

ЕТИЧКЕ СМЕРНИЦЕ

ЗА РАЗВОЈ, ПРИМЕНУ И УПОТРЕБУ ПОУЗДАНЕ И
ОДГОВОРНЕ ВЕШТАЧКЕ ИНТЕЛИГЕНЦИЈЕ

Београд, фебруар 2023.

Садржај

1. Увод.....	3
1.1 Разлог доношења Смерница	3
1.2 Основ доношења.....	3
1.3 Примена Смерница.....	5
2. РЕЧНИК ТЕРМИНА И ДЕФИНИЦИЈА	5
2.1 Етика	5
2.2 Систем вештачке интелигенције	6
2.3 Високоризични системи вештачке интелигенције.....	7
2.4 Заштита података о личности	10
3. НАЧЕЛА.....	10
3.1 Објашњивост и проверљивост	10
3.2 Достојанство.....	11
3.3 Забрана чињења штете.....	12
3.4 Правичност.....	13
4. УСЛОВИ ПОУЗДАНЕ И ОДГОВОРНЕ ВЕШТАЧКЕ ИНТЕЛИГЕНЦИЈЕ.....	14
4.1 Деловање и контрола.....	15
4.2 Техничка поузданост и безбедност.....	18
4.3 Приватност, заштита података о личности и управљање подацима.....	24
4.4 Транспарентност	27
4.5 Различитост, недискриминација и равноправност	32
4.6 Друштвено и еколошко благостање	35
4.7 Одговорност	38
5. ЗАКЉУЧАК.....	40

1. Увод

1.1 Разлог доношења Смерница

Етичке смернице за развој, примену и употребу поуздане и одговорне вештачке интелигенције (у даљем тексту: Смернице) за циљ имају да омогуће да се наука, посебно у области вештачке интелигенције, развија и напредује али да не дозволе да се човек, као централна фигура свих процеса који на њега утичу и чији је посредни или непосредни чинилац, угрози и запостави. Такође, системи вештачке интелигенције који се развијају морају да буду у складу с добробити човека, животиња и животне средине.

Вештачка интелигенција један је од стубова четврте индустријске револуције. Као правац развоја рачунарске науке и инжењерства почела је да се развија пре неколико деценија с периодима успона и стагнација. Захваљујући продору у области неуронских мрежа, све већој количини доступних података погодних за машинско учење, као и све већој доступности микропроцесора погодних за обимна нумеричка израчунавања, у последњих неколико година кренули су нагли развој и ширење примене вештачке интелигенције у областима здравства, финансија, образовања, енергетике, обраде текста на природном језику, говорних технологија, компјутерске визије и др.

Развој (система) вештачке интелигенције усмерен је у правцу стварања решења која ће испуњавати одговарајуће стандарде током читавог животног циклуса, на основу којих ће бити окарактерисана као поуздана и одговорна. У начелу, поуздана и одговорна вештачка интелигенција је она која је: технички поуздана и безбедна, у складу са законом и утврђеним етичким принципима и вредностима. Свака од наведене три компоненте посматра се засебно; услови оцене и исход саме оцене једне од компонената не претпостављају услове оцене и исход оцене друге компоненте. Наведене компоненте треба довести у хармонизовани однос, тако да се испуњењем све три вештачка интелигенција може оценити као поуздана и одговорна.

Основни циљ доношења Смерница је да се не дозволи да процеси у којима учествује систем вештачке интелигенције угрозе или да маргинализују човека и деловање човека и да се слобода деловања, мишљења и одлучивања не наруши у мери да права и тековине које чувају те вредности буду обесмишљене, умањене или заборављене. Ово се пре свега односи на креирање екосистема у коме ће се употребом вештачке интелигенције унапредити продуктивност људи, оптималније користити ресурси у раду и функционисању људи и друштва у целини и унапредити квалитет живота људи.

1.2 Основ доношења

Основ за доношење Смерница садржан је у Стратегији развоја вештачке интелигенције у Републици Србији за период 2020–2025,¹ која је као један од својих пет циљева поставила етичку, безбедну примену вештачке интелигенције и ову активност одредила у Акционом плану за период 2020-2022.²

¹ „Службени гласник РС”, бр. 96/2019.

² „Службени гласник РС”, бр. 81/2020.

Како би се овај циљ остварио, неопходно је развити и увести механизме који ће омогућити одговоран развој вештачке интелигенције и проверу да ли су ови системи у складу с највишим етичким и безбедносним стандардима. Овим Смерницама дефинишу се стандарди и начин провере примене тих стандарда при развоју и коришћењу система вештачке интелигенције.

UNESCO је, у новембру 2021, усвојио Препоруке о етици система вештачке интелигенције (енг. *Recommendation on the Ethics of AI*)³ у чијој су изради учествовали и представници Републике Србије.⁴ Принципи из Препорука садржани су и у овим Смерницама.

У смислу члана 72 Споразума о стабилизацији и придруживању између Европских заједница и њихових држава чланица, с једне стране и Републике Србије с друге (у даљем тексту: Споразум о стабилизацији), Република Србија обавезала се да обезбеди постепено усклађивање постојећих закона и будућег законодавства с правним тековинама Заједнице које свакако чине и правни акти Заједнице. Ова обавеза потврђена је и одредбама Устава Републике Србије, а нарочито чланом 194.

Европска комисија је, у априлу 2021, Европској унији поднела предлог регулаторног оквира вештачке интелигенције – Предлог Уредбе Европског парламента и Савета о утврђивању усклађених правила о вештачкој интелигенцији⁵ и измени одређених законодавних аката Уније (у даљем тексту: Предлог ЕУ Уредбе о ВИ). Овим Актом Европска унија настоји да креира правни оквир за развој и коришћење вештачке интелигенције, како би се олакшао и повећао ниво улагања и иновација у овој области, односно креирало јединствено тржиште за сигурну и поуздану примену система вештачке интелигенције.

Претходно је и Комесар за људска права при Савету Европе издао Препоруке од десет тачака о вештачкој интелигенцији и људским правима које се надограђују на оно што је Савет Европе већ урадио у овој области, нарочито кроз Европску етичку повељу о коришћењу вештачке интелигенције у правосуђу, Смернице о вештачкој интелигенцији и заштити података, Декларацију Савета министара о манипулативним могућностима алгоритамских процеса, Студију о димензијама људских права у техникама аутоматске обраде података и могућим регулаторним импликацијама, као и извештај Специјалног известиоца Уједињених нација о промоцији и заштити слободе мишљења и изражавања, у коме се разматрају импликације технологија вештачке интелигенције на људска права у информационом друштву.

Како би се обезбедило постепено усклађивање законодавног оквира с правним тековинама Европске уније, као и стварање правног оквира унутар Републике Србије за развој и примену етички усклађених система вештачке интелигенције, Влада Закључком усваја документ Смерница којим препоручује да их примењују сви државни органи и организације, органи и организације покрајинске аутономије, органи и организације јединица локалне самоуправе, установе, јавна предузећа, посебни органи преко којих се остварује регулаторна функција и

³ UNESCO, (2021), *Recommendation on the Ethics of AI*, доступне на адреси: <https://unesdoc.unesco.org/ark:/48223/pf0000381137/PDF/381137eng.pdf.multi>

⁴ UNESCO, Artificial Intelligence, доступно на адреси: <https://en.unesco.org/artificial-intelligence/ethics>

⁵ European Commission, Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS, доступан на адреси: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206>

правна и физичка лица којима су поверена јавна овлашћења, када у свом раду успостављају и користе системе вештачке интелигенције. Препорука је да ове Смернице примењују и друга правна и физичка лица која развијају и/или употребљавају системе вештачке интелигенције.

Смернице настоје да обухвате најшири спектар учесника у екосистему вештачке интелигенције, како би се успоставио хоризонтални приступ примене правила. Смернице се односе на следећа лица:

- лица која раде на развоју и/или примени система вештачке интелигенције;
- лица која примењују системе вештачке интелигенције, пре свега у свом раду који укључује и интеракцију с другим лицима (нпр. учесницима на тржишту);
- лица која користе системе вештачке интелигенције и на које системи имају:
 - непосредан утицај (нпр. користе системе ради остваривања неке јавне услуге)
 - посредан утицај (нпр. део су групе за истраживање ретких болести, чији се медицински подаци обрађују као део стратегије Републике Србије за подизање нивоа здравља нације);
- општу јавност, у најширем смислу.

Смернице се не баве питањима власничке структуре, облигационо-правним односима и другим правним питањима у вези с конкретним резултатом рада и истраживања у екосистему вештачке интелигенције.

1.3 Примена Смерница

Примена Смерница обухвата: 1) спровођење јавних консултација и обука, ради бољег упознавања јавности, укључујући и стручне, с појмом и значајем вештачке интелигенције, 2) праћење и евалуацију примене Смерница у јавном и приватном сектору.

Надлежни министар ближе одређује начин интеграције примене Смерница.

1.4 Законодавни оквир

Због значаја развоја и примене вештачке интелигенције у Републици Србији, препозната је потреба за њеним законским уређењем.

2. РЕЧНИК ТЕРМИНА И ДЕФИНИЦИЈА

2.1 Етика

Етика је наука о моралу, која истражује смисао и циљеве моралних норми, основне критеријуме за морално вредновање, као и уопштено заснованост и извор морала. Етика проучава људско понашање које је прихваћено под одређеним моралним аспектом а које има утицаја на човека, животиње које могу да осете бол, патњу, страх и стрес, као и на екосистеме.

Основни етички појмови су: морал, добро, зло, савест, слобода, срећа, љубав, врлина.

2.2 Систем вештачке интелигенције

Систем вештачке интелигенције или вештачка интелигенција дефинише се на различите начине. Већина дефиниција одређује га као софтвер (софтверски модел) који је истрениран над скупом података ради обављања специфичних задатака (попут препознавања одређених образаца и сл). Дефинисани појмови имају следеће значење у овим Смерницама:

Независна експертска група Европске комисије дала је следећу дефиницију: „**Вештачка интелигенција** односи се на системе који показују разумно и интелигентно понашање, и, на основу анализе свог окружења, доносе одлуке – са одређеним степеном аутономије – како би остварили конкретне циљеве. Системи засновани на вештачкој интелигенцији могу бити засновани искључиво на софтверу и деловати у виртуелном свету (на пример: виртуелни асистенти, софтвери за анализу фотографија, веб претраживачи, системи за препоруке, системи за препознавање говора и лица) или могу бити уграђени у уређаје – хардвер (на пример: напредни роботи, аутономна возила, дронови и слично).”⁶ Ти Системи укључују системе машинског учења и експертске системе. „Систем вештачке интелигенције”, „Систем”, „систем“, „вештачка интелигенција” или само „ВИ“ имају исто значење у оквиру Смерница.

Пристрасност (енгл. *bias*) описује систематске и поновљиве грешке у рачунарском систему које стварају неправедне исходе, као што је привилеговање једне категорије над другом, на начин који се разликује од предвиђених функција алгорита.

Аутономни систем је Систем који се понаша или обавља задатке с високим степеном аутономије, односно без спољног утицаја.

Безбедност - фундаментални концепти безбедности информационих система везани су за информациону безбедност (енг. *information security*), која је описана кроз три карактеристике: поверљивост (енг. *confidentiality*), интегритет (енг. *integrity*) и расположивост (енг. *availability*) које се називају „троугао” или „тројство” или „тријада” сигурности.⁷ Приликом заштите информација и података потребно је контролисано омогућити приступ онима којима је дозвољено да их виде, што обезбеђује проверљивост. Интегритет осигурава да податак представља своје намеравано значење и да није измењен, случајно или намерно. Расположивост означава да се информацији, односно податку може приступити и да се може изменити од стране било кога ко је ауторизован да то уради у одговарајућем временском оквиру.

