ЕТИКАТА, ИЗКУСТВЕНИЯТ ИНТЕЛЕКТ И ЕВОЛЮЦИЯТА НА ЧОВЕЧЕСТВОТО

ХРИСТО ХРИСТОВ

Институт за изследване на обществата и знанието, БАН

hristo2907@abv.bg

ETHICS, ARTIFICIAL INTELLIGENCE AND THE EVOLUTION OF HUMANITY

HRISTO HRISTOV

Institute for the Study of Societies and Knowledge, BAS

Abstract

The human intelligence usually has been presented as one of the greatest achievements of the evolution on Earth in last several millions of years. In this evolutionary dimension, on the one hand, the human intelligence appeared in the interaction with an environment of the insecure and uneasy world in the role of cognitive ability of foreseeing the scope of options and consequences of human behavior. In this sense the human intelligence appeared also as ethical cognizance of the effects of our actions on the well-being of others. Today, on the other hand, researchers as Nick Bostrom warn us for the possible arising of the artificial suprahuman intelligence as result of the achievements of contemporary and future scientific explorations in realms of machine learning, neuroscience, and psychology. Therefore, the main questions in that paper are following: Is it possible for us to think of artificial moral intelligence as counterpart of and complement to human ethical agency? May we hope for near arising of such an artificial moral intelligence in the light of the present scientific researches in realms of machine learning, neuroscience, and psychology? What we humans might know further about our evolutionary genesis as far as we would be cooperating with other moral agents that may be completely artificial and unhuman?

Keywords: Machine Learning, AI, Evolution of Humanity, Artificial Moral Agents

Човешкият интелект обикновено бива представян като едно от най-големите достижения на земната еволюция през последните няколко милиона години, макар това да звучи твърде неприемливо за личности като Тимон Атински. За хората се знае, че те са обществени и политически същества, базиращи своето битие на форма на колективна рационалност, която можем да определим като култура. Културата се различава от

натурата, защото предполага процеси, които излизат далеч извън рамките на роденото непосредствено тук. Културата е автономен и целеполагащ процес на взаимодействие с това, което редица изследователи мислят като "несигурен и динамичен свят" (Louie et al., 2012: 13–32). В еволюционен план интелектът се появява именно в такава среда, защото последната предполага голямата вероятност за възникването на горепосоченото взаимодействие. Несигурният природен свят налага на всеки организъм нуждата да се постига оптималното поведение в името на оцеляването. Оптималното поведение предполага от своя страна сложния процес на научаване кое е ценното в нашия живот, кое е най-доброто за нас през времето, което ни е отредено да изживеем. Търсенето на възможни отговори на този така фундаментален въпрос може да бъде наблюдавано в поведението на всички живи организми на Земята по един или друг начин. Общото между всички тях обаче е стремежът да се постигнат максимални ползи на най-ниски цени, а това става на бойното поле на ежедневното оцеляване. За лястовицата най-доброто място за отглеждане на ново поколение се оказва старата стреха на студентско общежитие. Последното е построено, за да даде подслон на неколцина млади люде, които са на път да спечелят първа награда на поредния младежки форум на науката, безспорно, място, където се срещат сякаш безсмъртните богове. Лястовицата строи гнездо за малки лястовичета, които трябва бързо да бъдат отгледани и научени да летят на големи разстояния, за да се избавят от зимата. Човекът строи дом за студенти, които едва са напуснали семейното гнездо, за да поемат по свой собствен път на жизнено развитие. В двата случая наблюдаваме процес, който можем да определим като ценностно ориентиран. Младите потомци на лястовиците и хората се научават сами да оцеляват в живота, ползвайки се от опита и знанията на по-старите членове на обществото, в което живеят. Този процес се характеризира с едно постепенно изграждане и развитие на знанието за полезното в живота на взаимодействие с природния свят (Louie et al., 2012: 14). По такъв начин живите организми са способни в по-малка или по-голяма степен да бъдат автономни деятели в своя жизнен свят. Те могат да избират едно или друго действие, да се ръководят от полезността, получена от неговите последици, доколкото същата тази полезност е запаметена, и да се учат в един по-широк смисъл на базата на опита си. Живите организми усвояват чрез учене кои и какви са ценните неща в техния жизнен свят, взаимодействайки непрекъснато с него.

