NEZÁVISLÁ

EXPERTNÁ SKUPINA NA VYSOKEJ ÚROVNI PRE UMELÚ INTELIGENCIU

ZRIADENÁ EURÓPSKOU KOMISIOU V JÚNI 2018

ETICKÉ USMERNENIA PRE DÔVERYHODNÚ UMELÚ INTELIGENCIU

ETICKÉ USMERNENIA PRE DÔVERYHODNÚ UMELÚ INTELIGENCIU

Expertná skupina na vysokej úrovni pre umelú inteligenciu

Tento dokument vypracovala expertná skupina na vysokej úrovni pre umelú inteligenciu (AI HLEG). Členovia expertnej skupiny na vysokej úrovni pre umelú inteligenciu menovaní v tomto dokumente podporujú všeobecný rámec pre dôveryhodnú umelú inteligenciu navrhnutý v týchto usmerneniach, hoci nemusia nevyhnutne súhlasiť s každým vyjadrením v tomto dokumente.

Zainteresované strany v rámci pilotnej fázy zavedú zoznam bodov na posudzovanie dôveryhodnej umelej inteligencie uvedený v kapitole III tohto dokumentu s cieľom získať praktickú spätnú väzbu. Revidovaná verzia zoznamu bodov na posudzovanie, v ktorej sa zohľadní spätná väzba získaná v pilotnej fáze, sa predloží Európskej komisii na začiatku roku 2020.

Expertná skupina na vysokej úrovni pre umelú inteligenciu je nezávislá expertná skupina, ktorú v júni 2018 zriadila Európska komisia.

Kontakt Nathalie Smuha – koordinátorka AI HLEG
E-mail CNECT-HLG-AI@ec.europa.eu

Európska komisia
B-1049 Brussels

Dokument zverejnený 8. apríla 2019.

Prvý návrh tohto dokumentu bol vydaný 18. decembra 2018 a bol predmetom otvorených konzultácií, prostredníctvom ktorých sa získala spätná väzba od vyše 500 prispievateľov. Chceli by sme výslovne a srdečne poďakovať všetkým, ktorí poskytli svoju spätnú väzbu k prvému návrhu dokumentu, ktorá bola zohľadnená pri príprave tejto revidovanej verzie.

Viac informácií o expertnej skupine na vysokej úrovni pre umelú inteligenciu je k dispozícii online (https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence).

OBSAH

ZHRNUTIE **2**

A) ÚVOD **5**

B) RÁMEC PRE DÔVERYHODNÚ UMELÚ INTELIGENCIU **7**

I. Kapitola I: Základy dôveryhodnej umelej inteligencie **11**

1. Základné práva ako morálny a právny nárok 11
2. Od základných práv k etickým zásadám 12

II. Kapitola II: Realizácia dôveryhodnej umelej inteligencie **16**

1. Požiadavky dôveryhodnej umelej inteligencie 16
2. Technické a netechnické metódy na realizáciu dôveryhodnej umelej inteligencie 24

III. Kapitola III: Posudzovanie dôveryhodnej umelej inteligencie **29**

C) PRÍKLADY PRÍLEŽITOSTÍ A VÁŽNYCH OBÁV VYVOLANÝCH UMELOU INTELIGENCIOU **40**

D) ZÁVER **44**

GLOSÁR **46**

ZHRNUTIE

1. Cieľom týchto usmernení je podpora dôveryhodnej umelej inteligencie. Dôveryhodná umelá inteligencia má **tri zložky**, ktorých sa treba pridržiavať počas celého životného cyklu systému: a) mala by byť **zákonná**, čiže by sa mala riadiť celým platným právom a právnymi predpismi; b) mala by byť **etická**, čiže by mala zabezpečiť súlad s etickými zásadami a hodnotami, a c) mala by byť **odolná**, a to z technického aj sociálneho hľadiska, keďže systémy umelej inteligencie môžu aj pri dobrých úmysloch spôsobiť neúmyselnú ujmu. Každá zložka je sama osebe potrebná, ale nestačí na dosiahnutie dôveryhodnej umelej inteligencie. V ideálnom prípade pôsobia všetky tri zložky vo vzájomnom súlade a prekrývajú sa vo svojom fungovaní. Ak medzi týmito zložkami v praxi vzniknú rozpory, spoločnosť by sa mala pokúsiť o ich zosúladenie.

2. V týchto usmerneniach sa stanovuje **rámec na dosiahnutie dôveryhodnej umelej inteligencie.** Tento rámec sa výslovne nezaoberá prvou zložkou dôveryhodnej umelej inteligencie (zákonná umelá inteligencia)[1]. Namiesto toho je jeho cieľom poskytnúť usmernenie na podporu a zabezpečenie etickej a odolnej umelej inteligencie (druhej a tretej zložky). Cieľom týchto usmernení určených všetkým zainteresovaným stranám je prekročiť rámec zoznamu etických zásad tak, že sa poskytne usmernenie o tom, ako možno tieto zásady zaviesť do praxe v sociálno-technických systémoch. Usmernenia sú rozdelené do troch úrovní podľa abstrakcie, od najabstraktnejšej v kapitole I až po najkonkrétnejšiu v kapitole III, a uzatvárajú ich príklady príležitostí, ktoré poskytujú systémy umelej inteligencie, a vážnych obáv, ktoré tieto systémy vyvolávajú.

   I. Na základe prístupu, ktorý vychádza zo základných práv, sa v kapitole I identifikujú **etické zásady** a s nimi súvisiace hodnoty, ktoré sa musia dodržiavať pri vývoji systémov umelej inteligencie, pri ich zavádzaní a používaní.

   Kľúčové usmernenia vyplývajúce z kapitoly I:

   - ✓ Vyvíjať systémy umelej inteligencie, zavádzať ich a používať tak, aby boli dodržané tieto etické zásady– *rešpektovanie ľudskej autonómie, prevencia ujmy, spravodlivosť a vysvetliteľnosť.* Uznať a riešiť prípadné rozpory medzi týmito zásadami.
   - ✓ Venovať osobitnú pozornosť situáciám, ktoré sa týkajú zraniteľných skupín, ako sú deti, osoby so zdravotným postihnutím a ďalšie skupiny, ktoré boli v minulosti znevýhodnené alebo ktorým hrozí vylúčenie, a situáciám, pre ktoré je charakteristická asymetria moci alebo dostupnosti informácií, napríklad vo vzťahoch medzi zamestnávateľmi a pracovníkmi alebo medzi podnikmi a spotrebiteľmi[2].
   - ✓ Uznať a myslieť na to, že hoci systémy umelej inteligencie sú veľkým prínosom pre jednotlivcov a spoločnosť, takisto predstavujú isté riziká a môžu mať negatívne následky vrátane účinkov, ktoré môže byť ťažké predvídať, určiť alebo zmerať (napr. vplyv na demokraciu, právny štát a spravodlivé rozdeľovanie alebo na samotnú ľudskú myseľ). V prípade potreby prijať náležité opatrenia na zmiernenie týchto rizík, ktoré budú primerané závažnosti rizika.

   II. Kapitola II vychádza z kapitoly I a obsahuje usmernenia o spôsobe, akým by sa dôveryhodná umelá inteligencia mohla realizovať, pričom sa v nej uvádza **sedem požiadaviek**, ktoré by systémy umelej inteligencie mali spĺňať. Na ich vykonanie sa môžu použiť technické aj netechnické metódy.

   Kľúčové usmernenia vyplývajúce z kapitoly II:

[1] Účelom všetkých normatívnych výrokov v tomto dokumente je zamerať usmernenia na dosiahnutie druhej a tretej zložky dôveryhodnej umelej inteligencie (etickej a odolnej umelej inteligencie). Ich účelom teda nie je poskytovať právne poradenstvo ani usmernenia týkajúce sa súladu s platným právom, hoci treba uznať, že veľa z týchto výrokov sa už do určitej miery zohľadňuje v existujúcich zákonoch. V tejto súvislosti pozri bod 21 a ďalej.

[2] Pozri články 24 až 27 Charty základných práv Európskej únie (ďalej aj „charta EÚ"), ktoré sú venované právam dieťaťa a starších osôb, integrácii osôb so zdravotným postihnutím a pracovným právam. Pozri aj článok 38 o ochrane spotrebiteľov.

- ✓ Zabezpečiť, aby vývoj systémov umelej inteligencie, ich zavádzanie a používanie spĺňali požiadavky na dôveryhodnú umelú inteligenciu: 1. ľudský faktor a dohľad; 2. technická odolnosť a bezpečnosť; 3. správa súkromia a údajov; 4. transparentnosť; 5. rozmanitosť, nediskriminácia a spravodlivosť; 6. environmentálny a spoločenský blahobyt a 7. zodpovednosť.
- ✓ Zvážiť technické a netechnické metódy na zabezpečenie vykonávania týchto požiadaviek.
- ✓ Podporovať výskum a inováciu s cieľom pomôcť pri posudzovaní systémov umelej inteligencie a pri presadzovaní plnenia požiadaviek. Zverejňovať výsledky a otvorené otázky širšej verejnosti a systematicky pripravovať novú generáciu odborníkov na etiku v oblasti umelej inteligencie.
- ✓ Jasným a iniciatívnym spôsobom poskytovať zainteresovaným stranám informácie o schopnostiach a obmedzeniach systému umelej inteligencie, čo im umožní vytvoriť si realistické očakávania, a o spôsobe, akým dochádza k plneniu požiadaviek. Nezakrývať skutočnosť, že majú do činenia so systémom umelej inteligencie.
- ✓ Uľahčiť vysledovateľnosť a kontrolovateľnosť systémov umelej inteligencie, najmä v kritických kontextoch alebo situáciách.
- ✓ Zapájať zainteresované strany počas celého životného cyklu systému umelej inteligencie. Podporovať odbornú prípravu a vzdelávanie tak, aby boli všetky zainteresované strany informované a vyškolené v oblasti dôveryhodnej umelej inteligencie.
- ✓ Mať na pamäti, že môžu existovať zásadné rozpory medzi jednotlivými zásadami a požiadavkami. Nepretržite identifikovať, vyhodnocovať a dokumentovať tieto kompromisy a ich riešenia a informovať o nich.

III. Kapitola III obsahuje konkrétny a neúplný zoznam bodov na posudzovanie dôveryhodnej umelej inteligencie, ktorého cieľom je uvádzanie požiadaviek stanovených v kapitole II do praxe. Tento **zoznam bodov na posudzovanie** sa musí upraviť podľa konkrétneho prípadu použitia systému umelej inteligencie[3].

Kľúčové usmernenia vyplývajúce z kapitoly III:

- ✓ Prijať zoznam bodov na posudzovanie dôveryhodnej umelej inteligencie, ktorý sa využije pri vývoji systémov umelej inteligencie, ich zavádzaní alebo používaní, a prispôsobiť ho konkrétnemu prípadu použitia, v ktorom sa systém uplatňuje.
- ✓ Pamätať si, že tento zoznam bodov na posudzovanie nikdy nebude úplný. Pri zabezpečení dôveryhodnej umelej inteligencie nejde o odškrtávanie políčok, ale o nepretržité identifikovanie a plnenie požiadaviek, vyhodnocovanie riešení a zabezpečovanie lepších výsledkov počas celého životného cyklu systému umelej inteligencie a o zapojenie zainteresovaných strán do týchto činností.

3. Účelom posledného oddielu tohto dokumentu je konkretizovať niektoré z otázok, ktoré boli v rámci spomenuté, a to uvedením príkladov prínosných príležitostí, ktoré by sa mali presadzovať, a vážnych obáv, ktoré vyvolávajú systémy umelej inteligencie a ktoré by sa mali starostlivo zvážiť.