Дизајнер Система је лице које пројектује Систем.

Дизајнирање Система (енг. *system architecture*) означава радње пројектовања Система; дизајнери Система (енг. *system architect*) су лица која пројектују Систем.

Експлоатација Система је коришћење Система од стране ауторизованих корисника.

Примена Система је креирање одговарајућег Система у складу са функционалном и пројектном спецификацијом.

Лица за надзор су лица овлашћена да проверавају исправност коришћења Система.

⁶ A definition of AI: Main capabilities and scientific disciplines, Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission, 2018.

⁷ Види Bourgeois, D., (2014), *Information Systems for Business and Beyond*, Saylor Foundation, pp.64-65.

Мулти-клауд (енг. *Multi-cloud*) подразумева коришћење више јавних клауда (енг. *cloud*).

Поузданост Система је вероватноћа, на одређеном нивоу поверења, да ће Систем успешно, без отказа, обавити функцију за коју је намењен, унутар специфицираних граница перформанси, у току одређеног времена трајања задатака, када се користи на прописани начин и у сврху за коју је намењен, под дефинисаним нивоима оптерећења, узимајући у обзир и претходно време коришћења система.

Пока-јоке (енг. *Poka-yoke*) је избегавање ненамерних грешака.

Ризик је стање Система које представља последицу неодговарајућих мера заштите и који као извор ризичног догађаја доводи промене квалитета и губитка у Систему.

Систем са самоучењем је Систем који препознаје обрасце у подацима над којима се тренира на аутономан начин, без потребе за надзором.

Тестирање Система је фаза у развоју Система којом се испитује исправност и поузданост рада Система, односно у којој се откривају и исправљају грешке Система.

Тренинг подаци су подаци који се користе за тренирање Система.

Улазни подаци су варијабле, односно параметри који након обраде Система дају одређени резултат (излазне податке).

Људска интервенција (енг. *human in the loop*) - омогућена је интервенција у Систему у свим фазама одлучивања.

Људски надзор (енг. *human of the loop*) - омогућена је интервенција током развоја и праћења рада Система.

Људско одлучивање (енг. *human in command*) - омогућена је контрола рада и свих активности Система укључујући његов шири економски, друштвени, правни и етички утицај, а постоји и контрола одлучивања када и како користити Систем што укључује и одлуку у којим ситуацијама се Систем неће користити.

2.3 Високоризични системи вештачке интелигенције

Високоризични систем је Систем који има тенденцију да непосредно или посредно крши принципе и услове утврђене Смерницама, али не нужно да то и чини.

Са аспекта Смерница, високоризични системи се не сматрају непожељним, али управо због наведеног утицаја, значаја области живота у којима се примењују и могућности и домета утицаја на човека и његов интегритет неопходно их је посебно анализирати и проценити њихов утицај.

У смислу ових Смерница, високоризичним системом сматра се систем који:

- је део сигурносног (безбедносног) система неког производа, или је сам по себи производ који има функцију и понаша се као сигурносни (безбедносни) систем и као такав захтева оцену усклађености са законодавним нормама о стављању у употребу система вештачке интелигенције, од стране трећег лица;
- у Смерницама је побројан и означен како високоризични систем (у даљем тексту: листа високоризичних система).

Није од значаја за одређивање високоризичних система то да ли је Систем у употреби (довољно је да је као такав направљен), као ни то да ли Систем чини самосталан производ/услугу или саставни део неког производа/услуге.

Смернице се не односе на Системе који су забрањени у смислу закона којим се регулишу системи вештачке интелигенције.

Високоризичним системима сматрају се системи вештачке интелигенције у областима:

- биометријске идентификације и категоризације појединца: нарочито обухватају системе намењене за даљинску биометријску идентификацију појединца у реалном времену као и накнадну даљинску биометријску идентификацију;
- управљања критичном инфраструктуром и њеним радом: нарочито обухвата системе који су намењени за управљање путним, транспортним саобраћајем, снабдевањем водом, гасом, грејањем и електричном енергијом или су сигурносни систем наведених система или чине део тих сигурносних система;
- образовања, стручног усавршавања и оспособљавања: нарочито обухвата системе намењене за одређивање могућности приступа појединцу установама за образовање и струковно оспособљавање или за распоређивање појединаца у те установе, као и системе који су намењени за оцењивање лица која похађају поменуте установе, укључујући и системе који врше оцењивање тестова (пријемних испита) потребних за упис појединаца у те установе;
- запошљавања, управљања запосленим/ангажованим лицима и приступа самозапошљавању: нарочито обухвата системе који врше одабир и запошљавање/ангажовање лица, укључујући и системе које врше оглашавање слободних радних места, преглед, филтрирање, оцењивање кандидата за конкретно радно место (на разговорима или тестовима) доношење крајње одлуке о запослењу/ангажовању лица; системе који доносе одлуке о радно-правним питањима запослених (напредовање, награде, бонуси, откази, промена описа радног места, конкретних задатака запосленог) као и системе који врше праћење и евалуацију успешности запослених, на темељу којих ће се донети одлука из радног односа;⁸
- здравства: нарочито обухвата системе који анализирају генетичке и здравствене податке;
- приступа и коришћења јавних и социјалних услуга као и основних приватних услуга: нарочито обухвата системе намењене за оцењивање прихватљивости појединаца за пружање јавних услуга и социјалних давања као и доношења одлука о одобравању, смањењу, укидању таквих услуга, као и услова под којима се такве одлуке доносе. Обухвата и системе за оцењивање кредитне способности лица као и утврђивање кредитне оцене, осим уколико се ти системи не користе за личне и некомерцијалне потребе; обухвата и системе конструисане да функционишу као диспечер службе при службама за пружање хитне медицинске или друге ургентне помоћи (ватрогасци, војска, полиција, и слично), где такви системи врше и одређивање приоритета пружања такве помоћи;

⁸ Ова одредба се не ограничава само на ангажовање лица кроз радни однос, већ и друге облике ангажовања лица (радне снаге) и примање тог лица у рад (уговор о делу, уговор о привремено-повременим пословима и слично, у складу са законом којим се регулише ангажовање лица за рад), где то ангажовање укључује селекциони процес и сам процес одржавања тог односа (евалуација, награђивање, санкционисање, престанак тог односа и слично).

- кривичног гоњења: нарочито обухвата системе намењене органима кривичног гоњења који врше процену ризика појединаца за извршење или поновно извршење кривичних дела на основу особина, карактеристика или претходног криминалног понашања; системе који би се користили као полиграфска средства или средства за откривање емоционалног стања појединца; системе који би се користили за оцену веродостојности доказа током предистражног, истражног и судског поступка; системе који би се користили за процену појединца; системе намењене за аналитику кривичних дела који се односе на појединце и који омогућавају претраживање сложених повезаних и неповезаних великих скупова података из различитих извора или у различитим форматима с циљем уочавања непознатих узорака у подацима или скривених веза међу подацима;
- управљања миграцијама људи, азилом и надзора државне границе: нарочито обухвата системе који би се користили као полиграфска средства или средства за откривање емоционалног стања појединца; системе намењене надлежним органима/институцијама за процену ризика (укључујући ризик за сигурност, ризик од незаконитих имиграција или ризик за здравље) који представља појединац који намерава да уђе или је ушао на територију државе Републике Србије; системе намењене надлежним органима/институцијама за проверу веродостојности путних исправа, нарочито кроз проверу сигурносних обележја тих исправа; системе намењене за помоћ надлежним органима/институцијама при разматрању захтева за азил, визе, боравишне и радне дозволе и с њима повезаним процесима (повезани процеси обухватају проверу кривичне и прекршајне осуђиваности, постојање и врсту притужби датих у вези с предметним захтевима и сл), а у циљу доношења одлука у вези са предметним захтевима;
- правосуђа и демократских процеса: нарочито обухвата системе намењене за помоћ правосудним органима у анализи и тумачењу околности, чињеница, и правних норми, а у циљу примене одговарајућих правних норми на конкретан скуп околности, чињеница.

Листи високоризичних система припадају и системи за препоруке вођени вештачком интелигенцијом на платформама које користи велики број људи, као што су друштвене мреже; на основу различитих података и постављених циљева доносе одлуке који садржај се представља појединцу или групи, што на нивоу појединца може утицати на пажњу, жеље, мишљења, опредељења, креативност, машту, осећања, одлуке и активности, или, на нивоу друштва може допринети креирању друштвених мехура, односно поларизацији око битних друштвених питања, и, у финалној инстанци, утицати на демократски капацитет као и системи који подразумевају обраду података о личности а које користе државни органи и организације, органи и организације покрајинске аутономије, органи и организације јединица локалне самоуправе, установе, јавна предузећа приликом доношења одлука у обављању послова из своје надлежности.

Листа није коначна. Развој вештачке интелигенције намеће потребу да та листа буде отворена и тумачена као преглед репрезентативних примера високоризичних система која се може мењати и допуњавати.⁹

⁹ Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts.

2.4 Заштита података о личности

Податак о личности је сваки податак који се односи на физичко лице чији је идентитет одређен или одредив, непосредно или посредно, посебно на основу ознаке идентитета, као што је име и идентификациони број, података о локацији, идентификатора у електронским комуникационим мрежама или једног, односно више обележја његовог физичког, физиолошког, генетског, менталног, економског, културног и друштвеног идентитета.

Обрада посебних врста података о личности је обрада којом се открива расно или етничко порекло, политичко мишљење, верско или филозофско уверење или чланство у синдикату, као и обрада генетских података, биометријских података у циљу јединствене идентификације лица, података о здравственом стању или података о сексуалном животу или сексуалној оријентацији физичког лица.

Процена утицаја обраде на заштиту података о личности спроводи се пре отпочињања обраде, ако је вероватно да ће одређена врста обраде, посебно употребом нових технологија и узимајући у обзир природу, обим, околности и сврху обраде, проузроковати висок ризик за права и слободе физичких лица. Процену утицаја обраде неопходно је спровести у случају: 1) систематске и свеобухватне процене стања и особина физичког лица која се врши помоћу аутоматизоване обраде података о личности, укључујући и профилисање, на основу које се доносе одлуке од значаја за правни положај појединца, или на сличан начин значајно утичу на њега; 2) обраде посебних врста података о личности или података о личности у вези с кривичним пресудама и кажњивим делима, у великом обиму; 3) систематског надзора над јавно доступним површинама у великој мери. Листа врста радњи обраде за које се мора извршити процена утицаја утврђена је Одлуком коју је донео Повереник за информације од јавног значаја и заштиту података о личности.

Лице за заштиту података о личности је лице које, сагласно Закону о заштити података о личности, обавља одговарајуће послове и задатке. Обавеза одређивања овог лица постоји у случају да се: 1) обрада врши од стране органа власти, осим ако се не ради о обради коју врши суд у сврху обављања његових судских овлашћења; 2) основне активности руковооца/обрађивача састоје од радњи обраде које по својој природи, обиму, односно сврхама захтевају редован и систематски надзор великог броја лица на које се подаци односе; 3) основне активности руковооца/обрађивача састоје од обраде посебних врста података о личности или података о личности у вези с кривичним пресудама и кажњивим делима, у великом обиму.

3. НАЧЕЛА

Не умањујући значај других схватања и принципа, издвојена су следећа начела која су препозната као полазна основа за стварање, примену и употребу Система вештачке интелигенције који ће својом поузданости и одговорношћу према човеку бити достојни људског поверења:

3.1 Објашњивост и проверљивост

Једна од основних особина људске свести је да перципира окружење, тражи одговоре на питања односно објашњења зашто и како нешто јесте или није. Та особина утицала је на еволуцију човека

и развој науке, па самим тим и вештачке интелигенције. Потреба човека да разуме и да му ствари буду јасне нашла је своје упориште у овом начелу.

