**Първа тема. Изкуствения интелект като философски проблем**

***В тази тема ще се запознаете с възникването на философските проблеми свързани с изкуствения интелект от древността до наши дни. На кратко ще бъде пояснено какво е това етика и какви са основните концепции в нея и защо на изкуствения интелект тя му е необходима***

Да се достигне до същността на даден проблем е да се разгледат всички източници по хронологичен ред от човешката култура влияещи върху сегашните нагласи спрямо Изкуствения интелект. Човечеството не за първи път си задава въпроса какво ще бъде неговото битие с технологиите.

В този поток от информация могат да бъдат открити изключително екзотични идеи и коментари по отношение на ИИ, които присъстват в колективното ни съзнание и най-вече в Западното културно пространство оставайки релевантни и в наше време.

Идеята да се създават живи същества от неодушевена материя може да бъде намерена в различни източници на шумерски, китайски, еврейски език и е свързана с всички религиозни конфесии.

При древногръцката митология можем да видим прото-идея за създаване на изкуствени човешки същества и по-точно жени. В Илиадата Hephaestus помага на слугите да бъдат направени от злато и да наподобяват жени. В митът за Пигмалион скулпторът се влюбва в създадената от него от слонова кост женска фигура и моли богинята Афродита да и вдъхне живот. Тук можем да видим препратка към съвременните. сексуални роботи.

В своята книга „Автоматите“ гръцкия математик Hero от Алесандрия описва машина която срещу заплащане в храмоветемогат да видят годишната динамика на боговете като предсказание. През 1901 година подобна машина бива изложена като археологическа находка наречена Antikythera и е прототип на древен аналогов компютър задвижван от часовников механизъм.

В древно еврейските текстове се описва създаването на „Голем“ - чудовище направено от глина създадено от равин през шестнадесети век, върху което хората загубват контрол. Тук за първи път се сблъскваме с проблема за контрола (митът за Прометей се тълкува по същият начин давайки знак, че някой неща трябва да останат недостъпни за простосмъртните).

Апогеят на темата идва при Мери Шели и нейният роман за Франкенщайн, тъй като при Голем в края на историята хората успяват да го унищожат, докато тук това не се случва. Въпреки, че създаден през Романтичната епоха и носи възторга спрямо науката и машините той е ясно послание към съвремието и изкуствения интелект.

По съвременните версии на романи и електронни игри екстраполират различни варианти на негативен сценарии с бунтове на роботи и деструктивни атаки от страна на Изкуствения интелект спрямо хората. Най-ярка е филмовата версия на Терминатор от 2015 година. Този страх пред машините известния писател фантаст Айзък Азимов нарича „комплекс Франкенщайн“.

Страхът бива подклаждан и от изказванията на учени като Стивън Хоукинс който през 2017 г. казва, че създаването на изкуствения интелект ще е фатална стъпка за човечеството.

Съществува обаче и друга гледна точка наричана Трансхуманизъм залегнат в западната религия и философско мислене (учението на Платон и юдео-християнската традиция). Тук религията и технологията хронологично се движат в Западната култура.

Теистичната идея за превъзходството на Бог независимо от материалния и физически свят е негово вътрешно присъщо качество. Той е създател на всичко съществуващо и то е негова вътрешна същност. Разкриването му се получава през неговия син Христос и Светия Дух. Когато разглеждаме Франкенщайн нишката между създател и създадено (между Homo deus и изкуствения интелект) се къса. Подобна надлиминарност е характерна за философската традиция на Запада. През второто столетие след новата ера в Средиземноморието се заражда течение наричано гностицизъм търсещо божествената искра в човешкото тяло. Преди това Платон вижда човешкото тяло като затвор за душата която иначе е безсмъртна. В неговата метафизика се разграничава формата на нещата в променящия се свят и надхвърлящата ги същност. Подобно нещо има и в трансхуманизма кудето се надхвърля човешката ограниченост приемайки предизвикателствата на Платон и гностицизма се постига безсмъртие като биологичното тяло трябва да се надгради посредством някакъв изкуствен агент. Извличайки от нечистия материален свят с помощта на математиката и техническите науки на чисти форми на разума се реализира нова форма на платонизма придобиващ качеството на технология- алгоритмите на изкуствения интелект стават Платоническа машина извличаща формата (модела) от света на проявленията. В християнската традиция това скъсява дистанцията между Бог и хората правейки ги равни на него и доказвайки богоподобието.

Трансхуманизма загатва за възможността за безсмъртието на човека, но във същият момент се появява и антипода – идеята за Страшния съд и Апокалипсиса.

