ЕТИЧКЕ СМЕРНИЦЕ

ЗА РАЗВОЈ, ПРИМЕНУ И УПОТРЕБУ ПОУЗДАНЕ И ОДГОВОРНЕ ВЕШТАЧКЕ ИНТЕЛИГЕНЦИЈЕ

Београд, децембар 2022

Садржај

1. Увод

1.1 Разлог доношења Смерница

Етичке смернице за развој, примену и употребу поуздане и одговорне вештачке интелигенције (у даљем тексту: Смернице) имају за циљ да омогуће да се наука, посебно у области вештачке интелигенције, развија и напредује али да не дозволи 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 и са добробити животиња.

Вештачка интелигенција је један од стубова четврте индустријске револуције. Као правац развоја компјутерске науке и инжењерства почела је да се развија пре неколико деценија са периодима успона и стагнација. Захваљујући продору у области неуронских мрежа, све веће количине доступних података погодних за машинско учење, као и све веће доступности микропроцесора погодних за обимна нумеричка израчунавања, у последњих неколико година кренуо је нагли развој и ширење примене вештачке интелигенције у областима здравства, финансија, образовања, енергетике и др.

Развој (система) вештачке интелигенције усмерен је у правцу стварања решења која ће испуњавати одговарајуће стандарде током читавог 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 у складу са утврђеним етичким принципима и вредностима. Свака од наведене три компоненте се посматра засебно; услови оцене и исход саме оцене једне од компонената, не претпоставља 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.

Циљ доношења Смерница је да се не дозволи да процеси у којима учествује систем вештачке интелигенције угрозе или 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.

1.2 Основ доношења

Основ за доношење Смерница садржан је у Стратегији развоја вештачке интелигенције у Републици Србији за период 2020–2025. године[1] која је као један од својих пет циљева поставила етичку, безбедну примену вештачке интелигенције и одредила ову активност у Акционом плану за период 2020-2022. године[2].

Како би се остварио овај циљ, неопходно је развити и увести механизме који ће омогућити одговоран развој вештачке интелигенције и проверу да ли су ови системи у складу са највишим

[1] „Службени гласник РС”, бр. 96/2019

[2] „Службени гласник РС”, бр. 81/2020

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, са једне стране и Републике Србије са друге стране (у даљем тексту: Споразум о стабилизацији), Република Србија се обавезала да обезбеди постепено усклађивање постојећих закона и будућег законодавства са правним тековинама Заједнице које свакако чине и правни акти Заједнице. Ова обавеза је потврђена и одредбама Устава Републике Србије, а нарочито чланом 194 .

Европска комисија је, у априлу 2021. године, поднела Европској унији предлог регулаторног оквира вештачке интелигенције – Предлог Уредбе Европског парламента и Савета о утврђивању усклађених правила о вештачкој интелигенцији[5] 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 у складу, 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 алгоритмичких процеса, као и Студију о димензији 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, где се разматрају импликације технологија вештачке интелигенције на људска права у информационом друштву.

Како би се обезбедило постепено усклађивање законодавног оквира са 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 налаже 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 правна и физичка лица којима су поверена јавна овлашћења, када у свом раду успостављају и користе системе

[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

вештачке интелигенције. Препоруке одређене Смерницама могу бити оквир и за 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 имплементацији система вештачке интелигенције;
- лица која примењују системе вештачке интелигенције, пре свега у свом раду који укључује и интеракцију са другим лицима (нпр. учесницима на тржишту);
- лица која користе системе вештачке интелигенције и на које системи имају:
  - директан утицај (нпр. користе системе ради остваривање неке јавне услуге)
  - индиректна 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 са конкретним резултатом рада и истраживања у екосистему вештачке интелигенције.

2. РЕЧНИК ТЕРМИНА И ДЕФИНИЦИЈА

2.1 Етика

Етика је наука о моралу, која истражује смисао и циљеве моралних норми, основне критерије 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 истребиран над групом података ради обављања специфичних задатака (попут препознавања одређених образаца и сл).