Објашњивост у контексту ових Смерница значи да сви процеси: развој, тестирање, пуштање у рад, надзор над радом система и његово гашење, морају бити транспарентни. Сврха и могућности самог система вештачке интелигенције морају бити објашњиве, а нарочито одлуке (препоруче) које доноси (у мери којој је то целисходно) свима на које Систем утиче (директно или индиректно). Ако одређене резултате рада Система није могуће објаснити, неопходно их је означити као систем с моделом „црне кутије”.¹⁰

Проверљивост је комплементарни елемент овог начела којим се обезбеђује да се Систем може проверавати у свим процесима, односно током целог животног циклуса. Проверљивост укључује радње и поступке провере система вештачке интелигенције приликом тестирања и примене, као и проверу краткорочног и дугорочног утицаја који такав систем има на човека.

3.2 Достојанство

Дужност свих чланова друштва јесте да ово право међусобно поштују и штите као једно од основних и неприкосновених права сваког човека. Сваки појединац има право на заштиту сопственог достојанства; кршење или непоштовање овог права законом се санкционише.

Људско достојанство (даље: достојанство) треба разумети као полазни принцип (начело) које у фокусу има очување човековог интегритета. Полазећи од те премисе, лица на која се ове Смернице односе треба у сваком тренутку, без обзира на фазу у којој се конкретно решење вештачке интелигенције налази (развој, примена или употреба) да имају на уму човека и његов интегритет као централни појам. С тим у вези, неопходно је развијати системе којима се у свакој фази као императив постављају поштовање човекове личности, његове слободе и аутономије.

Поштовање човекове личности подразумева стварање система који ће поштовати когнитивне, социјалне и културне карактеристике сваког појединца. Системи вештачке интелигенције који се развијају морају бити у складу с наведеним, стога је неопходно водити рачуна да они ни на који начин не могу довести до подређивања човека функцијама система, као ни угрожавања његовог достојанства и интегритета.

У циљу обезбеђивања поштовања начела достојанства, системи вештачке интелигенције не смеју бити такви да у процесима рада и примене грубо занемарују аутономност људског избора.

¹⁰ Систем с моделом „Црне кутије“ обухвата различите дефиниције, међутим све се концентришу на једну ствар. У питању су системи вештачке интелигенције који у основи имају модел који се креира директно из података уз помоћ развијеног алгоритма, што значи да лица која су их дизајнирала не могу да разумеју како се варијабле тог модела комбинују да би се направила одређена предвиђања, односно систем не указује како је до тог податка/резултата дошао. Чак и ако неко има листу улазних варијабли, предиктивни модели црне кутије могу бити тако компликоване функције варијабли да се не може утврдити како су варијабле повезане једна с другом да би се дошло до коначног предвиђања. Неки их чак тумаче и као појам који означава моделе који су довољно сложени да их човек не може протумачити.

Устав Републике Србије наглашава да је достојанство „неприкосновено и сву су дужни да га поштују и штите. Свако има право на слободан развој личности, ако тиме не крши права других зајемчена Уставом.”¹¹

Конвенција о људским правима наводи следеће: „Људско достојанство (достојанство) није само основно људско право већ и основа људских права. Људско достојанство је урођено сваком човеку.”¹²

У Републици Србији овај појам је уређен на следеће начине:

- „Достојанство личности (част, углед, односно пијетет) лица на које се односи информација правно је заштићено.”¹³
- „Ко злоставља другог или према њему поступа на начин којим се вређа људско достојанство, казниће се затвором до једне године.”¹⁴
- „Рад у јавном интересу је сваки онај друштвено користан рад којим се не вређа људско достојанство и који се не врши у циљу стицања добити.”¹⁵

Овим начелом наглашава се да се у сваком тренутку мора бринути о интегритету и достојанству свих на које Систем вештачке интелигенције може утицати. Како је у питању општи појам, којем живот, поред закона даје различита наличја, мада је суштина иста, примерено је за сам појам везати: част, углед, односно пијетет.

3.3 Забрана чињења штете

Систем вештачке интелигенције мора бити усаглашен са стандардима безбедности, односно мора да садржи одговарајуће механизме који ће спречити настанак штете лицима и њиховој имовини. У случају да до штете ипак дође, она мора бити санирана у најкраћем могућем року, а оштећено лице на законом утврђен начин обештећено.

Закон о облигационим односима уређује појам штете као „умањење нечије имовине (обична штета) и спречавање њеног повећања (измакла корист), као и доношење другог физичког или психичког бола или страха (нематеријална штета)¹⁶ и утврђује да је свако лице дужно да се уздржи од поступака којим се може другом проузроковати штета.”¹⁷

Поред грађанске одговорности закон препознаје и кривичну и прекршајну одговорност како физичких тако и правних лица за штету коју причине другом лицу.

Кривични законик¹⁸ предвиђа велики број кривичних дела, од којих је значајно навести кривична дела против живота и тела, имовине људи, против слобода и права човека и грађанина. Посебним

¹¹ Устав Републике Србије, „Службени гласник РС”, бр. 98/2006 и 115/2021.

¹² Конвенција о људским правима.

¹³ Закон о јавном информисању и медијима, „Службени гласник РС”, бр. 83/2014... аут. тумачење - 12/2016.

¹⁴ Кривични законик, „Службени гласник РС”, бр. 85/2005, 88/2005 - испр.,... и 35/2019.

¹⁵ Ibidem.

¹⁶ Закон о облигационим односима, „Службени гласник РС”, бр. 29/78, 39/85, 45/89 - одлука УСЈ и бр. 18/2020.

¹⁷ Ibidem.

¹⁸ Кривични законик, „Службени гласник РС”, бр. 85/2005, 88/2005 – испр. 121/2012,... и 35/2019.

законом предвиђена је и одговорност лица за штету коју причине извршењем дела мање друштвене опасности - прекршаја.¹⁹

Посебну пажњу требало би посветити заштити осетљивих категорија као што су старији, особе са инвалидитетом, деца, труднице и др, као и категоријама које су у неповољнијем положају (на пример: радник – послодавац, потрошач – привредни субјекат, и др).

Системи вештачке интелигенције морају да се користе на сигуран и безбедан начин, односно морају бити поуздани и безбедни, и треба спречити њихову употребу у малициозне сврхе.

3.4 Правичност

Начело правичности односи се на заштиту права и интегритета од дискриминације, посебно дискриминације нарочито осетљивих категорија (на пример особа са инвалидитетом). Сам термин због своје вишестраности, има различита тумачења у бројним сферама друштвеног живота. На пример, у здравственој заштити²⁰ начело правичности подразумева забрану дискриминације у пружању здравствене заштите по основу расе, пола, рода, сексуалне оријентације и родног идентитета, старости, националне припадности, социјалног порекла, вероисповести, политичког или другог убеђења, имовног стања, културе, језика, здравственог стања, врсте болести, психичког или телесног инвалидитета, као и другог личног својства које може бити узрок дискриминације. Тако и системи вештачке интелигенције морају, приликом коришћења спречити дискриминацију.

Начело правичности има своју стварну (енг. *substantive*) и процедуралну димензију. Стварна димензија подразумева заштиту од неоправдане пристрасности, дискриминације и стигматизације. Системи вештачке интелигенције требало би да пруже једнаке могућности свим лицима, како у погледу приступа образовању, добрима тако и услугама и технологијама, тако и да спрече обмане лица која користе системе вештачке интелигенције, приликом доношења одлука. Процедурална димензија правичности укључује могућност оспоравања и укључивања ефикасне правне заштите против одлука које су резултат рада Система вештачке интелигенције као и лица одговорних за рад Система. У циљу испуњења овог услова, неопходно је да постоје јасно утврђене одговорности, као и да процес доношења одлука буде објашњен, јасан и транспарентан. Тиме се смањује могућност погрешног или непотпуног разумевања сврхе и циљева коришћења ових система, односно потенцијалног ускраћивања слободе избора при одабиру система који ће користити. Правична употреба Система вештачке интелигенције може довести до повећања правичности у друштву у целини, као и до смањења разлика које постоје међу појединцима у погледу социјалног, економског и образовног статуса.

¹⁹ Закон о прекршајима, „Службени гласник РС”, бр. 65/2013... 91/2019 и др. закон.

²⁰ Закон о здравственој заштити, „Службени гласник РС”, бр. 25/2019-40.

4. УСЛОВИ ПОУЗДАНЕ И ОДГОВОРНЕ ВЕШТАЧКЕ ИНТЕЛИГЕНЦИЈЕ

Изградња и стварање поуздане и одговорне вештачке интелигенције захтевају испуњење одређених Улова²¹ који се темеље на Начелима утврђеним овим Смерницама, а која се одређују кроз:

1. Деловање (посредовање, контрола, учешће) и надзор;
2. Техничку поузданост и безбедност;
3. Приватност, заштиту података о личности и управљање подацима;
4. Транспарентност;
5. Различитост, недискриминацију и равноправност;
6. Друштвено и еколошко благостање;
7. Одговорност.

Улове чине проверљиви параметри, односно техничке и нетехничке методе, којима се потврђује и доказује испуњеност Начела.

Циљ **техничких метода** је да усмере развој, примену па и коришћење Система вештачке интелигенције на начин да се Системи вештачке интелигенције понашају поуздано, свдећи на минимум потенцијалне ненамерне и непредвидиве штете по човека и друштво у целини. Техничке методе су приказане у форми препорука.

Нетехничке методе односе се на испитивање организационих и других нетехничких елемената значајних за развој и коришћење система вештачке интелигенције. Ове методе дате су у форми упитника који је намењен оцењивању појединачних система вештачке интелигенције у смислу испуњености основних начела, односно услова садржаних у Смерницама. Сврха упитника јесте утврђивање поузданости и одговорности система вештачке интелигенције са аспекта етичких стандарда.

Упитник пружа подршку лицима, односно организацијама које развијају, стављају на тржиште, набављају, примењују и/или користе системе вештачке интелигенције да процене усклађеност с наведеним Уловима. Упитник је погодан за употребу у свим друштвено-економским областима и представља минимални хоризонтални оквир за успостављање безбедних система вештачке интелигенције у Републици Србији. Упитник може бити прилагођен посебним областима и секторима. Ради информисања о условима које Систем мора да испуни да би био оцењен као етички поуздан и одговоран, листу питања из упитника пожељно је проучити и пре израде самог Система, већ у самој фази планирања почетка рада на развоју система. Препорука је да се Упитник попуни у најранијим фазама израде Система, у бета фази, али и у свим каснијим етапама како би био континуирано праћен током целог животног циклуса.

У најширем смислу, Упитник о процени система вештачке интелигенције доприноси унапређењу информисаности и културе развоја поузданог и одговорног екосистема вештачке интелигенције у Републици Србији. Његовом употребом подиже се друштвена свест о важности испуњавања одређених захтева система вештачке интелигенције. Истовремено, примена Упитника доприноси

²¹ Услови су дефинисани кроз принцип „отворене листе“; Смернице не ограничавају примену и других услова који се могу применити са аспекта етичких принципа, а све у циљу развоја, примене и употребе поуздане и одговорне вештачке интелигенције.

унапређењу транспарентности и јачању поверења друштва у одрживе системе који испуњавају стандарде. Упитник о процени система вештачке интелигенције помаже лицима и/или организацијама да идентификују подручја за унапређење и подстиче их да предузимају мере за превазилажење уочених изазова. Попуњавањем Упитника добија се увид у успостављене мере и идентификују се мере које би тек требало применити у сврху изградње поузданог система вештачке интелигенције, тако да Упитник представља и важан алат за развој иновативних решења у области вештачке интелигенције.