4. Hoci cieľom týchto usmernení je poskytnutie pomoci pre aplikácie umelej inteligencie vo všeobecnosti, čím sa vytvára horizontálny základ na dosiahnutie dôveryhodnej umelej inteligencie, rôzne situácie vyvolávajú rôzne výzvy. Treba preto preskúmať, či okrem tohto horizontálneho rámca nie je potrebný sektorový prístup, s ohľadom na kontextovosť systémov umelej inteligencie.

5. Tieto usmernenia nemajú slúžiť ako náhrada za súčasnú či budúcu tvorbu politík alebo za súčasné či budúce právne predpisy a ich cieľom nie je ani odradiť od ich zavedenia. Mali by sa vnímať ako živý dokument, ktorý sa musí v priebehu času skúmať a aktualizovať, aby sa zabezpečila ich neustála relevantnosť, keďže technológia,

[3] V súlade s rozsahom rámca uvedeného v bode 2 tento zoznam bodov na posudzovanie neobsahuje nijaké odporúčania na zabezpečenie dodržiavania súladu s právnymi predpismi (zákonná umelá inteligencia), ale je obmedzený iba na ponuku usmernení na dosiahnutie druhej a tretej zložky dôveryhodnej umelej inteligencie (etickej a odolnej umelej inteligencie).

naše sociálne prostredia a naše poznatky sa vyvíjajú. Tento dokument sa považuje za východiskový bod pre diskusiu o „dôveryhodnej umelej inteligencii pre Európu“[4]. Za hranicami Európy sú tieto usmernenia zamerané aj na podporu výskumu, úvah a diskusií o etickom rámci pre systémy umelej inteligencie na globálnej úrovni.

[4] Tento ideálny stav má platiť pre systémy umelej inteligencie, ktoré vyvíjajú, zavádzajú a používajú členské štáty EÚ, ako aj pre systémy vyvíjané alebo vyrábané inde, ktoré sa však zavádzajú a používajú v EÚ. Výraz „Európa“ v tomto dokumente zahŕňa členské štáty EÚ. Cieľom týchto usmernení však je aj to, aby boli relevantné aj za hranicami EÚ. V tejto súvislosti možno poznamenať, že Nórsko aj Švajčiarsko sú súčasťou koordinovaného plánu v oblasti umelej inteligencie, na ktorom sa dohodla Komisia s členskými štátmi a ktorý bol uverejnený v decembri 2018.

A) ÚVOD

6. Európska komisia (ďalej len „Komisia“) vo svojich oznámeniach z 25. apríla 2018 a zo 7. decembra 2018 predstavila svoju víziu pre umelú inteligenciu, ktorá podporuje „etickú, bezpečnú a špičkovú umelú inteligenciu s pôvodom v Európe“[5]. Vízia Komisie je založená na troch pilieroch: i) zvýšenie verejných a súkromných investícií do umelej inteligencie na podporu jej využívania; ii) príprava na sociálno-ekonomické zmeny a iii) zabezpečenie vhodného etického a právneho rámca na posilnenie európskych hodnôt.

7. Na podporu realizácie tejto vízie Komisia zriadila expertnú skupinu na vysokej úrovni pre umelú inteligenciu, čo je nezávislá skupina poverená vypracovaním dvoch výstupov: 1. etických usmernení pre umelú inteligenciu a 2. politiky v oblasti umelej inteligencie a investičných odporúčaní.

8. Tento dokument obsahuje etické usmernenia pre umelú inteligenciu, ktoré boli revidované na základe hlbšej diskusie v našej skupine so zreteľom na spätnú väzbu z verejných konzultácií o návrhu zverejnenom 18. decembra 2018. Je pokračovaním práce Európskej skupiny pre etiku vo vede a v nových technológiách[6] a čerpá inšpiráciu z ďalších podobných snáh[7].

9. V posledných mesiacoch sa stretlo 52 členov expertnej skupiny na vysokej úrovni, aby spolu diskutovali a rokovali, odhodlaní presadzovať európske heslo: zjednotení v rozmanitosti. Veríme, že umelá inteligencia má potenciál podstatne transformovať spoločnosť. Umelá inteligencia nepredstavuje samotný cieľ, ale je skôr sľubným prostriedkom na zvýšenie prosperity ľudstva, čím zvyšuje blahobyt jednotlivcov a spoločnosti a spoločné blaho a prináša pokrok a inováciu. Systémy umelej inteligencie môžu najmä pomôcť uľahčiť dosiahnutie cieľov OSN v oblasti udržateľného rozvoja, ako je podpora rodovej rovnováhy a boj proti zmene klímy, racionalizácia nášho využívania prírodných zdrojov, zlepšenie nášho zdravia, mobility a výrobných procesov, a podpora nášho spôsobu monitorovania pokroku na základe ukazovateľov udržateľnosti a sociálnej súdržnosti.

10. Na tento účel musia byť systémy umelej inteligencie[8] **zamerané na človeka**, čo spočíva v záväzku používať ich v prospech ľudstva a spoločného blaha s cieľom zlepšiť dobré životné podmienky ľudí a ich slobodu. Hoci systémy umelej inteligencie ponúkajú veľké príležitosti, vyplývajú z nich aj isté riziká, ktoré sa musia riešiť vhodne a primerane. Teraz máme významnú príležitosť formovať ich vývoj. Chceme zabezpečiť, že budeme môcť dôverovať sociálno-technickým prostrediam, v ktorých sú zakorenené, a chceme, aby výrobcovia systémov umelej inteligencie získali konkurenčnú výhodu zakotvením dôveryhodnej umelej inteligencie do svojich výrobkov a služieb. To znamená úsilie o **maximalizáciu prínosov systémov umelej inteligencie a** zároveň to zahŕňa **prevenciu a minimalizáciu ich rizík**.

11. Sme presvedčení, že v kontexte rýchlej technologickej zmeny je nevyhnutné, aby dôvera ostala spojivom spoločností, spoločenstiev, hospodárstiev a trvalo udržateľného rozvoja. Identifikovali sme preto **dôveryhodnú umelú inteligenciu ako náš základný cieľ,** keďže jednotlivci a spoločenstvá budú môcť dôverovať technologickému rozvoju a jeho použitiam iba vtedy, keď bude existovať jasný a komplexný rámec na dosiahnutie ich dôveryhodnosti.

12. Veríme, že toto je cesta, ktorou by sa Európa mala uberať, aby sa stala domovom špičkových a etických

---

[5] COM(2018) 237 a COM(2018) 795. Upozorňujeme, že výraz „s pôvodom v Európe“ sa používa v celom oznámení Komisie. Rozsah týchto usmernení je však zameraný tak, aby nezahŕňal iba systémy umelej inteligencie s pôvodom v Európe, ale aj systémy vyvinuté inde, ktoré sa zavádzajú alebo používajú v Európe. V celom tomto dokumente sa preto snažíme podporovať dôveryhodnú umelú inteligenciu „pre“ Európu.

[6] Európska skupina pre etiku vo vede a v nových technológiách (EGE) je poradnou skupinou Komisie.

[7] Pozri oddiel 3.3 dokumentu COM(2018) 237.

[8] Systémy umelej inteligencie sa na účely tohto dokumentu vymedzujú v glosári na jeho konci. Toto vymedzenie sa podrobnejšie rozvádza v osobitnom dokumente, ktorý vypracovala expertná skupina na vysokej úrovni pre umelú inteligenciu a ktorý dopĺňa tieto usmernenia, s názvom „Vymedzenie pojmu umelej inteligencie: hlavné schopnosti a vedecké disciplíny“.

technológií a jednotkou v tejto oblasti. Práve prostredníctvom dôveryhodnej umelej inteligencie sa my, občania Európskej únie, budeme snažiť využívať jej výhody, a to spôsobom, ktorý je v súlade s našimi základnými hodnotami dodržiavania ľudských práv, demokracie a právneho štátu.

Dôveryhodná umelá inteligencia

13. Dôveryhodnosť je predpokladom toho, aby ľudia a spoločnosti vyvíjali systémy umelej inteligencie, zavádzali ich a používali. Bez systémov umelej inteligencie (a bez ľudí, ktorí za nimi stoja), ktoré sú preukázateľne hodné dôvery, môžu vzniknúť neželané následky a môže sa zbrzdiť jej využívanie, čo zabráni dosiahnutiu potenciálne obrovských sociálnych a hospodárskych prínosov vyplývajúcich zo systémov umelej inteligencie. Aby mohla Európa využívať tieto prínosy, našou víziou je použiť etiku ako základný pilier na zabezpečenie a rozšírenie dôveryhodnej umelej inteligencie.

14. Dôvera vo vývoj systémov umelej inteligencie, v ich zavádzanie a používanie sa netýka iba základných vlastností tejto technológie, ale aj vlastností sociálno-technických systémov, ktorých súčasťou sú aplikácie umelej inteligencie[9]. Obdobne ako v prípade otázok týkajúcich sa (straty) dôvery v bezpečnosť leteckej dopravy, jadrovej energie alebo potravín, nie sú to iba prvky systému umelej inteligencie, ale systém v celkovom kontexte, ktorý môže, ale nemusí vzbudiť túto dôveru. Úsilie o dosiahnutie dôveryhodnej umelej inteligencie sa preto netýka iba dôveryhodnosti samotného systému umelej inteligencie, ale vyžaduje si celostný a systematický prístup, ktorý zahŕňa dôveryhodnosť všetkých subjektov a procesov, ktoré sú súčasťou sociálno-technického kontextu systému počas jeho celého životného cyklu.

15. Dôveryhodná umelá inteligencia má **tri zložky,** ktorých sa treba pridržiavať počas celého životného cyklu systému:

    1. mala by byť **zákonná**, čím sa zabezpečí dodržiavanie celého platného práva a právnych predpisov;
    2. mala by byť **etická**, čím sa zabezpečí súlad s etickými zásadami a hodnotami, a
    3. mala by byť **odolná**, a to z technického aj sociálneho hľadiska, keďže systémy umelej inteligencie môžu aj pri dobrých úmysloch spôsobiť neúmyselnú ujmu.

16. Každá z týchto troch zložiek je potrebná, sama osebe však nestačí na dosiahnutie dôveryhodnej umelej inteligencie[10]. V ideálnom prípade pôsobia všetky tri zložky vo vzájomnom súlade a prekrývajú sa vo svojom fungovaní. V praxi však medzi týmito prvkami môžu existovať rozpory (napr. občas rozsah a obsah platného práva nemusí byť v súlade s etickými normami). Našou individuálnou a kolektívnou povinnosťou ako spoločnosti je usilovať sa o zabezpečenie toho, aby všetky tri zložky pomáhali zabezpečiť dôveryhodnú umelú inteligenciu[11].

17. Na zabezpečenie tzv. zodpovednej konkurencieschopnosti je kľúčový dôveryhodný prístup, ktorým sa vytvára základ, v rámci ktorého môžu všetky subjekty ovplyvnené systémami umelej inteligencie dôverovať tomu, že ich koncepcia, vývoj a používanie sú zákonné, etické a odolné. Tieto usmernenia sú určené na podporu zodpovednej a udržateľnej inovácie v oblasti umelej inteligencie v Európe. Ich cieľom je, aby sa etika stala základným pilierom vývoja jedinečného prístupu k umelej inteligencii, pilierom, ktorého cieľom je prospech, posilnenie a ochrana individuálnej prosperity ľudí, ako aj spoločného blaha spoločnosti. Domnievame sa, že toto Európe umožní stať sa svetovou jednotkou v oblasti špičkovej umelej inteligencie, ktorá je hodná našej individuálnej aj kolektívnej dôvery. Iba ak sa zabezpečí dôveryhodnosť, budú Európania plne využívať výhody systémov umelej inteligencie s vedomím, že existujú opatrenia, ktoré ich chránia proti možným rizikám.