За да въведем някаква философска универсалност в разсъжденията приемаме, че:

* Живота се възприема като процес притежаващ способността да се самовъзпроизвежда при запазване на сложността преминавайки през три етапа – *биологичен живот,* където на езика на информатиката можем да кажем, че живите организми са хардуера, а софтуера се развива по еволюционен път; *културата като форма на живот*, където софтуерът може да бъде проектиран на основа обучение и *технологичен живот* където и софтуера и хардуера могат да бъдат проектирани и човек получава контрола върху собствената си съдба. *Цивилизацията* е взаимодействието между разумните форми на живот. *Съзнанието* това съвсем на кратко са личните преживявания, а *квалия* техния индивидуален характер. *Телеология* е начинът нещата да бъдат обяснявани през намерения и цели, а не през причини и следствия.
* Изкуствения интелект е получен по небиологичен начин и може да се раздели на „слаб(ограничен) интелект“ – възможностите му се ограничават до постигане на определени цели и задачи; „силен интелект“ – можещ да постигне практически всякакви цели включващи и самообучение; „универсален интелект“ – способен да стане силен интелект при наличието на данни и ресурси; „силен изкуствен интелект“ (на човешко ниво)AGI – можещ да се справя със всякакви познавателни операции не по-лошо от човека; „интелектуален взрив“ – самоусъвършенстване довеждащо до появата на „свръх разум“ наричано „Сингурлярност“ „свръхразум“ – значително превъзхождащ човешкия;
* Изкуствения интелект дава възможност така формулирания *технологичен живот* да бъде постигнат още през този век, но от нас зависи кога и как това може да се случи. Полемиката по този въпрос се води между три основни лагера: „лудети“ (техноскептици), „цифрови утописти“ и движение за „Приятелски Изкуствен Интелект“
* Според „лудитите“[[1]](#footnote-1) създаването на свръхчовешки универсален ИИ е толкова сложно, че за това са необходимо повече от едно столетие и съществува опасност той да ни унищожи в момента в който започне да усеща някаква ограниченост.
* Цифровите утописти приемат, че постигането на свръх ИИ е възможно още през този век и те го очакват като естествена и желана стъпка към развитие на космическата ни еволюция.
* Движението за приятелски ИИ също предполага появата на свръх разум още през този век, но привържениците на тази идея не го приемат като гарантирано положителен ход – силно се акцентира върху контрола и безопасността.
* Основните митове относно Изкуствения интелект са свързани с:

1. Изкуствения Интелект е неизбежен до 2100 година – фактите говорят, че това може да се случи и в близките десетилетия или може изобщо да не се случи и никой не знае точният отговор;
2. За това какъв, ще е ИИ се тревожат само лудитите – фактите показват, че загриженост проявяват голяма част от специалистите в областта;
3. Митологичен повод за тревога е злонамереността на ИИ и това, че той ще придобие съзнание – реалният повод за тревога е, че целите на ИИ могат да станат несъвместими с нашите човешки такива;
4. Главната опасност е в роботите – фактите показват, че ако целите ни не съвпадат, то на ИИ не му е необходимо никакво тяло, а само интернет връзка
5. Изкуствения интелект не може да подчини хората – факт е че както човекът може да подчини тигри и лъвове благодарение на разумът си, така и ИИ ще може да го направи;
6. Машините не могат да имат цели – фактите показват, че както ракетите с термо ориентиране имат цел така и при ИИ може да се появи;
7. Митологичен повод за тревога е появата на ИИ след няколко години – реалността показва, че това ще се случи най-рано след няколко десетилетия и това е достатъчно време да се гарантира неговата безопасност;

Съвременната философия подема темата за изкуствения интелект преодолявайки митологичното ниво ***с Тестът на Тюринг.*** Този философски казус е предложен от британския математик Алан Тюринг през 1950 г. в статията му „Изчислителни машини и разум“ (Computing Machinery and Intelligence), публикувана в списание „Mind“ (сп. „Съзнание/Ум“). Тестът е за проверка дали компютърът има разум в човешкия смисъл на думата. Тюринг предложил тест, който да замени безсмисления според него въпрос „Може ли машината да мисли?“ с по-определен. Съвременната интерпретация на този тест изглежда по следния начин: Човек взаимодейства дистанционно с двама събеседници – компютър и човек; на основание на отговорите на въпросите, тестовият субект е длъжен да определи с кого разговаря – с машина (изкуствен интелект/компютърна програма) или с жив човек; задачата на компютърната програма е да въведе човека в заблуда, карайки го по този начин да направи грешен избор.

До 2008 година няма информация някоя от съществуващите компютърни системи да е преминала теста.

Тестът е на основата на игра с името „Imitation game“ (игра на имитация).

