Εργασία Project Ά Τετραμήνου(2017-2018):

**Μπορούν οι μηχανές να κατανοήσουν και να εφαρμόσουν την ανθρώπινη ηθική;**

Από τους μαθητές της ΄Β λυκείου:

Κωνσταντίνος Θωμόπουλος

Δανάη Κιλτσίκη

Εφραιμία Ρανάκη

Χαρά Χριστοδουλοπούλου

Περιεχόμενα

**Εισαγωγή**

**Κεφάλαιο Ά:** Ανθρώπινη ηθική-Βασικές ηθικές θεωρίες

**Κεφάλαιο ΄Β:** Μηχανές και Φιλοσοφία

**Κεφάλαιο ΄Γ:** ΤελικόΣυμπέρασμα-Απάντηση Ερωτήματος

Εισαγωγή

Τις προάλλες επισκέφθηκα έναν διαδικτυακό ιστότοπο, στον οποίο μπορούσα να επικοινωνήσω με ένα ρομπότ με το όνομα Evie. Αρχικά, συμπεριφέρθηκα με τρόπο με τον οποίο θα συμπεριφερόμουν σε έναν άγνωστο στο διαδίκτυο. Εδώ να σημειώσω πως μου επιτρεπόταν να μιλήσω μια φορά και να περιμένω απάντησή της. Παρόλο το προσεγμένο λεξιλόγιο που χρησιμοποιούσε, παρατήρησα πως οι απαντήσεις που μου έδινε, ήταν άσχετου περιεχομένου και δεν είχαν καμία σχέση με ότι την ρωτούσα. Κάποιες φορές, επαναλάμβανα τις ερωτήσεις μου ξανά και ξανά και οι απαντήσεις που δεχόμουν ήταν ακριβώς οι ίδιες. Στο σημείο αυτό να τονίσω πως ένα πραγματικός άνθρωπος θα συμπεριφερόταν διαφορετικά και ίσως να με αγνοούσε ή να σχολίαζε τη συμπεριφορά μου. Οι περισσότερες συζητήσεις που κάναμε με το ρομπότ δεν κατέληγαν σε κανένα συμπέρασμα. Τελικά από τη συζήτηση αυτή το συμπέρασμα στο οποίο κατέληξα εγώ προσωπικά είναι ότι τα ρομπότ δεν θα μπορούσαν ποτέ να αντικαταστήσουν τους ανθρώπους και ότι η τεχνητή νοημοσύνη διαφέρει σημαντικά από την ανθρώπινη. Αυτό το γεγονός μου δημιούργησε έντονο προβληματισμό για το μέλλον, όπου οι μηχανές θα αποτελούν αναπόσπαστο κομμάτι της ζωής μας.

Κεφάλαιο Ά

Ηθική είναι ένα σύνολο κανόνων μιας κοινωνίας, ένας κώδικας αξιών που μελετά και ενδεχομένως καθορίζει τη συμπεριφορά των μελών της κοινωνίας αυτής με κριτήριο τι είναι αποδεκτό και τι όχι. Σε μια απλούστερη προσέγγιση: τι είναι "καλό" και τι "κακό".

Είναι γνωστό οτι σε “ζητήματα ηθικής” δεν υπάρχει ομοφωνία. Διαφορετικές απόψεις μπορούν να διατυπωθούν και αντίστοιχα να τεκμηριωθούν. Οι ηθικές αξίες ενδέχεται να μεταβάλλονται στην πορεία του χρόνου. Συχνά αποτελούν αυτό που λέμε “άγραφο νόμο”.

***Α. Η ΘΕΩΡΙΑ ΤΟΥ ΩΦΕΛΙΜΙΣΜΟΥ (Ιδρυτές της οι Τζέρεμυ Μπένθαμ και Τζον Στιούαρτ Μιλ)***

Σύμφωνα με τη θεωρία αυτή, ηθικά ορθή χαρακτηρίζεται μια πράξη αν παράγει τη μεγαλύτερη δυνατή ωφέλεια για το μεγαλύτερο δυνατό αριθμό ανθρώπων. Άρα για να αποφασίσουμε αν θα κάνουμε μια πράξη, πρέπει να υπολογίσουμε τις συνέπειες που θα έχει η πράξη αυτή στην ανθρωπότητα.

***Β. Η ΘΕΩΡΙΑ ΤΗΣ ΚΑΤΗΓΟΡΙΚΗΣ ΠΡΟΣΤΑΓΗΣ (του Ιμμάνουελ Καντ)***

Σύμφωνα με τον Καντ, δεν πρέπει να αξιολογούμε μια πράξη ανάλογα με τα αποτελέσματά της (ωφελιμισμός), αλλά κρίνοντας τον τρόπο με τον οποίο σκέφτεται αυτός που εκτέλεσε την πράξη.

