Presentation for the 6th Panhellenic Conference in Philosophy of Science | Department of History and Philosophy of Science, NKUA, Athens, Greece, 03-05 December 2020[1]

# Can we literally talk about artificial moral agents?

**Alkis Gounaris**
National and Kapodistrian University of Athens, Department of Philosophy

**Abstract:** The current debate regarding the moral status of intelligent machines includes a wide range of arguments. In one direction, we tend to attribute anthropomorphic properties to machines (Nyholm, 2020), due to our interaction with them and possibly due to limitations or the metaphorical use of language. There are many who argue that what should actually concern us is how to make machines safer, not "ethical" (Yampolskiy, 2012). On the contrary, according to the other direction, it is sufficient to evaluate the results and consequences of the machines' actions, regardless of whether and how they think or operate (Dennet, 1997).

However, the main body of arguments considers the moral status of artificial agents subject to their degree of autonomy (Tzafestas, 2016) and consequently, to the behavior that machines display towards humans and other machines (Anderson, 2007). In this presentation, adopting the assumption that in the near future, fully autonomous artificial intelligence systems will coexist and interact with humans, animals and other systems in almost all aspects of personal and social life (Tegmark, 2017), I ask the question whether indeed: a) full autonomy in decision-making and b) these systems behaving in line with a framework of principles and rules, is enough to literally talk about artificial moral agents. Making the distinction between morality and ethics, I will argue that a fully autonomous machine demonstrating sincere, polite, honest, consistent, tolerant, protective, etc. behavior for example, or obeying the laws and respecting the current social and cultural conditions or rules of coexistence, is not sufficient to call it literally a moral machine.

Distinguishing between a quasi moral agent and a literally moral agent, I will attempt to describe those conditions beyond autonomy and behavior that must be met, in order to attribute the traits of a moral agent to an artificial intelligence system. Such a system, in addition to duties, could potentially have rights, obligations and responsibilities, coexisting with other intelligent beings or systems in a possibly revised form of social fabric. An investigation towards this direction can aim at highlighting information regarding the type, characteristics and "personality" of the moral agent and lay the theoretical foundations that could possibly lead to the description of the technical specifications required for its realization.

**Keywords:** Ethics of AI, Machine Ethics, Artificial Moral Agents, Philosophy of AI

---

[1] Gounaris, A. (2020). 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. Retrieved [25/12/2020] from https://alkisgounaris.gr/en/archives

