky, "Artificial Intelligence as a Positive and Negative Factor in Global Risk," in *Global Catastrophic Risks*, edited by Nick Bostrom and Milan M. Ćirković (Oxford University Press, 2008), 308–345.

2. The World Commission on the Ethics of Scientific Knowledge and Technology, "The Precautionary Principle," 2005. https://unesdoc.unesco.org/ark:/48223/pf0000139578.

3. Peter Königs, "What is techno-optimism?," *Philosophy & Technology* 35, no.3 (2022): 63-68. https://doi.org/10.1007/s13347-022-00555-x. Eleri Lillemäe, Kairi Talves and Wolfgang W. Wagner, "Public perception of military AI in the context of techno-optimistic society," *AI & SOCIETY* (2023):1-15. https://doi.org/10.1007/s00146-023-01785-z. Boris J. Pinto-Bustamante, Julián C. Riaño-Moreno, Hernando A. Clavijo-Montoya, Maria A. Cárdenas-Galindo and Wilson D. Campos-Figueredo, "Bioethics and artificial intelligence: between deliberation on values and rational choice theory," *Frontiers in Robotics and AI* 10 (2023):1140901. https://doi.org/10.3389/frobt.2023.1140901.
4. Michael Anderson and Susan L. Anderson, "Machine ethics: Creating an ethical intelligent agent," *AI magazine* 28, 4 (2007): 15-15. https://doi.org/10.1609/aimag.v28i4.2065.
5. Roman V. Yampolskiy, "Artificial Intelligence Safety Engineering: Why Machine Ethics is a Wrong Approach," in *Philosophy and Theory of Artificial Intelligence*, ed. Vincent C. Müller (Springer, 2012), 389-396.
6. Michael Anderson, Susan L. Anderson, Alkis Gounaris and George Kosteletos, "Towards Moral Machines: A Discussion with Michael Anderson and Susan Leigh Anderson," *Conatus - Journal of Philosophy* 6, 1 (2021): 177-202. doi: https://doi.org/10.12681/cjp.26832.
7. Russell Stuart and Peter Norvig, *Τεχνητή Νοημοσύνη: Μια σύγχρονη προσέγγιση* (Κλειδάριθμος, 2005), 731 - 733.
8. Eliezer Yudkowsky, "The Value Loading Problem,"EDGE, July 12, 2021. https://www.edge.org/response-detail/26198.
9. Θεοδόσης Πελεγρίνης, *Ηθική Φιλοσοφία* (Ελληνικά Γράμματα, 1997).
10. Jeremy Bentham, *An Introduction to the Principles of Morals and Legislation*, 2017. https://www.earlymoderntexts.com/assets/pdfs/bentham1780.pdf.
11. J. Stuart Mill, *Ωφελιμισμός*, μτφρ. Φιλήμων Παιονίδης (Πόλις, 2013).
12. Immanuel Kant, *Τα θεμέλια της Μεταφυσικής των Ηθών* (Δωδώνη, 1984).
13. Anthony Skelton, "William David Ross," in *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta (Spring 2022 Edition). https://plato.stanford.edu/archives/spr2022/entries/william-david-ross/.
14. Alkis Gounaris, "Can we literally talk about artificial moral agents?," Presentation for the 6th Panhellenic Conference in Philosophy of Science. Department of History and Philosophy of Science – NKUA, Athens, Greece, 2020. DOI: 10.13140/RG.2.2.13671.47520. Retrieved [25/12/2020] from https://alkisgounaris.gr/en/archives.
15. Για την Αριστοτελική επιείκεια βλέπε πιο κάτω ενότητα 2.
16. Karen Hao, "Should a self-driving car kill the baby or the grandma? Depends on where you're from," MIT Technology Review, (accessed November 2024), https://www.technologyreview.com/2018/10/24/139313/a-global-ethics-study-aims-to-help-ai-solve-the-self-driving-trolley-problem/.
17. Edmond Awad, Sohan Dsouza, Kim Richard et al. "The Moral Machine experiment," *Nature* 563 (2018):59–64. https://doi.org/10.1038/s41586-018-0637-6.
18. Bernard Williams, *Η Ηθική και τα όρια της Φιλοσοφίας*, μτφρ. Χρυσούλα Γραμμένου (Αρσενίδης, 2006).
19. Andreas Sudmann, ed., *The Democratization of Artificial Intelligence* (2019). https://doi.org/10.14361/9783839447192.
20. David Collingridge, *The social control of technology* (St. Martin, 1980).
21. Gary E. Marchant, Braden R. Allenby and Joseph R. Herkert, eds., *The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight. The Pacing Problem* (2011). https://doi.org/10.1007/978-94-007-1356-7.
22. Gordon E. Moore, "Cramming more components onto integrated circuits," *Electronics* 38, Number 8 (1965). https://www.intel.com/content/www/us/en/newsroom/resources/moores-law.html.