Мъж и жена се намират в една стая и си разменят бележки с водещия, който е в друга стая. Задачата на водещия е да определи кой от двамата е жената. Задачата на мъжа е да обърка водещия. Задачата на жената е да помогне на водещия да направи правилен избор. Тюринг задава въпроса какво ще стане, ако мъжът се замени с машина. Ще може ли водещият да направи правилен избор и ще се поддаде ли на машината да го излъже? Отговорът на този въпрос показва може ли машината да се явява като мислеща.

Тестът е необходимо да протече по следния начин: съдия (човек) разменя съобщения на обикновен език (естествен, човешки език, напр. български или английски) с двама събеседници. Единият от тях е машина, а другият е човек. Предполага се, че всеки от събеседниците ще се стреми да бъде възприет като човек. За да се направи тестът универсален, е нужно да се разменят текстови съобщения (напр. посредством програма за чат разговори)[[2]](#footnote-2).

Някои оспорват идеята, че разумът е равнозначен на разумно поведение. Най-известният контрааргумент е мисловният експеримент на Джон Сърл (John Searle), наречен „Китайска стая“. Сърл говори за експеримент, при който човек, който не владее китайски, е затворен в стая. Отвън стои човек, който може през процеп за писма да пуска вътре бележки, написани на китайски. Човекът в стаята разполага с голям наръчник с подробни инструкции, които да му помогнат да отговаря на получаваните отвън бележки.

Сърл твърди, че дори човекът отвън да остане с впечатлението, че води разговор с някой, който владее езика, човекът в стаята не разбира китайски. Сърл добавя, че по същата логика дори ако една машина се държи разумно и премине теста на Тюринг, от това не следва, че тя е разумна или че има „разум“ в човешкия смисъл на думата. Същото е в сила, ако заменим думата „разумен“ със „съзнателен“.

**Етика – теоретични аспекти**

*Преди да заговорим за етични проблеми възникващи като последствие от цифровизацията е добре да поясним основните етични понятия и концепции имащи отношение към нея.*

Наименованието Етика е предложено от Аристотел с идеята тя да бъде науката за правилното поведение. В последствие се появяват нови и различни концепции, но етиката като цяло присъства в основите на много науки насочени в изучаването на живота и поведението на хората.

И така:

***Етиката е философска дисциплина изучаваща поведението на хората на основата на нравствени мотиви***.

Тя показва какви могат да бъдат последствията ако дадена конкретна идея бъде възприета от отделен човек или социум представяйки всички нейни предпоставки и перспективи. По този начин етиката може да бъде определена като рефлексия на морала на дадено общество. От другите средства за социална регулация като правото, традицията и обичаите, моралната норма се отличава с това, че винаги дава свобода на избор и преимуществено регулира вътрешните чувства на срам, дълг, угризения на съвестта. Нормите на морала съдържат идеята за изискване на дължимото и важното, апел към самоусъвършенстване и креативно преобразуване на действителността.

|  |  |
| --- | --- |
| **Обект на отражение** | **Категория** |
| Оценка на действителността от гледна точка на желателност (или нежелателност) за човека | Добро, зло, справедливост, щастие, смисъл на живота |
| Начини за подреждане на съвместния живот на хората | Норма, принцип, оценка, идеал |
| Индивидуални механизми на работа на нравственото самосъзнание | Дълг, съвест, срам, чест и достойнство, морални чувства |

Сложността на въвличането на предмета етика като наука за морала е свързана с това, че нравствените отношения включват широк кръг от обществени отношения. Явленията в обществения живот могат да бъдат класифицирани както следва:

* Морални (положителни спрямо морала);
* Аморални (отрицателни спрямо морала);
* Неморални;

Неморалните явления се определят трудно. Разглеждайки конкретно човешко действие включително действия които влияят на характера на производствените отношения ние откриваме, че те носят определена морална тежест. Примерно:

*Закриването на даден рудник е решение от производствен характер. Реално работодателя спестява средства от поддържането на нерентабилно производство, но в същия момент хиляди работници се оказват без препитание. Подобно дейсторална оценка и предизвиква етични въпроси.*

**Основни етични концепции:**

**Хедонизъм – признава наслаждението като основна цел на живота и висше благо.**

**Евдемонизъм – позиционира стремежа към щастие като критерий за нравственост и основен поведенчески модел.**

**Абсолютизма провъзгласява морала като основа на човешкото битие,а стремежа да бъдеш морален като смислообразуващ мотив.**

**Уталитаризъм – разглежда морала само като средство за постигане на щастието, а добродетелите като път към щастието.**

**Релативизъм – твърди, че оценката на ситуацията може да варира в зависимост от културата, историческия и личен контекст. Правилността на всяко нравствено твърдение подлежи на проверка през практиката**

**Формализъм – основен принцип на етическа теория в която формалнологичния момент превалира над анализа на съдържанието и социалната природа на морала. Характерен е за етиката на Кант, приемаща, че от някакво твърдение (категоричен императив) може да се изведат съдържателни морални принципи и решения приложими в различен контекст.**

Етиката в своето развити лежи върху нравствените нагласи които във времето са били многократно преразглеждани, а също така и на постиженията на другите науки като естествознание, биология и психология. Етическите концепции натрупани от човечеството са безкрайно разнообразни.