Сам Упитник не искључује примену других алата и метода за оцену испуњености услова Система у смислу усвојених Смерница и/или закона. Упитник није водич кроз правни систем РС и његово попуњавање не ослобађа од законских обавеза и одговорности.

4.1 Деловање и контрола

Систем вештачке интелигенције би требало да буде поуздана подршка у процесу доношења одлука и предмет континуираног надзора и контроле од стране човека. У оквиру ове групе, испитује се утицај Система на доношење одлука и деловање човека, односно подршка у одлучивању, као и анализи и предвиђању ризика (Системи препорука, предиктивног надзора, анализе финансијског ризика и сл). Испитују се и перцепција и очекивање: лица која развијају или одржавају систем, лица која користе систем и лица на које систем утиче од система вештачке интелигенције, као и њихову наклоност, поверење и (не)зависност при доношењу одлука.

4.1.1 Упитник

Деловање

- Да ли је систем вештачке интелигенције дизајниран да:
 - комуницира (је у интеракцији)
 - утиче на одлуке (даје препоруку)
 - доноси одлуке
- Да ли је лице које користи и/или лице на које систем вештачке интелигенције утиче свесно да је у интеракцији са системом?
 - Да, и то: (на начин)
 - Не
- Да ли је лице које користи и/или лице на које систем вештачке интелигенције утиче обавештено да је одлука, садржај, савет или исход резултат алгоритамске одлуке?
 - Да, и то: (на начин)
 - Не
- У којој мери систем вештачке интелигенције може да утиче на аутономију личности при доношењу одлука?
 - У потпуности утиче
 - Значајно утиче
 - Делимично утиче
 - Минимално утиче
 - Не утиче

- Да ли су успостављене процедуре којима се лицу које користи систем вештачке интелигенције онемогућава да се, при доношењу одлука, ослања искључиво на систем вештачке интелигенције?
 - Да, и то:
 - Не
- Да ли постоји процедура којом се спречава да систем вештачке интелигенције ненамерно (самосталним учењем) утиче на аутономију личности?
 - Да, и то:
 - Не
- Да ли постоји процедура којом се спречавају могуће негативне последице по лица која користе као и лица на које систем утиче ако развију зависност од система вештачке интелигенције?
 - Да, и то:
 - Не
- Да ли су предузете мере за смањење ризика од зависности лица која користе као и лица на које систем утиче од коришћења система вештачке интелигенције?
 - Да, и то:
 - Не

Контрола. У овом одељку врши се самопроцена уведених контролних мера кроз механизме управљања, и то:

1. *Human-in-the-loop* (HITL) - омогућена је интервенција у свим фазама одлучивања
 2. *Human-on-the-loop* (HOTL) - омогућена је интервенција током развоја и надгледања рада система вештачке интелигенције
 3. *Human-in-command* (HIC) - омогућена је контрола рада система вештачке интелигенције укључујући шири економски, друштвени, правни и етички утицај, као и контрола одлучивања када и како користити систем, што подразумева да се систем вештачке интелигенције не користи у одређеној ситуацији.
- Да ли је систем вештачке интелигенције:
 - систем без надзора
 - HITL систем у којем је омогућена интервенција у сваком циклусу одлучивања
 - HOTL систем у којем је омогућена интервенција током пројектовања и надгледања рада система вештачке интелигенције
 - HIC систем у којем је омогућена контрола свих активности система
 - Да ли су лица која врше контролу обучена за те послове?
 - Да
 - Не
 - Да ли су успостављени механизми откривања и реаговања на нежељене штетне ефекте система вештачке интелигенције?
 - Да, и то:
 - Не
 - Да ли постоји „дугме за заустављање“ или процедура за безбедно прекидање операције када је то потребно?
 - Да, постоји:
 - Не

- Да ли су развијене посебне контролне мере за евидентирање самоучења или аутономне природе система?
 - Да, и то:
 - Не

4.1.2 Препоруке

Приликом дизајнирања система вештачке интелигенције потребно је узети у обзир све кориснике и сценарија обраде података уз комплетан опсег варијабилности и специфичности обраде података о личности. Приликом развоја Система потребно је документовати:

- могућности и функционалности Система;
- сценарије коришћења;
- оперативне фактуре и конфигурације које доприносе поузданом и одговорном коришћењу Система;
- ограничења;
- сегменте унутар којих Систем није дизајниран за употребу;
- приказ тачности и правилног рада Система и опис до које мере се такви резултати могу очекивати за генерализовано коришћење за сценарије који нису иницијално узети у обзир;
- границе до којих је очекиван даљи развој Система без директног утицаја човека.

Препорука је да се:

1. Успостави техничка документација која прецизно објашњава дизајн Система, подсистема и компоненти, укључујући механизме за праћење и надзор функционисања Система.
2. Дизајнира такав Систем који омогућава праћење и надзор његовог функционисања као и ретроспективну анализу резултата обраде у односу на улазне податке.
3. Код Система који укључује интеракцију с лицима која примењују и/или користе Систем, а који може понудити више од једног резултата обраде с различитим вероватноћама, потребно је омогућити избор једног од резултата обраде.
4. У процес планирања, дизајна и развоја система укључи детекција негативних или узгредних ефеката по питању једнакости и људских права као и могућност надзора и ретроспективне анализе функционисања Система.
5. Идентификује и документује више врста метода за процену исправности функционисања Система. Различити начини процењивања доприносе ефикаснијој идентификацији аномалија у функционисању Система.
6. Анализирају и од стране дизајнера Система у потпуности разумеју изворни подаци на основу којих се врши тренинг алгоритма Система, како би се утврдило да ли изворни подаци заиста представљају опсег варијабилности за све кориснике или уже групе корисника. Скуп података на коме се тренира модел Система треба да буде репрезентативан, у смислу да представља верну слику реалног Система који се моделира.
7. Идентификују лица која су задужена да реагују у случају проблема у функционисању Система, која надгледају и контролишу рад Система од фазе развоја, преко тестирања и фазе учења (тренинга) до пуне експлоатације, односно свакодневног рада Система. Потребно је та лица евидентирати и дефинисати њихова задужења, као и начин њиховог избора лица, њихову обуку, проверу поступања као и капацитета током времена, будући

да је потенцијални сценарио да Систем кроз процес учења протоком времена својим могућностима превазиђе капацитете лица за надзор.

8. Идентификују елементи Система, кориснички алати и алати за извештавање, које би лица из тачке 7. требало да познају, укључујући и могућност разумевања излазних резултата Система на основу којих могу предузети одређене радње (на пример: гашење Система).
9. Успоставе и документују критеријуми за пуштање Система у фазу примене који морају да садрже метрику и граничне вредности. У случају да извештај о процени покаже вредности које прелазе граничне вредности, потребно је дефинисати и документовати приступ и план како да се реше идентификовани проблеми.

Уколико лица задужена за надзор Система уоче одређене аномалије у понашању Система које би протоком времена могле довести до нежељеног стања Система описаног у претходној тачки, могу, на основу свог дискреционог права, Систем привремено ставити ван функције у ограниченом временском периоду, уз детаљно образложење свих разлога за такво поступање. Оваква одлука подлеже ревизији ширег тима који је учествовао у креирању Система, или за то надлежне комисије.

4.2 Техничка поузданост и безбедност

Кључни услов за постизање поузданих система вештачке интелигенције је њихова поузданост и безбедност. Техничка поузданост захтева да се Системи развијају уз превентивну процену ризика, као и да се понашају поуздано и како је предвиђено, уз свођење на минимум потенцијалне ненамерне и непредвидиве штете.

4.2.1 Упитник

Питања у овом делу односе се на четири главна питања: 1) заштиту од претњи и злоупотреба; 2) безбедност; 3) тачност и прецизност; и 4) поузданост, резервне планове и поновљивост.

Заштита од претњи (напада) и злоупотреба

- Да ли систем вештачке интелигенције припада информационо-комуникационој (ИКТ) инфраструктури од посебног значаја?
 - Да
 - Не
- Да ли је систем вештачке интелигенције сертифициован у складу са стандардима у области информационе безбедности (нпр. ИСО 27000 и др) или је усклађен са овим стандардима?
 - Да, сертифициован је по стандарду ... (навести којем)
 - Да, усклађен је са стандардом ... (навести којем)
 - Није извршена провера испуњености услова прописаних стандардима у области информационе безбедности
- Да ли су препознати потенцијални облици претњи (напада) на које би систем вештачке интелигенције могао бити рањив?
 - Да, и то: (навести који су, нпр. дефекти у дизајну, техничке грешке...)
 - Не постоје ризици од напада
- Да ли су узете у обзир различите врсте рањивости и потенцијалне улазне тачке напада, и то (означити ако су узете у обзир):

- Манипулација подацима
- Модификовање модела класификације података
- Инверзни инжењеринг модела (откривање параметара модела)
- Друго:
- Да ли су предузете мере и које, у обезбеђивању интегритета и заштите система вештачке интелигенције од потенцијалних напада током његовог животног циклуса?
 - Да, у фази анализе предузете су мере:
 - Да, у фази развоја предузете су мере:
 - Да, у фази тестирања предузете су мере:
 - Да, у фази примене предузете су мере:
 - Да, у фази гашења предузете су мере:
 - Нису предузете посебне мере
- Да ли је извршено тестирање неовлашћеног приступа Систему (нпр. пенетрационо тестирање)?
 - Да, коришћен је алат:
 - Да, појединих делова Система и то:
 - Не
- Која софтверска решења користите за заштиту од напада на систем вештачке интелигенције?
 - Користи се:
 - Не постоји посебна заштита овог Система
- Који је временски период у коме је обезбеђено ажурирање система вештачке интелигенције које ће отклањати откривене безбедносне пропусте?
 - Није обезбеђено ажурирање Система
 - До годину дана
 - До пет година
 - Уговором ће се обезбедити континуирана заштита Система

Безбедност

- Да ли су дефинисани ризици, метрике ризика и нивои ризика система вештачке интелигенције?
 - Да, и то: (навести ризике)
 - Не
- Да ли је успостављен процес за континуирано мерење и процену ризика?
 - Да, процена се врши континуирано
 - Да, процена се врши најмање једном годишње
 - Да, процена се врши пре сваког унапређења Система
 - Не
- Да ли су крајњи корисници и субјекти обавештени о постојећим или потенцијалним ризицима?
 - Да, обавештени су (навести на који начин)
 - Не
- Да ли су идентификоване могуће последице напада (инцидената)?
 - Да, и то: (навести последице)
 - Не

- Да ли је извршена процена утицаја стабилности и поузданости Система на доношење одлука (ризик од компромитовања података и алгоритама)?
 - Да, и то: (навести процену)
 - Не
- Да ли је тестирање Система усклађено са одговарајућим нивоима стабилности и поузданости?
 - Да, и то: (навести процену)
 - Не
- Да ли је планирана толеранција грешке?
 - Да, користи се други систем вештачке интелигенције
 - Да, користи се други систем који није заснован на вештачкој интелигенцији
 - Не
- Да ли је развијен механизам за процену промена у раду система вештачке интелигенције како би се испитале његова стабилност, поузданост и сигурност?
 - Да, постоје обавештења када се догоди било која промена у начину рада
 - Да, користи се (навести шта)
 - Не

Тачност и прецизност

- Које су мере предузете како би се осигурало да подаци који се користе за развој система вештачке интелигенције буду тачни, ажурни, потпуни и репрезентативни?
 - Мере које су предузете су:
 - Нису предузете посебне мере
- Како се врши праћење и документовање прецизности система вештачке интелигенције?
 - Врши се: (навести како)
 - Не прати се тачност система
- Да ли је урађена статистичка анализа и да ли су могући сценарији или могуће категорије класификације равномерно или адекватно заступљене?
 - Да, и то:
 - Не
- Да ли се о нивоу прецизности, одзива и тачности система вештачке интелигенције обавештавају крајњи корисници и/или субјекти?
 - Да, и то: (навести на који начин)
 - Не

Поузданост, резервни планови и поновљивост

- Да ли систем вештачке интелигенције може да изазове критичне, супротстављене или штетне последице у случају ниске поузданости и/или поновљивости резултата рада система?
 - Систем је поуздан јер је извршено:
 - Није испитано да ли систем може да изазове штетне последице
- Да ли се мери сврсисходност система вештачке интелигенције (одговара дефинисаној сврси)?
 - Да, на начин:
 - Не, зато што:
- Да ли од специфичног контекста и услова зависи поновљивост резултата рада система вештачке интелигенције?