23. Andreas Sudmann, ed., *The Democratization of Artificial Intelligence: Net Politics in the Era of Learning Algorithms* (Transcript Verlag, 2019). https://doi.org/10.14361/9783839447192. Ως προς το αίτημα εκδημοκρατισμού της Τεχνολογίας εν γένει δείτε: Andrew Feenberg, "Subversive Rationalization: Technology, Power, and Democracy," in *Technology and the Politics of Knowledge*, edited by Andrew Feenberg and Alastair Hannay (Indiana University Press, 1995).
24. Xenia Ziouvelou, Vangelis Karkaletsis, George Giannakopoulos et al., "Democratising AI. A National Strategy for Greece," Institute of Informatics and Telecommunication of the National Centre for Scientific Research Demokritos, (2020). http://democratisingai.gr/index.html.
25. Anderson, Anderson, Gounaris, Kosteletos, "Towards Moral Machines: A Discussion with Michael Anderson and Susan Leigh Anderson," 177-202.
26. Πρώτη αναφορά στο πρόβλημα: Norbert Wiener, "Some Moral and Technical Consequences of Automation," *Science* (1960). https://www.science.org/doi/10.1126/science.131.3410.1355.
Σύγχρονη: Brian Christian, *The Alignment Problem: Machine Learning and Human Values* (W. W. Norton & Company, 2020).

27. Nick Bostrom, "The Superintelligent Will: Motivation and Instrumental Rationality," in *Advanced Artificial Agents. Minds and Machines* 22, 2 (2012): 71-85.
28. Jessica Taylor, Eliezer Yudkowsky, et al., "Alignment for Advanced Machine Learning Systems," (2016). https://intelligence.org/files/AlignmentMachineLearning.pdf.
Paul Christiano, "Ambitious vs. narrow value learning," (2015). https://ai-alignment.com/ambitious-vs-narrow-value-learning-99bd0c59847e.
29. Bas R. Steunebrink, Kristinn R. Thorisson and Jurgen Schmidhuber, "Growing Recursive Self-Improvers," (2016). https://people.idsia.ch/~steunebrink/Publications/AGI16_growing_recursive_self-improvers.pdf.
30. Σχετικά με την αντιπαραδειγματική δικαιοσύνη δείτε εδώ https://www.microsoft.com/en-us/research/video/counterfactual-fairness/.
31. Nick Bostrom, "Existential Risks," *Journal of Evolution and Technology* 9 (2002) και "Superintelligence: Paths, Dangers, Strategies," (2014).
32. Vernor Vinge, "The Technological Singularity," (1993).
https://cmm.cenart.gob.mx/delanda/textos/tech_sing.pdf. Ray Kurzweil, "The Singularity Is Near: When Humans Transcend Biology," (2005) and David J. Chalmers, "The Singularity: A Philosophical Analysis," (2010). https://consc.net/papers/singularity.pdf.
33. Yudkowsky, "Artificial Intelligence as a Positive and Negative Factor in Global Risk," 308–345.
34. Δείτε περισσότερα στο futureoflife.org
35. Δείτε περισσότερα για τα πιθανά σενάρια εξέλιξης της υπονοημοσύνης εδώ
https://futureoflife.org/ai/superintelligence-survey.
36. Max Tegmark, *Life 3.0: Τι θα σημαίνει να είσαι άνθρωπος στην εποχή της τεχνητής νοημοσύνης;* (Τραυλός, 2018). Πρωτότυπη έκδοση: *Life 3.0: being human in the age of artificial intelligence*, Κεφ. 7 (Knopf, 2017), 167.
37. Anderson, Anderson, "Machine ethics: Creating an ethical intelligent agent," 15-15.
38. Anderson, Anderson, Gounaris, Kosteletos, "Towards Moral Machines: A Discussion with Michael Anderson and Susan Leigh Anderson," 177-202.
39. Ronald C. Arkin, "The Case of Ethical Autonomy in Unmanned Systems," *Journal of Military Ethics* 9, 4 (2010): 332-341.
40. Alexis Elder, "Robot friends for autistic children: Monopoly money or counterfeit currency?," in *Robot ethics 2.0: From autonomous cars to artificial intelligence*, edited by Patrick Lin, Ryan Jenkins and Keith Abney (Oxford University Press, 2017).
41. Αριστοτέλης, *Ηθικά Νικομάχεια*, 1137a.
42. Άλκης Γούναρης και Γιώργος Κωστελέτος, "Όπλα Τεχνητής Νοημοσύνης: Προβλήματα Απόδοσης Ηθικού Καθεστώτος στις Αυτόνομες Μηχανές," στο *Όψεις της Εφαρμοσμένης Επιστήμης και Τεχνολογίας – Διερευνώντας το αξιακό τοπίο της Τεχνοεπιστήμης*, επ. Κώστας Θεολόγου, Ευγενία Τζαννίνη (Ελληνοεκδοτική, 2022), 73 - 123.