Същите организми съхраняват в паметта си и постоянно актуализират това познание през целия си живот. Тук най-вече става дума за познание, което дава възможност за оценка на полезността в дадена ситуация, която е налагала правенето на избор между един или друг начин на действие. Естеството на това познание се оказва първостепенен предмет в изследванията на учени, които са посветили своите усилия на работа в междинната област между психологията, невронауката и икономиката (Louie et al., 2012: 13). Трябва да си дадем сметка за високата степен на комплицираност, заложена в основите на това интердисциплинарно поле. От една страна, на лице е значимостта на традиционните нормативни теории за избора, които основно можем да опишем като задаващи нормативни модели за начина, по който хората трябва да направят своя избор според очакваната пределна полезност на последиците от него. Поради това тези нормативни модели най-общо може да бъдат определени като модели на очакваната полезност (Johnson et al., 2014: 36). От друга страна, днес ставаме свидетели на развитието на изследвания в невронауката, които имат за цел хвърлянето на светлина относно проблема за възпроизвеждането на очакваната полезност като кодирана в активността на невроните и мрежите, които ги свързват. Същата тази информация в нормативните теории за очакваната полезност е представена на абстрактно теоретично равнище (Louie et al., 2012: 13). Трето, психологията на свой ред ни подсказва, че това кодиране на информацията за очакваната полезност, при правенето на избор, в различни части от човешкия мозък, процес, който може да се проследи посредством наблюдение на активността на редицата мозъчни дялове и да се картографира чрез технологии като функционалната магнитно-резонансна томография, е кодиране, което се намира в пряка зависимост от контекста, свързан с предишния опит на постигната полезност и настоящия набор опции. Нашият мозък по различен начин кодира това, което в езика си понятизираме като стойност, в различните свои сензорни дялове, като по този начин ни предлага една своеобразна рамка за разбирането на този процес. Тя може да се окаже сериозно оръжие в ръцете на специалистите по маркетинг, например (Louie et al., 2012: ibid.).

Изложеното по-горе принципно трябва да ни провокира да се замислим сериозно в една или друга насока. Ценни указания по този въпрос според мен ни предлага известният шведски философ и основател на Института за бъдещето на човечеството в

Оксфордския университет Ник Бостръм (Никлас Бострьом). Всичките тези сложни процеси на възпроизвеждане и кодиране на информация за очакваната полезност, произлизаща от нашите постъпки, съобразно сложността както на нервната ни система, така и на системата от символи, която най-общо включва в себе си историята на човешките общества, езици и култури, могат да бъдат включени в рамките на това, което определяме като еволюция. Понятието за еволюция ни принуждава да мислим, че е нужно да се приеме за вярно твърдението, че развитието на живота на Земята в геологически и чисто исторически план по необходимост свидетелства за развитие, започващо от фазата на възможно най-прости организми и достигащо фаза на съществуването на сложни организми, най-вече тези, които притежават съзнание, език и разум като хората в един късен етап от планетарната ни история (Bostrom, 2016: 212-216). Тук е мястото, настоява Бостръм, да си дадем сметка за съществуването на един голям предразсъдък в нашето мислене, който се свързва с просвещенския идеал за прогреса в човешката история (Bostrom, 2016: 213). Нека си представим за момент картина, която заимства в една или друга степен художествената изразност на Платон. Представяме си идеалния наблюдател, една ангелическа интелигенция, например, която от вечността съзерцава развитието на живота в нашите земни селения. Пред взора на тази интелигенция по необходимост се разгръща картината на зараждане и гибел, но винаги така, че продължават напред онези, които живеят, ползвайки се в най-голяма степен от разум, сложно познание, съзнание и живот, подчинен на благото на общностен целеполагащ живот. По този начин тази интелигенция ще наблюдава свидетелство за присъствието на един божествен принцип на прогреса в иначе безрадостното и жалко земно съществуване, подчинено на преходност и гибел. Тази интелигенция би трябвало да забележи, че на планетата Земя по необходимост се появяват други разумни същества, които удивително я наподобяват, поели по своя път към съзнанието и познаването на вечните космически принципи. Земната еволюция по този начин в очите на подобна вечно съзерцаваща интелигенция би трябвало да е подчинена на разума и благото, водеща справедливо до съществуването на общности на разумни и освободени в голяма степен от плена на природната необходимост същества, които, бидейки такива, по необходимост трябва да представляват венеца на същата тази земна еволюция, а и на еволюцията във всеки възможен свят във вселената,