Κριτήριο για να θεωρηθεί η πράξη μας ορθή είναι να αναλογιστούμε, πριν την εκτελέσουμε, αν θα θέλαμε το κίνητρο και ο κανόνας της πράξης μας αυτής να ισχύει ως καθολικός νόμος.

Η θεωρία αυτή ονομάστηκε «κατηγορική προσταγή» επειδή έχει κατηγορικό, δηλαδή απόλυτο χαρακτήρα.

Το κριτήριο, σύμφωνα με τον Καντ δεν είναι οι συνέπειες μιας πράξης, αλλά το αν  ένας ορθολογικός άνθρωπος θα ήθελε να ζει σε μια τέτοια κοινωνία όπου καταργούνται σημαντικοί κοινωνικοί και οικονομικοί θεσμοί.

Ο ορθολογικός άνθρωπος είναι σε θέση να κατανοήσει ότι μόνο με την απαρέγκλιτη τήρηση του ηθικού νόμου είναι δυνατό να συμβιώσουμε αρμονικά και να εξασφαλίσουμε την ελευθερία μας, προστατευόμενοι από τα ένστικτα των άλλων.

***ΜΙΑ ΔΕΥΤΕΡΗ ΔΙΑΤΥΠΩΣΗ ΤΗΣ ΚΑΤΗΓΟΡΙΚΗΣ ΠΡΟΣΤΑΓΗΣ***

Πρέπει κάθε υποκείμενο να πράττει με τέτοιον τρόπο, ώστε να μεταχειρίζεται πάντα τους άλλους (όπως και τον εαυτό του) ως σκοπούς  και όχι ως μέσα των πράξεών του. Αυτό σημαίνει ότι πρέπει πάντα να σεβόμαστε την ελευθερία και την αξιοπρέπεια του άλλου. Όταν τους εξαπατούμε, τους φερόμαστε σαν όντα κατώτερα από εμάς.

***Γ. Η ΘΕΩΡΙΑ ΤΩΝ ΗΘΙΚΩΝ ΑΡΕΤΩΝ (Αριστοτέλης)***

Επειδή δεν είναι πάντα εύκολο να υπολογίσουμε τις συνέπειες της πράξης μας, ούτε να τηρούμε μια αφηρημένη ηθική αρχή που την ονομάζουμε ηθικό νόμο,  το καλύτερο είναι να προσπαθήσουμε να διαμορφώσουμε έναν τέτοιο ηθικό χαρακτήρα, έτσι ώστε να αποκτήσουμε την ικανότητα να δρούμε σωστά σε περιπτώσεις ηθικών διλημμάτων.

Σύμφωνα με τον Αριστοτέλη, με την κατάλληλη εκπαίδευση μπορεί να καλλιεργηθεί η ηθική αρετή, η οποία είναι μια μεσότητα και βρίσκεται ανάμεσα σε δύο άκρα: την υπερβολή και την έλλειψη. Την κατακτά εκείνος που έχει αναπτύξει επιπλέον τη διανοητική αρετή της φρόνησης, την ικανότητα να αναγνωρίζει το καλό για τον ίδιο και για τους συνανθρώπους του.

Για τον Αριστοτέλη η κατάκτηση της ηθικής αρετής οδηγεί στην ευδαιμονία, η οποία στην αριστοτελική θεωρία είναι η πλήρη άνθιση της ανθρώπινης προσωπικότητας, η πλήρης ανάπτυξη των ικανοτήτων του κάθε ανθρώπου και δεν έχει καμία σχέση με αυτό που αποκαλούμε ευτυχία, δηλαδή μια ψυχική κατάσταση ευφορίας.

Κεφάλαιο΄Β

**Δυϊσμός**

Η θέση του Δυϊσμού στη Φιλοσοφία του Νου αναγνωρίζει την ύπαρξη δυο διαφορετικών οντολογικών κατηγοριών: της νοητικής και της υλικής / φυσικής.

Μέσα στην ίδια τη σχολή του Δυϊσμού μπορούμε να διακρίνουμε ανάμεσα στο Δυϊσμό Υπόστασης και το Δυϊσμό Ιδιοτήτων.

 **Δυϊσμός Υπόστασης (Καρτεσιανός Δυϊσμός)**

Ο René Descartes (1596-1650) αναλύει τις δύο βασικές οντολογικές κατηγορίες, Νου και Ύλη (‘Σώμα’), ως ακολούθως:

Οι νόες είναι άυλες, μη εκτατές υποστάσεις, το ουσιώδες χαρακτηριστικό των οποίων είναι η σκέψη.