# Μπορούμε να μιλάμε κυριολεκτικά για ηθικές μηχανές;

**Άλκης Γούναρης**
Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών, Τμήμα Φιλοσοφίας

**Abstract:** Η σύγχρονη συζήτηση σχετικά με το ηθικό status των έξυπνων μηχανών περιλαμβάνει ένα ευρύ φάσμα επιχειρημάτων. Σύμφωνα με τη μία κατεύθυνση, έχουμε την τάση να προσδίδουμε ανθρωπομορφικές ιδιότητες στις μηχανές (Nyholm, 2020), λόγω της αλληλεπίδρασής μας με αυτές και πιθανόν λόγω των λεξιλογικών περιορισμών ή της μεταφορικής χρήσης της γλώσσας. Πολλοί είναι αυτοί που υποστηρίζουν ότι αυτό που θα πρέπει να μας απασχολεί πραγματικά, είναι το πώς θα κάνουμε τις μηχανές πιο ασφαλείς και όχι πιο «ηθικές» (Yampolskiy, 2012). Αντίθετα, η άλλη κατεύθυνση αρκείται στην αξιολόγηση των αποτελεσμάτων και των συνεπειών της δράσης των μηχανών, ανεξάρτητα από το αν και πώς αυτές σκέφτονται ή λειτουργούν (Dennet, 1997). Ωστόσο το κυρίως σώμα των επιχειρημάτων συναρτά την ηθική των μηχανών με τον βαθμό πραγματικής αυτονομίας τους (Tzafestas, 2016) και συνεπεία αυτού, με τη συμπεριφορά που οι μηχανές επιδεικνύουν απέναντι στους ανθρώπους και στις άλλες μηχανές (Anderson, 2007). Στην παρουσίαση αυτή, υιοθετώντας την παραδοχή ότι στο άμεσο μέλλον, πλήρως αυτόνομα συστήματα τεχνητής νοημοσύνης θα συνυπάρχουν και θα αλληλεπιδρούν με ανθρώπους, ζώα και άλλα συστήματα σε σχεδόν όλες τις εκφάνσεις της προσωπικής και κοινωνικής ζωής (Tegmark, 2017), θέτω το ερώτημα αν όντως α) η πλήρης αυτονομία στη λήψη αποφάσεων και β) η ευθυγραμμισμένη με ένα πλαίσιο αρχών και κανόνων συμπεριφορά των συστημάτων αυτών, αρκεί για να μιλάμε κυριολεκτικά για ηθικές μηχανές. Κάνοντας τη διάκριση μεταξύ ηθικής και ηθικότητας, θα υποστηρίξω, ότι το να επιδεικνύει μια πλήρως αυτόνομη μηχανή ειλικρινή, ευγενική, έντιμη, συνεπή, ανεκτική, προστατευτική κτλ. συμπεριφορά για παράδειγμα, ή το να τηρεί τους νόμους και να σέβεται τις εκάστοτε κοινωνικές και πολιτισμικές συνθήκες ή τους κανόνες συμβίωσης, δεν την κάνει κυριολεκτικά ηθική μηχανή. Κάνοντας τον διαχωρισμό ανάμεσα σε quasi ηθικό πράκτορα και κυριολεκτικά ηθικό πράκτορα, θα επιχειρήσω να περιγράψω τις προϋποθέσεις εκείνες, πέρα από την αυτονομία και τη συμπεριφορά, οι οποίες θα πρέπει να πληρούνται ώστε να μπορούμε να αποδώσουμε χαρακτηριστικά ηθικού προσώπου σε ένα σύστημα τεχνητής νοημοσύνης. Ένα τέτοιο σύστημα δυνητικά, εκτός από καθήκοντα, θα μπορεί να έχει δικαιώματα, υποχρεώσεις και ευθύνες και να συνυπάρχει με τα άλλα νοήμονα όντα ή συστήματα σε μια αναθεωρημένη ενδεχομένως μορφή κοινωνικού ιστού. Μια διερεύνηση προς αυτή την κατεύθυνση μπορεί να στοχεύσει στην ανάδειξη στοιχείων για το είδος, τα

χαρακτηριστικά και την «προσωπικότητα» του ηθικού πράκτορα και να θέσει τις θεωρητικές βάσεις εκείνες που ενδεχομένως να οδηγήσουν στην περιγραφή των τεχνικών προδιαγραφών που απαιτούνται για την πραγμάτωσή του.

**Keywords:** Ethics of AI, Machine Ethics, Artificial Moral Agents, Philosophy of AI

❖

Σε αυτήν την παρουσίαση θέτω ένα ερώτημα που αφορά το ηθικό status ευφυών μηχανών, των οποίων η λειτουργία, τους επιτρέπει να λαμβάνουν αποφάσεις και να προβαίνουν σε πράξεις οι οποίες έχουν ηθική βαρύτητα.

Το ερώτημα είναι: Μπορούμε να μιλάμε κυριολεκτικά για ηθικές μηχανές;

Κάποιος που γνωρίζει λίγα πράγματα από ηθική φιλοσοφία θα μπορούσε να απαντήσει – και θα έβρισκε πολλούς να συμφωνούν μαζί του – ότι όχι, δεν μπορούμε να μιλάμε κυριολεκτικά για ηθικές μηχανές. Οι μηχανές είναι μηχανές και η ηθική είναι ηθική –και όπως πολύ σωστά μας διδάσκει ο Hume, η ηθική είναι ανθρώπινη υπόθεση. Και η συζήτηση θα τελείωνε εκεί.

Δεν είναι άλλωστε λίγοι αυτοί που επιχειρηματολογούν ότι λόγω της ολοένα αυξανόμενης αλληλεπίδρασής μας με τις μηχανές έχουμε την τάση να τους αποδίδουμε ανθρωπομορφικές ιδιότητες (Nyholm, 2020) και μιλώντας για ηθικές μηχανές κάνουμε απλώς μεταφορική χρήση της γλώσσας.