където е възможно да се зароди живот. Ако се доверим на един подобен мит, а такъв под една или друга форма стои все още в основите на западното разбиране за природата на човешкия интелект, ние успешно бихме наподобили оптимизма на Кандид. Това е така, доколкото си представяме еволюцията като прогрес, който по необходимост е довел до, или би трябвало да доведе, до съществуването на разумна форма на живот, която е, или наподобява човешката. Историята на човешките общества, особено от началото на индустриалната революция насетне, също спомага за затвърждаването на подобна нагласа, защото културният и технологичен прогрес сякаш ни внушава, че хората са се превърнали в качествено нов по-висш вид в сравнение с това, което са били в зората на каменната епоха (Bostrom, 2016: 213). Все пак, както отбелязва и самият Бострьм, тази твърде оптимистична еволюционна визия по никакъв начин не може да бъде подкрепена от редица факти, пред които се изправяме, когато решаваме да оценим същата в строго нормативен план. Еволюцията на планетата Земя със сигурност представлява развитие, което обаче не е в никакъв случай единствено с положителни черти. Напротив, хората са живели, живеят и по всяка вероятност завинаги ще продължат да живеят в свят на голямо страдание. Страдание тук трябва да означава страдание в чисто природен детерминистичен план, т.е. по природа в живота си се сблъскваме със страдание, доколкото живеем в относителна несигурност относно нашето благосъстояние и това на най-близките ни. Страдание също така обаче трябва да означава и страдание в социокултурен аспект, т.е. символният свят на човешките култури и общества генерира страдание, което не е природно детерминирано, а по-скоро в голяма степен може да бъде ситуирано в областта на чистия практически разум, т.е. то е последица от един акт на чиста мисловна спонтанност, когато извършителят свободно е избрал точно това поведение. Тези два основни смисъла на понятието за страдание може да ни подскажат твърде много за приемливостта на гореизложеното оптимистично еволюционно гледище за прогреса на разумността, манифестиращ в човешката история. Еволюцията е наложила най-вероятно необходимостта от появата на разумен биологичен вид като човешкия на планетата Земя, но тази необходимост трудно може да се мисли като отражение на някакъв Божествен план, който я оправдава като справедлива и добра в строг етически план. Напротив, тази необходимост ни подсказва, че всеки напредък по пътя към изграждането на високоразвит обществен

живот на рационално целеполагане се заплаща с цената на страдание във физически и душевен смисъл. Историята на това страдание е история на взаимодействието между човешките общества и природната среда, в която те са потопени, както и на взаимодействието между отделните човешки общества и между отделните обществени групи и формации. Всичко това дава основание и на философи като Ник Бостръм да твърдят, че достиженията на еволюцията не са предмет на уважение в етически план, а по-скоро са ценени чисто естетически в най-добрия случай (Bostrom, 2016: 213). Поради тази причина не сме в състояние да мислим нашия свят като постоянно прогресиращ в етически смисъл, доколкото същият е сцената, на която се появява развитието на сложни разумни човешки общества, които се радват на етически оправдано благосъстояние. Поне не сме в състояние да сторим това с помощта на едно оптимистично еволюционно гледище за зараждането и прогреса на разумността на планетата Земя, каквото в едри щрихи очертах по-горе.

Човешкият интелект като еволюционен продукт най-вероятно не е нито нещо уникално, нито нещо недостижимо, но най-вече, същият сам по себе си не е и нещо добро в етически смисъл, доколкото в голяма степен неговото развитие се е дължало на принципи, които може да противоречат на това, което намираме в една или друга от познатите ни днес базови етически секуларни системи в съвременния свят като консеквенциалистката или деонтологическата да речем. Хората развиват своя интелект, например, за да увеличават материалните ресурси, с които да разполагат в бъдеще, като това по необходимост става за сметка на другите. Това неравенство е подкрепяно в името на общата стабилност на социалния живот. Силният лидер властва, но по необходимост защитава онези, които подкрепят властта и богатството му. Това може да го прави филантроп в очите на другите днес, или цар с божествен дар в миналото, но той не е добър a priori, следвайки алгоритмичната динамика на очакваната полезност, базирана на чисто разсъдъчни правила, защото същата може да изисква лишаване от живота и свободата на неговите опоненти, например, което само по себе си не е етически оправдан акт. Разумът в качеството си на мощно оръжие за кодиране на най-прекия път за постигане на максималната полезност не е тъждествен на етически ценното, мислено често като подчинение на принципа на самопожертването в името на моралния дълг в общочовешки смисъл. Етическото често поради това се мисли като

нямащо голяма стойност, защото не е в състояние да се изправи пред категоричността на природния свят на невроните и техните алгоритми на очакваната полезност, които са довели до появата на човешкия интелект в света най-общо.