---

[9] Tieto systémy tvoria ľudia, štátne subjekty, spoločnosti, infraštruktúra, softvér, protokoly, normy, správa vecí verejných, platné zákony, mechanizmy dohľadu, stimulačné štruktúry, audítorské postupy, nahlasovanie najlepších postupov a ďalšie prvky.

[10] To nevylučuje skutočnosť, že môžu byť potrebné ďalšie podmienky (alebo že sa môžu stať potrebnými).

[11] To okrem iného znamená, že zákonodarcovia alebo tvorcovia politík budú možno musieť preskúmať vhodnosť existujúceho práva, ak by mohlo byť v rozpore s etickými zásadami.

18. Práve tak, ako sa používanie systémov umelej inteligencie nekončí na hraniciach, hranice nezastavia ani ich vplyv. Globálne príležitosti a výzvy, ktoré vyplývajú z umelej inteligencie, si teda vyžadujú globálne riešenia. Vyzývame preto všetky zainteresované strany, aby sa usilovali o vytvorenie globálneho rámca pre dôveryhodnú umelú inteligenciu, dosahovali medzinárodný konsenzus, a zároveň podporovali a presadzovali náš prístup založený na základných právach.

Cieľová skupina a rozsah pôsobnosti

19. Tieto usmernenia sú určené všetkým zainteresovaným stranám v oblasti umelej inteligencie, ktoré navrhujú, vyvíjajú, zavádzajú, uplatňujú, používajú umelú inteligenciu alebo ktoré sú ňou ovplyvňované, okrem iného vrátane podnikov, organizácií, výskumných pracovníkov, verejných služieb, vládnych agentúr, inštitúcií, organizácií občianskej spoločnosti, jednotlivcov, pracovníkov a spotrebiteľov. Zainteresované strany, ktoré sa zaviazali dosiahnuť dôveryhodnú umelú inteligenciu, sa môžu dobrovoľne rozhodnúť využívať tieto usmernenia ako spôsob realizácie svojho záväzku, a to najmä použitím praktického zoznamu bodov na posudzovanie v kapitole III v procese vývoja a zavádzania systémov umelej inteligencie. Tento zoznam môže takisto dopĺňať existujúce postupy posudzovania, a tak sa môže stať ich súčasťou.

20. Cieľom týchto usmernení je poskytnutie pomoci pre aplikácie umelej inteligencie vo všeobecnosti, čím sa vytvára horizontálny základ na dosiahnutie dôveryhodnej umelej inteligencie. **Rôzne situácie** však **vyvolávajú rôzne výzvy**. Pri systémoch umelej inteligencie, ktoré slúžia na odporúčanie hudby, nevznikajú rovnaké etické obavy ako v prípade systémov umelej inteligencie, ktorých účelom je navrhovať liečbu kritických stavov. Podobne rôzne príležitosti a výzvy vyplývajú zo systémov umelej inteligencie, ktoré sa používajú v rámci vzťahov medzi podnikom a koncovým zákazníkom, medzi podnikmi, medzi zamestnávateľom a zamestnancom a medzi verejnosťou a občanom, alebo všeobecnejšie povedané, v rozličných sektoroch alebo prípadoch použitia. Vzhľadom na kontextovosť systémov umelej inteligencie sa preto uznáva, že vykonávanie týchto usmernení sa musí upraviť podľa konkrétnej aplikácie umelej inteligencie. Okrem toho by sa mala preskúmať potreba dodatočného sektorového prístupu, ktorý bude dopĺňať všeobecnejší horizontálny rámec navrhovaný v tomto dokumente.

V záujme lepšieho porozumenia možného vykonávania týchto usmernení na horizontálnej úrovni a porozumenia záležitostí, ktoré si vyžadujú sektorový prístup, vyzývame všetky zainteresované strany, aby skúšobne používali zoznam bodov na posudzovanie dôveryhodnej umelej inteligencie (kapitola III), ktorým sa tento rámec zavádza do praxe, a aby nám poskytli spätnú väzbu. Na základe spätnej väzby získanej v pilotnej fáze do začiatku roku 2020 prepracujeme zoznam bodov na posudzovanie uvedený v týchto usmerneniach. Pilotná fáza sa začne do leta 2019 a potrvá do konca roku. Všetky zaujímajúce sa zainteresované strany sa budú môcť zúčastniť tak, že svoj záujem prejavia prostredníctvom Európskej aliancie pre umelú inteligenciu.

B) RÁMEC PRE DÔVERYHODNÚ UMELÚ INTELIGENCIU

21. Týmito usmerneniami sa vytvára rámec na dosiahnutie dôveryhodnej umelej inteligencie založenej na základných právach zakotvených v Charte základných práv Európskej únie (ďalej aj „charta EÚ“) a v príslušnom medzinárodnom práve v oblasti ľudských práv. V ďalšej časti sa stručne opisujú tri zložky dôveryhodnej umelej inteligencie.

Zákonná umelá inteligencia

22. Systémy umelej inteligencie nefungujú vo svete bez zákonov. Na vývoj systémov umelej inteligencie, ich zavádzanie a používanie sa už dnes vzťahuje alebo sa ich týka viacero právne záväzných pravidiel na európskej, vnútroštátnej a medzinárodnej úrovni. Medzi právne zdroje relevantnosti patrí okrem iného primárne právo EÚ (zmluvy Európskej únie a Charta základných práv), sekundárne právo EÚ (napríklad všeobecné nariadenie o ochrane údajov, antidiskriminačné smernice, smernica o strojových zariadeniach, smernica o zodpovednosti

za výrobky, nariadenie o voľnom toku iných ako osobných údajov, právne predpisy o ochrane spotrebiteľov a smernice o bezpečnosti a ochrane zdravia pri práci), ale aj zmluvy OSN týkajúce sa ľudských práv a dohovory Rady Európy (napríklad Európsky dohovor o ľudských právach) a množstvo právnych predpisov členských štátov EÚ. Okrem horizontálne uplatňovaných pravidiel existujú rozličné pravidlá špecifické pre jednotlivé oblasti, ktoré sa uplatňujú na konkrétne aplikácie umelej inteligencie (ako napríklad nariadenie o zdravotníckych pomôckach v sektore zdravotnej starostlivosti).

23. V práve sa stanovujú pozitívne aj negatívne povinnosti, čo znamená, že právo by sa nemalo vykladať iba s odkazom na to, čo sa *nesmie*, ale aj s odvolaním sa na to, čo by sa *malo* robiť. Právne predpisy neslúžia len na zákaz istých činností, ale aj iné činnosti umožňujú. V tejto súvislosti možno upozorniť na to, že charta EÚ obsahuje články o „slobode podnikania“ a „slobode umenia a vedeckého bádania“ popri článkoch venovaných oblastiam, ktoré sa viac týkajú našej snahy o zabezpečenie dôveryhodnosti umelej inteligencie, ako je napríklad oblasť ochrany osobných údajov a nediskriminácie.

24. Usmernenia sa výslovne nezaoberajú prvou zložkou dôveryhodnej umelej inteligencie (zákonnou umelou inteligenciou), namiesto toho je ich cieľom poskytnúť usmernenie na podporu a zabezpečenie druhej a tretej zložky (etickej a odolnej umelej inteligencie). Hoci druhá a tretia zložka sa už do určitej miery odzrkadľujú v existujúcich právnych predpisoch, ich úplná realizácia môže presahovať rámec platných právnych povinností.

25. Žiadne z ustanovení tohto dokumentu sa nesmie chápať ani vykladať ako právne poradenstvo alebo usmernenia týkajúce sa spôsobu, ako možno dosiahnuť súlad s platnými existujúcimi právnymi normami a požiadavkami. Zo žiadnych z ustanovení tohto dokumentu nevyplývajú žiadne vymáhateľné práva ani sa nimi neukladajú právne povinnosti voči tretím stranám. Pripomíname však, že každá fyzická alebo právnická osoba má povinnosť dodržiavať súlad s právnymi predpismi, či už ide o predpisy platné v súčasnosti alebo predpisy, ktoré budú prijaté v budúcnosti v závislosti od vývoja umelej inteligencie. Tieto usmernenia vychádzajú z predpokladu, že **všetky vymáhateľné práva a právne povinnosti, ktoré sa vzťahujú na postupy a činnosti súvisiace s vývojom umelej inteligencie, jej zavádzaním a používaním sú naďalej záväzné a musia sa riadne dodržiavať.**

Etická umelá inteligencia

26. Na dosiahnutie dôveryhodnej umelej inteligencie nie je potrebný iba súlad s právnymi predpismi – ten tvorí len jednu z jej troch zložiek. Právne predpisy nie vždy držia krok s technologickým vývojom, občas môžu byť v rozpore s etickými normami alebo jednoducho nemusia byť primerané na riešenie určitých otázok. Na to, aby boli systémy umelej inteligencie dôveryhodné, mali by byť teda aj etické, čiže by mal byť zabezpečený súlad s etickými normami.

Odolná umelá inteligencia

27. Aj keď sa zabezpečí etický účel, jednotlivci a spoločnosť musia mať aj istotu, že systémy umelej inteligencie nespôsobia žiadnu neúmyselnú ujmu. Tieto systémy by mali fungovať bezpečným, chráneným a spoľahlivým spôsobom, a s cieľom zabrániť neúmyselným nepriaznivým účinkom by sa mali navrhnúť ochranné opatrenia. Je preto dôležité zabezpečiť odolnosť systémov umelej inteligencie. Toto je potrebné z technického hľadiska (zabezpečenie technickej odolnosti systému podľa potreby v danom kontexte, ako je oblasť použitia alebo fáza životného cyklu) a zo sociálneho hľadiska (náležité zohľadnenie kontextu a prostredia, v ktorých sa systém prevádzkuje). Etická a odolná umelá inteligencia sú teda úzko prepojené a navzájom sa dopĺňajú. Zásady predložené v kapitole I a požiadavky, ktoré sú z nich odvodené v kapitole II, sa týkajú obidvoch zložiek.

Rámec

28. Usmernenia uvedené v tomto dokumente sú rozdelené do troch úrovní podľa abstrakcie, od najabstraktnejšej v kapitole I až po najkonkrétnejšiu v kapitole III:

I. Základy dôveryhodnej umelej inteligencie. V kapitole I sa vysvetľujú základy dôveryhodnej umelej inteligencie, čím sa stanovuje jej prístup založený na základných právach[12]. Uvádzajú a opisujú sa v nej etické zásady, ktoré sa musia dodržiavať v záujme zabezpečenia etickej a odolnej umelej inteligencie.

