Ана ФРИЦХАНД
Билјана БЛАЖЕВСКА-СТОИЛКОВСКА

УДК: 17.021:159.947.2]:004.8
Изворен научен труд

ИНТЕГРИРАЊЕ НА МОРАЛНОСТА ВО ИНТЕЛИГЕНТНИТЕ МАШИНИ – МОЖЕ ЛИ ВЕШТАЧКАТА ИНТЕЛИГЕНЦИЈА САМОСТОЈНО ДА НОСИ МОРАЛНИ ОДЛУКИ?

***Кратка содржина**:*

Со експанзијата на вештачката интелигенција и напредните технологии, светот во 21 век бргу се менува и наметнува нова динамика на живеење. Иако ваквите промени ги засегаат сите возрасти, помладите генерации побрзо ги прифаќаат и реагираат попозитивно на нив. Новите кохорти - генерација Z и алфа - живеат во дигитален свет кој влијае врз нивниот начин на живот, интерперсоналните релации, квалитетот на менталното здравје, психолошката благосостојба и секојдневните предизвици со кои се соочуваат. Присуството на таканаречениот „Франкенштајн ефект“ кај некои возрасни, предизвикан од брзиот развој на вештачката интелигенција и роботиката, ја одразува позицијата „луѓето против машините“, сфаќајќи ја вештачката интелигенција како закана за човештвото. Сепак, реалноста е дека дигиталниот и човековиот свет не се во конфликт, бидејќи многумина веќе секојдневно ги користат алатките на вештачката интелигенција. Таа е присутна во одредени аспекти на медицината, образованието, бизнисот, правото, земјоделието, индустријата, вселенската технологија и во многу други области. Имајќи го ова предвид, доменот на моралот се наметнува како многу важен. Едно често поставувано прашање е: Дали вештачката интелигенција има капацитет самостојно да донесува морални одлуки? Затоа, интегрирањето на моралот во алгоритмите на вештачката интелигенција е еден од приоритетите на кои интензивно работат интердисциплинарни тимови од областа на инженерството и компјутерските науки, психологијата, филозофијата, социологијата, правото итн. Овој труд го разгледува ова прашање преку презентирање на наоди од неодамнешни релевантни истражувања кои ги дискутираат предизвиците и можностите за интегрирање на димензиите на моралот и воведувањето на човечките вредности во автономните системи кои извршуваат сложени задачи.

***Клучни зборови**: моралност, морално расудување, морално одлучување, вештачка интелигенција.*

Вовед

Вештачката интелигенција (ВИ) денес е длабоко навлезена во различни аспекти на човековото секојдневие. Со нејзина помош автономните системи (како што се, на пример, автономните возила или роботите асистенти) многукратно влијаат врз општествениот развој. Тие извршуваат сложени задачи, а некои од нив влегуваат и во интеракција со луѓето. Нејзината способност брзу да учи и да користи огромни бази на податоци ѝ овозможува, на пример, експедитивно и ефикасно да ги процесира природните јазици, да анализира податоци и да биде ефикасна во симулациите, во пишувањето финансиски извештаи, изведбата на репетитивни задачи во подолг временски период (значајни во индустриските процеси), препознавањето на шеми и детали (особено важни во медицинската дијагностика), симултаното извршување на повеќе задачи (доколку е тренирана за тоа) и во многу други аспекти.

Ваквата експанзија и моќ на ВИ некому делува застрашувачки, поттикнувајќи го т.н. Франкенштајн ефект – страв дека човековите креации ќе се свртат против и ќе го уништат човештвото, доколку науката падне во раце на несовесни поединци и биде злоупотребена. Една од најчестите грижи поврзани со развојот на ВИ денес, е дека нејзиното засилено користење ќе доведе до затворање на многу работни места, а со самото тоа и до губење на потребата од одредени професии. Така, на пример, Kelly (2024)[1] во текстот објавен на 28-ми февруари годинава во списанието Форбс, пишува дека работните места поврзани со анализа на податоци, пишување финансиски извештаи, рутински закажувања (состаноци, лекарски прегледи итн.), потоа работните места кои бараат меморирање напамет, основни услуги кон клиенти, како и репетитивните административни работи, се меѓу најзасегнатите од растечкиот тренд на користење на генеративната ВИ, затоа што се мошне подложни на автоматизација. Од друга страна, професионалните улоги кои вклучуваат значителна социјална и емоционална компонента (на пример, психотерапијата, психолошкото советување, носењето сложени бизнис-одлуки, социјалната работа, наставата, работата во секторот на продажба каде вработените влегуваат во директни интеракции со купувачите, понатаму менаџерите, адвокатите итн.), се помалку засегнати заради нужноста на човечкиот фактор во овие процеси.