Άλλοι πάλι θεωρούν ότι το ερώτημα είναι εσφαλμένο εν τη διατυπώσει του και αυτό που θα πρέπει να μας απασχολεί πραγματικά, είναι όχι το αν μπορούμε να κάνουμε τις μηχανές κυριολεκτικά ηθικές, αλλά πώς μπορούμε να τις κάνουμε κυριολεκτικά ασφαλείς (Yampolskiy, 2012).

Παρά τις εκ πρώτης όψεως εύλογες αντιρρήσεις λοιπόν, εξακολουθώ να θέτω το ερώτημα και θα επιχειρήσω μάλιστα μια καταφατική απάντηση, δηλαδή ότι ναι –υπό προϋποθέσεις θα μπορούμε να μιλάμε κυριολεκτικά για ηθικές μηχανές.

Και θεωρώ ότι το να θέτει κανείς ξανά και ξανά αυτό το ερώτημα έχει μεγάλη σημασία – όχι μόνο σε στενό φιλοσοφικό πλαίσιο – κυρίως για τρεις λόγους:

1. Επειδή ήδη σήμερα αυτόνομα συστήματα τεχνητής νοημοσύνης συνυπάρχουν και αλληλεπιδρούν με ανθρώπους και η σχέση αυτή, ανθρώπων και μηχανών, θα γίνει πιο άμεση και πιο περίπλοκη τα επόμενα χρόνια, καθώς μηχανές θα λαμβάνουν αποφάσεις και θα προβαίνουν σε πράξεις όπου θα κρίνονται ζητήματα ζωής και θανάτου, όπως για παράδειγμα στις στρατιωτικές επιχειρήσεις (Tegmark, 2017).
2. Διότι ο όρος «ηθική» στον σχεδιασμό και στην κατασκευή ευφυών μηχανών, όπως χρησιμοποιείται σήμερα από τους τεχνικούς για να περιγράψει τις αντίστοιχες εφαρμογές και λειτουργίες των μηχανών, δεν αναφέρεται κυριολεκτικά στην ηθική και

3. Γιατί η προοπτική της ύπαρξης κυριολεκτικά ηθικών μηχανών, εκτός από φιλοσοφικές και τεχνολογικές, θα έχει κυρίως πολιτικές, κοινωνικές και νομικές προεκτάσεις. Αυτό ενδεχομένως να σημαίνει ότι τέτοιες ευφυείς μηχανές εκτός από την εκτέλεση των καθηκόντων και των εργασιών τους, θα μπορούν να έχουν δικαιώματα πολιτικά ή νομικά, να λαμβάνουν αξιώματα, να συμβάλλονται, να κατέχουν περιουσιακά στοιχεία, και παράλληλα μέσα από ένα πλέγμα υποχρεώσεων και ευθυνών, να συνυπάρχουν με τα άλλα νοήμονα όντα ή συστήματα σε μια αναθεωρημένη ενδεχομένως μορφή κοινωνικού ιστού.

Αυτό που αρχικά θα ήθελα να υπογραμμίσω σχετικά με τη μη κυριολεκτική χρήση του όρου ηθική μηχανή – που αφορά στο σύνολο των εφαρμογών όπου σήμερα αποδίδεται– είναι ότι το να επιδεικνύει μια μηχανή ειλικρινή, ευγενική, έντιμη, συνεπή, ανεκτική, προστατευτική κτλ. συμπεριφορά, ή το να τηρεί τους νόμους και να σέβεται τις εκάστοτε κοινωνικές και πολιτισμικές συνθήκες ή τους κανόνες συμβίωσης, ή το να ακολουθεί ένα ορισμένο πρωτόκολλο ή κάποιες αρχές σύμφωνα με τις οποίες έχει προγραμματιστεί, δεν την κάνει κυριολεκτικά ηθική μηχανή (Gounaris, 2019).

Είναι σαν να λέμε ότι ένας αυτόματος τηλεφωνητής που απαντάει όταν τον καλούμε είναι ευγενικός, ή ότι το σύστημα που υπολογίζει με ακρίβεια το τραπεζικό επιτόκιο του δανείου μας είναι έντιμο.