Интелектът в своето еволюционно развитие сам по себе си може и да не се придържа към етическите ценности, които през вековете са се появили в рамките на модерните човешки общества. Трябва ли тогава да мислим съществуването на радикално зло в сферата на интелектуалното a priori? Трябва ли в частност да мислим човешкия интелект като сравнително предразположен към етически осъдими актове дотолкова, доколкото това е една логика на развитие, заложена в еволюцията на човешкия вид? Тези въпроси може да бъдат определени най-вече като изграждащи ядрото на етическия дебат в интердисциплинарното поле, в което работят учени от сферите на когнитивната психология, невронауката, етиката, еволюционната биология и пр. през последните няколко десетилетия. Стивън Пинкър, известен отдавна не само в академичните среди, но и сред значително по-широката публика, канадско-американски когнитивен психолог, предлага отговори на тези въпроси, които са приемани противоречиво, но аз се спирам на тях, защото мисля, че те може да ни предложат известна яснота, най-малкото защото са базирани на дългогодишни, грижливо провеждани изследвания. През 1997 г. е публикувана една негова книга, която впоследствие ще бъде причината той да е номиниран за наградата "Пулицър". Става дума за "Как умът работи?". В рамките на тази монография Пинкър убедително изгражда своята аргументация по въпросите, които тревожно свързват темите за нашите гени, злото, еволюцията и моралната отговорност в човешкия избор (Pinker, 1997: 53). Интелигентните същества, като тук освен хората можем да мислим за всяко разумно същество, както и за изкуствения интелект, за който предполагаме, че е възможно да се появи в бъдещето, по-близко или по-далечно, имат своето поведение, обусловено от определени причини. Последните според Пинкър по необходимост ни въвеждат в рамките на дебата за свободната воля и отговорността (Pinker, 1997: ibid.). Важното в случая за Пинкър е наличието на необходимостта да се прави строго разграничение между понятието за обяснение на поведението, продиктувано от един разум, и моралната оценка на същото това поведение като оправдано. Поради това и трябва да се прави строго разграничение между световете на съвременната наука и етиката, настоява

Пинкър (Pinker, 1997: 55). Науката може да предостави единствено обяснения как дадено поведение на дадена разумна същност възниква, като се базира на описанието на природни и културни закономерности и процеси. Човешкото поведение, например, може да бъде обяснено като продукт от взаимодействието между нашите гени, анатомията на мозъка ни, неговото биохимично състояние, семейното ни възпитание, социалните ни роли и редицата стимули, които имат своето решаващо влияние в един или друг момент (Pinker, 1997: 53). Днешната наука, за Пинкър, в общи линии третира хората предимно като материални същности, чието поведение има своите обяснения в областите на естествения подбор на гените и неврофизиологичното функциониране (Pinker, 1997: 55). Статията на Кенуей Луи, цитирана по-горе, ни подсказва твърде ясно за тази парадигма. Етиката, от друга страна, според Пинкър определя хората като имащи еднаква ценност, надарени с чувства, разум и свободна воля същества, чието поведение може да бъде оценявано на базата на морални правила, които почиват на определено разбиране за естеството на моралното поведение и неговите последици (Pinker, 1997: ibid.). По този начин излиза, че законодателството на една морална максима за разумното поведение представлява идеализация, която за Пинкър се равнява на идеализацията в Евклидовата геометрия, която почива на предпоставката, че има идеални окръжности и безкрайни прави в пространството (Pinker, 1997: ibid.). В природния свят не срещаме нито идеални окръжности, нито изрично доказателство за свободната воля, но допускането в идеален план на съществуването и на двете е целесъобразно, защото светът се доближава достатъчно много до тези състояния така, че това да ни върши работа и ние да можем да прилагаме на практика както геометрията на Евклид, така и началата на моралната философия в познавателния си опит, доколкото разполагаме с обосновани и полезни заключения както за физическото пространство, така и за човешкото поведение, мислено в рамките на нравственото (Pinker, 1997: ibid.).