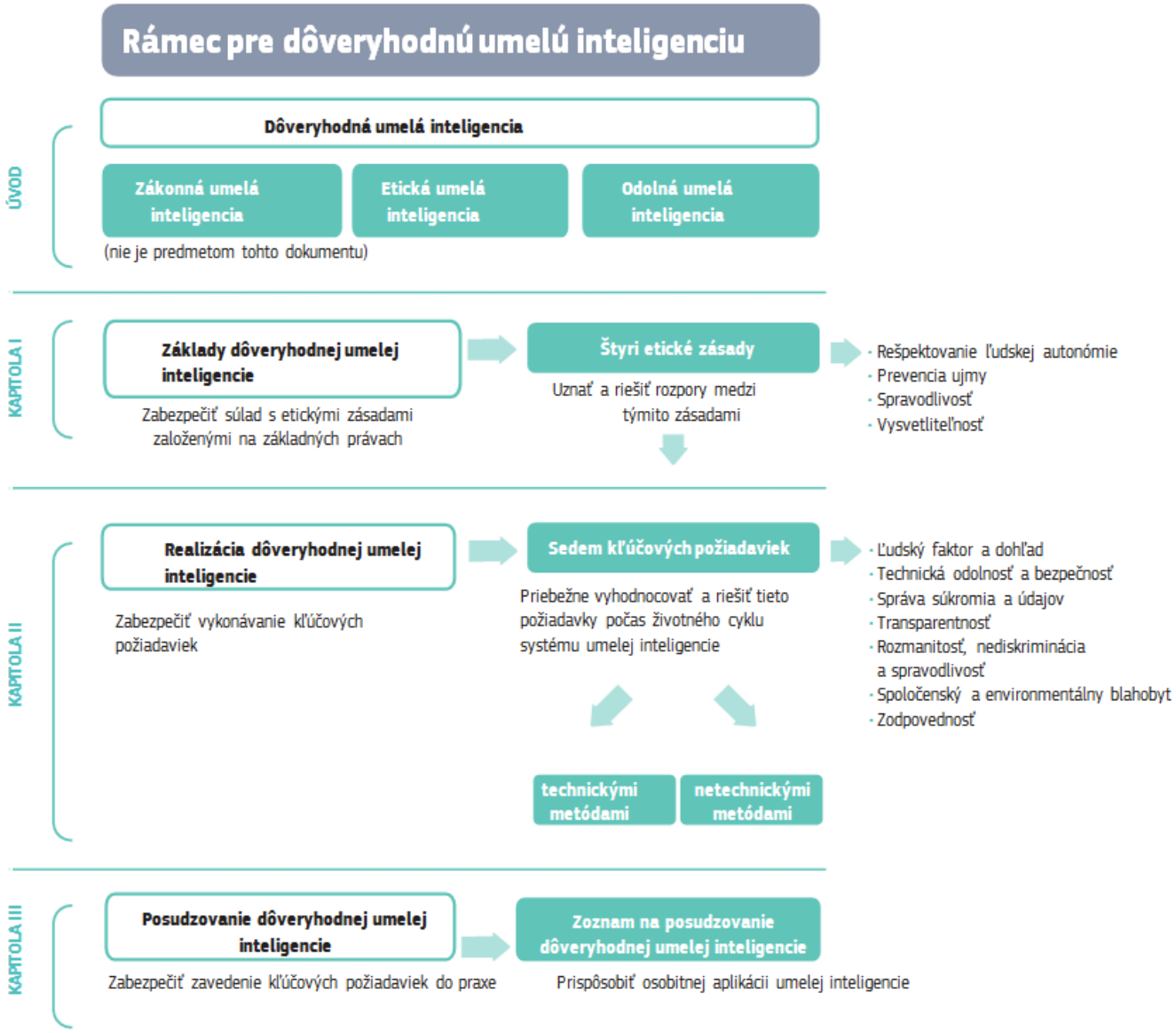
II. Realizácia dôveryhodnej umelej inteligencie. V kapitole II sa tieto etické zásady premietajú do siedmich požiadaviek, ktoré by systémy umelej inteligencie mali realizovať a spĺňať počas celého svojho životného cyklu. Táto kapitola navyše obsahuje opis technických a netechnických metód, ktoré sa môžu použiť na ich realizáciu.

III. Posudzovanie dôveryhodnej umelej inteligencie. Špecialisti na umelú inteligenciu očakávajú konkrétne usmernenia. V kapitole III sa preto uvádza predbežný a neúplný zoznam bodov na posudzovanie dôveryhodnej umelej inteligencie s cieľom uviesť požiadavky z kapitoly II do praxe. Toto posúdenie by sa malo prispôsobiť konkrétnemu použitiu systému.

29. V poslednom oddiele dokumentu sa uvádzajú prospešné príležitosti a vážne obavy súvisiace so systémami umelej inteligencie, ktoré treba zvážiť a o ktorých by sme chceli vyvolať ďalšiu diskusiu.

30. Štruktúra týchto usmernení je znázornená na *obrázku 1*.

[12] Základné práva sú základom medzinárodného aj európskeho práva v oblasti ľudských práv a podporujú právne vymožiteľné práva zaručené zmluvami EÚ a Chartou základných práv EÚ. Vzhľadom na ich právnu záväznosť tak dodržiavanie základných práv patrí do prvej zložky dôveryhodnej umelej inteligencie, čiže do zákonnej umelej inteligencie. Základné práva sa však takisto môžu chápať tak, že odzrkadľujú osobitné morálne nároky všetkých jednotlivcov, ktoré vyplývajú z ich príslušnosti k ľudstvu, a to bez ohľadu na ich právnu záväznosť. V tomto zmysle tvoria súčasť aj druhej zložky dôveryhodnej umelej inteligencie, čiže etickej umelej inteligencie.



Obrázok 1: Usmernenia ako rámec pre dôveryhodnú umelú inteligenciu

I. Kapitola I: Základy dôveryhodnej umelej inteligencie

31. V tejto kapitole sa vysvetľujú základy dôveryhodnej umelej inteligencie zakotvené v základných právach, ktoré sa odzrkadľujú v štyroch etických zásadách, ktoré by sa mali dodržiavať v záujme zabezpečenia etickej a odolnej umelej inteligencie. Táto kapitola značne čerpá z oblasti etiky.

32. Etika umelej inteligencie je podoblasťou aplikovanej etiky zameranou na etické otázky, ktoré vznikajú pri vývoji umelej inteligencie, jej zavádzaní a používaní. Jej hlavným záujmom je určiť, ako môže umelá inteligencia zlepšiť život jednotlivcov alebo vyvolať obavy o ich dobrý život, či už z hľadiska kvality života, alebo z hľadiska ľudskej autonómie a slobody, ktoré sú potrebné pre demokratickú spoločnosť.

33. Etické úvahy o technológii umelej inteligencie môžu poslúžiť viacerým účelom. Po prvé, môžu podnietiť úvahy o potrebe chrániť jednotlivcov a skupiny na najzákladnejšej úrovni. Po druhé, môžu podnietiť nové druhy inovácií, ktorých cieľom je podporiť etické hodnoty. Ide napríklad o inovácie, ktoré pomáhajú dosahovať ciele OSN v oblasti udržateľného rozvoja[13], ktoré sú pevne zakotvené v nadchádzajúcej európskej Agende 2030[14]. Hoci sa tento dokument zaoberá prevažne prvým účelom, význam, ktorý by etika mohla predstavovať pri druhom účele, by sa nemal podceňovať. Dôveryhodná umelá inteligencia môže zvýšiť prospech jednotlivcov a spoločný blahobyt, a to vytvorením prosperity, tvorbou hodnôt a zväčšením bohatstva. Môže prispieť k vytvoreniu spravodlivej spoločnosti prostredníctvom zlepšenia zdravia a blahobytu občanov spôsobmi, ktoré podporujú rovnosť pri rozdeľovaní hospodárskych, sociálnych a politických príležitostí.

34. Je preto nevyhnutné, aby sme chápali, ako čo najlepšie podporiť vývoj umelej inteligencie, jej zavádzanie a používanie s cieľom zabezpečiť, aby zo sveta umelej inteligencie mohli mať prospech všetci, a s cieľom vybudovať lepšiu budúcnosť, a pritom si uchovať konkurencieschopnosť na celosvetovej úrovni. Podobne ako pri iných výkonných technológiách, aj používanie systémov umelej inteligencie v našej spoločnosti vyvoláva viaceré etické otázky, napríklad v súvislosti s ich vplyvom na ľudí a spoločnosť, schopnosti rozhodovania a bezpečnosť. Ak začneme v čoraz väčšej miere využívať pomoc systémov umelej inteligencie alebo na ne prenášať rozhodovanie, musíme zabezpečiť, aby tieto systémy boli spravodlivé, pokiaľ ide o ich vplyv na ľudské životy, aby boli v súlade s principiálnymi hodnotami a mohli podľa nich konať, a aby to všetko mohli zabezpečiť vhodné procesy vyvodzovania zodpovednosti.

35. Európa musí stanoviť, ktorú normatívnu víziu budúcnosti s umelou inteligenciou chce uskutočniť, a potom musí porozumieť, ktorá predstava o umelej inteligencii by sa mala v Európe študovať, vyvíjať, zaviesť a používať na dosiahnutie tejto vízie. Cieľom tohto dokumentu je pomôcť tomuto úsiliu zavedením predstavy o dôveryhodnej umelej inteligencii, o ktorej sme presvedčení, že je správnym spôsobom budovania budúcnosti s umelou inteligenciou. Budúcnosť, v ktorej demokracia, právny štát a základné práva tvoria základ systémov umelej inteligencie a v ktorej tieto systémy nepretržite zlepšujú a chránia kultúru demokracie, takisto umožní vznik prostredia, v ktorom sa môže dariť inovácii a zodpovednej konkurencieschopnosti.

36. Etický kódex špecifický pre jednotlivé oblasti (akokoľvek dôsledné, prepracované a podrobné môžu byť jeho budúce verzie) nikdy nebude môcť fungovať ako náhrada za samotné etické uvažovanie, ktoré musí vždy ostať citlivé voči skutočnostiam vyplývajúcim z kontextu, ktoré nie je možné zachytiť vo všeobecných usmerneniach. Okrem vypracovania súboru pravidiel od nás zabezpečenie dôveryhodnej umelej inteligencie vyžaduje, aby sme vytvorili a udržiavali etickú kultúru a postoje prostredníctvom verejnej diskusie, vzdelávania a praktického učenia sa.

1. Základné práva ako morálny a právny nárok

37. Dôverujeme prístupu k etike umelej inteligencie, ktorý je založený na základných právach zakotvených

13 https://ec.europa.eu/commission/publications/reflection-paper-towards-sustainable-europe-2030_sk.

14 https://sustainabledevelopment.un.org/?menu=1300.

v zmluvách EÚ[15], Charte základných práv EÚ (charte EÚ) a v medzinárodnom práve v oblasti ľudských práv[16]. Dodržiavanie základných práv v demokratickom rámci a v rámci právneho štátu predstavuje najsľubnejší základ na určenie abstraktných etických zásad a hodnôt, ktoré sa môžu zaviesť do praxe v kontexte umelej inteligencie.

38. V zmluvách EÚ a v charte EÚ sa stanovuje niekoľko základných práv, ktoré sú pre členské štáty a inštitúcie EÚ právne záväzné pri vykonávaní práva Únie. Tieto práva sú opísané v charte EÚ prostredníctvom odkazu na dôstojnosť, slobody, rovnosť a solidaritu, občianske práva a spravodlivosť. Spoločný základ, ktorý spája tieto práva, možno chápať tak, že má korene v úcte k ľudskej dôstojnosti. Odzrkadľuje preto to, čo charakterizujeme ako „prístup zameraný na človeka“, v ktorom má človek jedinečné a nescudziteľné morálne postavenie prvenstva v občianskej, politickej, hospodárskej a sociálnej oblasti[17].

39. Hoci sú práva stanovené v charte EÚ právne záväzné[18], je dôležité uznať, že základné práva nie vždy zaručujú komplexnú právnu ochranu. V prípade charty EÚ treba napríklad poukázať na to, že jej oblasť uplatňovania je obmedzená na oblasti, v ktorých sa uplatňuje právo Únie. Medzinárodné právo v oblasti ľudských práv, a najmä Európsky dohovor o ľudských právach, sú pre členské štáty EÚ právne záväzné, a to aj v oblastiach, ktoré nepatria do pôsobnosti práva Únie. Zároveň treba zdôrazniť, že základné práva sa udeľujú aj osobám a (do istej miery) skupinám na základe ich morálneho postavenia ako ľudských bytostí, nezávisle od ich právnej sily. Základné práva, chápané ako právne vymožiteľné práva, preto patria do prvej zložky dôveryhodnej umelej inteligencie (zákonná umelá inteligencia), čo zaručuje súlad s právnymi predpismi. Chápané ako práva každého človeka, zakotvené v ich základnom morálnom postavení ako ľudských bytostí, sú takisto oporou druhej zložky dôveryhodnej umelej inteligencie (etická umelá inteligencia), v rámci ktorej sa týkajú etických noriem, ktoré nemusia byť nevyhnutne právne záväzné, sú však kľúčové pre zabezpečenie dôveryhodnosti. Keďže cieľom tohto dokumentu nie je poskytnúť usmernenia o prvej zložke, na účely týchto nezáväzných usmernení odkazy na základné práva súvisia s druhou zložkou.