Hatzius et al. (2023)[2] во текстот објавен на 26-ти март минатата година во *Economic Research* на инвестиционата банка Голдман Саш,

[1] https://www.forbes.com/sites/jackkelly/2024/02/28/what-white-collar-jobs-are-safe-from-ai-and-which-professions-are-most-at-risk/

[2] https://www.key4biz.it/wp-content/uploads/2023/03/Global-Economics-Analyst_-The-Potentially-Large-Effects-of-Artificial-Intelligence-on-Economic-Growth-Briggs_Kodnani.pdf

користејќи податоци за работните места во САД и во Европа, посочуваат дека во 2023 година околу две третини од постоечките работни места се изложени на одреден степен на автоматизација со ВИ и дека генеративната ВИ може да замени околу ¼ од нив. Со екстраполација на проценките, авторите прогнозираат дека на глобален план ВИ би можела да подложи на автоматизација околу 300 милиони работни места. Дополнително, анализата на *Resume Builder*[3] од 8-ми ноември 2023, во кои биле интервјуирани 750 бизнис-лидери на компании кои веќе користат или во 2024 година планираат да користат ВИ, покажува дека 53 % од компаниите користат ВИ, а 24 % планираат да почнат со користење на ВИ во 2024 година. Воедно, 37 % од компаниите кои користат ВИ изјавуваат дека таа заменила дел од работните места во 2023 година. Во 2024 година, 44 % од компаниите кои користат или планираат да почнат да користат ВИ изјавуваат дека, дефинитивно или веројатно, дел од вработените ќе бидат отпуштени поради користењето на ВИ.

Сепак, ВИ не е омнипотентна. Таа има низа ограничувања. Познато е, на пример, дека кај неа не постојат свест и совест. Вештачката интелигенција, исто така, не поседува теорија на ум, ниту емоционална интелигенција. Кога станува збор за моралот и ВИ, најновите истражувања покажуваат дека таа сè уште не може самостојно да донесува морални одлуки. Нема ниту способност за адаптирање на сосема нови ситуации кои бараат флексибилност каква што ја има човековиот ум, а за што ВИ не е тренирана. Слаба страна ѝ се вистинската креативност и иновативност каква што поседуваат луѓето, која подразбира размислување надвор од рамката на постоечките податоци итн. Оттука, човекот и покрај рапидно брзиот развој на ВИ, сè уште останува супериорен во поглед на емоционалната длабочина (вклучително и моралните чувства), интуицијата, способноста за морално расудување и морално одлучување, креативното решавање на проблемите и поседувањето богато животно искуство кое му овозможува да го разбира и динамично да се прилагодува на променливиот контекст во кој живее.

Вештачката интелигенција и формите во кои постои

Кога се зборува за вештачката интелигенција од аспект на способноста да учи и да го применува наученото, според експертскиот тим на IBM[4], најчесто се реферира на три типа: слаба (тесна, ограничена), силна (генерална) и вештачка суперинтелигенција (за сега, само во теорија). Првиот тип се среќава, на пример, кај автономните возила, софтверите за препознавање слики, или пак кај виртуелните асистенти. Процесирањето

[3] https://www.resumebuilder.com/1-in-3-companies-will-replace-employees-with-ai-in-2024/

[4] https://www.ibm.com/think/topics/artificial-intelligence-types

на природните јазици, е исто така еден ваков вид ВИ. Терминот „слаба“ укажува на тоа дека нејзините можности се ограничени на извршување специфични активности или команди. Со развојот на големите јазични модели (Large Language Models – LLMs), отворено е ново поглавје во напредувањето на генеративната ВИ и нејзината способност да креира различни содржини слични на оние од реалниот свет. Иако како концепт е присутна уште од 50-тите години на минатиот век, усовршувањето на ВИ забрзува по 2012-та година со пробивот на вештачките невронски мрежи (најчестиот тип на генеративна ВИ), кои им овозможуваат на интелигентните машини забрзано да учат симулирајќи го начинот на кој мозокот ги процесира информациите[5]. Најновите податоци кои ги соопштуваат од McKinsey & Company (2024)[6], покажуваат дека во 2024 година генеративната ВИ сè повеќе се применува од страна на организациите (односно 65 % од вклучените во анкетата редовно ја користат во своето работење). Денес веќе, како што истакнуваат Bonnefon, Rahwa & Shariff (2024), напредокот на ВИ се мери во месеци, а не во години (стр. 654).

Вториот тип - генералната или силна ВИ - е сè уште теоретски систем во развој, кој би можел да учи без надзор и да извршува широк спектар на мултифункционални активности парирајќи им во способностите на луѓето, што најверојатно би овозможило да биде првата ВИ што успешно би го поминала Туринговиот тест[7]. Експертите се со поделени мислења околу брзината со која овој тип на ВИ би можел да се достигне. Според Berruti, Nel & Whiteman (2020), некои автори сметаат дека постои можност ваквиот тип на ВИ да се развие до 2030-та или 2040-та година, додека пак други се на ставот дека човештвото не е ни блиску до негово развивање (на пример, Rodney Brooks – поранешен професор на MIT и коосновач на iRobot - смета дека генералната ВИ нема да пристигне пред 2300-та година).

Конечно, вештачката суперинтелигенција е веројатно тоа што влева најмногу страв и недоверба кај луѓето, придонесувајќи кон т.н. Франкенштајн ефект. Имено, овој супериорен тип ВИ (доколку некогаш се достигне), ќе учи толку брзу и толку самостојно, што ќе ги надмине когнитивните капацитети и способностите на човекот, станувајќи многукратно помоќен, а со тоа и неконтролабилен. Тој ќе може автономно да мисли, расудува, донесува одлуки итн. Се претпоставува дека овој тип на ВИ ќе биде клучен дел од целосно самосвесната ВИ и од останатите видови автономни хуманоидни роботи, станувајќи најмоќната форма на интелигенција на планетава[8]. Сепак, сето ова е сè уште на ниво на шпекулација.