Το ουσιώδες χαρακτηριστικό των υλικών σωμάτων είναι ότι είναι εκτατά, δηλ. ότι καταλαμβάνουν χώρο.

Κάθε άνθρωπος είναι ένωση μιας νοητικής και μιας σωματικής υπόστασης, ένωση η οποία είναι ενδεχομενική, δηλ. που θα μπορούσε και να μην ισχύει (στην περίπτωση διαχωρισμού του νου από το σώμα). Κατά συνέπεια, η νοητική υπόσταση του ανθρώπου είναι αυτόνομη, φέρει δηλαδή τις νοητικές ιδιότητες ανεξάρτητα από την ύπαρξη ή όχι ύλης (σώματος).

 **Δυϊσμός Ιδιοτήτων**

Σύμφωνα με το Δυϊσμό Ιδιοτήτων, ο νους δεν αποτελεί ξεχωριστή υπόσταση. Οι μόνες υποστάσεις που υπάρχουν στον κόσμο είναι οι υλικές υποστάσεις, δηλ. τα βασικά σωματίδια που αναγνωρίζουν οι φυσικές επιστήμες, καθώς και οι συναθροίσεις αυτών. Δεν υπάρχουν νοητικές υποστάσεις όπως πνεύματα, ψυχές, ζωτικές δυνάμεις, κλπ.

Αυτό όμως είναι συμβατό με τη θέση ότι υπάρχουν γνήσια μη φυσικές ιδιότητες, όπως για παράδειγμα οι νοητικές. Η συνύπαρξη του δυϊσμού ιδιοτήτων με το μονισμό υπόστασης μπορεί να εξηγηθεί μέσω της σχέσης *ανάδυσης* που υποτίθεται ότι ισχύει μεταξύ νοητικών και φυσικών ιδιοτήτων.

.**Λογικός (Αναλυτικός) Συμπεριφορισμός**

Η φιλοσοφική θέση του Λογικού η Αναλυτικού Συμπεριφορισμού αφορά κατά κύριο λόγο στο σημασιολογικό περιεχόμενο των εκφράσεων που περιέχουν νοητικούς/ψυχολογικούς όρους. Συγκεκριμένα, αρνείται ότι οι ψυχολογικοί όροι και οι έννοιες που χρησιμοποιούνται στη Ψυχολογία του Κοινού Νου και στην επιστήμη της Ψυχολογίας (όπως π.χ. ‘επιθυμία’, ‘πεποίθηση’, ‘πόνος’, ‘θυμός’, ‘λύπη’, ‘αισθητηριακή εμπειρία’, κλπ) αναφέρονται σε εσωτερικές ψυχολογικές καταστάσεις, και τους αναλύει με όρους που αφορούν στην παρατηρήσιμη συμπεριφορά ενός ατόμου και στα εξωτερικά ερεθίσματα τα οποία δέχεται.

Με βάση το παραπάνω σκεπτικό, οι Λογικοί Συμπεριφοριστές προτείνουν την εννοιολογική ανάλυση της κάθε ψυχολογικής δήλωσης με βάση τα σχετικά φυσικά ερεθίσματα και την επακόλουθη συμπεριφορά.

**Θεωρία Ταυτότητας**

Η Θεωρία Ταυτότητας είναι η πιο απλή μορφή Φυσικαλισμού (Υλισμού). Αποτελεί μορφή *Φυσικαλισμού* διότι, σύμφωνα με αυτήν, ο νους δεν αποτελεί ξεχωριστό είδος υπόστασης, όπως υποστηρίζει ο δυϊστής υπόστασης, αλλά ούτε οι νοητικές ιδιότητες διαφέρουν από τις φυσικές, όπως υποστηρίζει ο δυϊστής ιδιοτήτων. Είναι η πιο *απλή* μοφή Φυσικαλισμού διότι, σύμφωνα με αυτήν, οι νοητικές καταστάσεις απλά ταυτίζονται με (δηλ. είναι) φυσικές καταστάσεις.