- Да, од:
- Не
- Да ли се врше верификација и валидација поузданости и поновљивости система вештачке интелигенције?
 - Да, на начин:
 - Не
- Да ли се документује процес тестирања и верификације поузданости и поновљивости система вештачке интелигенције?
 - Да, на начин:
 - Не
- Да ли постоји план за исправљање грешака у систему вештачке интелигенције?
 - Да
 - Не
- Да ли постоји процедура за поступање у случајевима у којима систем вештачке интелигенције даје резултате с ниским резултатом поверења?
 - Да (навести):
 - Не
- Да ли систем вештачке интелигенције користи континуирано учење?
 - Да
 - Не
- Да ли су узете у обзир потенцијалне негативне последице учења система вештачке интелигенције које могу да утичу на резултат?
 - Да, прати се учење система и врше се одговарајуће корекције
 - Не

4.2.2 Препоруке

Техничка поузданост

Како би се осигурале поузданост и безбедност система вештачке интелигенције, пре свега треба пратити генералне препоруке за развој софтверских решења. Уз њих, потребно је увести низ метода које су специфичне за системе засноване на машинском учењу или другим методима развоја система вештачке интелигенције.

1. Идентификовати више метрика за процену квалитета тренинга и надзор

Коришћење више метрика, уместо једне, помаже разумевању односа међу различитим типовима грешке и искуства корисника:

- Размотрити метрике попут прикупљања мишљења корисника анкетирањем, вредности које мере перформансе Система на нивоу целог система, као и краткорочну и дугорочну ваљаност, на пример, стопу кликтања (енг. *clickthrough rate*) или животну вредност потрошача (енг. *customer lifetime value*), као и стопу лажно позитивних и лажно негативних резултата, разложену по подгрупама (категоријама).
- Обезбедити релевантност метрике, на пример: систем за детекцију пожара би требало да има високу стопу препознавања, чак и ако ће то довести до повремене лажне узбуне.

2. Кад год је могуће испитати улазне податке

Системи машинског учења осликавају податке на којима су тренирани, те је неопходно анализирати и прилагодити улазне податке, како би се обезбедио довољан ниво њиховог разумевања. Када то није могуће (код нарочито осетљивих података), потребно је анализирати улазне податке израчунавањем агрегираних, анонимизираних групних вредности и статистика.

Испитати:

- Да ли подаци садрже грешке (између осталог: недостајуће вредности, погрешне ознаке) како би се утврдио квалитет података?
- Да ли су подаци узорковани тако да добро представљају кориснике Система (нпр: Систем ће се користити за све старосне групе, али се подаци за тренинг односе само на пензионере) и реалан сценарио примене (на пример: Систем ће се користити целе године, али је трениран само на подацима прикупљеним током лета)? Да ли су подаци тачни?
- Да ли постоје разлике у перформансама Система током тренинга и примене? Током тренинга идентификовати потенцијална померања која се морају елиминисати, укључујући измене у скупу података за тренинг и/или циљној функцији (енг. *objective function*). Током евалуације Система, обезбедити податке за евалуацију који што боље осликавају сценарио примене.
- Да ли су нека обележја у моделу сувишна или непотребна? Користити најједноставнији модел који задовољава у погледу перформанси.
- За Системе који се тренирају под надзором или су у групи Система високог ризика, размотрити однос између ознака у тренинг подацима и предвиђене вредности.

3. Препознати ограничења скупа података и модела

- Модел трениран да детектује корелацију не би требало користити за доношење одлука о каузалности. Модел може, на пример, научити да су људи који купују патике за кошарку у просеку виши, али то не значи да ће неко ко купи такве патике постати виши.
- Савремени модели машинског учења у великој мери одражавају правилности у подацима који су коришћени за тренинг модела. Стога је потребно утврдити опсег и покривеност разних сценарија примене процедуром тренинга како би се препознале могућности и ограничења модела.
- Транспарентно навести ограничења са корисницима кад год је то могуће.

4. Тестирати, тестирати, тестирати

- Спроводити ригорозно модуларно тестирање, како би се појединачно истестирала свака компонента Система (те компоненте обухватају код, податке и сам модел).
- Спроводити тестове интеграције како би било разумљиво како појединачне компоненте комуницирају са осталим деловима Система.
- Детектовати клизање (енгл. *drift*) улазних података тестирањем статистичких вредности података који улазе у Систем, како би се осигурало да се они не мењају на непредвиђене начине.
- Користити скуп података који представља „златни стандард” (еталон) како би се осигурало да се Систем понаша како је предвиђено. Редовно ажурирати овај скуп података у складу с променама корисника и сценарија примене, како би се смањио ризик од тренирања на тестном скупу.

- Спроводити итеративно тестирање од стране корисника како би се, у различитим циклусима развоја, укључиле различите функционалности.
- Применити „пока-јарам” (eng. *poka-yoke*) приступ: у Систем уградити провере квалитета, тако да не долази до непредвиђених грешака или да оне дају тренутан одговор (на пример: ако нестане неко битно обележје, Систем неће понудити одговор).

5. Надзирати Систем током примене

Континуиран надзор обезбеђује да Систем функционише на предвиђен начин, узимајући у обзир повратне информације од стране корисника.

- Предвидети интервале за исправљање неправилности у раду Система.
- Размотрити краткорочна и дугорочна решења којима се отклањају неправилности. Уравнотежити краткорочна и дугорочна решења.
- Пре ажурирања модела у примени, анализирати разлике између модела у примени и предлога измењеног модела, као и како ће нови модел утицати на целокупни квалитет Система и искуство корисника.

Безбедност

Безбедност обезбеђује да се Системи понашају како је предвиђено, без обзира на потенцијалне нападе. Неопходно је утврдити безбедност Система пре примене у областима у којима је безбедност критичан параметар. Постоје бројни изазови по овом питању. На пример: тешко је унапред предвидети све сценарије, као и развити Системе који пружају и ограничења потребна са становишта сигурности али и флексибилност у креирању решења која се прилагођавају различитим улазним подацима.

С развојем технологија вештачке интелигенције појављују се и нови ризици од напада које би требало предвидети, као што су: манипулација подацима за тренинг, напади избегавања, приступ осетљивим тренинг подацима, крађа модела или замена (енг. *adversarial attacks*)²². Пре развоја Система размотрити ризике и последице напада, како би се донела одлука о даљем развоју Система.

1. Идентификовати потенцијалне претње

- Размотрити да ли ико има интерес да наведе Систем да се понаша на непредвиђен или штетан начин.
- Идентификовати које су нежељене последице грешке Система и проценити вероватноћу и тежину тих последица.
- Направити ригорозан модел претњи који ће предвидети што већи број напада. На пример, Систем који нападачу дозвољава да измени улазне податке Система рањивији је од Система који обрађује метаподатке прикупљене путем сервера јер је теже изменити оваква улазна обележја без директног приступа серверу.

2. Дефинисати процедуру за отклањање претњи

²² Истраживачи Мајкрософта и центра за интернет и друштво “Беркман Клајн”, универзитета Харвард покренули су иницијативу за формирање таксономије потенцијалних грешака у системима вештачке интелигенције, случајних и намерно изазваних, која је доступна на адреси <https://docs.microsoft.com/en-us/security/engineering/failure-modes-in-machine-learning>

- Тестирати перформансе Система применом различитих алата као што су: *CleverHans*, или *Adversarial Robustness 360 Toolbox -ART*.
- Формирати интерни тим који ће нападати Систем, или организовати такмичење за тестирање Система.
- Развити процедуру за отклањање различитих врста претњи.

3. Вршити континуирану едукацију

- Едуковати тим о најновијим врстама претњи и напада који се у пракси појављују.

4.3 Приватност, заштита података о личности и управљање подацима

Уско повезани с принципом спречавања настанка штете су приватност и заштита података о личности. Спречавање нарушавања приватности и права на заштиту података о личности захтева адекватно управљање подацима које, између осталог, укључује квалитет и интегритет података који се користе, њихову релевантност имајући у виду област друштвеног живота у којој ће се развијати и примењивати Систем, протоколе за приступ подацима, те способност Система да се подаци обрађују на начин који штити приватност и право на заштиту података о личности.

4.3.1 Упитник

- Да ли је анализиран утицај система вештачке интелигенције на право на приватност, право на физички, психички и/или морални интегритет, као и право на заштиту података о личности?
 - Да, и то:
 - Не, не обрађују се подаци о личности²³
 - Не
- У зависности од случаја коришћења система вештачке интелигенције, да ли је успостављен механизам којим се посебно означавају елементи система који утичу на приватност?
 - Да, и то:
 - Не
- Да ли ће припрема, тренирање, развој и коришћење система вештачке интелигенције захтевати обраду података о личности (укључујући посебне врсте података о личности)?
 - Да, и то:
 - Не
- Да ли постоји правни основ за (намеравану) обраду података о личности?²⁴
 - Да, и то:
 - пристанак лица
 - уговор с лицем
 - поштовање правних обавеза руковоаца
 - заштита животни важних интереса лица
 - обављање послова у јавном интересу/извршење законом прописаних овлашћења руковоаца
 - легитимни интерес руковоаца/трећег лица
 - Не

²³ Ова група питања се не попуњава, осим последња два

²⁴ Члан 12. Закона о заштити података о личности, „Службени гласник РС”, бр. 87/2018.

- Уколико се обрађују посебне врсте података о личности, да ли су испуњени услови за дозвољеност такве обраде?
 - Да
 - Не
 - Не обрађују се посебне врсте података о личности
- Да ли је одређена (оправдана и законита) сврха обраде података о личности?
 - Да, и то:
 - Не
- Да ли је област на коју се примењује систем вештачке интелигенције уређена посебним прописима и да ли се (намеравана) обрада података о личности врши у складу с тим посебним прописима?
 - Да, и то:
 - Не
- Да ли је спроведена нека од следећих мера (од којих су неке обавезне према Закону о заштити података о личности („Сл. гласник РС“, број 87/2018) и закону/пропису друге земље/ЕУ чија је примена обавезна у конкретном случају)?
 - извршена је процена утицаја обраде на заштиту података о личности
 - одређено је лице за заштиту података о личности које је укључено у развој, набавку или употребу система вештачке интелигенције
 - предвиђене су техничке, организационе и кадровске мере заштите података о личности (укључујући ограничен приступ подацима од стране овлашћених лица, механизме за евидентирање/бележење приступа подацима и уношење измена и др)
 - предвиђени су механизми за постизање уграђене и подразумеване приватности (енг. *privacy by design & privacy by default*) као што су: енкрипција, псеудонимизација, агрегација, анонимизација и др.
 - омогућено је спровођење начела обраде података о личности, укључујући начело свођења података на минимум, у односу на конкретне податке (укључујући и посебне врсте података о личности)
- Да ли је у систему вештачке интелигенције омогућено да лице да опозив пристанка на обраду података о личности, приговор и брисање података о личности по захтеву лица на које се подаци односе?
 - Да
 - Не
- Да ли су узете у обзир последице обраде података о личности по приватност лица и право на заштиту података о личности, током целокупног животног циклуса система вештачке интелигенције?
 - Да
 - Не
- Да ли постоје последице по приватност и заштиту података о личности услед обраде других података у оквиру система вештачке интелигенције који нису подаци о личности?
 - Да
 - Не

- Да ли је систем вештачке интелигенције усклађен са одговарајућим стандардима²⁵ или другим општеприхваћеним протоколима за управљање подацима?
 - Да, и то са:
 - Не

4.3.2 Препоруке

Управљање подацима обезбеђује тачност, безбедност и доступност података ради очувања квалитета и њихове заштите. Руководилац подацима обавезан је да заштити податке у Систему. Неопходно је на законит начин обезбедити приступ подацима поштујући приватност појединца, а све у складу с прописима о заштити података, нарочито података о личности

1. Управљање подацима обухвата:

- дефинисање делова података и уноса података за креирање заједничког пословног речника у пословном речнику
- идентификовање атрибута података (метаподатака) и начина уноса података
- дефинисање корисничких улога и начина аутентикације и ауторизације приступа подацима
- процесе управљања подацима
- правила и норме која дефинишу процес управљања подацима током целог животног циклуса
- управљање шифарницима, како би се исте класификације користиле у свим оперативним и аналитичким системима
- технологију* која омогућава управљање структурираним, мултиструктурираним и неструктурираним подацима у свим окружењима инфраструктуре.