---

[5] Исто.

[6] https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

[7] https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-artificial-general-intelligence-agi#/

[8] https://www.ibm.com/think/topics/artificial-intelligence-types

Гледано од аспект на нејзината функционалност, ВИ се јавува во четири форми[9]: реактивни машини, ограничена меморија, ВИ со теорија на ум и самосвесна ВИ. Така, на пример, кога играме шах со суперкомпјутер или кога алгоритмите на социјалните мрежи ни предлагаат што следно да гледаме врз основа на нашата историја на прегледувања, станува збор за реактивни машини кои се системи на ВИ дизајнирани за извршување на специфични задачи. Вештачката интелигенција со ограничена меморија, за разлика од реактивните машини кои немаат меморија, може да користи минати и актуелни податоци со цел да даде насоки како да се достигне посакуваниот исход. Вештачката интелигенција која поседува теорија на ум, а уште повеќе самосвесната ВИ, денес се сè уште нереализирани форми. Доколку ВИ поседува теорија на ум, таа би можела да ги разбере мислите и чувствата на луѓето, како и да ги предвидува постапките врз основа на познатите преференции на другиот ентитет. Оттука, би можела да симулира релации слични на оние помеѓу луѓето. Вградувањето на емоции во ВИ со теорија на ум е следен чекор на кој се насочени истражувачите, надевајќи се дека таквата ВИ ќе може да ги разбере и соодветно да одговори на различните емоционални состојби кај луѓето, анализирајќи ја бојата на гласот, фацијалната експресија и другите релевантни податоци. Најдалечната и сè уште недостижна форма на ВИ – самосвесната ВИ – карактеристична за суперинтелигенцијата, покрај мислите и емоциите на луѓето би можела да ги разбира и сопствените внатрешни состојби и црти, па дури и самата би поседувала сопствени мисли, емоции, потреби и верувања. Денес, ВИ сè уште нема капацитет да ги разбере, ниту да одговори соодветно на човековите емоционални состојби и чувства.[10]

Вештачката интелигенција и социјалната роботика

Вештачката интелигенција и роботиката - како гранка на инженерските и компјутерските науки каде машините се програмирани да извршуваат многу специјализирани задачи - се две сосема различни дисциплини. Постојат роботи со ВИ, но и такви кои воошто не вклучуваат ВИ (кои во моментов се поброjни). Се претпоставува дека во иднина, како што ВИ ќе станува сè пософистицирана, сè почести ќе бидат дизајните каде таа ќе биде интегрирана во роботите. Денес, примената на ВИ во роботиката е видлива во индустриското производство, различните видови бизниси, медицината, земјоделието, вселенската технологија, воената индустрија итн. (Martin, 2021)[11]. Доменот во кој овие две дисциплини исто

[9] Исто.

[10] https://www.ibm.com/think/topics/artificial-intelligence-types

[11] https://aibusiness.com/verticals/robotics-and-artificial-intelligence-t he-role-of-ai-in-robots

така се испреплетуваат, е оној на социјалната роботика каде се истражува интеракцијата човек - робот (Human-Robot Interaction). Социјалните роботи се всушност автономни системи со вештачка интелигенција кои се дизајнирани да влегуваат во интеракција со луѓето и со други роботи. Според дизајнот може да наликуваат на луѓе, животни или на потполно нови суштества; можат да зборуваат и да слушаат, да одат на две или на четири нозе (или воопшто да не се движат); да комуницираат на различни начини со луѓето – преку светлина, звук или движења. Главната цел на нивното креирање е да бидат лични асистенти, тутори, негователи, рецепционери итн. (Clark & Fisher, 2023). Имајќи ја предвид растечката вклученост на роботите и автономните системи во секојдневниот живот, станува сè покруцијална потребата да се разбере и оптимизира оваа интеракција. Оттука, во последните децении експанзијата на ваквите истражувања е придружена со дизајнирање на социјални роботи кои се мошне умешни во вклучувањето во заеднички активности со луѓето (Belhassein et al., 2022).

Зголемувањето на соработката помеѓу луѓето и роботите е актуелен предизвик кој го истражуваат интердисциплинарни тимови научници. Во таа насока, психологијата на роботиката има за цел да го премости јазот помеѓу луѓето и роботите, овозможувајќи увид во специфичностите на оваа интеракција (Stock & Nguyen, 2019). Еден иновативен интердисциплинарен научен приод помеѓу областите на роботиката (социјална), социјалната психологија (поконкрено, просоцијалното однесување) и психологијата на моралот, ја реконструира традиционалната динамика во интеракцијата човек - робот и наместо да се насочи само кон тоа како роботот му асистира на човекот, ги истражува начините на кои роботите бараат помош од корисниците и од пасивните набљудувачи. Ваквото менување на парадигмата отвора нови можности за зголемување на перформансата на роботите и на соработката помеѓу нив и луѓето, обезбедувајќи двонасочен процес на учење.