2. Od základných práv k etickým zásadám

2.1 Základné práva ako základ dôveryhodnej umelej inteligencie

40. Spomedzi komplexného súboru nedeliteľných práv stanovených v medzinárodnom práve v oblasti ľudských práv, v zmluvách EÚ a v charte EÚ sú tieto skupiny základných práv obzvlášť vhodné pre systémy umelej inteligencie. Mnohé z týchto práv sú v stanovených okolnostiach právne vymožiteľné v EÚ, takže súlad s ich podmienkami je právne záväzný. Ale aj keď bol dosiahnutý súlad s právne vymožiteľnými základnými právami, etické úvahy nám môžu pomôcť porozumieť, ako vývoj umelej inteligencie, jej zavedenie a používanie môžu zahŕňať základné práva a hodnoty, na ktorých sú tieto práva založené, a môžu nám pomôcť vytvoriť presnejšie usmernenia, keď sa budeme usilovať určiť, čo by sme *mali* robiť namiesto toho, čo (v súčasnosti) *môžeme* robiť s technológiami.

41. ***Úcta k ľudskej dôstojnosti***. Ľudská dôstojnosť zahŕňa myšlienku, že každý človek má „vnútornú hodnotu“, ktorá by sa nikdy nemala zmenšovať, ohrozovať alebo potláčať inými osobami – ani novými technológiami, ako sú systémy umelej inteligencie[19]. Rešpektovanie ľudskej dôstojnosti v kontexte umelej inteligencie znamená,

---

15 EÚ je založená na ústavnom záväzku chrániť základné a nedeliteľné práva ľudí, zabezpečiť dodržiavanie zásad právneho štátu, podporovať demokratické slobody a presadzovať spoločné blaho. Tieto práva sú zachytené v článkoch 2 a 3 Zmluvy o Európskej únii a v Charte základných práv EÚ.

16 Tieto záväzky sú zachytené a spresnené v ďalších právnych nástrojoch, napríklad v Európskej sociálnej charte Rady Európy alebo v osobitných právnych predpisoch, akým je všeobecné nariadenie EÚ o ochrane údajov.

17 Treba poznamenať, že záväzok voči umelej inteligencii zameranej na človeka a jej ukotvenie v rámci základných práv si vyžaduje kolektívny spoločenský a ústavný základ, podľa ktorého je individuálna sloboda a úcta k ľudskej dôstojnosti prakticky možná aj zmysluplná, namiesto toho, aby sa iba naznačil neprimerane individualistický opis človeka.

18 Podľa článku 51 charty sa charta týka inštitúcií EÚ a členských štátov EÚ pri vykonávaní práva Únie.

19 McCrudden, C., *Human Dignity and Judicial Interpretation of Human Rights* (Ľudská dôstojnosť a súdny výklad ľudských práv), *EJIL*, roč. 19, č. 4, 2008.

že so všetkými ľuďmi sa zaobchádza s úctou, ktorá im prislúcha ako morálnym *subjektom*, a nie iba ako s *objektmi*, ktoré sa majú skúmať, usporiadať, hodnotiť, u ktorých sa majú vytvoriť vzorce stádovitého a podmieneného správania alebo s ktorými sa má manipulovať. Systémy umelej inteligencie by sa preto mali vyvíjať tak, aby rešpektovali telesnú a duševnú nedotknuteľnosť ľudí, ich zmysel pre osobnú a kultúrnu identitu a uspokojovanie ich základných potrieb, aby im slúžili a chránili ich[20].

42. ***Sloboda jednotlivca***. Ľudia by mali ostať slobodní, aby mohli sami za seba prijímať životné rozhodnutia. To zahŕňa zabránenie zasahovaniu do suverenity, ale na druhej strane si to vyžaduje aj zásahy zo strany vlády a mimovládnych organizácií, ktorých cieľom je zabezpečiť, že jednotlivci alebo ľudia ohrozovaní vylúčením budú mať rovnaký prístup k výhodám a príležitostiam vyplývajúcim z umelej inteligencie. V kontexte umelej inteligencie si sloboda jednotlivca vyžaduje zmiernenie (ne)priameho protiprávneho nátlaku, hrozieb pre duševnú autonómiu a duševné zdravie, neoprávneného sledovania, podvodného konania a nečestnej manipulácie. Sloboda jednotlivca vlastne znamená záväzok umožniť jednotlivcom, aby mali ešte väčšiu kontrolu nad svojimi životmi, vrátane (okrem iných práv) ochrany slobody podnikania, slobody umenia a vedeckého bádania, slobody prejavu, práva na súkromný život a na súkromie a slobody zhromažďovania a združovania**.**

43. ***Dodržiavanie demokracie, spravodlivosti a právneho štátu***. Všetka moc vlády v ústavných demokraciách musí byť zákonne povolená a obmedzená zákonom. Systémy umelej inteligencie by mali slúžiť na udržiavanie a podporu demokratických procesov a dodržiavanie plurality hodnôt a životných volieb jednotlivcov. Systémy umelej inteligencie nesmú pôsobiť proti demokratickým procesom, ľudskému uvažovaniu či demokratickým hlasovacím systémom. V systémoch umelej inteligencie musí byť zakotvený aj záväzok zabezpečiť, aby nepôsobili spôsobom, ktorý by oslaboval základné záväzky, na ktorých je založený právny štát, kogentné zákony a právne predpisy, a zabezpečiť riadny proces a rovnosť pred zákonom.

44. ***Rovnosť, nediskriminácia a solidarita*** *– vrátane práv osôb ohrozených vylúčením.* Musí sa zabezpečiť rovnaká úcta k morálnej hodnote a dôstojnosti všetkých ľudí. Presahuje to rámec nediskriminácie, ktorý pripúšťa rozlišovanie medzi rozdielnymi situáciami na základe objektívneho odôvodnenia. V kontexte umelej inteligencie rovnosť znamená, že pri prevádzke systému nemôžu vznikať nespravodlivo zaujaté výstupy (napr. údaje použité na výcvik systémov umelej inteligencie by mali byť čo najviac inkluzívne a mali by zastupovať rôzne skupiny obyvateľstva). To si takisto vyžaduje primerané rešpektovanie potenciálne zraniteľných osôb a skupín[21], ako sú pracovníci, ženy, osoby so zdravotným postihnutím, etnické menšiny, deti, spotrebitelia alebo iné skupiny ohrozené vylúčením.

45. ***Práva občanov***. Občania majú široký okruh práv vrátane práva voliť, práva na dobrú správu vecí verejných alebo na prístup k verejným dokumentom a právo obrátiť sa s petíciou na verejnú správu. Systémy umelej inteligencie ponúkajú značný potenciál na zlepšenie rozsahu a efektivity vlády pri poskytovaní verejných statkov a služieb spoločnosti. Zároveň by na práva občanov mohli mať aplikácie umelej inteligencie aj negatívny vplyv, a tieto práva by sa mali chrániť. Vždy, keď sa v tomto dokumente použije výraz „práva občanov“, neznamená to popretie alebo zanedbanie práv štátnych príslušníkov tretej krajiny a osôb, ktoré sú v EÚ neoprávnene (alebo nelegálne). Aj tieto osoby majú práva podľa medzinárodného práva, a teda aj v oblasti umelej inteligencie.

2.2 Etické zásady v kontexte systémov umelej inteligencie[22]

[20] Pokiaľ ide o chápanie „ľudskej dôstojnosti“ v tomto zmysle, pozri Hilgendorf, E., *Problem Areas in the Dignity Debate and the Ensemble Theory of Human Dignity* (Problémové oblasti v diskusii o dôstojnosti a súhrnná teória ľudskej dôstojnosti), in: Grimm, D., Kemmerer, A., Möllers, C. (editori), *Human Dignity in Context. Explorations of a Contested Concept* (Ľudská dôstojnosť v súvislostiach. Štúdie o spornom pojme), 2018, s. 325 a nasl.

[21] Vymedzenie pojmu v zmysle, v akom sa používa v tomto dokumente, sa nachádza v glosári.

[22] Tieto zásady sa uplatňujú aj na vývoj, zavádzanie a používanie iných technológií, a teda sa netýkajú iba systémov umelej inteligencie. V tejto časti sme sa snažili vysvetliť ich relevantnosť konkrétne v kontexte umelej inteligencie.

46. Veľa verejných, súkromných a občianskych organizácií pri vytváraní etických rámcov pre umelú inteligenciu čerpalo inšpiráciu zo základných práv[23]. Európska skupina pre etiku vo vede a v nových technológiách pre EÚ navrhla súbor deviatich základných zásad vychádzajúcich zo základných hodnôt stanovených v zmluvách EÚ a v Charte základných práv EÚ[24]. Staviame na tejto práci, pričom uznávame väčšinu zásad, ktoré navrhli rôzne skupiny, a zároveň vysvetľujeme ciele, ktoré sa týmito zásadami majú rozvíjať a podporovať. Tieto etické zásady môžu inšpirovať nové a osobitné regulačné nástroje, môžu pomôcť pri výklade základných práv (keďže naše sociálno-technické prostredie sa v priebehu času vyvíja) a môžu usmerniť dôvody pre vývoj systémov umelej inteligencie, ich používanie a zavádzanie – pričom sa dynamicky prispôsobujú vývoju samotnej spoločnosti.

47. Systémy umelej inteligencie by mali zlepšiť blahobyt jednotlivcov a spoločnosti. V tomto oddiele sa uvádzajú **štyri etické zásady**, zakotvené v základných právach, ktoré sa musia dodržiavať s cieľom zabezpečiť, aby sa systémy umelej inteligencie vyvíjali, zavádzali a používali dôveryhodným spôsobom. Tieto zásady boli určené ako **etické požiadavky**, ktoré by sa špecialisti na umelú inteligenciu mali vždy usilovať dodržiavať. Poradie v zozname nepredstavuje poradie podľa dôležitosti. Zásady sa v ňom uvádzajú spôsobom odzrkadľujúcim poradie, v akom sa v charte EÚ[25] uvádzajú základné práva, z ktorých tieto zásady vychádzajú.

48. Ide o tieto zásady:

i) rešpektovanie ľudskej autonómie;
ii) prevencia ujmy;
iii) spravodlivosť;
iv) vysvetliteľnosť.

49. Mnohé z týchto zásad sa už do značnej miery odzrkadľujú v existujúcich právnych požiadavkách, pre ktoré sa vyžaduje povinný súlad, a preto patria aj do rozsahu „zákonnej umelej inteligencie“, ktorá tvorí prvú zložku dôveryhodnej umelej inteligencie[26]. Zároveň, ako už bolo spomenuté, hoci sa v mnohých právnych povinnostiach odzrkadľujú etické zásady, dodržiavanie týchto zásad presahuje formálny súlad s existujúcim právom[27].

- <u>Zásada rešpektovania ľudskej autonómie</u>

50. Základné práva, na ktorých je založená EÚ, smerujú k zabezpečeniu rešpektovania slobody a autonómie ľudí. Ľudia, ktorí prichádzajú do styku so systémami umelej inteligencie, musia byť schopní zachovať si úplné a účinné sebaurčenie a zúčastňovať sa na demokratickom procese. Systémy umelej inteligencie by nemali neoprávnene podriaďovať, donucovať, podvádzať, manipulovať ľudí a vytvárať v nich vzorce podmieneného alebo stádovitého správania. Mali by sa namiesto toho navrhnúť tak, aby zvyšovali, dopĺňali a posilňovali ľudské kognitívne, sociálne a kultúrne zručnosti. Rozdelenie funkcií medzi ľudí a systémy umelej inteligencie by sa malo riadiť zásadami návrhu zameraného na človeka a zmysluplné príležitosti ponechať na ľudské

[23] Spoliehanie sa na základné práva takisto pomáha znížiť regulačnú neistotu, keďže môže stavať na základe desaťročí praxe s ochranou základných práv v EÚ, čím zabezpečuje jasnosť, zrozumiteľnosť a predvídateľnosť.