Независна експертска група Европске комисије дала је следећу дефиницију: „**Вештачка интелигенција** се односи 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 базирани искључиво на софтверу и деловати у виртуелном свету (на пример: виртуелни асистенти, софтвери за анализу фотографија, интернет претраживачи, системи за препознавање говора и лица) или могу бити уграђени у уређаје – хардвер (на пример: напредни роботи, аутономна возила,

дронови и слично*)*.”[6] „Систем вештачке интелигенције”, „Систем”, „систем“ или само „вештачка интелигенција” имају исто значење у оквиру Смерница.

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

Архитекта Система је лице које пројектује Систем.

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

Безбедност Према *Bourgeois*[7] фундаментални концепти безбедности информационих система су везани за информациону безбедност (енг. *information security*), која је описана преко три карактеристике: поверљивост (енг. *confidentiality*), интегритет (енг. *integrity*) и раположивост (енг. *availability*) које се називају „троугао” или „тројство” или „тријада” сигурности. Приликом заштите информација и података потребно је контролисано омогућити приступ онима којима је дозвољено да их виде, што обезбеђује поверљивости. Интегритет осигурава да информација, односно податак представља њено намеравано значење и да није измењена, 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.

Експлоатација Система је коришћење Система од стране ауторизованих корисника.

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

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

Multi-cloud подразумева коришћење више јавних облака (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 (eng. *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 у којима 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 учење 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] Bourgeois, D., (2014), *Information Systems for Business and Beyond*, Saylor Foundation, pp.64-65

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

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

Човек у управљању (енг. *human of the loop*) - омогућена је интервенција током развоја и праћења рада Система.

Човек у управљању (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 Високо-ризични системи вештачке интелигенције

Високо-ризични систем је Систем који има тенденцију да директно или индиректно крши принципе и услове утврђене Смерницама, али не нужно да то и чини.

Са аспекта Смерница, високо-ризични 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, високо-ризичним системом се сматра 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;
- побројан и означен како високо-ризични систем у Смерницама (у даљем тексту: листа високо-ризичних система).

Није од значаја за одређивање високо-ризичних 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.

Важно је нагласити да се из Смерница искључују забрањени Системи: који су супротни начелима Смерница, негативно утичу на било ког учесника у екосистему вештачке интелигенције и као такви не могу се развијати нити користити.

Високо-ризични системима се сматрају системи вештачке интелигенције у областима:

- биометријске идентификације и категоризације појединца: нарочито обухвата системе намењене за даљинску биометријску идентификацију појединца у реалном времену као и накнадну даљинску биометријску идентификацију;
- управљања критичном инфраструктуром и њен рад: 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, управљање радницима лицима и приступа самозапошљавању: нарочито обухвата системе који врше одабир и запошљавање 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 лица; системе који доносе одлуке о радно-правним питањима запослених (напредовање, награде, бонуси, отказ, промена описа радног места, конкретних задатака запосленог) као и системи 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 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.
- кривичног гоњења: нарочито обухвата системе намењене органима кривичног гоњења који врше процену ризика појединаца за извршење или поново 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 ући или је ушао на територију државе Републике Србије; системе намењене надлежним органима/институцијама за проверу веродостојности путних исправе 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 са њима повезаним процесима (повезани процеси обухватају проверу кривичне и прекршајне осуђиваности,

[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).

постојање и врсту притужби датих у вези са предметних 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 високо-ризичних 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 може на нивоу појединца утицати на пажњу, жеље, мишљења, опредељења, креативност, машту, осећања, одлуке и активности, или на нивоу друштва може допринети креирању друштвених мехура, односно поларизацију око битних друпштвених питања, и у финалној инстанци утицати на демократски капацитет.

Листа није коначна. Развој вештачке интелигенције намеће потребу да изложена листа буде отворена и тумачена као преглед репрезентативних примера високо-ризичних система која се може мењати и допуњавати.

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 са 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

ако се: 1) обрада врши од стране органа власти, осим ако се ради о обради коју врши суд у сврху обављања његових судских овлашћења; 2) основне активности руковаоца/обрађивача састоје у радњама обраде које по својој природи, обиму, односно сврхама захтевају редован и систематски надзор великог броја лица на које се подаци односе; 3) основне активности руковаоца/обрађивача састоје у обради посебних врста података о личности или података о личности у вези са кривичним пресудама и кажњивим делима, у великом обиму.