Морално расудување и морално одлучување при различни морални дилеми: улогата на невронауката

Психологијата на моралот е сложена и интердисциплинарна област која наликува на сложувалка со многу парчиња, каде некои допрва треба да се откријат. Таа е поврзана со невронауката, економијата на однесувањето, социологијата, антропологијата, правото и други научни дисциплини (Bloom, 2023). Мултифакторијалната условеност на моралното однесување е аспект кој дополнително придонесува оваа област да биде една од најтешките за истражување. Просоцијалното однесување, на пример, според Decety & Steinbeis (2020) вклучува голем број механизми засновани врз различна мотивација и однесувања кои за социјалната кохезија и соработка претставуваат значајни адаптивни елементи. Во психологијата

на моралот јазот помеѓу моралното знаење и моралното однесување не може едноставно да се премости фокусирајќи се само врз еден домен на моралноста (пр. моралното расудување, моралните чувства, моралната мотивација и слично).

Како што уште на почетокот на 21-виот век истакнале Green & Haidt (2002), истражувањата од психологијата и од когнитивната невронаука покажуваат дека токму моралното расудување е честопати прашање на афективни и интуитивни одговори, а не резултат на продлабочено и намерно расудување. Оттука, сознанијата од когнитивната и афективната невронаука се мошне корисни, бидејќи даваат значајни информации за функционирањето на мозокот во услови кога поединецот е соочен со ситуации кои се различни од аспект на нивната морална релевантност. Така, на пример, денес веќе со сигурност се знае дека воопшто не е сеедно дали поединецот се соочува со лична или нелична, конкретна или апстрактна (хипотетска) морална дилема, или пак морално неутрална дилема, затоа што при решавање на различни видови дилеми се активираат различни мозочни региони.

Кога се зборува за моралното расудување треба да се има предвид дека во него и емоциите играат многу значајна улога, како што е погоре споменато. Покрај тоа, човечкиот ум има изразена способност да ги решава проблемите автоматски и несвесно (што важи и за оние проблеми кои потекнуваат од сложениот социјален контекст). Ова е посебно нагласено во приодот на социјалниот интуиционизам кој ги обединува сознанијата од истражувањата на автоматското реагирање, невронауката и еволуциската психологија. Според овој модел, моралните судови се носат слично како и естетските, односно поединецот веќе при иницијалното соочување со одредена морална дилема интимно во себе чувствува дека однесувањето е правилно или погрешно, прифатливо или неприфатливо. Ваквите чувства (кои можат да бидат позитивни или негативни) се јавуваат во свесноста одеднаш и без когнитивен напор, при што поединецот многупати дури и не може јасно да го аргументира својот морален суд – едноставно на длабоко емотивен план „знае“ дека нешто е морално или неморално. Нив, авторот на овој модел – Џонатан Хајд - ги нарекува интуиции, а истите се создадени низ природната селекција и силите на културата (Green & Haidt, 2002).

Истражувањата од невронауката покажуваат дека кога се донесува морална одлука потребна е интеракција меѓу неколку одделни, но поврзани мозочни региони. Така, на пример, во две експериментални студии на Green et al. (2001) користена е батерија од 60 практични дилеми поделени во две категории – морални и неморални. Дополнително, моралните дилеми биле поделени во две групи – лични (емоционално поинтензивни) и нелични (емоционално неинтензивни). Очекувано, наодите покажале дека личните морални дилеми претежно ангажирале региони од мозокот поврзани со емоциите (меѓу кои MFG, PCG и ANG сите билатерално),

додека неличните морални дилеми предизвикале зголемена активност во регионите поврзани со когнитивната контрола и работната меморија (посебно во DLPFC). Понатаму, истражувањето на Green et al. (2004) покажало дека мозочните региони поврзани со апстрактното мислење и со когнитивната контрола (пред сè rDLPFC и ACC), се активираат кога треба да се разреши сложена (лична) морална дилема во која поединецот носејќи утилитарен суд, истовремено треба да прекрши некои свои лични вредности и принципи. Воедно, делови од фронталниот и од париеталниот кортекс биле повеќе активни при утилитарните судови, предвидувајќи разлики во однесувањето на поединецот врз основа на донесениот морален суд. Од друга страна, структурите на MPFC биле одговорни за поинтуитивни емоционални реакции.

Според Green (2023), постојат силни докази дека моралните дилеми поттикнуваат натпреварувачки одговори поддржани од различни когнитивни системи, при што едниот одговор се опишува како повеќе емоционален, а другиот како повеќе рационален. Од позиција на теоријата на дуални процеси на моралното расудување, типичниот деонтолошки суд на веројатно најчесто наведуваниот пример за лична морална дилема – дилемата на пешачкиот мост, каде испитаникот треба да одговори дали е прифатливо да се турне прилично крупен човек од мостот за да се сопре вагонот кој се движи по шините под него и на тој начин да се спасат петмина работници на пругата – дека такво нешто е потполно неприфатливо, е поддржан од интуитивна негативна емоционална реакција при самата помисла на таквата постапка. Од друга страна, типично утилитаристичкиот одговор – прифатливо е да се жртвува еден, за да се спасат петмина – е поддржан од рационална проценка врз база на трошоци–и–добивки, која испитаниците многу добро ја разбираат.