[24] Prednedávnom osobitná skupina AI4People uskutočnila prieskum uvedených zásad Európskej skupiny pre etiku vo vede a v nových technológiách, ako aj 36 ďalších doteraz predložených etických zásad, a začlenila ich do štyroch spoločných zásad: Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., Vayena, E. J. M., (2018), *AI4People – An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations* (AI4People – Etický rámec pre dobrú spoločnosť s umelou inteligenciou: príležitosti, riziká, zásady a odporúčania), *Minds and Machines*, roč. 28, č. 4, s. 689 – 707.

[25] Rešpektovanie ľudskej autonómie je výrazne spojené s právom na ľudskú dôstojnosť a slobodu (uvedené v článkoch 1 a 6 charty). Prevencia ujmy úzko súvisí s ochranou telesnej alebo duševnej nedotknuteľnosti (článok 3). Spravodlivosť je úzko prepojená s právami na nediskrimináciu, solidaritu a spravodlivosť (uvedené v článku 21 a nasledujúcich článkoch). Vysvetliteľnosť a zodpovednosť úzko súvisia s právami, ktoré sa týkajú spravodlivosti (ako sa uvádza v článku 47).

[26] Pripomíname napríklad všeobecné nariadenie o ochrane údajov alebo právne predpisy EÚ o ochrane spotrebiteľov.

[27] Viac informácií o tejto téme sa nachádza napríklad v Floridi, L., *Soft Ethics and the Governance of the Digital* (Mäkká etika a správa digitálnej sféry), *Philosophy & Technology,* marec 2018, roč. 31, č. 1, s. 1 – 8.

rozhodnutie. To znamená zabezpečenie ľudského dohľadu[28] a kontroly nad pracovnými procesmi v systémoch umelej inteligencie. Systémy umelej inteligencie môžu takisto zásadne zmeniť oblasť práce. Mali by podporovať ľudí v pracovnom prostredí a ich cieľom by malo byť vytvorenie zmysluplnej práce.

- Zásada prevencie ujmy

51. Systémy umelej inteligencie by nemali spôsobovať ujmu ani ju zväčšovať[29], ani by nemali inak nepriaznivo pôsobiť na ľudí[30]. To so sebou prináša aj ochranu ľudskej dôstojnosti, ako aj duševnej a telesnej nedotknuteľnosti. Systémy umelej inteligencie a prostredia, v ktorých fungujú, musia byť bezpečné a chránené. Musia byť technicky odolné a malo by sa zabezpečiť, že nebudú náchylné na zneužitie. Väčšia pozornosť by sa mala venovať zraniteľným osobám, ktoré by sa mali začleniť do vývoja a zavádzania systémov umelej inteligencie. Osobitná pozornosť sa musí venovať aj situáciám, v ktorých by systémy umelej inteligencie mohli spôsobovať alebo zhoršovať nepriaznivé účinky, a to z dôvodu asymetrie moci alebo dostupnosti informácií, napríklad vo vzťahoch medzi zamestnávateľmi a zamestnancami, medzi podnikmi a spotrebiteľmi alebo medzi vládami a občanmi. V rámci zásady prevencie ujmy sa musí brať zreteľ aj na prírodné prostredie [a] na všetky živé bytosti.

- Zásada spravodlivosti

52. Vývoj, zavádzanie a používanie systémov umelej inteligencie musia byť spravodlivé. Hoci uznávame, že existuje mnoho rôznych výkladov spravodlivosti, veríme, že spravodlivosť má hmotnoprávny a procesný rozmer. Hmotnoprávny rozmer so sebou prináša záväzok zabezpečiť rovnaké a spravodlivé rozdelenie prínosov aj nákladov a zabezpečiť, že jednotlivci a skupiny nebudú vystavení nespravodlivej zaujatosti, diskriminácii a stigmatizácii. Ak je možné zabrániť nespravodlivej zaujatosti, systémy umelej inteligencie by dokonca mohli posilniť spoločenskú spravodlivosť. Podporovať by sa mali aj rovnaké príležitosti, pokiaľ ide o prístup k vzdelávaniu, tovaru, službám a technológiám. Používanie systémov umelej inteligencie by okrem toho nikdy nemalo viesť k oklamaniu (koncových) používateľov alebo k oslabeniu ich slobodnej voľby. Navyše spravodlivosť znamená, že špecialisti na umelú inteligenciu by mali dodržiavať zásadu proporcionality medzi prostriedkami a cieľmi a starostlivo zvažovať, ako vyvážiť protikladné záujmy a zámery[31]. Procesný rozmer spravodlivosti zahŕňa schopnosť namietať proti rozhodnutiam systémov umelej inteligencie a ľudí, ktorí ich prevádzkujú, a požadovať účinnú nápravu[32]. Na tento účel musí byť možné identifikovať subjekt zodpovedný za rozhodnutie a procesy rozhodovania by mali byť vysvetliteľné.

- Zásada vysvetliteľnosti

53. Vysvetliteľnosť je kľúčová na vytvorenie a udržiavanie dôvery používateľov voči systémom umelej inteligencie. To znamená, že procesy musia byť transparentné, o schopnostiach a účele systémov umelej inteligencie sa musí informovať otvorene a rozhodnutia, pokiaľ to je možné, musia byť vysvetliteľné osobám, ktorých sa priamo alebo nepriamo dotýkajú. Bez týchto informácií nie je možné proti rozhodnutiu riadne namietať. Vysvetlenie, prečo model vytvoril konkrétny výstup alebo rozhodnutie (a aká kombinácia vstupných faktorov k nim prispela), nie je vždy možné. Tieto prípady sa označujú ako algoritmy tzv. čiernej skrinky a vyžadujú si

---

28 Koncepcii ľudského dohľadu je podrobnejšie venovaný bod 65 uvedený ďalej v tomto dokumente.

29 Ujma môže byť individuálna alebo kolektívna a môže ísť aj o nehmotnú ujmu pre sociálne, kultúrne a politické prostredie.

30 To sa týka aj spôsobu života jednotlivcov a sociálnych skupín, napríklad vyhýbanie sa vzniku kultúrnej ujmy.

31 Toto súvisí so zásadou proporcionality (ako je zachytená v porekadle, že by sa nemalo „chodiť s kanónom na vrabce“). Opatrenia prijaté na dosiahnutie cieľa (napr. opatrenia na extrakciu údajov zavedené s cieľom realizovať funkciu optimalizácie umelej inteligencie) by sa mali obmedziť na to, čo je skutočne nevyhnutné. Znamená to takisto, že ak na uspokojenie cieľa existuje viacero konkurujúcich si opatrení, prednosť by malo dostať to opatrenie, ktoré je najmenej nepriaznivé voči základným právam a etickým normám (napr. vývojári umelej inteligencie by vždy mali uprednostňovať údaje verejného sektora pred osobnými údajmi). Spomenúť možno aj proporcionalitu medzi používateľom a prevádzkovateľom vzhľadom na práva spoločností (vrátane duševného vlastníctva a dôvernosti) na jednej strane a na práva používateľov na druhej strane.

32 A to aj prostredníctvom využitia práva na združovanie a na členstvo v odboroch na pracovisku, ako sa stanovuje v článku 12 Charty základných práv EÚ.

osobitnú pozornosť. V týchto prípadoch môžu byť potrebné iné opatrenia súvisiace s vysvetliteľnosťou (napr. vysledovateľnosť, kontrolovateľnosť a transparentná komunikácia o schopnostiach systému) za predpokladu, že systém ako celok rešpektuje základné práva. Miera, do akej je vysvetliteľnosť potrebná, značne závisí od situácie a závažnosti následkov v prípade, ak je daný výstup chybný alebo inak nepresný[33].

2.3 Rozpory medzi zásadami

54. Medzi uvedenými zásadami môžu vznikať rozpory a na ich odstránenie neexistuje žiadne konkrétne riešenie. Na riešenie týchto rozporov by sa v súlade so základným záväzkom EÚ týkajúcim sa demokratickej angažovanosti, riadneho procesu a otvorenej politickej účasti mali stanoviť postupy zodpovedných diskusií. Napríklad v rôznych oblastiach použitia môžu existovať rozpory medzi *zásadou prevencie ujmy* a *zásadou rešpektovania ľudskej autonómie*. Príkladom môže byť použitie systémov umelej inteligencie na „prediktívne vykonávanie policajných funkcií“, ktoré môžu pomôcť znížiť mieru kriminality, ale spôsobom, ktorého súčasťou je sledovanie, ktoré sa dotýka slobody a súkromia jednotlivcov. Okrem toho by celkový prínos systémov umelej inteligencie mal značne prevyšovať predvídateľné individuálne riziká. Hoci tieto zásady celkom isto predstavujú pomoc pri hľadaní riešení, stále ide o abstraktné etické nariadenia. Od špecialistov na umelú inteligenciu sa tak nemôže očakávať, že nájdu správne riešenie založené na uvedených zásadách, k etickým dilemám a kompromisom by však mali pristupovať skôr na základe logickej úvahy podloženej dôkazmi než na základe intuície alebo náhodnej úvahy. Môžu sa však vyskytnúť situácie, v ktorých nie je možné určiť žiadne eticky prijateľné kompromisné riešenia. Určité základné práva a súvisiace zásady sú absolútne a nemožno ich podriadiť hľadaniu kompromisu (napr. ľudská dôstojnosť).

Kľúčové usmernenia vyplývajúce z kapitoly I:

- ✓ Vyvíjať systémy umelej inteligencie, zavádzať ich a používať tak, aby boli dodržané etické zásady: *rešpektovanie ľudskej autonómie, prevencia ujmy, spravodlivosť a vysvetliteľnosť.* Uznať a riešiť prípadné rozpory medzi týmito zásadami.
- ✓ Venovať osobitnú pozornosť situáciám, ktoré sa týkajú zraniteľnejších skupín, ako sú deti, osoby so zdravotným postihnutím a ďalšie skupiny, ktoré boli v minulosti znevýhodnené, ohrozované vylúčením, a/alebo situáciám, pre ktoré je charakteristická asymetria moci alebo dostupnosti informácií, napríklad vo vzťahoch medzi zamestnávateľmi a pracovníkmi alebo medzi podnikmi a spotrebiteľmi[34].
- ✓ Uznať a myslieť na to, že hoci aplikácie umelej inteligencie majú potenciál prinášať jednotlivcom aj spoločnosti množstvo podstatných výhod, niektoré aplikácie môžu mať aj negatívne následky vrátane účinkov, ktoré môže byť ťažké predvídať, identifikovať alebo zmerať (napr. vplyv na demokraciu, právny štát a spravodlivé rozdeľovanie alebo na samotnú ľudskú myseľ). V prípade potreby prijať náležité opatrenia na zmiernenie týchto rizík, ktoré budú primerané závažnosti rizika.

II. Kapitola II: Realizácia dôveryhodnej umelej inteligencie

55. Táto kapitola obsahuje pomoc pri vykonávaní a realizácii dôveryhodnej umelej inteligencie prostredníctvom siedmich požiadaviek, ktoré by sa mali splniť, pričom sa vychádza zo zásad uvedených v kapitole I. Navyše sa v nej predstavujú v súčasnosti dostupné technické a netechnické metódy na vykonávanie týchto požiadaviek počas celého životného cyklu systému umelej inteligencie.