Εδώ θα μπορούσε να αντιτείνει κανείς ότι μια αυτόνομη μηχανή που λαμβάνει αποφάσεις βασισμένη σε ένα πλαίσιο αρχών και κανόνων αποτελεί ηθική μηχανή, ή ένα «διαφανές» σύστημα ΤΝ είναι ηθικότερο από ένα «μαύρο κουτί».

Το ότι μια συμπεριφορά ή το αποτέλεσμα της λειτουργίας μιας μηχανής μπορεί να αξιολογηθεί ηθικά, δεν σημαίνει απαραίτητα ότι έχουμε να κάνουμε κυριολεκτικά με έναν ηθικό πράκτορα. Οι μηχανές που έχουν προγραμματιστεί για να λειτουργούν σύμφωνα με κάποιους ηθικούς κανόνες είναι φορείς μιας έμμεσης ή υπαγορευμένης ηθικής (Moor, 2006) αφού λειτουργούν σε ένα οριοθετημένο από τον προγραμματιστή τους πλαίσιο αρχών. Θα μπορούσαμε να τις ονομάσουμε quasi ηθικές μηχανές και σίγουρα είναι χρήσιμες μηχανές – αλλά όχι κυριολεκτικά ηθικές.

Η πραγματική πρόκληση για τους ερευνητές και τους φιλοσόφους της ΤΝ είναι η προοπτική της δημιουργίας ηθικών μηχανών με την κυριολεκτική σημασία του όρου –δηλαδή μηχανών που θα εκδηλώνουν ηθική συμπεριφορά η οποία θα προέρχεται μεν από ένα σύνολο αρχών το οποίο θα καθοδηγεί τις αποφάσεις και τις πράξεις τους, οι αρχές όμως αυτές δεν θα είναι τοποθετημένες εκεί από κάποιον προγραμματιστή ή σχεδιαστή, αλλά οι ίδιες μόνες τους με τη γνώση και την υπολογιστική τους ικανότητα θα προβαίνουν σε αποφάσεις σχετικά με ηθικά διλήμματα.

Συνήθως η συζήτηση για το αν μια μηχανή μπορεί να θεωρηθεί ηθική ή όχι, περιστρέφεται γύρω από το ζήτημα της αυτονομίας (Tzafestas, 2016). Μεγαλύτερος βαθμός αυτονομίας ή πλήρης αυτονομία, συνεπάγεται δυνατότητα εκδήλωσης ηθικής συμπεριφοράς. Πρόκειται για μια συζήτηση που έχει καντιανή καταγωγή, υπό την έννοια ότι η αυτονομία αποτελεί προϋπόθεση για την καντιανή ελεύθερη βούληση, δηλαδή την καθαρή βούληση σύμφωνα με την οποία πράττει κανείς ηθικά και συνεπώς έχει την ευθύνη για τις πράξεις του.

Όμως όπως δείξαμε πρόσφατα με τον συνάδελφο και φίλο Γιώργο Κωστελέτο (Gounaris, Kosteletos, 2020) σε ένα paper που δημοσιεύθηκε πριν από λίγες μέρες, αφενός η χρήση του όρου «αυτονομία» στην ΤΝ μπορεί θα θεωρηθεί μεταφορική, δεδομένου ότι η «βούληση» μιας μηχανής είναι (με καντιανούς όρους) δεσμευμένη από κάποιον τελικό σκοπό (έχουμε να κάνουμε δηλαδή με μια υποθετική και όχι μια κατηγορική προσταγή),

αφετέρου –και κυρίως αυτό– η ίδια η έννοια της αυτονομίας είναι προβληματική και η επίκλησή της προϋποθέτει απαντήσεις σε μια σειρά από δυσεπίλυτα οντολογικά και επιστημολογικά ερωτήματα. Θα πρέπει δηλαδή να απαντήσουμε πρώτον στο ποια είναι τα κριτήρια εκείνα που καθιστούν μια οντότητα αυτόνομη και δεύτερον στο πώς μπορούμε εμείς ως παρατηρητές να γνωρίζουμε ότι η συμπεριφορά αυτής της οντότητας είναι πράγματι αυτόνομη.