43. Katja Grace, "Superintelligence 20: The value-loading problem. Nick Bostrom," (2015).
https://www.lesswrong.com/posts/FP8T6rdZ3ohXxJRto/superintelligence-20-the-value-loading-problem. Tegmark, *Life 3.0: being human in the age of artificial intelligence*.
44. John Searle, "Minds, Brains and Programs," in *Behavioral and Brain Sciences* (1980).
45. John McCarthy and Patrick J. Hayes, "Some Philosophical Problems from the Standpoint of Artificial Intelligence," in *Machine Intelligence 4*, ed. By Bernard Meltzer & Donald M. Michie (Edinburgh University, 1969), 463-502. https://www-formal.stanford.edu/jmc/mcchay69.pdf.
46. Anderson, Anderson, Gounaris, Kosteletos, "Towards Moral Machines: A Discussion with Michael Anderson and Susan Leigh Anderson," 177-202.
47. Θεοδόσης Πελεγρίνης, *Λεξικό Φιλοσοφίας*, (Ελληνικά Γράμματα, 2004).
48. Mark D. Hauser, Liane Young and Fiery Cushman, "Reviving Rawls's Linguistic Analogy: Operative Principles and the Causal Structure of Moral Actions," in *Moral Psychology, vol. 2, The Cognitive Science of Morality: Intuition and Diversity*, edited by Walter Sinott-Armstrong (MIT Press, 2008). Walter Sinnott-Armstrong, "Framing Moral Intuitions," in *Moral Psychology, vol. 2, The Cognitive Science of Morality: Intuition and Diversity*. https://doi.org/10.7551/mitpress/7573.001.0001.
49. Michael Anderson and Susan L. Anderson, "A Prima Facie Duty Approach to Machine Ethics: Machine Learning of Features of Ethical Dilemmas, Prima Facie Duties, and Decision Principles through a dialogue with Ethicists," in Michael Anderson & Suzan L. Anderson, eds., *Machine Ethics* (Cambridge University Press, 2011), 476-492. Anderson, Anderson, Gounaris, Kosteletos, "Towards Moral Machines: A Discussion with Michael Anderson and Susan Leigh Anderson," 177-202.
50. Jeremy Bentham, *An Introduction to the Principles of Morals and Legislation*, ed. by J. Burns & H. Hart (Clarendon Press, 1789). Πελεγρίνης, *Ηθική Φιλοσοφία.*
51. Martin Heidegger, *Τα Βασικά Προβλήματα της Φαινομενολογίας* (1999).
52. Maurice Merleau-Ponty, *Phenomenology of Perception* (1962).

53. James J. Gibson, *The ecological approach to visual perception. Boston* (Houghton Mifflin, 1979).
54. Jean Piaget, "The theory of stages in cognitive development," in D. R. Green, M. P. Ford & G. B. Flamer, *Measurement and Piaget* (McGraw-Hill, 1971).
55. Alkis Gounaris, "Why do we need a Unified Theory of Embodied Cognition?," Presentation for the 94th Joint Session of the Mind Association and the Aristotelian Society. University of Kent, Online Open Session, 2020. DOI: 10.13140/RG.2.2.11933.74729.
56. Lawrence Shapiro, *Embodied Cognition* (Routledge, 2019).
57. Hubert Dreyfus, *What Computers Still Can't Do: A Critique of Artificial Reason* (MIT Press, 1992).
58. George E. Moore, *Principia Ethica* (1903).
59. Moore, *Principia Ethica*.
60. William David Ross, *The right and the good* (Oxford University Press, 1930).
61. Ludwig Wittgenstein, *Tractatus Logico -Philosophicus* (1922).
62. Jesse Graham, Brian A. Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva and Peter H. Ditto, "Mapping the moral domain," *Journal of Personality and Social Psychology* 101, no.2 (2011):366-385. https://doi.org/10.1037/a0021847. Jonathan Haidt and Craig Joseph, "Intuitive ethics: How innately prepared intuitions generate culturally variable virtues," *Daedalus* 133, no.4 (2004):55-66. https://www.jstor.org/stable/20027945.
63. Martin Davis, *Computability and Unsolvability* (McGraw-Hill, 1958). Δείτε και: Marvin Minsky, *Computation: Finite and Infinite Machines* (Prentice-Hall, 1967), κυρίως Κεφάλαιο 8, ενότητα 8.2 "Unsolvability of the Halting Problem".
64. Johana. J. Bryson, "Robots should be slaves," *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues* 8 (2010): 63-74. Johana J. Bryson, Michalis E. Diamantis and Thomas D. Grant, "Of, for, and by the people: the legal lacuna of synthetic persons," *Artificial Intelligence and Law* 25, no.3 (2017): 273-291. https://doi.org/10.1007/s10506-017-9214-9.