Сепак, смета Green (2023), прашањето дали деонтолошкиот одговор е брз, а утилитаристичкиот бавен се дискутира од различни автори, бидејќи се зголемува бројот на истражувачки наоди според кои утилитаристичкиот одговор и не е така бавен, иако постои голем корпус на истражувачки податоци кои покажуваат дека е побавен. Податоците на низа студии јасно ја потврдуваат теоријата на дуално процесирање на моралното расудување. Така, на пример, повеќе истражувања (пр. Ciaramelli et al. 2007; Koenigs et al., 2007; Moretto et al., 2010; Thomas et al., 2011; според Shenhav & Green, 2014) покажуваат дека пациенти со оштетувања во VMPFC се склони почесто да носат утилитаристички судови. Оние, пак, со оштетување во пределот на хипокампусот значително почесто донесуваат деонтолошки судови (McCormick, Rosenthal, Miller, & Maguire, 2016; според Green, 2023), исто како и пациентите кај кои е оштетена базолатералната амигдала важна во целно насоченото носење одлуки (van Honk et al., 2022; според Green, 2023). Ваквиот ефект овде, како и во случајот со пациентите со оштетувања на хипокампусот, е поврзан со доминирачките интуитивни емоционални реакции кои ги придружуваат деонтолошките судови. Оттука, според

Green (2023), двата вида судови – деонтолошки и утилитаристички – се водени од различни процеси. Воедно, бихевиоралните податоци упатуваат на тоа дека ниту еден од овие процеси не е побрз од другиот. Но, останува и силниот впечаток дека деонтолошките судови се сепак поинтуитивни и проследени со посилни чувства на правилно или погрешно, во споредба со утилитаристичките, кои се видливо порационални.

Истражувањата, исто така, покажуваат дека постојат потенцијални разлики помеѓу тоа што некој расудува дека е морално правилно и како тој некој реално се однесува кога треба да избере помеѓу алтернативните постапки (т.е. да донесе морална одлука). Во таа насока, на пример, одат резултатите од истражувањето на Tassy et al. (2013), кои упатуваат на заклучокот дека во основата на моралното расудување и на носењето морална одлука, најверојатно се наоѓаат различни психолошки процеси. На вкупно 240 испитаници поделени во 8 групи во споменатото истражување им биле зададени 15 морални дилеми и 9 морално неутрални, контролни дилеми. За неутралните дилеми, голем дел од испитаниците дале соодветен одговор што, според авторите, укажува на нивната способност да носат соодветни одлуки. Сепак, кај моралните дилеми се забележало дека генерално одлуките за постапките биле поутилитарни отколку судовите, што во превод значи дека испитаниците биле склони да прифатат да постапат на начин за кој претходно донеле суд дека е морално неприфатлив.

Добиените резултати покажале и дека кога бројот на спасени животи бил голем, поголем број испитаници имале тенденција да даваат утилитаристички судови и да избираат такво однесување. Притоа, веројатноста за давање утилитаристички одговори била константно повисока за изборот на постапка (морално одлучување), отколку за моралното расудување. Конечно, одлуките на испитаниците биле помалку утилитарни кога потенцијалната жртва била некој кој им е близок. Ова било најдено и за моралното расудување и за моралното одлучување, но сепак регистрираниот ефект бил значително поголем за изборот на постапките, отколку за моралниот суд. Притоа, веројатноста за давање утилитаристички одговори била поголема за избор на постапка, отколку за расудување во случај кога афективната близина со потенцијалната жртва била мала, но ситуацијата била сосема спротивна кога афективната близина со потенцијалната жртва била голема (лична морална дилема со висок степен на конфликт).

Психологија на моралот на вештачката интелигенција – може ли ВИ самостојно да носи морални одлуки?

Bonnefon, Rahwan & Shariff (2024), сметаат дека со развојот на ВИ интелигентните машини се појавуваат како уште една категорија со која психологијата на моралот треба да се справи. Обидот да се интегрираат

човечките вредности во интелигентните машини со цел да носат издржани морални судови и на нив соодветни морални одлуки, е сè уште неуспешен. Во оваа смисла, веројатно во најнапредна фаза се истражувањата на автономните возила и можноста да се програмираат во насока да донесат одлука дали при видлива пречка на патот, ќе скршнат кон група (од петмина, на пример) со цел да ја избегнат препреката и да го спасат животот на возачот и ќе ги усмртат или ќе скршнат кон sид (повторно за да ја избегнат препреката) и ќе го усмртат возачот. Истражувањата на Bonnefon et al. (2016; според Green, 2016) покажуваат дека кога ја решаваат оваа дилема испитаниците се генерално согласни дека е подобро да се спасат пет животи, жртвувајќи еден (што е, патем, класичен утилитаристички суд). Но, што доколку испитаникот е возачот кој се наоѓа во автономното возило? Во тој случај испитаниците ги менуваат судовите и не би сакале да се возат во таков автомобил. Со други зборови, не им е прифатливо да загинат за да се спасат петмина други животи.

Оттука, јасно се гледа сложеноста и важноста од вклучување на моралноста во инженерството. Програмерите ќе се соочат со многу голем предизвик, како што истакнува Green (2016), да ја интегрираат моралноста во алгоритмите со што би ги програмирале автономните возила да бидат доблесни и праведни, земајќи ги предвид човековите права и вредности. Според авторот, такво нешто би било можно кога би постоеле доволно прецизни морални теории врз основа на кои би се изработиле стриктни критериуми и протоколи врз основа на кои би можело со сигурност да се каже кои точно доблести овие интелигентни машини треба да ги поседуваат и следат, кои точно човекови права треба да ги земаат предвид при носењето морална одлука (притоа, како правата се подредени по приоритет и дали воопшто се подредени по приоритет), но и како да прават праведни компромиси.