Οι καθιερωμένες προσπάθειες θεμελίωσης των κριτηρίων που καθιστούν μια οντότητα αυτόνομη είναι ως επί τω πλείστον συναφειοκρατικές και τελικά όπως αποδεικνύεται αλυσιτελείς, καθώς προσφεύγουν στην επίκληση εσωτερικών καταστάσεων όπως η συνείδηση, τα συναισθήματα, οι πεποιθήσεις κτλ. –δηλαδή σε όρους, η χρήση των οποίων όπως γνωρίζουμε από τη φιλοσοφία του νου, γεννά περισσότερα προβλήματα απ’ όσα τελικά λύνει.

Η θέση μου είναι ότι η ιντερναλιστική αυτή οπτική είναι παρούσα σε όλες τις προσπάθειες θεώρησης μιας ηθικής μηχανής και αυτή η οπτική μας καταδικάζει να περιστρεφόμαστε συνεχώς γύρω από τα ίδια προβλήματα.

Και θα εξηγήσω τι εννοώ:

Ας υποθέσουμε ότι θέλαμε να σχεδιάσουμε ένα σύστημα, του οποίου οι αποφάσεις θα βασίζονται στην παραδοχή ενός ύψιστου αγαθού εκφρασμένο σε ένα σύστημα αρχών, τότε, ή

α) το σύστημα αυτό θα ήταν quasi ηθική οντότητα, καθώς θα λειτουργούσε με υπαγορευμένη ηθική επειδή κάποιος το προγραμμάτισε με αυτόν τον τρόπο ή και

β) αν ήταν πράγματι νοήμον σύστημα, θα έβρισκε από μόνο του μια ύψιστη αρχή, ως συμπέρασμα της παρατήρησης της ανθρώπινης συμπεριφοράς και θα υιοθετούσε το συμπέρασμα αυτό ως γεγονός της πραγματικότητας, αλλά τότε θα βρισκόταν αργά ή γρήγορα σε υπολογιστικό και λογικό αδιέξοδο, καθώς δεν θα μπορούσε να παράγει συμπεράσματα του ΠΡΕΠΕΙ από προτάσεις του ΕΙΝΑΙ -δηλαδή να τελεί υπό καθεστώς φυσιοκρατικής πλάνης.

Αν πάλι θέλαμε να δημιουργήσουμε έναν καντιανού τύπου ηθικό πράκτορα ο οποίος θα επέλεγε ελεύθερα – από μόνος του - τις αρχές του (επί τη βάσει της δυνάμει καθολικοποίησης των αρχών αυτών), θα έπρεπε να επιλύσουμε το πρόβλημα της εννοιολόγησης και της οντολογίας της καθαρής βούλησης και θα έπρεπε να ξεπεράσουμε τη δεσμευμένη από την υποθετική προστακτική συλλογιστική του.

Τέλος, αν αποφασίζαμε ότι το σύστημά μας θα λειτουργεί ωφελιμιστικά, θα έπρεπε πρώτον να ξεπεράσουμε το πρόβλημα των δεδομένων αρχών και της φυσιοκρατικής πλάνης που

είδαμε παραπάνω και επί πλέον θα έπρεπε να επιλύσουμε το πρόβλημα της νοηματοδότησης και συνεπακόλουθα το πρόβλημα του πλαισίου στη συγκέντρωση, επεξεργασία και αξιολόγηση των πληροφοριών εκείνων που απαιτούνται κάθε φορά για τον υπολογισμό του συνόλου των συνεπειών μιας πράξης. Εννοείται ότι αν τα κριτήρια αυτά ήταν υπαγορευμένα από τον προγραμματιστή της μηχανής, θα οδηγούμασταν πάλι σε μια quasi ηθική οντότητα.

Οι Andersons (Anderson, 2007) γνωρίζοντας το αδιέξοδο της ενδοσκοπικής εξέτασης του ηθικού πράκτορα, προτείνουν ένα είδος «Ηθικού Τεστ Turing», για να διακριβώσει κανείς αν μια μηχανή είναι όντως φορέας ηθικών αξιών και αν βάσει αυτών λαμβάνει αποφάσεις και προβαίνει σε πράξεις ηθικής βαρύτητας, ανεξάρτητα από το τι ακριβώς συμβαίνει εντός της, ανεξάρτητα δηλαδή από τον αλγόριθμο που την οδηγεί κάθε φορά σε μια ορισμένη απόφαση και συμπεριφορά.