1. Požiadavky dôveryhodnej umelej inteligencie

56. Aby sa dosiahla dôveryhodná umelá inteligencia, zásady opísané v kapitole I sa musia premeniť na konkrétne

[33] Napríklad z nepresných odporúčaní týkajúcich sa nakupovania, ktoré poskytol systém umelej inteligencie, môžu vyplývať zanedbateľné etické obavy na rozdiel od systémov umelej inteligencie, ktoré posudzujú podmienečné prepustenie osoby odsúdenej za trestný čin.

[34] Pozri články 24 až 27 charty EÚ, ktoré sú venované právam dieťaťa a starších osôb, integrácii osôb so zdravotným postihnutím a základným pracovným právam. Pozri aj článok 38 o ochrane spotrebiteľov.

požiadavky. Tieto požiadavky sa týkajú rôznych zainteresovaných strán, ktoré sa podieľajú na životnom cykle systémov umelej inteligencie: vývojári, prevádzkovatelia a koncoví používatelia, ako aj širšia spoločnosť. Výrazom „vývojári“ označujeme osoby, ktoré sa zaoberajú výskumom, navrhovaním a/alebo vývojom systémov umelej inteligencie. Výrazom „prevádzkovatelia“ označujeme verejné alebo súkromné organizácie, ktoré používajú systémy umelej inteligencie vo svojich pracovných postupoch a na účely ponuky výrobkov a služieb ostatným. Koncoví používatelia sú osoby, ktoré prichádzajú do kontaktu so systémom umelej inteligencie, a to priamo alebo nepriamo. A napokon, širšia spoločnosť zahŕňa všetky ostatné subjekty, ktorých sa systémy umelej inteligencie priamo alebo nepriamo dotýkajú.

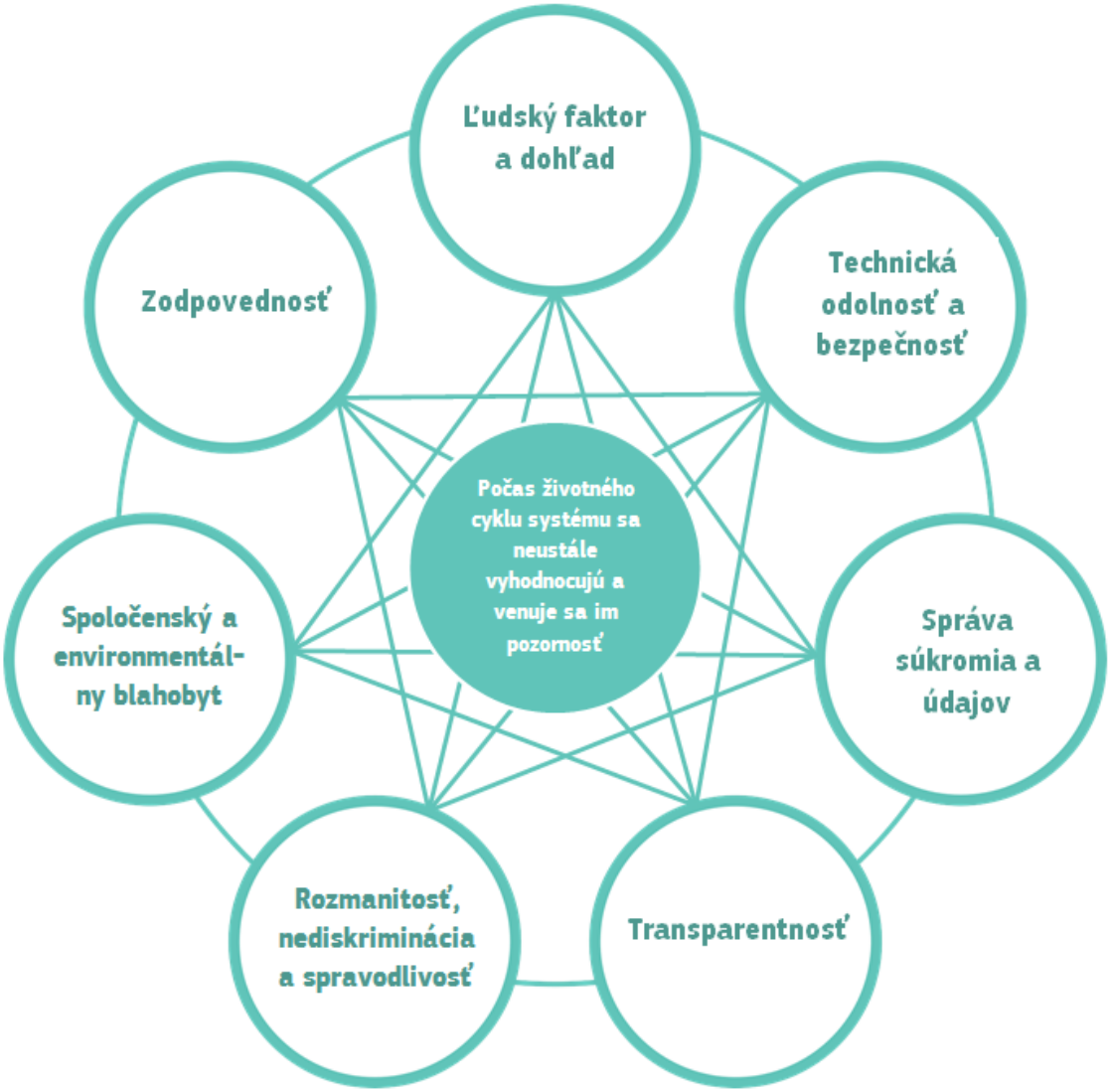
57. Jednotlivé triedy zainteresovaných strán zohrávajú pri zabezpečovaní plnenia požiadaviek rozličné úlohy:

a) vývojári by mali vykonávať a uplatňovať požiadavky v rámci postupov navrhovania a vývoja;
b) prevádzkovatelia by mali zabezpečiť, že systémy, ktoré používajú, a výrobky a služby, ktoré ponúkajú, spĺňajú tieto požiadavky;
c) koncoví používatelia a širšia spoločnosť by mali byť informovaní o týchto požiadavkách a mali by byť schopní požadovať ich dodržiavanie.

58. Tento zoznam požiadaviek nie je úplný[35]. Obsahuje systémové, individuálne a spoločenské aspekty:

1. **Ľudský faktor a dohľad**
   Zahŕňa základné práva, ľudský faktor a ľudský dohľad
2. **Technická odolnosť a bezpečnosť**
   Zahŕňa odolnosť voči útokom a bezpečnostnú ochranu, záložný plán a všeobecnú bezpečnosť, presnosť, spoľahlivosť a reprodukovateľnosť
3. **Správa súkromia a údajov**
   Zahŕňa rešpektovanie súkromia, kvalitu a integritu údajov a prístup k údajom
4. **Transparentnosť**
   Zahŕňa vysledovateľnosť, vysvetliteľnosť a komunikáciu
5. **Rozmanitosť, nediskriminácia a spravodlivosť**
   Zahŕňa zabránenie nespravodlivej zaujatosti, prístupnosť a dizajn pre všetkých a účasť zainteresovaných strán
6. **Spoločenský a environmentálny blahobyt**
   Zahŕňa udržateľnosť a šetrnosť k životnému prostrediu, sociálny vplyv, spoločnosť a demokraciu
7. **Zodpovednosť**
   Zahŕňa kontrolovateľnosť, minimalizáciu negatívneho vplyvu a jeho oznamovanie, kompromisy a nápravu.

[35] Poradie v zozname nepredstavuje poradie podľa dôležitosti. Zásady sa v ňom uvádzajú spôsobom odzrkadľujúcim poradie, v akom sa v charte EÚ uvádzajú zásady a práva, s ktorými tieto zásady súvisia.



Obrázok 2: Vzájomné vzťahy medzi siedmimi požiadavkami: všetky požiadavky majú rovnakú dôležitosť, navzájom sa podporujú a mali by sa vykonávať a vyhodnocovať počas celého životného cyklu systému umelej inteligencie

59. Hoci sú všetky požiadavky rovnako dôležité, pri ich uplatňovaní v rozličných oblastiach a odvetviach sa bude musieť zohľadniť kontext a prípadné konflikty medzi nimi. Plnenie týchto požiadaviek by malo prebiehať počas celého životného cyklu systému umelej inteligencie a závisí od konkrétneho uplatňovania. Hoci sa väčšina požiadaviek týka všetkých systémov umelej inteligencie, osobitná pozornosť sa venuje systémom, ktoré majú priamy alebo nepriamy vplyv na jednotlivcov. Niektoré aplikácie (napríklad v priemyselnom prostredí) preto môžu byť menej relevantné.

60. Uvedené požiadavky obsahujú prvky, ktoré sa v niektorých prípadoch už odzrkadľujú v existujúcich právnych predpisoch. Opakujeme, že v súlade s prvou zložkou dôveryhodnej umelej inteligencie je povinnosťou vývojárov a prevádzkovateľov systémov umelej inteligencie, aby zabezpečili súlad s právnymi povinnosťami, pokiaľ ide o horizontálne uplatňované pravidlá, ako aj o predpisy špecifické pre jednotlivé oblasti.

61. Ďalej uvedené odseky obsahujú podrobnejší opis každej požiadavky.

1. Ľudský faktor a dohľad

62. Systémy umelej inteligencie by mali podporovať ľudskú autonómiu a rozhodovanie, ako to vyplýva zo zásady

rešpektovania ľudskej autonómie. To znamená, že systémy umelej inteligencie by mali pomáhať zabezpečiť demokratickú, prosperujúcu a spravodlivú spoločnosť podporou činnosti používateľov a rozvoja základných práv, ako aj umožňovať ľudský dohľad.

63. ***Základné práva***. Tak ako v prípade mnohých technológií systémy umelej inteligencie môžu v rovnakej miere pomôcť posilniť základné práva alebo im brániť. Ľuďom môžu priniesť prospech napríklad tak, že im pomáhajú pri sledovaní ich osobných údajov alebo že zvyšujú dostupnosť vzdelávania, čím podporujú ich právo na vzdelanie. Vzhľadom na dosah a kapacitu systémov umelej inteligencie však môžu ovplyvniť základné práva aj negatívne. V situáciách, keď existuje takéto riziko, by sa malo uskutočniť posúdenie vplyvu na základné práva. Toto posúdenie by malo prebehnúť pred vývojom týchto systémov a jeho súčasťou by malo byť hodnotenie toho, či je možné tieto riziká zmierniť alebo odôvodniť ich ako nevyhnutné v demokratickej spoločnosti v záujme dodržania práv a slobôd ostatných. Okrem toho by sa mali zaviesť mechanizmy na získavanie vonkajšej spätnej väzby v súvislosti so systémami umelej inteligencie, ktoré by mohli porušovať základné práva.

64. ***Ľudský faktor.*** Používatelia by mali mať možnosť prijímať samostatné informované rozhodnutia, pokiaľ ide o systémy umelej inteligencie. Mali by dostať poučenie a nástroje, aby v dostatočnej miere pochopili systémy umelej inteligencie a aby s nimi mohli uspokojivo zaobchádzať, a ak je to možné, malo by sa im umožniť primerané sebahodnotenie alebo umožniť poukázať na nedostatky systému. Systémy umelej inteligencie by mali podporovať jednotlivcov pri prijímaní lepších a informovanejších rozhodnutí v súlade s ich cieľmi. Systémy umelej inteligencie sa občas môžu zaviesť do prevádzky s cieľom utvárať a ovplyvňovať ľudské správanie prostredníctvom mechanizmov, ktoré môže byť ťažké odhaliť, pretože môžu využívať podvedomé procesy vrátane rozličných foriem nečestnej manipulácie, podvodného konania, vytvárania vzorcov podmieneného a stádovitého správania v ľuďoch. Všetky tieto procesy môžu ohrozovať osobnú autonómiu. Celková zásada používateľskej autonómie musí tvoriť ústredný pilier funkčnosti systému. Kľúčové tu je právo, aby sa na používateľov nevzťahovalo rozhodnutie založené výlučne na automatizovanom spracovaní údajov, z ktorého vyplývajú právne účinky na používateľov alebo ktoré ich ovplyvňuje podobne významným spôsobom[36].