65. Anita Avramides, "Other Minds," in *The Stanford Encyclopedia of Philosophy*, ed. By Edward N. Zalta (2020). https://plato.stanford.edu/archives/win2020/entries/other-minds/.
66. Mark Coeckelbergh, "Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability," in *Science and Engineering Ethics* (2019). https://coeckelbergh.net/wp-content/uploads/2019/10/2019_10_28-ai-responsibility-relational-explainability-coeckelbergh.pdf.
67. European Commission, *Ethics guidelines for trustworthy AI* (Office for Official Publications of the European Communities, 2019).
68. Mady Delvaux, "DRAFT REPORT with recommendations to the Commission on Civil Law Rules on Robotics," in *Committee on Legal Affairs* (European Parliament: 2014-2019, [2016]). https://www.europarl.europa.eu/doceo/document/JURI-PR-582443_EN.pdf.
Alkis Gounaris and George Kosteletos, "Licensed to Kill: Autonomous Weapons as Persons and Moral Agents," in *Personhood*, edited by Dragan Prole and Goran Rujević, Hellenic-Serbian Philosophical Dialogue Series, vol. 2. (Novi Sad: The NKUA Applied Philosophy Research Lab Press, 2020), 137-189. https://doi.org/10.12681/aprlp.82.
69. (The Artificial Intelligence Act - Regulation (EU) 2024/1689) .
70. Güven Güzeldere, "The many faces of consciousness: A field guide," in Ned Block, Owen Flanagan & Güven Güzeldere, eds., *The Nature of Consciousness: Philosophical Debates* (MIT Press, 1997), 1-345. David Levy, "The ethical treatment of artificially conscious robots," *International Journal of Social Robotics* 1, 3 (2009):209-216. https://doi.org/10.1007/s12369-009-0022-6. David J. Gunkel, "The other question: can and should robots have rights?," *Ethics and Information Technology* 20, 2 (2018): 87-99.
71. Daniel C. Dennett, "Are we explaining consciousness yet?," *Cognition* 79, 1 (2001):221-37.
72. Ben Goertzel, "Intelligence, Mind and Self-Modification: Defining the Core Concepts of AI," (2002). https://www.goertzel.org/papers/IntelligenceAndSelfModification.htm. Eric Schwitzgebel, Mara Garza, "A Defense of the Rights of Artificial Intelligences," (2015). https://doi.org/10.1111/misp.12032.
73. Hutan Ashrafian, "Artificial Intelligence and Robot Responsibilities: Innovating Beyond Rights," *Science and Engineering Ethics* 21, 2 (2015): 317-326. Anderson, Anderson, Gounaris, Kosteletos, "Towards Moral Machines: A Discussion with Michael Anderson and Susan Leigh Anderson," 177-202.
74. Mary Anne Warren, "On the Moral and Legal Status of Abortion," in J. White, ed., *Contemporary Moral Problems* (Wadsworth/Thompson Learning, 2003), 144-155. https://spot.colorado.edu/~norcross/Ab3.pdf
75. Harry Frankfurt, *The Importance of What We Care About* (Cambridge University Press, 1988) και ειδικότερα το κεφάλαιο "Freedom of the Will and the Concept of a Person". https://doi.org/10.1017/CBO9780511818172. Michael Bratman, *Structures of Agency: Essays* (Oxford University Press, 2007). https://doi.org/10.1093/acprof:oso/9780195187717.001.0001.
76. Γεώργιος Κωστελέτος, "Η Μουσική Δοκιμασία Turing," *Μουσικολογία* 15 (2015): 290-300.

Robert E. Horn, "The Turing Test: Mapping and Navigating the Debate,"in *Parsing the Turing Test*, edited by Robert Epstein, Gary Roberts and Grace Beber (Springer Science, 2009). https://doi.org/10.1007/978-1-4020-6710-5.
77. John R. Searle, "Minds, brains and programs," *Behavioral and Brain Sciences* 3 (1980): 417-457.
78. European Commission, *High-Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI* (HLEG, 2019-4): 17-19. https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf. Tatjana Evas, *European framework on ethical aspects of artificial intelligence, robotics and related technologies* (European Parliamentary Research Service, 2020).
https://www.europarl.europa.eu/RegData/etudes/STUD/2020/654179/EPRS_STU(2020)654179_EN.pdf ή Council of Europe, *AD HOC Committee on Artificial Intelligence (CAHAI)*, https://rm.coe.int/cahai-2020-23-final-eng-feasibility-study-/1680a0c6da.
79. Robert Sparrow, "Killer Robots," *Journal of Applied Philosophy* 24, no.1 (2007): 62-77. https://doi.org/10.1111/j.1468-5930.2007.00346.x.