Μετατοπίζουν δηλαδή το κέντρο βάρους σε εξτερναλιστικά κριτήρια προκειμένου κανείς να αποφανθεί αν κάποια μηχανή είναι κυριολεκτικά ή όχι ηθικός πράκτορας.

Μια εξτερναλιστική οπτική της ηθικής αποτίμησης σύμφωνα με τη θέση μου, θα μπορούσε να λειτουργήσει αντίστροφα τόσο στον σχεδιασμό όσο και στην κατασκευή ενός κυριολεκτικά ηθικού πράκτορα, απαλλάσσοντάς μας παράλληλα από τα ιντερναλιστικά αδιέξοδα.

Αντί δηλαδή να προσπαθούμε να απαντήσουμε στο ερώτημα για το πώς ένας πράκτορας πρέπει να λαμβάνει τις αποφάσεις του και να πράττει κυριολεκτικά ηθικά, μπορούμε να διερωτηθούμε αν υπάρχει κάποιος παράγοντας Χ που θα προσδίδει στον πράκτορα αναγνωρίσιμα χαρακτηριστικά ηθικού προσώπου (ή έστω ηθικού δράστη).

Και σε αυτό το ερώτημα η απάντηση είναι ήδη δοσμένη υπό το πρίσμα μιας ηθικής θεωρίας που δεν είναι προσανατολισμένη σε κάποιο ύψιστο αγαθό, ή στα κριτήρια βάσει των οποίων αξιολογείται κάθε απόφαση και πράξη ενός πράκτορα – αλλά στον συνολικό βίο του.

Για να γίνει κατανοητό αυτό θα πρέπει να πάμε πίσω στον Αριστοτέλη, σύμφωνα με τον οποίο το να είναι κάποιος ηθικό πρόσωπο δεν προϋποθέτει μια συνταγή, σύμφωνα με την οποία το πρόσωπο αυτό αποφασίζει στις διάφορες καταστάσεις. Το να είναι κάποιος ηθικό πρόσωπο σημαίνει να καλλιεργεί τις αρετές του και να αποκτά ενάρετο χαρακτήρα σε όλη τη διάρκεια της προσωπικής και κοινωνικής του ζωής.

Αν ονομάσουμε δηλαδή τον παράγοντα Χ = Αρετή και επιχειρήσουμε μια αριστοτελικής κατεύθυνσης επίλυση του προβλήματος των ηθικών μηχανών, τότε απεμπλέκουμε τη συζήτηση από τις ιντερναλιστικές ατραπούς και τη φέρνουμε έξω στην κοινωνική σφαίρα, εκεί όπου ούτως ή άλλως ανήκει η ηθική διερώτηση.

Μια τέτοια αντιστροφή, υποστηρίζω ότι μπορεί να μας υποδείξει εν τέλει και τα κριτήρια που θα καθορίσουν τη δημιουργία ενός κυριολεκτικά Ενάρετου Πράκτορα (Virtuous Agent).

Πρόκειται εν ολίγοις για την υιοθέτηση μιας οντολογίας βασισμένης στις εξής παραδοχές:

Ο ενάρετος πράκτορας θα έχει έναν κοινωνικό σκοπό και έναν εξατομικευμένο σκοπό και η τελολογική αυτή σκοποθεσία δεν θα αντιβαίνει την ηθική θεώρηση (επειδή παραβιάζει την ελεύθερη βούληση ή την κατηγορική προσταγή) – αντιθέτως θα την οριοθετεί και θα είναι συμβατή με την ίδια την ουσία των μηχανών.

Η επίτευξη του κοινωνικού σκοπού θα συνεπάγεται την πλήρωση του εξατομικευμένου σκοπού καθώς θα υπάρχει μια αναγκαστική εσωτερική σχέση μεταξύ των δύο, υπό την έννοια ότι δεν θα μπορεί να υπάρχει το Α χωρίς το Β και αντιστρόφως.