3. НАЧЕЛА

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

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 објашњене, а нарочито одлуке (препоруке) које доноси (у мери којој је то целисходно) свима на које Систем утиче (директно или индиректно). Ако није могуће објаснити одређене резултате рада неопходно их је означити као систем са моделом „црне кутије”[9]. Овакве Системе потребно је избегавати јер се сматрају непожељним.

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

3.2 Достојанство

Људско достојанство (даље: достојанство) треба разумети као полазни принцип (начело) које у фокусу има човеков интегритет. 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.”[10]

[9] Систем са 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 са 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.

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

Конвенција о људским правима наводи да „Људско достојанство (достојанство) није само основно људско право већ и основа људских права. Људско достојанство је урођено сваком човеку."[11]

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."[12]
- „Ко злоставља другог или према њему поступа на начин којим се вређа људско достојанство, казниће се затвором до једне године."[13]
- „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."[14]

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 наношење другоме физичког или психичког бола или страха (нематеријална штета)[15] и утврђује да је свако лице дужно да се уздржи од поступака којим се може другом проузроковати штета[16].

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

Кривични законик[17] предвиђена 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 законом предвиђена је и одговорност лица за штету коју причине извршењем дела мање друштвене опасности - прекршаја[18].

[11] Конвенција о људским правима

[12] Закон о јавном информисању и медијима, „Службени гласник РС", бр. 83/2014… аут. тумачење - 12/2016.

[13] Кривични законик, „Службени гласник РС", бр. 85/2005, 88/2005 - испр.,… и 35/2019

[14] Ibidem

[15] Закон о облигационим односима, „Службени гласник РС", бр. 29/78, 39/85, 45/89 - одлука УСЈ …. и бр. 18/2020

[16] Ibidem

[17] Кривични законик, „Службени гласник РС", бр. 85/2005, 88/2005 – испр. 121/2012,… и 35/2019

[18] Закон о прекршајима, „Службени гласник РС", бр. 65/2013… 91/2019 и др. закон

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.).

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 заштити[19] 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 било кога по свим основама.

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 потенцијално ускраћивање 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.

4. УСЛОВИ ПОУЗДАНЕ И ОДГОВОРНЕ ВЕШТАЧКЕ ИНТЕЛИГЕНЦИЈЕ

Изградња и стварање поуздане и одговорне вештачке интелигенције захтева испуњење одређених Услова[20] који се темеље на Начелима утврђеним овим Смерницама, који 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, недискриминација и равноправност;
6. Друштвено и еколошко благостање;

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

[20] 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.

7. Одговорност.

Услови чине проверљиве параметре, односно техничке и нетехничке методе, којима се потврђује и доказује испуњеност Начела.