McKendrick & Thurai (2022)[12] во написот „ВИ не е подготвена да носи несупервизирани одлуки“, објавен во Harvard Business Review, пишуваат дека ВИ е дизајнирана да помогне во донесувањето одлуки кога вклучените податоци, параметри и варијабли се надвор од човечкото разбирање. Сепак, авторите нагласуваат дека таа не успева да ги долови или да одговори на нематеријалните човечки фактори кои влегуваат во донесувањето одлуки во реалниот живот. Pazzanese (2020)[13] во својот стручен осврт „Големо ветување, но потенцијална опасност“, објавен во The Harvard Gazette, нагласува дека ВИ поинтензивно почнува да се користи, со што се зголемуваат и етичките грижи поврзани со нејзината (не)способност да носи морални одлуки. Присутна е, на пример, во

[12] https://hbr.org/2022/09/ai-isnt-ready-to-make-unsupervised-decisions

[13] https://news.harvard.edu/gazette/story/2020/10/ethical-concerns-mount-as-ai-takes-bigger-decision-making-role

медицината (пр. за поставување дијагноза или за проценка на кој пациент поприоритетно треба да му биде пресаден орган итн.), во банкарството (пр. кому да се додели кредит), во воената индустријата (пр. производство на автономно летално оружје), во судството (пр. кој да оди во затвор) итн.

Истражувањето на Zhang, Chen, & Xu (2022) во врска со перцепцијата за носењето морални одлуки од страна на ВИ, во кое се прикажани наоди од 4 експериментални студии и вкупно 804 испитаници, соопштува дека испитаниците ја перципираат ВИ како посклона кон носење утилитаристички одлуки при морални дилеми, во споредба со луѓето. Тие, исто така, ја перципираат ВИ како покомпетентна од луѓето, но истовремено и помалку емоционално „топла". Воедно, поединците може да се однесуваат помалку морално и може да бидат повеќе подготвени да ги измамат другите кога комуницираат преку ВИ. Понатаму, истражувањето на Zhang et al. (2023) во врска со моралните судови на луѓето наспроти оние на ВИ при морални дилеми (дилемата на пешачкиот мост и дилемата на трамвајот), во кое се соопштени резултати од три експерименти и 626 испитаници, покажува дека во дилемата на трамвајот, испитаниците ја оцениле ВИ како понеморална и сметале дека нејзините одлуки заслужуваат поголема осуда, отколку однесувањето на луѓето. Во дилемата на пешачкиот мост, испитаниците го оцениле утилитарното однесување (без оглед дали е од ВИ или од човек) како понеморално, помалку дозволиво и попогрешно, отколку деонтолошкото однесување (непреземање ништо). Заклучокот е дека во различни видови морални дилеми, луѓето применуваат различни начини на морално расудување при проценката на однесувањето на ВИ.

Резултатите од експерименталното истражување на Tolmeijer et al. (2022), со вкупно 428 испитаници, во кое се разгледувало како експертското ниво (човек наспроти ВИ) и нивото на експертска автономија (советник наспроти одлучувач) влијае врз довербата, перципираната одговорност и потпирањето врз изворот, покажуваат дека испитаниците перципирале дека на човекот експерт повеќе може да му се верува (од морален аспект), но дека е помалку способен од ВИ. Исто така, почесто ги прифаќале препораките и одлуките на ВИ-експертот, отколку на човекот експерт, додека пак ВИ-експертот го перципирале како помалку одговорен од луѓето.

Bonnefon, Rahwan & Shariff (2024), ги разгледуваат интелигентните машини како морални актери и тоа имплицитни и експлицитни, зависно од целта за која се програмирани. Кога зборуваат за имплицитни морални машини, авторите мислат на машини кои доколку згрешат можат да предизвикаат штета, иако не се иницијално програмирани за експлицитно шифрирање на моралните вредности (пр. ако постават погрешна дијагноза или згрешат во идентификувањето на некое лице, погрешно препознавајќи го како баран криминалец). Експлицитните морални машини, пак, се или

програмирани да решаваат морални дилеми (пр. при проценување кој пациент од листата треба да има приоритет за пресадување на орган) или постои можност да се соочат со морална дилема во одредена ситуација, поради што треба да бидат способни да се справат со истата (пр. при фаворизирање на брзината и прецизноста во работењето, на штета на емпатијата и долгорочната психолошка благосостојба на вработените во организацијата).