Говорейки за етическите принципи не може да не се обърне внимание на такава философска система като хуманизма. Тъкмо принципите на хуманизма са ръководили човечеството при съставянето на важни международни документи (примерно „Декларацията за правата на човека“[[3]](#footnote-3). Идеите на хуманизма са претърпели многократна трактовка в различните исторически периоди и въпреки това могат да бъдат фиксирани някой общи положения. Хуманизма предполага грижа за конкретния човек и стремеж на обществото да създаде условия за удовлетворяването на индивидуалните му потребности и лична самореализация. Би било погрешно, обаче хуманизма да бъдат разглеждан и сравняван като някакво съвършенно състояние на обществото. Хората се сблъскват с постоянно с проблеми от морален и практичен характер и това изисква внимание не само от страна на обществените структури, но и инидивидуални усилия.

**Основни принципи на хуманизма:**

* Гарантиране на основните човешки права като условие за истинско индивидуално съществуване;
* Поддържане на видимо слаби излизащи извън рамките на обичайните за даденото общество представи за справедливост;
* Създаване на социални и нравствени качества, позволяващи на личността да се самореализира, потребявайки обществените ценности.
* Съвременната етика е крайно приложна в различни сфери:
* *Етика на отворените морални проблеми* призвана да решава проблеми които до настоящия момент не са намерили своя отговор (примерно въпроса за евтаназията).
* *Приложна етика* занимаваща се с анализа на обществените заплахи (като екологични кризи, ядрена война и пр.) правейки социална експертиза и възможни последствия.

Приложната етика има следните области:

* *Професионална етика* използваща предимно диентологични принципи[[4]](#footnote-4) и другите елементи на съзнанието за ценност оформящи нашата действителност.
* *Практична етика* търсеща съответните процедури за решаване на етични въпроси
* *Корпоративна етика*  даваща възможност на фирмите да формулират етични формулировки за дейността си.

Трябва да отбележим, че съвременната етика в крайна сметка не се свежда до формулировката на приложната, защото съществуват и теоретични проблеми свързани със свободната воля, обосноваването на морала, различните етични концепции, въпросите свързани с историята на морала и характеристиките на моралния субект. Очевидно е, че в наше време се наблюдава усилено разширяване на приложната етика в търсене на решения за конкретни приложни проблеми.

Взаимодействието на човека с техниката е област на разглеждане от много време за философията, но сега се наблюдава появата на нови технологии и свързаните с тях проблеми. Всяко постижение на науката носи със себе си както огромни перспективи за обществото така и рискове. Разглеждането на етични въпросите свързани с цифровите технологии е фактор намаляващ потенциалните рискове и създаващ платформа за предначертаване на пътищата за развитието на технологията. В момента цифровия потенциал на обществото е свързан само с разпространението на информацията и използването на голяма база от данни. Развитието на технологиите включително и тези на Изкуствения интелект (от тук нататък маркирано като **ИИ**) създават все повече поводи за етическото им осмисля.

1. Лудитите са социално движение от XIX век на английските производители на текстил, които протестират, често унищожавайки механизирани тъкачни станове, обявявайки се срещу промените, предизвикани от промишлената революция. Смятат, че тези промени ги оставят без работа и променят начина им на живот. Движението е наречено на генерал Нед Лъд или на крал Лъд – митична фигура, за която се е смятало, че живее в Шеруудската гора, като Робин Худ. Днес „лудит“ се употребява за тези, които се противопоставят на индустриализацията, автоматизацията, компютризацията или на новите технологии изобщо [↑](#footnote-ref-1)
2. https://bg.wikipedia.org/wiki/%D0%A2%D0%B5%D1%81%D1%82\_%D0%BD%D0%B0\_%D0%A2%D1%8E%D1%80%D0%B8%D0%BD%D0%B3 [↑](#footnote-ref-2)
3. URL: https://www.un.org/ru/universal-declaration-human-rights/index.html [↑](#footnote-ref-3)
4. **Деонтология** – област на етиката изучаваща проблемите на морала, нравствеността и дълга . проекцията и’ се проявява в лекарската и юридическа диентология. [↑](#footnote-ref-4)