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

Нетехничке методе односе се на испитивању организационих и других нетехничких елемената значајних при развоју и коришћењу 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 са наведеним Условима. 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 је и адаптабилан, што значи да 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 сви каснијим етапама како би континуирано био 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. Његовим употребом подиже се друштвена свест о важности испуњавања одређених захтева система вештачке интелигенције. Истовремено, примена Упитника доприноси 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 имплементирати 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 искључују 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 требало би 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.). Испитује 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 свестан да је у интеракцији са системом?
  - Да, на који начин
  - Не
- Да ли је лице које користи и/или лице на које систем вештачке интелигенције утиче обавештен 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 лице које користи систем вештачке интелигенције спречава 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* се (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 који самостално учи
  - 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 активност система
- Да ли су лица која врше контролу обучена за те послове?
  - Да, и то:
  - Не, не постоје лица која врше контролу
  - Не, нису посебно обучавана
- 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 са лицима која примењују и/или користе Систем, а који може понудити више од једног резултата обраде са различитим вероватноћама, потребно је омогућити избор једног од резултата обраде.
4. У процес планирања, дизајна и развоја система укључити детекцију 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 архитекте 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.
7. Идентификују лица која су задужена да реагују у случају проблема у функционисању Система, који 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 имплементације 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.
10. 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 привремено ставити Систем ван функције у ограничено 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 минимизацију 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 ИКТ инфраструктури од посебног значаја?
  - Да
  - Не
- Да ли је систем вештачке интелигенције сертификован у складу са стандардима у области информационе безбедност (нпр. ИСО 27000 и др) или је усклађен са овим стандардима?
  - Да, сертификован 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 систем 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 имплементације предузете су мере:
  - Да, у фази гашења предузете су мере:
  - Нису предузимане посебне мере
- Да ли је извршено тестирање од 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 који 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, метрика 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, користи се
  - 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 базиран на вештачкој интелигенцији
  - Не
- Да ли је развијен механизам за процену промена у раду система вештачке интелигенције како би се испитала 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 мере су 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 се о нивоу прецизности 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 систем?
  - 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 врши верификација и валидација поузданости и поновљивости система вештачке интелигенције?
  - Да
  - Не
- Да ли се документује процес тестирања и верификације поузданости и поновљивости система вештачке интелигенције?
  - Да, на начин:
  - Не
- Да ли постоји план за решавање грешака у систему вештачке интелигенције?
  - Да
  - Не
- Да ли постоји процедура за поступање у случајевима у којима систем вештачке интелигенције даје резултате са ниским резултатом поверења?
  - Да
  - 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 осигурала 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 ситема 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 стопа кликтања (енг. *click trough rate*) или животна вредност потрошача (енг. *customer lifetime value*), као и стопу лажно позитивних и лажно негативних резултата, разложену по подгрупама инстанци.
- Обезбедити релевантност метрике, на пример: систем за детекцију пожара требало би да има висопу стопу препознавања, чак и ако ће то довести до повремене лажне узбуне.

2. *Кад год је то 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), анализирати улазне податке израчунавањем агрегираних, анонимизираних групних вредности и статистика. Испитати:

- Да ли подаци садрже грешке (нпр. недостајуће вредности, погрешне ознаке) како би се утврдио квалитет података?
- Да ли су подаци узорковани тако да добро репрезентују кориснике Система (на пример: 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 су у великој мери рефлексија 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.
- Спроводити тестове интеграције како би било разумљиво како појединачне компоненте комуницирају са осталим деловима Система.
- Детектовати клизање 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" како би се осигурало да се Систем понаша како је предвиђено. Редовно ажурирати овај скуп података у складу са променама корисника и сценарија примене, како би се смањио ризик од тренирања на тестном скупу.
- Спроводити итеративно тестирање од стране корисника како би се инкорпорирали различите функционалности у различитим циклусима развоја.
- Применити „пока-јоке" (eng. *Poka-yoke*) приступ: уградити провере квалитета у Систем, 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 решавање 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

предвидети све сценарије унапред, као и развити Системе који пружају и ограничења потребна са становишта сигурности али и флексибилност у креирању решења која се прилагођавају различитим улазним подацима.

Са 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*)[21]. Пре развоја Система размотри ризике и последице напада, како би се донела одлука о даљем развоју Система.

1. *Идентификовати потенцијалне претње*
    - Размотрити да ли ико има интерес да натера 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 дозвољава нападачу да измени улазне податке Система је рањивији 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*
    - Тестирати перформасе 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 са 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:

[21] Истраживачи Мајкрософта и Беркман Клајн центра за интернет и друштво, Харвард универзитета дали 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

- Не, не обрађују се подаци о личности[22]
- 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 личости)?
  - Да, и то:
  - Не
- Да ли постоји правни основ за (намеравану) обраду података о личности?[23]
  - Да, и то:
    - пристанак лица
    - уговор са лицем
    - поштовање правних обавеза руковаоца
    - заштита животно важних интереса лица
    - обављање послова у јавном интересу/извршење законом прописаних овлашћења руковаоца
    - легитимни интерес руковаоца/треће стране
  - Не, није релевантно
- 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 са 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 уношења измена и др)
  - предвиђени су механизми за постизање уграђене и подразумеване приватности (*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 минимизације података, у односу на конкретне податке (укључујући и посебне врсте података о личности)

[22] Ова група питања се не попуњава, сем последња два

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

- 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 података система вештачке интелигенције који нису подаци о личности?
  - Да
  - Не
- Да ли је систем вештачке интелигенције усклађен са одговарајућим стандардима[24] или другим опште прихваћеним 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. Руковалац подацима је обавезан да заштити податке у Систему. Приступ подацима неопходно је обезбедити на законит начин поштујући приватност појединца.