**Λειτουργισμός**

Η άνοδος του Λειτουργισμού στη Φιλοσοφία του Νου ακολούθησε την επανάσταση των ηλεκτρονικών υπολογιστών των δεκαετιών του 1950 και 1960. Ο Λειτουργισμός βασίζεται στην παρατήρηση ότι οι υπολογιστικές λειτουργίες μπορούν να οριστούν και να μελετηθούν ανεξάρτητα από τη φυσική σύσταση και δομή των μηχανημάτων που τις διεκπεραιώνουν. Για παράδειγμα, η ‘Αναλυτική Μηχανή’ που σχεδίασε ο Charles Babbage (1792-1871) και για την κατασκευή της οποίας προβλεπόταν η χρήση μεταλλικών κυλίνδρων, χάλκινων εργαλείων, μοχλών κλπ, θα μπορούσε να επιτελέσει μερικούς από τους υπολογισμούς τους οποίους επιτελεί ένας σύγχρονος υπολογιστής που είναι φτιαγμένος από τρανζίστορ από πυρίτιο και άλλα εξαρτήματα νεώτερης τεχνολογίας. Δηλαδή, οι υπολογιστικές λειτουργίες μηχανημάτων όπως τα ανωτέρω ορίζονται με αφαιρετικό τρόπο, ως λειτουργίες που συντελούνται σε ένα επίπεδο διαφορετικό από αυτό των υλικών υποστάσεων που τις πραγματώνουν. Οι υποστηρικτές του Λειτουργισμού αναλύουν τις νοητικές ιδιότητες με τρόπο ανάλογο αυτού με τον οποίο αναλύουμε τις ιδιότητες του λογισμικού ενός υπολογιστή. Όπως οι υπολογιστικές δράσεις ενός υπολογιστή πραγματώνονται στην υλική (φυσική) υπόσταση του, χωρίς όμως να ταυτίζονται με αυτήν, έτσι και οι νοητικές διαδικασίες μπορούν να εκληφθούν ως υπολογιστικές δράσεις που υλοποιούνται από φυσικές διαδικασίες (οι οποίες λαμβάνουν χώρα στις φυσικές δομές του εγκεφάλου), χωρίς όμως να μπορούν να ταυτιστούν με αυτές.

Κεφάλαιο Γ

**Τελικά μπορούν;**

 Έχοντας διαβάσει και λάβει υπόψη τα παραπάνω η απάντηση που δίνουμε στο αρχικό ερώτημα είναι ότι οι μηχανές δεν μπορούν να κατανοήσουν και να εφαρμόσουν την ανθρώπινη ηθική. Συνειδητοποιήσαμε ότι παρόλο που οι μηχανές έχουν τη δυνατότητα να εκτελέσουν πράξεις και να πάρουν αποφάσεις σύμφωνα με την προαίρεση του κατασκευαστή τους δεν έχουν τη δύναμη της αυτόνομης βούλησης και έτσι της αυτοσυνειδησίας. Δεν μπορούν να αυτοπρογραμματιστούν και να κρίνουν ή να ελέγξουν τη νοημοσύνη τους. Είναι μέσα που χρησιμοποιεί ο άνθρωπος, ο οποίος και τις κατασκευάζει, προκειμένου να εξυπηρετούν δικές του ανάγκες και συμφέροντα. Δεν μπορούν δηλαδή να αποκτήσουν ηθική κρίση. Οι μηχανές δεν μπορούν να βιώσουν τις ψυχικές καταστάσεις που μπορεί να βιώσει ο άνθρωπος ούτε έχουν τις αισθήσεις και την αντίληψή του και δεν έχουν δικές τους προσωπικές εμπειρίες. Επομένως δεν μπορούν να μπουν στη διαδικασία να λύσουν ηθικά διλήμματα, κάτι που εμείς ως άνθρωποι αντιμετωπίζουμε συνεχώς στη καθημερινότητά μας. Αν και οι μηχανές έχουν τη δυνατότητα να χειριστούν και να διασταυρώσουν πληθώρα πληροφοριών και να εκτελέσουν πολυσύνθετα υπολογιστικά προβλήματα σε μικρό χρονικό διάστημα εξαιτίας της τεχνητής τους νοημοσύνης δεν παρουσιάζουν τη δημιουργικότητα και την επινοητικότητα που χαρακτηρίζει εμάς τους ανθρώπους. Εμείς έχουμε επίγνωση των διανοητικών μας δυνατοτήτων για αυτό τον λόγο μπορούμε να δημιουργήσουμε τεχνολογικά μέσα που μπορούν να τη μιμηθούν, όχι ακόμη όμως να τη φτάσουν ή να την κατανοήσουν. Εξάλλου το ζήτημα της ηθικής δεν έχει να κάνει μόνο με τη λογική αλλά κυρίως με το συναίσθημα, την ελεύθερη βούληση και το σύστημα αξιών βάση των οποίων δομείται κατεξοχήν η ανθρώπινη κοινωνία.