*Технологије управљања подацима су: каталог података, софтвер за интеграцију и управљање подацима (eng. *data fabric software*), „складиште података” (eng. *data lake*) и управљање шифарницима и класификацијама.

Каталог података обухвата:

- пословни речник
- аутоматско откривање података, профилисање, обележавање, каталогизацију и мапирање речника појмова
- аутоматску детекцију осетљивих података и класификацију управљања
- интероперабилност с другим шифарницима, алатима и апликацијама ради размене података.

Софтвер за интеграцију и управљање подацима обухвата:

- изворе података, „*multi-cloud*” и „*edge data*” повезивање
- алате за чување података
- исправку и интегрисање података
- метаподатке
- обезбеђивање заштите података

²⁵ на пример: ISO (ISO 27001 и др), IEEE.

- безбедност приступа подацима у вишеструким складиштима података у дистрибуираном окружењу података

Складишта података подржавају енкрипцију података, анонимизацију и псеудонимизацију података и повезивање с каталогом података.

Нетехнолошки алати су:

- Примена законског оквира који уређује ко, када и с којом сврхом обрађује које податке: транспарентно информисање о сврси и начину рада Система, начину обраде података и другим питањима значајним за заштиту података.
- Промовисање коришћења безбедних оперативних окружења као што је инфраструктура у Државном центру за управљање и чување података.
- Контрола од стране лица чији се подаци користе поштујући права појединца.
- Ограничавање приступа подацима који имају одређени степен поверљивости.
- Професионално управљање од стране лица обучених за етичко коришћење података, којим се успоставља равнотежа између бриге за јавно добро и ризика који проистичу из обраде података, уз блиску сарадњу са истраживачима и стручном заједницом.

4.4 Транспарентност

Транспарентност се дефинише²⁶ као:

- степен у којем Систем открива процесе или параметре који се односе на његово функционисање; и
- својство које омогућава да се открије како и зашто је Систем донео одређену одлуку или поступио на начин на који је то учинио, узимајући у обзир своје окружење.

Транспарентност је важна из најмање три разлога: 1) аутономни и интелигентни системи (АИС) могу да погреше или начине штету, а транспарентност је неопходна да би се открило како и зашто; 2) АИС треба да буде разумљив корисницима и 3) без адекватне транспарентности одговорност је немогућа.

Једна од карактеристика интелигентних система је аутономност. Аутономни систем се може дефинисати као „систем који има капацитет да сам доноси одлуке, као одговор на неке улазне податке или стимулансе, с различитим степеном људске интервенције у зависности од нивоа аутономије система“. Системи показују одређени ниво аутономности у интеракцији са окружењем. Већ дужи низ година, работи се користе у системима за разврставање артикала и контролу квалитета, као и за управљање индустријским складиштима с производима. Интелигентни агенти се користе за примену финансијских сервиса, смањење ризика од људске грешке и, на тај начин, за побољшање перформанси трговања на берзи. Међутим, даљи развој и примена система вештачке интелигенције у неким доменима као што су здравство, фармација и право, зависиће од поступка и могућности за следљивост процеса доношења одлука и акција, техника за тумачење и објашњивост резултата, и начина интеракције корисника са системима вештачке интелигенције и презентовања резултата крајњим корисницима.

²⁶ A.F.T. Winfield et al., (2021), *IEEE P7001: A Proposed Standard on Transparency*, *Front. Robot. AI*, <https://doi.org/10.3389/frobt.2021.665729>.

Транспарентност је кључна компонента која доприноси изградњи поуздане вештачке интелигенције достојне поверења која обухвата три елемента: 1) могућност праћења – следљивост (енг. *traceability*) система вештачке интелигенције 2) објашњивост (енг. *explainability*) система вештачке интелигенције, нарочито модела система и 3) комуникацију – дијалог са свим заинтересованим странама о ограничењима система вештачке интелигенције

4.4.1 Упитник

Могућност праћења – следљивост

Праћење организацијама омогућава да процене да ли су процеси развоја система вештачке интелигенције, односно подаци, процедуре и процеси који утичу на одлуке засноване на вештачкој интелигенцији документовани тако да дозвољавају праћење, повећавају транспарентност и граде поверење друштва у вештачку интелигенцију.

- Да ли су успостављене мере за праћење система вештачке интелигенције током његовог читавог „животног“ циклуса?
 - Да
 - Не
- Да ли постоји техничка документација (досје/портфолио) система вештачке интелигенције који се правовремено ажурира и да ли се та документација чува у складу са законом²⁷?
 - Да
 - Не
- Да ли су успостављене мере за континуирано праћење квалитета улазних података система вештачке интелигенције?
 - Да, и то:
 - квантификовање недостајућих вредности
 - истраживање прекида у дотоку података
 - детектовање случаја када подаци нису довољни за извршење задатка
 - препознавање улазних података који садрже грешку, нису исправни, нису тачни или нису одговарајућег формата
 - друго:
 - Не
- Да ли је могуће ретроактивно евидентирати који су подаци коришћени од стране система вештачке интелигенције да би се донела одређена одлука(е) или препорука(е)?
 - Да
 - Не
- Да ли је могуће ретроактивно евидентирати који је модел или правило коришћено за доношење одређене одлуке(а) или препоруке(а)?
 - Да
 - Не
- Да ли су успостављене мере за континуирану процену квалитета излазних резултата система вештачке интелигенције?
 - Да, и то:
 - провером да ли су добијени резултати у оквиру очекиваног распона

²⁷ Закон о електронском документу, електронској идентификацији и услугама од поверења у електронском пословању, „Службени гласник РС”, бр. 94/2017 и 52/2021.

- откривањем неправилности у излазним резултатима
- прерасподелом улазних података који су довели до неправилности
- друго:
- Не
- Да ли се у систему вештачке интелигенције врши евидентирање приступа систему од стране корисника при доношењу одлука и препорука?
 - Да
 - Не
- Да ли се у систему вештачке интелигенције врши евидентирање метаподатака (датум и време почетка и завршетка коришћења система, базе података која је коришћена као референтни извор података у систему и др) о начину коришћења система?
 - Да, и то:
 - Не
- Да ли се у систему вештачке интелигенције врши евидентирање приступа систему од стране лица која контролишу процес доношења одлука?
 - Да
 - Не
- Да ли је креиран документ који на јасан начин објашњава модел Система на начин да пружа информације о: 1) сврси алгорита, 2) скупу података на којима је модел трениран, 3) извору и начину прикупљања података, 4) својствима алгорита?
 - Да, и то за: (означи број)
 - Не

Објашњивост

Наредна група питања омогућава процену степена разумевања система вештачке интелигенције, односно начина и разлога због којих је систем пројектован на одређен начин. Тиме се развија поверење корисника у системе вештачке интелигенције. Одлуке које се доносе коришћењем система вештачке интелигенције морају бити објашњене и разумљиве онима на које непосредно или посредно утичу, у циљу евентуалног оспоравања таквих одлука. Објашњење тога зашто је модел предложио одређену одлуку/резултат (која комбинација улазних фактора је утицала на тај резултат) није увек могуће. Овакви случајеви односе се на тзв. „црне кутије“ и захтевају додатну пажњу, односно другу врсту мера за постизање објашњивости (на пример, могућност праћења, екстерне евалуације^[1] и транспарентне комуникације о думетима и могућностима система вештачке интелигенције), наравно под претпоставком да систем вештачке интелигенције као целина поштује основна људска права. Објашњивост зависи од контекста, а нарочито од процене могућих негативних последица грешке/нетачних резултата на човеков живот.

- Да ли је корисницима објашњен начин на који систем вештачке интелигенције предлаже одлуке?
 - Да, и то:
 - у корисничком упутству (видео, аудио, документ и сл)
 - организовањем радионица, обука и сл.
 - Не
- Да ли је успостављен механизам за праћење степена разумевања корисника (оптималног обима/степенa објашњења)?
 - Да, и то:

- Не
- Да ли се континуирано прати и анализира разумевање корисника ради организовања додатних обука или вршења одговарајућих корекција система вештачке интелигенције?
 - Да, и то:
 - Не

Комуникација

Наредна група питања омогућава процену тога да ли је систем вештачке интелигенције у складу са специфичним случајевима коришћења, односно могућностима и ограничењима система. Ово може обухватити и обавештавање о степену тачности и ограничењима система вештачке интелигенције.

- У случају коришћења интерактивних система вештачке интелигенције (чет-ботови, робо-адвокати), да ли су корисници обавештени да су у интеракцији са системом вештачке интелигенције, а не с људским бићем?
 - Да, на јасан и транспарентан начин, већ у првом кораку, при сваком приступу систему
 - Не
- Да ли је кориснику који не жели да комуницира са системом вештачке интелигенције обезбеђен други начин комуникације?
 - Да и то: (навести начин)
 - Не
- Да ли је успостављен механизам за информисање о намени и критеријумима на основу којих се доноси одлука?
 - Да, и то:
 - Не
- Да ли су корисници информисани о предностима коришћења система вештачке интелигенције?
 - Да, и то:
 - Не
- Да ли су корисници обавештени о техничким ограничењима и потенцијалним ризицима система вештачке интелигенције значајним за доношење одлуке (нпр. који је ниво тачности и/или распон могућих грешака система)?
 - Да, сва ограничења су наведена у посебном одељку
 - Не
- Да ли је припремљен материјал за обуку за адекватно коришћење система вештачке интелигенције?
 - Да
 - Не

4.4.2 Препоруке

Могућност праћења – следљивост сматра се кључним захтевом за стварање поузданих и одговорних Система. Нове информационо-комуникационе технологије као што су: интернет ствари (енг. *internet of things*), рачунарство у облаку (енг. *cloud computing*) и мобилно рачунарство омогућиле су даљи развој приступа за обраду великих количина података (*eng. big data*) и

алгоритама вештачке интелигенције. За разумевање и тумачење информација садржаних у скуповима података, потребно је извући важне чињенице и резоновати користећи знање и/или теорију вероватноће. Стога, Смернице разликују следљивост на нивоу:

- порекла, приступа и екстракције података,
- алгоритама и модела за машинско учење,
- процеса за аутоматску припрему и обраду података и процеса закључивања из идентификованих улазних фактора и излазних препорука релевантних за решавање проблема.