Για την επίτευξη των σκοπών αυτών, απαραίτητη προϋπόθεση αποτελεί η καλλιέργεια των αρετών. Όχι μόνο των υπολογιστικών ή χρηστικών δεξιοτήτων του πράκτορα, αλλά και της εκδήλωσης ηθικών αρετών – ή τουλάχιστον όπως ο παρατηρητής θα τις ερμηνεύει ως τέτοιες – όπως η πρακτική σοφία (φρόνηση), η δικαιοσύνη, η ωφέλεια και μη βλάβη, η πραότητα, η αξιοπιστία, η ειλικρίνεια, η ευγνωμοσύνη, ο σεβασμός της ελευθερίας, κ.α.

Η καλλιέργεια των αρετών αυτών γίνεται εντός του κοινωνικού ιστού -υπό την έννοια ότι δεν μπορεί να θεωρηθεί ένας ηθικός πράκτορας ξεκομμένος από τον κοινωνικό ρόλο του. Αυτός είναι ένας περιορισμός που συνεπάγεται τον οργανικό - πέραν του εργαλειακού - ρόλο του πράκτορα εντός ενός συμβιωτικού συστήματος.

Εντός του συμβιωτικού αυτού συστήματος επιτυγχάνεται πρακτική μηχανική μάθηση - κατακτάται η γνώση και δημιουργείται μια ανατροφοδοτική αντιστοιχία της ενάρετης συμπεριφοράς με τις κατά τα άλλα ασαφείς αξιολογικές έννοιες.

Μια τέτοια αρετολογική ηθική προσέγγιση υπερβαίνει τα ιντερναλιστικά οντολογικά και επιστημολογικά προβλήματα -είναι μια συμπεριφοριστική προσέγγιση και όχι μια λειτουργιστική θεώρηση και σε θεωρητικό επίπεδο αφήνει ανοιχτό το ενδεχόμενο να μπορούμε να μιλάμε κυριολεκτικά για ηθικές μηχανές.

Ηθικές μηχανές υπό την έννοια των ενάρετων πρακτόρων που μαθαίνουν να συμπεριφέρονται με τέτοιο τρόπο, ώστε το συμβιωτικό περιβάλλον τους να τις αναγνωρίζει ως τέτοιες και οι οποίες θα αξιολογούνται τελικά επί τη βάσει της επίτευξης του κοινωνικού και ατομικού τους σκοπού και του συνολικού βίου τους.

**Ενδεικτικές βιβλιογραφικές αναφορές:**

Anderson, M., Anderson, S. L. (2007). Machine Ethics: Creating an Ethical Intelligent Agent. *AI Magazine*. 28(4), 15

Dennett, D.C. (1997). When Hal Kills, Who's to Blame? Computer Ethics. In Stork, D. (ed.) *Hal's Legacy: 2001's Computer as Dream and Reality*. MIT Press. pp. 351-365

Gounaris, A. (2019). Who's to blame when HAL kills again? *Ithiki*. V.12. pp. 4-10

Gounaris, A., Kosteletos, G. (2020). Licensed to Kill: Autonomous Weapons as Persons and Moral Agents. In Prole, D. and Rujević, G. (ed.). *Personhood*. Novi Sad, Filozofski Fakultet &

The NKUA Applied Philosophy Research Lab Press. DOI: https://doi.org/10.12681/aprlp.49.916.

Moor, J. H. (2006). The Nature, Importance, and Difficulty of Machine Ethics. *IEEE Intelligent Systems* 21(4): 18–21.

Nyholm, S. (2020). *Humans and Robots: Ethics, Agency, and Anthropomorphism*. Rowman & Littlefield

Tegmark, M. (2017). *Life 3.0: Being Human in the Age of Artificial Intelligence*. Alfred A. Knopf

Tzafestas, S. (2016). *Roboethics*: A Navigating Overview. Springer

Yampolskiy, R.V. (2012). Artificial Intelligence Safety Engineering: Why Machine Ethics is a Wrong Approach. In Müller, V.C. (Ed.). *Philosophy and Theory of Artificial Intelligence*. Springer. pp. 389-396