Според овие автори, дополнително на сето ова се надоврзуваат и прашањата колку грешки можат да ѝ се дозволат на ВИ и од каква природата можат да бидат тие грешки? Каква би била реакцијата на луѓето на грешките на ВИ? Дали и колку би ја обвиниле ВИ кога греши самостојно, а колку кога грешката ја споделува со човек? Дали човештвото можеби треба да почека додека не се изгради верзија на интелигентни машини кои ќе бидат совршени во проценката и ќе прават 0 % грешки, пред да бидат ставени во функција? Нема ли тоа истовремено да значи дека ќе се жртвува можноста, на пример во медицината, да се спасат многу пациенти со рано дијагностицирање на одредена патологија? И уште многу други познати и непознати прашања и дилеми. За некои од овие прашања веќе постојат најнови истражувања кои покажуваат дека кога имплицитна морална машина греши, луѓето имаат многу посилни негативни реакции (бидејќи имаат многу високи очекувања од неа), отколку кога човек ќе направи слична грешка. Ова е особено видливо во истражувањата на автономните возила, каде испитаниците се многу посклони да ја осудат како потешка и помалку прифатлива сообраќајната несреќа предизвикана од автономно возило префрлајќи му многу поголема одговорност и вина, отколку кога иста таква несреќа би била предизвикана од човек (пр. Franklin et al., 2021; Hidalgo et al, 2021; Hong et al., 2020; Liu & Du, 2022; Liu et al., 2019; според Bonnefon, Rahwan & Shariff, 2024). Сепак, резултатите се менуваат кога во задачата што ја извршува интелигентната машина е вклучен и човек. Кога автономното возило и неговиот возач ќе направат сообраќајна несреќа (пр. ќе удрат пешак), испитаниците се посклони да го обвинуваат возачот. Ваквата разлика во наодите сè уште не е доволно јасна и продолжува да се испитува.

Оттука, одговорот на прашањето дали ВИ може самостојно да носи издржани морални одлуки е: сè уште не. Овој заклучок произлегува од сето досега наведено. Имајќи ја предвид комплексноста на моралното расудување и одлучување, како и фактот дека моралните дилеми кои се јавуваат во различни видови содржат конфликт помеѓу две морални вредности кој интелигентните машини немаат капацитет да го решат; потоа дека нивното решавање истовремено вклучува и когнитивни и афективни процеси, но и автоматски и интуитивни реакции (што повторно изостанува кај ВИ), активирајќи различни мозочни региони и невронски мрежи; дека вештачките невронски мрежи врз кои се потпираат алатките на генеративната ВИ сè уште не се така сложени како што е

човечкиот мозок (иако се обидуваат да ја симулираат неговата активност); како и дека моралното расудување во голем процент се потпира врз специфични процеси на социјалната когницијa и врз претставувањето на менталните состојби на другите луѓе, т.е. теоријата на умот која, заедно со емоционалната интелигенција и самосвеста, е недостижна за ВИ, јасно е дека остануваат уште многу отворени прашања и скалила што треба да се искачат пред ВИ да стане способна самостојно да носи издржани морални одлуки, со 0 % на грешки.

На крај, повеќе од очигледно е дека човештвото е исправено пред сериозен предизвик да одговори на напредокот на ВИ и на нејзината колизија со, како што ќе напишат Bonnefon, Rahwan & Shariff (2024) „... моралните интуиции на луѓето исковани од културата и еволуцијата низ милениумите.“ (стр. 669). Тоа подразбира големо знаење, голема способност за адаптација, отворен ум, визионерство, соработливост, совесност и одговорност на сите релевантни чинители вклучени во процесот на креирање и програмирање на алгоритмите, автономните системи и во интегрирање на моралноста во интелигентните машини. До каде човештвото ќе стигне на тој план и дали навистина еден ден напредните автономни системи со интелигенција иста како човечката ќе „завладеат“ со светот, времето ќе покаже. Сепак, сè додека постои ред и контрола над алгоритмите, а ВИ совесно и одговорно се користи како алатка за да го унапреди квалитетот на животот на луѓето, работите нема да излезат надвор од контрола. Во таа насока, веќе се преземени конкретни чекори од страна на Европската Унија со донесување на Актот за ВИ од 19-ти април 2024 година, со кој се воведуваат нови легислативи и се поставуваат основите за нејзино користење во границите на Унијата.

БИБЛИОГРАФИЈА:

- Belhassein, K., Fernandez-Castro, V., Mayima, A., Clodic, A., Pacherie, E., Guidetti, M., Alami, R., Cochet, H. (2022). Addressing joint action challenges in HRI: Insights from psychology and philosophy. *Acta Psychologica* 222, pp. 1-14 Достапно на: https://doi.org/10.1016/j.actpsy.2021.103476
- Berruti, F., Nel, P., Whiteman, R. (29 April, 2020). *An Executive Primer on Artificial General Intelligence*. McKinsey & Company. Достапно на: https://www.mckinsey.com/capabilities/operations/our-insights/an-executive-primer-on-artificial-general-intelligence (Пристапено на 20.6.2024)
- Bloom, P. (2023). *The Human Mind. A Brief Tour of Everything We Know*. Penguin Random House, UK
- Bonnefon, J.F., Rahwan, I., Shariff, A. (2024). The Moral Psychology of Artificial Intelligence. *Annual Review of Psychology*, 75, pp. 653-75 https://doi.org/10.1146/annurev-psych-030123-113559
- Clark, H.H. & Fisher, K. (2023). Social robots as depictions of social agents. *Behavioral and Brain Sciences* 46, e21, pp. 1–65 Достапно на: https://doi.org/10.1017/S0140525X22000668
- Decety, J., Steinbeis, N. (2020). Multiple Mechanisms of Prosocial Development. In: Decety, J. (ed.) (2020). *The Social Brain. A Developmental Perspective*, The MIT Press, pp. 219-246
- Green, J.D. (2023). Dual-process moral judgment beyond fast and slow. Commentary. *Behavioral and Brain Sciences*, 46 e123, pp. 35-36 https://doi.org/10.1017/S0140525X22003193
- Green, J.D. (2016). Our driverless dilemma: When should your car be willing to kill you? *Science 352*, pp. 1514-1515 doi: 10.1126/science.aaf9534
- Green, J.D., Haidt, J. (2002). How (and where) does moral judgment work? *TRENDS in Cognitive Sciences* 6 (2), pp. 517-523 https://doi.org/10.1016/S1364-6613(02)02011-9
- Green, J.D., Nystrom, L.E., Engell, A.D., Darley, J.M., Cohen, J.D. (2004). The Neural Bases of Cognitive Conflict and Control in Moral Judgment. *Neuron*, 44 (2), pp. 389-400 https://doi.org/10.1016/j.neuron.2004.09.027
- Green, J.D., Sommerville, R.B., Nystrom, L.E., Darley, J.M., Cohen, J.D. (2001). An fMRI Investigation of Emotional Engagement in Moral Judgment. *Science* 293, pp. 2105-2108 doi: 10.1126/science.1062872
- Hatzius, J., Briggs, J., Kodnani, D., Pierdomenico, G. (26 March, 2023). The Potentially Large Effects of Artificial Intelligence on Economic Growth (Briggs/Kodnani). *Economics Research*, Global Economics Analyst. 1-20 Достапно на: https://www.key4biz.it/wp-content/uploads/2023/03/