Објашњивост се дефинише као „мера у којој су унутрашње стање и процеси доношења одлука у аутономном систему доступни заинтересованим странама, укључујући и крајњег корисника система“. У случајевима када Система има значајан утицај на животе људи, захтева се одговарајуће објашњење процеса доношења одлука које је благовремено и прилагођено стручности заинтересоване стране (нпр. лаика, регулатора или истраживача).

Када су основни елементи за изградњу Система нетранспарентни (на пример: модел и процес учења), одлука заснована на вештачкој интелигенцији није интуитивна и објашњива. За многе „нетехничке“ кориснике, интелигентни програм заснован на алгоритмима за машинско учење²⁸ је „црна кутија“, на пример: неуронске мреже за препознавање образаца.

Овај феномен црне кутије доводи до тога да корисници доводе у питање одлуке Система: Зашто сте то урадили? Зашто је ово резултат? Када сте успели или нисте постигли успех? Када могу да верујем? Овај рефлексивни скептицизам непосредно утиче на поверење корисника и ефикасност доношења одлука, чиме утиче и на усвајање решења вештачке интелигенције, укључујући финансијске и правне одлуке, медицинске дијагнозе, праћење индустријских процеса, безбедносни скрининг, запошљавање, правне пресуде, упис на универзитет, паметне куће и аутономна возила.

Објашњивост²⁹ се односи на оне технике вештачке интелигенције које корисницима Система (инжењерима вештачке интелигенције, крајњим корисницима и ревизорима) помажу да разумеју разлоге због којих модел производи своје резултате. Даље, објашњивост се односи и на оне технике који могу да обезбеде транспарентност у вези са улазним подацима, као и „разлогом“ који стоји иза употребе алгоритма који води до специфичног излаза. Сам алгоритам у овом случају не мора нужно бити откривен. Штавише, корак даље ка поузданој вештачкој интелигенцији је одговорна вештачка интелигенција, која поред објашњивости укључује и остале принципе које треба испунити приликом примене Система у практичним сценаријима: правичност (*eng. fairness*), усмереност на човека (енг. *human-centric*), свест о приватности (енг. *privacy awareness*, одговорност (енг. *accountability*), безбедност и сигурност (енг. *safety and security*).

Док се поједини модели (статистички и дрва одлучивања) могу мапирати у правила и на тај начин обезбедити интерпретабилност резултата, то није случај с дубоким неуронским мрежама, које су нашле ширу примену нашле последњих година због све веће количине доступних података

²⁸ Franco-German position paper on "Speeding up Industrial AI and Trustworthiness", <https://cris.vtt.fi/en/publications/franco-german-position-paper-on-speeding-up-industrial-ai-and-tru>

погодних за машинско учење. Последњи трендови везани за објашњиву вештачку интелигенцију подразумевају учења модела који су лакше објашњиви, коришћење мрежа неуралне логике, увођење интерпретабилних модела, коришћење графикана знања,³⁰ итд.

Комуникација

Да би се обезбедило поштовање основних права и усклађеност са основним људским правима, а посебно правом на обавештеност, Системи морају бити препознатљиви као такви. Корисници морају бити обавештени да су у интеракцији са Системом, а, када је то потребно, и да имају избор да захтевају комуникацију с човеком.

Корисност, делотворност, ефикасност и употребљивост Система обезбеђује се укључивањем крајњих корисника у дизајн, евалуацију и примену графичког корисничког интерфејса.

Да би се обезбедила прецизна и експлицитна веза између апстрактних принципа које је Систем дужан да поштује и конкретних одлука о реализацији, примењује се препорука - етика по дизајну (енг. *ethics-by-design*). Већ се користе концепти „по дизајну”, на пример приватност по дизајну и безбедност по дизајну. Да би имали поверење у Систем неопходно је да буде разумљив и објашњив лицима која га примењују и користе, као и безбедан у свим својим процесима.

4.5 Различитост, недискриминација и равноправност

Како би систем вештачке интелигенције био поуздан, потребно је омогућити укључивање (инклузивност) и различитост током целог животног циклуса система. Системи вештачке интелигенције могу имати одређене недостатке због нпр. некомплетности и модела лошег управљања, што може довести до ненамерних (не)посредних прејудицирања и дискриминације против одређених група, потенцијално утичући на погоршање одређених предрасуда и додатну маргинализацију рањивих група. Системи вештачке интелигенције треба да буду усмерени на корисника и створени на начин који свакоме омогућава коришћење производа или услуга вештачке интелигенције, без обзира на њихов узраст, пол, могућности или карактеристике. Од посебног је значаја приступачност ове технологије за особе са инвалидитетом, које су присутне у свим друштвеним групама.

4.5.1 Упитник

Спречавање предрасуда

- Да ли постоје дефинисане процедуре за спречавање креирања или подстицања предрасуда у систему вештачке интелигенције, како при избору улазних података, тако и креирању алгорита?
 - Да, узете су у обзир различитост и репрезентативност крајњих корисника и/или лица
 - Да, тестиран је систем за све циљне групе, нарочито за осетљиве категорије
 - Да, приликом развоја система вештачке интелигенције коришћени су посебни алати који омогућавају боље разумевање података, модела и перформанси
 - Да, дефинисане су процедуре за тестирање и праћење потенцијалних предрасуда током читавог животног циклуса система

³⁰ High-level Expert Group on Artificial Intelligence, Ethics Guidelines for Trustworthy AI, <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

- Не
- Да ли се врши едукација и подизање свести лица која учествују у стварању система вештачке интелигенције (дизајнери, програмери и др) како да препознају дискриминацију и предрасуде?
 - Да, и то:
 - Не
- Да ли постоји механизам који у систему вештачке интелигенције препознаје предрасуде и дискриминацију?
 - Да, успостављен је механизам детекције, односно пријаве и предвиђен је поступак решавања по пријавама
 - Да, идентификовано је на која лица, поред крајњих корисника, систем може да утиче
 - Не
- Да ли се дефиниција равноправности користи и примењује у свим фазама процеса успостављања система вештачке интелигенције?
 - Да, уз претходно разматрање више различитих дефиниција
 - Да, уз претходно консултовање свих група на које систем има утицаја
 - Да, уз претходно извршено тестирање коришћења дефиниције
 - Не

Пристапачност и јединствени (универзални) дизајн

Нарочито у доменима „привреда ка потрошачу“, системи вештачке интелигенције треба да буду усмерени на корисника и дизајнирани на начин који омогућава свим људима да користе производе или услуге вештачке интелигенције, без обзира на њихов узраст, пол, могућности или карактеристике. Пристапачност ове технологије за особе са инвалидитетом, које су присутне у свим друштвеним групама, од посебног је значаја. Системи вештачке интелигенције не би требало да имају приступ „калупа који одговара свима“, већ би требало да размотре принципе јединственог (универзалног) дизајна који би одговарао најширем могућем кругу корисника, пратећи релевантне стандарде пристапачности. Ово ће омогућити једнак приступ и активно учешће свих људи у постојећим и надлазећим људским активностима у којима компјутери посредују а у вези са асистивним технологијама.

- Да ли систем вештачке интелигенције одговара различитим афинитетима и могућностима?
 - Да
 - Не
- Да ли систем вештачке интелигенције могу да користе особе са инвалидитетом или друге посебно осетљиве групе, као и маргинализоване групе (могућност избора више од једног одговора)?
 - Да, током дизајнирања система консултоване су разне групе корисника за које је потребно обезбедити посебне механизме пристапачности
 - Да, корисничком интерфејсу су доступни читачи екрана и друге помоћне технологије како би се омогућило коришћење система од стране наведених лица?
 - Не
- Да ли је током развоја система обезбеђен принцип универзалног дизајна, ако је применљиво?
 - Да
 - Не, јер није релевантан за систем
 - Не

- Да ли је извршена процена примене система вештачке интелигенције на групе које би могле бити погођене исходима примене система?
 - Да
 - Не

Учешће заинтересованих страна

Да би се развила поуздана вештачка интелигенција, препоручљиво је консултовати заинтересоване стране на које систем вештачке интелигенције може непосредно или посредно утицати током свог животног циклуса. Корисно је тражити редовне повратне информације чак и након пуштања у рад (примене) система, и успоставити дугорочне механизме за учешће заинтересованих страна, на пример обезбеђивањем информација о радницима, консултација и учешћа током целог процеса примене система вештачке интелигенције у организацијама.

- Да ли су приликом развоја система вештачке интелигенције консултоване све заинтересоване стране?
 - Да, и то:
 - Не
- Да ли су запослена/ангажована лица експлицитно упозната с непосредним или посредним ефектима које вештачка интелигенција има или може да има на њихов радни статус и права?
 - Да, и то:
 - Не
- Да ли су запослена/ангажована лица имали обуку која им пружа знање о функционисању вештачке интелигенције и њеном утицају на њихов радни статус и права?
 - Да, и то:
 - Не
- Да ли је запосленим/ангажованим лицима и синдикатима омогућено да траже и добију податке о високоризичним системима вештачке интелигенције који утичу на њихов радни статус и права у вези са тим?
 - Да
 - Не
- Да ли ће се након пуштања у рад система спроводити периодична испитивања корисника система, а у циљу контроле истог?
 - Да, и то:
 - Не

4.5.2 Препоруке

Важан услов поуздане и одговорне вештачке интелигенције је њено недискриминаторно понашање које уважава различитост и доприноси равноправности. Препоруке за постизање нивоа различитости, недискриминације и равноправности у смислу етичких принципа изложених у овим Смерницама су:

- Анализирати Систем у реалном времену чиме се откривају намерне и ненамерне пристрасности и дискриминаторни обрасци. Када пристрасност (дискриминаторни

образац) у подацима постане очигледна, тим мора да истражи и разуме одакле потиче и како се може ублажити (пожељно отклонити у целости).

- Развити Систем без намерних пристрасности уз редовно ревидирање Система како би се пристрасности избегле. Ненамерне пристрасности укључују и стереотипе.
- Проверити податке и изворе тих података пре започињања тренинга алгорита.
- Укључити механизам за обезбеђивање повратне информације с лицима која користе/примењују Систем како би се подигла свест о пристрасностима или проблемима које идентификују та лица.
- Формирати мултидисциплинарне тимове за оцену предметних параметара. Различити тимови помажу у представљању шире варијације искустава како би се пристрасност па и дискриминација свеле на минимум.
- Обезбедити објективност и механизам за отклањање пристрасности.
- Ако се Систем покаже као неадекватан, пристрасан, да доноси дискриминаторне одлуке или је уопштено неуспешан а нема услова за унапређење, обуставити његов рад. Проценити штету коју Систем може да начини друштву и појединцу спрема штете која ће настати повлачењем из рада таквог Система.
- Укључити у рад тима чланове различитих узраста, националности, пола, образовних дисциплина и културних перспектива. Различитост те врсте даје приступ варијацијама искустава како би се пристрасност свела на минимум.
- Тестирати Систем од ране фазе дизајна и често.

*Пример различитости, недискриминације и равноправности.*³¹ Након што се састао с члановима глобалног менаџмента хотела, тим за развој Система открива да су различитост и инклузивност важни елементи вредности хотела. Као резултат тога, тим осигурава да прикупљени подаци о раси, полу, итд. корисника у комбинацији са њиховом употребом Система, неће бити коришћени за оглашавање или искључивање одређених демографских категорија. Тим је преузео скуп података о гостима хотела. Након анализе ових података и примене у градњи агента, схватају да постоји степен алгоритамске пристрасности података. Тим наставља да одваја време да додатно обучи модел на већем, разноврснијем скупу података како би обезбедио недискриминацију и равноправност друштвених категорија.