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 структуираним, мулти-структуираним и неструктуираним подацима у свим окружењима инфраструктуре.

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

Каталог података обухвата:
- пословни речник

[24] на пример: ISO25, IEEE26, ISO27001

- аутоматско откривање података, профилисање, маркирање, каталогизација и мапирање речника појмова
- аутоматска детекција осетљивих података и класификација управљања
- интероперабилност са другим шифарницима, алатима и апликацијама ради размене података.

Софтвер за интеграцију и управљањем подацима обухвата:
- изворе података, „*multi-cloud*" и „*edge data*" повезивање
- алате за чување података
- исправку и интегрисање података
- метаподатке
- обезбеђивање заштите података
- безбедност приступа подацима у вишеструким складиштима података у дистрибуираном окружењу података

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

Не-технолошки алати су:
- Примена законског оквира који уређује ко, када и са 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 дефинише[25] 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*

[25] 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

податке или стимулансе, са различитим степеном људске интервенције у зависности од нивоа аутономије система“. Системи показују одређени ниво аутономности у интеракцији са окружењем. Већ дужи низ година, роботи се користе у системима за разврстање артикала и контролу квалитета, као и управљање индустријских складишта са производима. Интелигентни агенти се користе за имплементацију финансијских сервиса, смањење ризика 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 интерпретацију и објашњивост резултата, и начина интеракције корисника са системима вештачке интелигенцијеи презентовања резултата крајњим корисницима.

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 – разумевање (енг. *explainability*) система вештачке интелигенције и 3) комуникација – дијалог са свим заинтересованим странама о ограничењима система вештачке интелигенције.

4.4.1 Упитник

<u>Могућност праћења – следљивост</u>

Праћење омогућава организацијама 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 (досије/портфолио) система вештачке интелигенције који се правовремено ажурира и да ли се та документација чувају у складу са законом[26]?
  - 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

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

- 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
    - откривањем неправилности у излазним резултатима
    - прерасподелом улазних података који су довели до неправилности
    - друго:
  - Не
- Да ли је у систему вештачке интелигенције врши евидентирање приступа систему од стране корисника при доношењу одлука и препорука?
  - Да
  - Не
- Да ли је у систему вештачке интелигенције врши евидентирање мета података (датум и време почетка и завршетка коришћења система, база података која се користила као референтни извор података у систему и др) о начину коришћења система?
  - Да, и то:
  - Не
- Да ли је у систему вештачке интелигенције врши евидентирање приступа систему од стране лица која контролишу процес доношења одлука?
  - Да
  - Не

Објашњивост (Разумевање)

Наредна група питања омогућава процену степена разумевања система вештачке интелигенције, односно начина и разлога због којих је системпројектован 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 директно или индиректно утичу, у циљу евентуалног оспоравања таквих одлука. Објашњење, зашто је модел предложио одређену одлуку/резултат (која комбинација улазних фактора је утицала на тај резултат) није увек могуће. Овакви случајеви се односе 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 дометима и могућностима система вештачке интелигенције), наравно под претпоставком да систем вештачке интелигенције као целина поштује основна људска права. Степен објашњивости зависи од контекста, а нарочито од процене могућих негативних последица грешке/нетачности 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/степена 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

<u>Комуникација</u>

Наредна група питања омогућава процену организација 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 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 који 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 значајних 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, сва ограничења су наведена у посебном одељку
  - Не

4.4.2 Препоруке

Могућност праћења – следљивост се сматра 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 интерпретацију информација садржаних у скуповима података, потребно је екстраховати важне чињенице и резоновати користећи знање и/или теорије вероватноћа. 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 креирања закључака 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 нестручним заинтересованим странама*“. У случајевима када Систем 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 учење[27] је „црна кутија”, на пример: неуронске мреже за препознавање образаца у дубоком учењу.