Човешкият чист разум, изглежда, е законодателстващ единствено в пределите на своя собствен опит на подреждане на феномените, свързани с историята на земното му развитие и битие, по силата на категории и постулати, които той намира единствено като мислени от самия него. Ще можем ли да мислим по същия начин за разума, който би принадлежал на едни изкуствени, или машинни, морални деятели? Терминът “изкуствен морален деятел” е предложен на широката публика от изследователите

Уендъл Валах и Колин Алън в тяхната монография за моралните машини (Wallach et al., 2009: 4). Флориди и Сандърс (Floridi & Sanders, 2004: 1) предлагат още през 2004 г. свое изследване относно понятието за нравственост на изкуствените деятели, каквито срещаме в киберпространството и на много други места, където все по-често същности, които в близкото минало сме считали за обикновени машини, сега са включени като деятели в известен смисъл в редица ситуации, налагащи взимането на морални решения. Тези изкуствени машинни дигитални същности може според Флориди и Сандърс да бъдат разглеждани като притежаващи в една или друга степен морална природа в две измерения. Първо, те може да претърпяват извършени върху тях действия, които са характеризирани като добри или лоши. Второ, те самите са морални агенти в една ограничена степен, доколкото са в състояние да извършват действия, които също така са определяни като добри или лоши (Floridi & Sanders, 2004: ibid.). Валах и Алън са повлияни от тази аргументативна линия, която извежда на преден план понятието за изкуствен морален агент. За яснота те се спират на някои по-конкретни характеристики, които са разглеждани по-рано от Флориди и Сандърс, като интерактивност, автономия и адаптивност (Wallach et al., 2009: 60). Изкуствените агенти, които се доближават до идеята за притежаване на нравствена природа, най-вече са интерактивни, т.е. отговарят на даден стимул, сменяйки своето състояние и взаимодействайки със средата си. Второ, те са автономни, т.е. могат да променят своето състояние без наличието на стимул в тяхната среда, което предполага независимост от същата. Трето, тези изкуствени агенти притежават способността да се учат как сами да променят своя начин на функциониране и взаимодействие с околната среда, почивайки на наученото от опита си, като последното представлява същностен елемент в идеята за машинното алгоритмично обучение. Последното е причина мнозина да хранят надеждата, че изкуствените агенти може да бъдат изградени като развити морални деятели (Wallach et al., 2009: 60-61). Валах и Алън мислят понятието за развит морален деятел най-вече като включващо представата за деятел, който е в състояние да различава различни перспективи на своята деятелност, които предполагат различно подреждане на предпочитанията, неговите и тези на останалите деятели, включени в ситуацията. Тази ситуация мислим като морална, т.е. ситуация, която предполага съзнание за намиращи се в конфликт цели, следвани от отделните морални деятели,

както и съзнание за възможните лоши последици от постъпките на всеки един деятел върху благосъстоянието на останалите (Wallach et al., 2009: 61). Моралният деятел според Валах и Алън притежава развита когнитивна способност да предвижда обхвата на своето етическо действие през призмата на редица възможности за избор и последици от взетите решения. Тази способност, твърдят те с право, е развита от естествения човешки интелект в еволюционен план, като тук трябва да мислим и за безкрайното множество от социални и междуличностни взаимодействия. Етиката е онази област на свободата, настояват те (Wallach et al., 2009: ibid.), която предполага свободен избор между различни модели на поведение и различни степенувания на предпочитанията относно техните крайни резултати, които, естествено, не са неутрални, т.е. те засягат по различен начин благосъстоянието на участниците в дадена ситуация. Поради тази причина етическият дизайн трябва, както настояват и цитираните по-горе изследователи, да предполага идеята за подобна етическа свобода в рамките на все по-динамично развиващата се област на машинното обучение. Това би ни дало възможността да мислим, че не само земната, продължила еони, еволюция е спомогнала да се появи интелект с нравствена природа, каквато ни е позната днес, но и че е възможно съществуването на изкуствени морални деятели в бъдещето, които да разширят общността на автономните морални създания в нашия свят по начин, който ще помогне на самото човечество да преосмисли по по-добър начин своя произход и хода на еволюционното си развитие.

ЛИТЕРАТУРА

Bostrom, N. (2016). *Superintelligence. Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Floridi, L. & J. Sanders. (2004).*MindsandMachines*.14:349.https://doi.org/10.1023/B:MIND.0000035461.63578.9d.

Johnson, E. J. et al. (2014). *Computational and Process Models of Decision Making in Psychology and Behavioral Economics*. In: Neuroeconomics. Decision Making and the Brain. Eds. Paul W. Glimcher and Ernst Fehr. Amsterdam: Elsevier, Ch. 3.

Louie, K. et al. (2012). Efficient coding and the neural representation of value. *Annals of New York Academy of Sciences* 1251, 13–32.

Pinker, S. (1997). *How the Mind Works?* New York: W. W. Norton and Company.

Wallach, W. et al. (2009). *Moral Machines. Teaching Robots Right from Wrong*. Oxford: Oxford University Press.