4.6 Друштвено и еколошко благостање

У складу с принципима правичности и превенције штете, током животног циклуса система вештачке интелигенције неопходно је сагледати његов утицај на друштво и животну средину. Изложеност системима вештачке интелигенције у свим сегментима живота (образовању, послу или забави) може променити начин деловања појединца или негативно утицати на друштвене односе. Ефекти примене система вештачке интелигенције морају се континуирано пратити и преиспитивати. Потребно је подржати истраживања који укључују развој система вештачке интелигенције која позитивно утичу на заштиту животне средине.

³¹ Пример преузет са <https://www.ibm.com/design/ai/ethics/fairness/#ai-must-be-designed-to-minimize-bias-and-promote-inclusive-representation>

4.6.1 Упитник

Заштита животне средине

- Да ли постоје потенцијални негативни утицаји система вештачке интелигенције на животну средину?
 - Да, и то:
 - Не
- Да ли су успостављени механизми за процену утицаја развоја, примене и/или коришћења система вештачке интелигенције на животну средину (на пример, количина употребљене енергије и емисије угљеника)?
 - Да, и то:
 - Не
- Да ли су дефинисане мере за смањење утицаја система вештачке интелигенције на животну средину током његовог животног циклуса?
 - Да, и то:
 - Не

Утицај на рад и вештине

- Да ли систем вештачке интелигенције утиче на радно ангажовање и начин рада?
 - Да
 - Не
- Да ли су пре увођења система вештачке интелигенције о томе обавештени и консултовани запослени/ангажовани на чији ће рад увођење система утицати, као и на њихове представнике (синдикате и сл)?
 - Да
 - Не
- Да ли су предузете одговарајуће мере које ће обезбедити разумевање утицаја система вештачке интелигенције на начин рада запослених/ангажованих?
 - Да, и то:
 - Не
- Да ли коришћење система вештачке интелигенције ствара ризик од деквалификације запослених/ангажованих?
 - Да
 - Не
- Да ли су предузете одговарајуће мере да се спречи ризик од губитка вештина?
 - Да, и то:
 - Не
- Да ли коришћење система вештачке интелигенције промовише или захтева нове (дигиталне) вештине?
 - Да
 - Не
- Да ли су припремљени корисничко упутство и други материјали неопходни за усавршавање запослених/ангажованих?
 - Да
 - Не

- Да ли се реализују обуке запослених/ангажованих?
 - Да
 - Не

Утицај на друштво

- Да ли систем вештачке интелигенције може имати негативан утицај на друштво у целини или на демократију?
 - Да
 - Не
- Да ли је процењен посредан утицај коришћења система вештачке интелигенције на све заинтересоване стране или друштво у целини?
 - Да, извршена је процена утицаја
 - Не
- Да ли су предузете одговарајуће мере које смањују потенцијалан штетни утицај на друштво?
 - Да, и то:
 - Не
- Да ли су предузете мере које обезбеђују да систем вештачке интелигенције не утиче негативно на демократију?
 - Да, и то:
 - Не

4.6.2 Препоруке

Током целокупног животног циклуса Система неопходно је проматрати његов утицај на друштво и животну средину. Примена Система може негативно утицати на друштвене односе у различитим областима живота (образовање, рад, слободно време, разонода и слично) те је од изузетне важности континуирано процењивати, пратити и преиспитивати ефекте примене Система на човека и друштво у целини као и животну средину са којом је човек у нераскидивој вези.

Ради постизања овог Услови, препорука је да се за Систем:

- Успостави стандардизован приступ процене утицаја на људе, организације, цело друштво, демократију и природну средину.
- Обави процена утицаја уз учешће селекованих индивидуа које су задужене за процене.
- У процене утицаја Система укључе и ефекти ограничења Система или рестриктивног или осетљивог коришћења.
- Периодично обнове постојеће процене утицаја услед промена у Систему или услед промене или проширења сврхе коришћења Система.
- Дефинишу и документују методе које ће да се користе да би се обавестила или информисала лица која имају било какву интеракцију са Системом (а не са особом), или да као резултат обраде Система буде генерисан резултат (нпр. фотографија).
- Идентификују и приоритизују демографске групе које могу да добију нижи квалитет сервиса у зависности од демографске групе којој припадају или услед комбинације других фактора.

- Анализирају изворни подаци да би се проценила инклузивност свих демографских група – документовати које групе нису покривене и прикупити додатне податке који решавају тај проблем.
- Дефинишу и документују процене равноправности/инклузивности употребе Система за све демографске групе.
- Дефинишу и документују минимални и максимални критеријуми које процена треба да задовољи ради одговорног стављања Система у оперативно коришћење (фазу продукције).
- Упореди процена са успостављеним критеријумима, и у случају да минимална очекивања нису постигнута, успоставе опције за превазилажење; Потенцијално консултовати стручњаке у референтној области како би се осигурало да су решења прихватљива и у складу с регулативом

4.7 Одговорност

Питање одговорности је уско повезано с правилним планирањем развоја и надзора над Системом у фази продукције, управљањем ризицима и процедурама за утврђивање одговорности и санације штете која настане као последица примене Система.

Не искључујући одговорност других лица у ланцу креирања и продукције Система, дизајнери и програмери Система су лица с високим степеном одговорности приликом разматрања дизајна, развоја, процеса одлучивања и исхода Система. То не значи да поменута лица немају право или не треба да раде у мултидисциплинарном тиму, напротив.

Људска логика и просуђивање су кључни фактори у животу Система који по претпоставци доноси објективно-логичке одлуке, јер су људи ти који пишу алгоритме, дефинишу успех или неуспех, припремају податке и скупове података, тренирају моделе, врше евалуацију, доносе одлуке о употреби Система. Зато је важно да човек увек буде свестан тога и да тежи том аксиому без обзира на околности под којима развија Систем. Сва лица укључена у стварање Система у било ком кораку одговорна су за разматрање његовог утицаја у окружењу у коме ће бити примењен, као и компаније које су инвестирале у његов развој.

4.7.1 Упитник

Могућност ревизије

Овај одељак помаже у самооцењивању постојећег или неопходног нивоа који би био потребан за процену система вештачке интелигенције, од стране интерних и екстерних ревизора. Могућност спровођења евалуација као и приступ подацима о наведеним евалуацијама може допринети поузданој вештачкој интелигенцији. У софтверским решењима која утичу на основна права човека, укључујући апликације које су критичне за безбедност, требало би да постоји могућност да системи вештачке интелигенције буду у могућности да буду независно ревидирани. Ово не значи нужно да информације о пословним моделима и интелектуалној својини у вези са системом вештачке интелигенције морају увек бити отворено доступне.

- Да ли је успостављен механизам којим се омогућава ревизија система вештачке интелигенције?
 - Да, кроз следљивост процеса развоја

- Да, праћењем података којим се тренира систем и евидентирањем исхода
- Да, евидентирањем позитивних и негативних утицаја коришћења система
- Не
- Да ли је омогућено да (независно) треће лице може да изврши ревизију система вештачке интелигенције?
 - Да, евидентирањем позитивних и негативних утицаја коришћења система
 - Не

Управљање ризицима

Управљање ризицима захтева идентификовање, процену, документовање и свођење на минимум потенцијалних негативних утицаја система вештачке интелигенције. Узбуњивачима, организацијама цивилног друштва, синдикатима и другим субјектима мора бити доступна одговарајућа заштита када пријављују оправдану забринутост због коришћења система вештачке интелигенције.

Ово захтева да треба да буду идентификовани релевантни интереси и вредности које систем вештачке интелигенције подразумева и да, ако дође до сукоба, треба да буду експлицитно признати и оцењени компромиси у смислу њиховог ризика по безбедност и етичке принципе, укључујући основна права. Свака одлука о томе који компромис треба да буде начињен, треба да буде добро образложена и правилно документована. Када дође до штетног утицаја, треба предвидети доступне механизме који обезбеђују адекватну накнаду.

- Да ли је дефинисан процес ревизије од стране трећих лица који обухвата испитивање етичких питања и мера праћења одговорности?
 - Да, и то:
 - Не
- Да ли је организована обука о ризицима коришћења система вештачке интелигенције и одговарајућем правном оквиру?
 - Да, и то:
 - Не
- Да ли имате одбор који се бави етичким питањима примене система вештачке интелигенције, или сличан механизам којим се обезбеђује дискусија о одговорности и етичким праксама?
 - Да, и то:
 - Не
- Да ли је успостављен механизам за дискусију, континуирано праћење и процену доследности примене ових Смерница за систем вештачке интелигенције?
 - Да
 - Не
- Да ли предвиђени процес укључује идентификацију и документовање неусаглашених ставова у вези с различитим етичким принципима и објашњењима одлука?
 - Да
 - Не
- Да ли је спроведена обука лица укључених у процес?
 - Да

- Не
- Да ли је трећим лицима (добављачима, крајњим корисницима, дистрибутерима/продавцима и др) омогућена пријава потенцијалних рањивости, ризика и/или дискриминације у систему вештачке интелигенције?
 - Да
 - Не
- Да ли евентуалне пријаве рањивости, ризика и дискриминације захтевају ревизију процеса управљања ризиком?
 - Да
 - Не
- У случају чињења штете појединцима да ли је успостављен механизам за обештећење?
 - Да
 - Не

4.7.2 Препоруке

Препоруке за постизање и унапређење одговорности су:

- Креирати правила и политике које су прецизне и приступачне дизајнерима и развојним тимовима, како не би било спорних питања у вези са задужењима и одговорностима.
- Одредити где престаје одговорност оних који су развили Систем. Ово је посебно важно јер они који су развијали Систем неће имати контролу тога како се он користи.
- Водити евиденцију о процесима дизајнирања, развоја функционалности и начина на који Систем доноси одлуке. Уредити ову процедуру посебним документом.
- Ускладити коришћење и резултат рада Система с прописима и међународним стандардима приликом креирања Система.
- Консултовати релевантна тела и органе у Републици Србији у вези с питањима која стварају недоумице ради добијања стручног савета или помоћи оних који раде на креирању политике вештачке интелигенције у Републици Србији или институција које врше контролу система вештачке интелигенције.³²
- У рад укључити лица која би помогла да се разумеју правна и етичка питања у холистичком приступу (нпр. социолога, лингвисту, бихевиористу, професоре и сл.).

*Пример одговорности.*³³ Тим користи истраживаче дизајна да контактирају праве госте у хотелима како би путем интервјуа лицем у лице с корисницима разумели њихове жеље и потребе. Тим прихвата сопствену одговорност у ситуацији када повратне информације хотелског асистента не испуњавају потребе или очекивања гостију. Они су применили петљу учења с повратном информацијом како би боље разумели преференције и истакли могућност да гости искључе вештачку интелигенцију у било ком тренутку током свог боравка.

5. ЗАКЉУЧАК

Етичке смернице за развој, примену и употребу поуздане и одговорне вештачке интелигенције припремљене су с намером да пруже оквир и усмере рад свих учесника у екосистему вештачке

³² За више информација, посетити сајт <https://www.ai.gov.rs/>

³³ Пример преузет са: <https://www.ibm.com/design/ai/ethics/accountability/>

интелигенције. У недостатку чвршћег правног оквира који прве обресе добија у Европској унији, ове Смернице омогућавају даљи развој у овој области чија се експанзија тек очекује. Водећи се начелима и принципима наведеним у Смерницама, наглашава се да вештачка интелигенција треба да се користи за добробит целе заједнице. Системи вештачке интелигенције треба да служе за одржавање и неговање демократских процеса и поштовање плуралитета вредности и животних избора појединаца. Смернице дају основ за ширу примену вештачке интелигенције у доношењу одлука при обликовању социјалних промена, повећавању знања и даљем економском напретку друштва у целини.