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 неуспели? Када могу да верујем? Овај рефлексивни скептицизам директно 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, правна пресуда, упис на универзитет, паметне куће и аутономна возила.

Објашњивост[28] се односи на оне технике вештачке интелигенције које помажу корисницима Система (инжењерима вештачке интелигенције, крајњим корисницима и ревизорима) 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 (*eng. Human-centric*), свест о приватности (*eng.*

[27] 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

Privacy Awareness,, одговорност (*eng. Accountabilitys*),, безбедност и безбедност (*eng. Safety and Security*).

Док се поједини модели (статистички и дрва одлучивања) могу мапирати у правила и на тај начин обезбедити интерпретабилост резултата, то није случај са дубоким неуронским мрежама, који су нашли ширу примену у последњим годинама због све веће количине доступних података 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 графова знања[29], итд.

Комуникација

Да би се обезбедило поштовање основних права и усклађеност са основним људским правима[30] а посебно право 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 са човеком.

Корисност, ефективност, ефикасност и употребљивост Система обезбеђује се укључивањем крајних корисника у дизајн, евалуацију и имплементацију графичког корисничког интерфејса.

Да би обезбедила прецизна и експлицитне веза између апстрактних принципа које је Систем дужан да поштује и конкретних одлука о имплементацији, примењује се препорука - Етика по дизајну (*eng. Ethics-by-design)*. Различити концепти „по дизајну" се већ користе, 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 и разноликост током целог животног циклуса система. Системи вештачке интелигенције могу имати одређене недостатке због нпр. некомплетности и модела лошег управљања, што може довести до ненамерних (ин)директних 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 омогућава свакоме коришћење производа или услуга вештачке интелигенције, без обзира на њихов узраст, пол, могућности или карактеристике. Приступачност овој технологији за особе са инвалидитетом, које су присутне у свим друштвеним групама, је од посебног значаја.

4.5.1 Упитник

Спречавање предрасуда

- Да ли постоје дефинисане процедуре за спречавање креирања или подстицања предрасуда у систему вештачке интелигенције при избору улазних података тако и креирању алгоритма?
  - Да, узете су у обзир различитост и репрезентативност крајњих корисника и/или лица

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

- 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
- Не

- Да ли се врши едукација и подизање свести лица која учествују у стварању система вештачке (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 устављања 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

<u>Приступачност и јединствени (универзални) дизајн</u>

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. Приступачност овој технологији за особе са инвалидитетом, које су присутне у свим друштвеним групама, је од посебног значаја. Системи вештачке интелигенције не би требало да имају приступ „калупа који свима одговара“, већ би требали да размотре принципе јединственог (универзалног) дизајна који би одговарао најширем могућем кругу корисника, пратећи релевантне стандарде приступачности. Ово ће омогућити једнак приступ и активно учешће свих људи у постојећим и надолазећим људским активности 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?
  - Да, током дизајнирања система консултоване су разне групе корисника за које је потребно обезбедити посебне механизме приступачности
  - Да, корисничком интерфејсу су доступни читачи екрана и друге помоћне технологије како би се омогућило коришћење система од стране слабовидих 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 директно или индиректно 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 имплементације система вештачке интелигенције у организацијама.

- Да ли је приликом развоја система вештачке интелигенције консултоване све заинтересоване стране?
  - 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 дискриминаторне обрасце. 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 алгоритма.
- Укључити механизам за обезбеђивање повратне информације са лицима која користе/примењују Систем како би се подигла свест о пристрасностима или проблемима које идентификују та лица.
- Формирати мултидисиплинарне тимове за оцену предметних параметара. Различити тимови помажу у представљању шире варијације искустава како би се пристрасност па и дискриминација свела 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 спрам 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 различитост, недискриминација и равноправност[31]. Након што се састао са члановима глобалног менаџмента хотела, тим за развој Система открива да су разноликост 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 из хотела. Након анализе ових података и имплементације у градњу агента, схватају да има степен алгоритамске пристрасности података. Тим наставља да одвоји 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 са 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 која укључују развој система вештачке интелигенције која позитивно утичу на заштиту животне средине.