Global-Economics-Analyst_-The-Potentially-Large-Effects-of-Artificial-Intelligence-on-Economic-Growth-Briggs_Kodnani.pdf (Пристапено на 20.6.2024)

- Kelly, J. (28 February, 2024). *What White-Collar Jobs Are Safe From AI—And Which Professions Are Most At Risk?* Forbes. Достапно на: https://www.forbes.com/sites/jackkelly/2024/02/28/what-white-collar-jobs-are-safe-from-ai-and-which-professions-are-most-at-risk/ (Пристапено на 20.6.2024)
- Martin, A. (26 November, 2021). *Robotics and Artificial Intelligence: The Role of AI in Robots*. AI Business. Достапно на: https://aibusiness.com/verticals/robotics-and-artificial-intelligence-the-role-of-ai-in-robots (Пристапено на 20.6.2024)
- McKendrick, J., Thurai, A. (15 September, 2022). *AI Isn't Ready to Make Unsupervised Decision*, Harvard Business Review, Достапно на: https://hbr.org/2022/09/ai-isnt-ready-to-make-unsupervised-decisions (Пристапено на 20.6.2024)
- Pazzanese, Ch. (26 October, 2020). *Great Promise but Potential for Peril*, The Harvard Gazette, Достапно на: https://news.harvard.edu/gazette/story/2020/10/ethical-concerns-mount-as-ai-takes-bigger-decision-making-role/ (Пристапено на 20.6.2024)
- Shenhav, A., Green, J.D. (2014). Integrative Moral Judgment: Dissociating the Roles of the Amygdala and Ventromedial Prefrontal Cortex. *The Journal of Neuroscience*, 34 (13), pp. 4741-4749 DOI: 10.1523/JNEUROSCI.3390-13.2014
- Stock, R.M. & Nguyen, M.A. (2019). Robotic Psychology: What Do We Know about Human–Robot Interaction and What Do We Still Need to Learn? *Proceedings of the 52nd Hawaii International Conference on System Sciences*, pp. 1936-1945 Достапно на: https://scholarspace.manoa.hawaii.edu/items/7ab2be7a-3a3a-463e-89b3-9753041f7e19
- Tassy, S., Oullier, O., Mancini, J., & Wicker, B. (2013). Discrepancies between judgment and choice of action in moral dilemmas. *Frontiers in Psychology*, 4, pp. 1-8 https://doi.org/10.3389/fpsyg.2013.00250
- Tolmeijer, S., Christen, M., Kandul, S., Kneer, M., & Bernstein, A. (2022). Capable but Amoral? Comparing AI and Human Expert Collaboration in Ethical Decision Making. In: *ACM CHI Conference on Human Factors in Computing Science (CHI'22)*, New Orleans, LA, USA, 29, April 2022-5 May 2022, ACM Press https://doi.org/10.1145/3491102.3517732
- Zaixuan, Z., Chen, Z., & Xu, L. (2022). Artificial intelligence and moral dilemmas: Perception of ethical decision-making in AI. *Journal of Experimental Social Psychology*, 101, pp. 1-8 DOI:10.1016/j.jesp.2022.104327

- Zhang, Y., Wu, J., Yu, F., & Xu, L. (2023). Moral Judgments of Human vs. AI Agents in Moral Dilemmas. *Behavioral Science*, 13, 181, pp. 1-14 https://doi.org/10.3390/bs13020181

Интернет-ресурси:

1. https://www.ibm.com/think/topics/artificial-intelligence-types (Пристапено на 10.6.2024)
2. https://www.resumebuilder.com/1-in-3-companies-will-replace-employees-with-ai-in-2024/ (Пристапено на 10.6.2024)
3. https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-artificial-general-intelligence-agi#/ (Пристапено на 10.6.2024)
4. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai (Пристапено на 10.6.2024)
5. https://artificialintelligenceact.eu/ai-act-explorer/ (Пристапено на 20.6.2024)