4.6.1 Упитник

<u>Заштита животне средине</u>

- 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

<u>Утицај на рад и вештине</u>

- Да ли систем вештачке интелигенције утиче на радно ангажовање и начин рада?

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

- Да
- Не

- Да ли су пре увођења система вештачке интелигенције о томе обавештени и консултовани запослени на чији је рад увођење система утицати, као и на њивове представнике (синдикате и сл)?
  - Да
  - Не
- Да ли су предузете одговарајуће мере које ће обезбедити разумевање утицаја система вештачке интелигенције на начин рада запослених?
  - Да, и то:…
  - Не
- Да ли коришћење система вештачке интелигенције ствара ризик од де-квалификације запослених?
  - 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 је припремљено корисничко упутство и други материјали неопходни за усавршавање запослених?
  - Да
  - Не
- Да ли се реализују обуке запослених?
  - Да
  - Не

<u>Утицај на друштво</u>

- Да ли систем вештачке интелигенције може имати негативан утицај на друштво у целини или на демократију?
  - Да
  - Не
- Да ли је процењен индиректан 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 потенцијални 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 друштву 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 Услова, 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 селектованих 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 резултат процесирања Система није рефлексија стварности већ генерисан резултат (пример: фотографије).
- Идентификују и приоритетизују 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 са регулативом.

4.7 Одговорност

Питање одговорности је уско повезано са 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 са високим степеном одговорности приликом разматрања дизајна,

развоја, процеса одлучивања и исхода Система. То не значи да поменута лица немају право или не требају да раде у мултидисциплинарном тиму, напротив.

Људска логика и просуђивање је кључни фактор животу Система који по претпоставци доноси објективно-логичке одлуке, јер су људи ти који пишу алгоритме, дефинишу успех или неуспех, 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 ододељак 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 које утичу на основна права човека, укључујући апликације које су критичне за безбедност, системи вештачке интелигенције би требало да 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 (независна) трећа страна 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 минимизирање 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 релевантни интереси и вредности које имплицира систем вештачке интелигенције треба да буду идентификовани и да, ако дође до сукоба, компромиси треба да буду експлицитно признати и оцењени у смислу њиховог ризика по безбедност и етичке принципе, укључујући основна права. Свака одлука о томе који компромис треба да буде начињен, треба бити 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 слични 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 овог упитника за систем вештачке интелигенције?
  - Да
  - Не
- Да ли предвиђени процес укључује идентификацију и документовање неусаглашених ставова у вези са различитим етичким принципима и објашњењима одлука?
  - Да
  - Не
- Да ли је извршена обука лица укључених у процес?
  - Да
  - Не
- Да ли трећим странама (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 како се он користи.

- Водити евиденцију о процесима дизајнирања, развоја функционалности и начина на који Систем доноси одлуке. Уредити посебним документом ову процедуру.
- Ускладити коришћење и резултат рада Система са прописима и међународним стандардима приликом креирања Система.
- Консултовати релевантна тела и органе у Републици Србији у вези са питањима која стварају недоумице ради добијања стручног савета или помоћ оних који раде на креирању политике вештачке интелигенције у Републици Србији или институције које раде контролу система вештачке интелигенције[32].
- Укључити у рад 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 разумели њихове жеље и потребе путем интервјуа са корисницима лицем у лице. Тим прихвата сопствену одговорност у ситуацији када повратне информације хотелског асистента не испуњавају потребе или очекивања гостију. Они су имплементирали петљу учења са 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 са намером да пруже оквир и усмере рад свих учесника у екосистему вештачке интелигенције. У недостатку чвршћег правног оквира који прве обрисе 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 моделовању социјалних промена, повећању знања и даљем економском напретку друштва у целини.

[32] За више информација посетити сајт https://www.ai.gov.rs/

[33] Пример преузет са: https://www.ibm.com/design/ai/ethics/accountability/