65. ***Ľudský dohľad***. Ľudský dohľad pomáha zabezpečiť, aby systém umelej inteligencie neoslaboval ľudskú autonómiu ani nespôsoboval iné nepriaznivé účinky. Dohľad sa môže dosiahnuť prostredníctvom mechanizmov riadenia, ako je zabezpečenie prístupu human-in-the-loop (HITL), human-on-the-loop (HOTL) alebo human-in-command (HIC). Human-in-the-loop (HITL) znamená možnosť ľudského zásahu v každom rozhodovacom cykle systému, ktorý v mnohých prípadoch nie je možný ani žiaduci. Human-on-the-loop (HOTL) znamená možnosť ľudského zásahu počas cyklu navrhovania systému a monitorovania prevádzky systému. Human-in-command (HIC) znamená možnosť dohliadať na celkovú činnosť systému umelej inteligencie (vrátane jeho širšieho hospodárskeho, spoločenského, právneho a etického vplyvu) a schopnosť rozhodnúť, kedy a ako používať systém v akejkoľvek konkrétnej situácii. Môže sem patriť rozhodnutie nepoužívať systém umelej inteligencie v konkrétnej situácii, stanoviť úrovne ľudského rozhodovania počas používania systému alebo zabezpečiť schopnosť zrušiť rozhodnutie systému. Okrem toho sa musí zabezpečiť, aby verejné orgány na presadzovanie práva boli schopné vykonávať dohľad v súlade so svojím mandátom. Mechanizmy dohľadu môžu byť v rozličnej miere potrebné na podporu ostatných bezpečnostných a kontrolných opatrení, a to v závislosti od oblasti uplatňovania systému umelej inteligencie a od jeho potenciálneho rizika. Ak sa nezmenia ostatné okolnosti, platí, že čím menej dohľadu môže človek nad systémom umelej inteligencie vykonávať, tým viac je potrebné rozsiahlejšie testovanie a prísnejšie riadenie.

2. <u>Technická odolnosť a bezpečnosť</u>

66. Kľúčovou zložkou dosahovania dôveryhodnej umelej inteligencie je technická odolnosť, ktorá úzko súvisí so *zásadou prevencie ujmy*. Technická odolnosť si vyžaduje, aby sa systémy umelej inteligencie vyvíjali

[36] Ako príklad možno uviesť článok 22 všeobecného nariadenia o ochrane údajov, v ktorom je toto právo už zakotvené.

s preventívnym prístupom k rizikám, a tak, že sa budú spoľahlivo správať zamýšľaným spôsobom, pričom sa minimalizuje výskyt neúmyselnej a neočakávanej ujmy a zabráni sa neprijateľnej ujme. To by sa malo vzťahovať aj na možné zmeny v ich prevádzkovom prostredí alebo na prítomnosť iných faktorov (ľudského a umelého), ktorých interakcia so systémom môže byť nepriateľská. Okrem toho by sa mala zabezpečiť telesná a duševná nedotknuteľnosť ľudí.

67. ***Odolnosť voči útokom a bezpečnostná ochrana***. Systémy umelej inteligencie, podobne ako všetky softvérové systémy, by mali byť chránené proti chybám, ktoré môžu umožniť zneužitie systému nepriateľskými subjektmi, napr. v prípade hekingu. Útoky sa môžu zameriavať na údaje (tzv. data poisoning), na model (tzv. model leakage) alebo na príslušnú infraštruktúru, a to softvérovú aj hardvérovú. V prípade napadnutia systému umelej inteligencie, napr. pri nepriateľských útokoch, je možné zmeniť údaje, ako aj správanie systému. Následkom toho systém prijíma odlišné rozhodnutia alebo sa úplne vypne. Systémy a údaje sa môžu poškodiť aj v dôsledku zlomyseľného zámeru alebo vystavenia neočakávaným situáciám. Nedostatočné bezpečnostné procesy môžu takisto viesť k chybným rozhodnutiam či dokonca k fyzickej ujme. Na to, aby sa systémy umelej inteligencie mohli považovať za bezpečné[37], treba zohľadniť možné neúmyselné použitia umelej inteligencie (napr. aplikácie s dvojakým použitím) a potenciálne zneužitie systému umelej inteligencie škodlivými aktérmi, a na zabránenie možnosti výskytu takýchto prípadov a na ich zmiernenie by sa mali prijať opatrenia[38].

68. ***Záložný plán a všeobecná bezpečnosť***. Systémy umelej inteligencie by mali mať bezpečnostné záruky, ktoré v prípade problémov aktivujú záložný plán. To môže znamenať, že systémy umelej inteligencie sa prepnú zo štatistického postupu na postup založený na pravidlách alebo že si pred pokračovaním v činnosti vyžiadajú zásah ľudského operátora[39]. Musí sa zabezpečiť, že systém bude robiť to, čo má bez toho, aby spôsoboval ujmu živým bytostiam alebo životnému prostrediu. To zahŕňa minimalizáciu neúmyselných následkov a chýb. Okrem toho by sa mali zaviesť postupy na objasnenie a posúdenie možných rizík spojených s používaním systémov umelej inteligencie v rôznych oblastiach použitia. Úroveň potrebných bezpečnostných opatrení závisí od rozsahu rizika, ktoré systém umelej inteligencie predstavuje, čo zase závisí od schopností systému. Ak je možné predvídať, že proces vývoja alebo samotný systém budú znamenať obzvlášť vysoké riziká, je rozhodujúce, aby sa bezpečnostné opatrenia vyvíjali a skúšali iniciatívne.

69. ***Presnosť.*** Presnosť sa týka schopnosti systému umelej inteligencie utvárať správne úsudky, napríklad v záujme správneho roztriedenia informácií do správnych kategórií, alebo jeho schopnosti robiť správne predpovede, odporúčania alebo rozhodnutia na základe údajov alebo modelov. Otvorený a náležite utvorený proces vývoja a hodnotenia môže podporiť nápravu neúmyselných rizík vyplývajúcich z nepresných predpovedí, zmierniť ich alebo viesť k ich náprave. Ak nie je možné zabrániť výskytu príležitostných nepresných predpovedí, je dôležité, aby systém mohol naznačiť pravdepodobnosť týchto chýb. Vysoká úroveň presnosti je obzvlášť kľúčová v situáciách, keď má systém umelej inteligencie priamy vplyv na ľudské životy.

70. ***Spoľahlivosť a reprodukovateľnosť.*** Je nevyhnutné, aby výsledky systémov umelej inteligencie boli reprodukovateľné aj spoľahlivé. Spoľahlivý systém umelej inteligencie je taký, ktorý funguje správne s viacerými vstupmi a vo viacerých situáciách. Je to potrebné na kontrolu systému umelej inteligencie a na to, aby sa predišlo vzniku neúmyselnej ujmy. Pojmom reprodukovateľnosť sa opisuje to, či sa pokus v oblasti umelej inteligencie prejavil rovnakým správaním pri jeho opakovaní za rovnakých podmienok. Vedcom a tvorcom politík to umožňuje presne opísať, čo robia systémy umelej inteligencie. Súbory týkajúce sa

---

37 Pozri napr. úvahy v bode 2.7 koordinovaného plánu Európskej únie v oblasti umelej inteligencie.

38 Pokiaľ ide o bezpečnosť systémov umelej inteligencie, môže existovať naliehavá požiadavka na vytvorenie pozitívneho kolobehu v rámci výskumu a vývoja medzi porozumením útokov, vývojom uspokojivých ochranných opatrení a zlepšením metodík hodnotenia. Na tento účel by sa malo podporiť zbližovanie medzi komunitou zaoberajúcou sa umelou inteligenciou a bezpečnostnou komunitou. Navyše je povinnosťou všetkých zapojených subjektov, aby vytvorili spoločné cezhraničné normy pre bezpečnostnú ochranu a bezpečnosť a vybudovali prostredie vzájomnej dôvery, ktorá podporuje medzinárodnú spoluprácu. Možné opatrenia sa nachádzajú v dokumente *Malicious Use of AI* (Škodlivé využitie umelej inteligencie) (Avin, S., Brundage, M., a kol., 2018).

39 Zvážiť by sa mali aj scenáre pre prípad, že nie je možný okamžitý ľudský zásah.

opakovateľnosti[40] môžu uľahčiť testovací proces a správanie pri reprodukovaní.

3. Správa súkromia a údajov

71. So *zásadou prevencie ujmy* úzko súvisí súkromie, základné právo, ktorého sa systémy umelej inteligencie osobitne dotýkajú. Prevencia ujmy voči súkromiu si takisto vyžaduje primeranú správu údajov, ktorá sa týka kvality a integrity použitých údajov, jej relevantnosť v porovnaní s oblasťou, v ktorej systémy umelej inteligencie zavedú prístupové protokoly, a schopnosť spracúvať údaje spôsobom chrániacim súkromie.

72. ***Ochrana súkromia a údajov***. Systémy umelej inteligencie musia zaručovať ochranu súkromia a údajov počas celého životného cyklu systému[41]. Patria sem informácie, ktoré pôvodne poskytol používateľ, ako aj informácie, ktoré o používateľovi vznikli v priebehu jeho komunikácie so systémom (napr. výstupy, ktoré systém umelej inteligencie vytvoril v súvislosti s konkrétnymi používateľmi, alebo spôsob, akým používatelia reagovali na konkrétne odporúčania). Digitálne záznamy ľudského správania môžu systémom umelej inteligencie umožniť vyvodiť nielen preferencie jednotlivcov, ale aj ich sexuálnu orientáciu, vek, pohlavie, náboženské alebo politické názory. Na to, aby jednotlivci mohli dôverovať procesu získavania údajov, sa musí zabezpečiť, aby sa získané údaje o nich nepoužili s úmyslom protiprávne alebo nespravodlivo ich diskriminovať.

73. ***Kvalita a integrita údajov***. Kvalita používaných dátových súborov je pre výkon systémov umelej inteligencie rozhodujúca. Údaje môžu pri ich získavaní obsahovať spoločenskú zaujatosť, nepresnosti, chyby a omyly. To sa musí riešiť pred výcvikom s akýmkoľvek daným dátovým súborom. Okrem toho sa musí zabezpečiť integrita údajov. Vloženie škodlivých údajov do systému umelej inteligencie môže viesť k zmene jeho správania, čo platí najmä v prípade samoučiacich sa systémov. Použité procesy a dátové súbory sa musia testovať a dokumentovať pri každom kroku, ako je plánovanie, výcvik, testovanie a zavádzanie. To by sa malo vzťahovať aj na systémy umelej inteligencie, ktoré neboli vyvinuté interne, ale boli získané inde.

74. ***Prístup k údajom***. V každej organizácii, ktorá nakladá s údajmi jednotlivcov (bez ohľadu na to, či ide o používateľa systému alebo nie), by sa mali zaviesť dátové protokoly, ktorými sa bude riadiť prístup k údajom. V týchto protokoloch by sa malo stanoviť, kto môže mať prístup k údajom a za akých okolností. Prístup by sa mal povoliť iba náležite kvalifikovaným zamestnancom, ktorí majú oprávnenie a potrebu pristupovať k údajom jednotlivcov.

4. Transparentnosť

75. Táto požiadavka úzko súvisí so *zásadou vysvetliteľnosti a* zahŕňa transparentnosť prvkov dôležitých pre systém umelej inteligencie: model údajov, model systému a obchodný model.

76. ***Vysledovateľnosť***. Dátové súbory a procesy, ktorých výsledkom je rozhodnutie systému umelej inteligencie, vrátane tých zo získavania údajov a z označovania údajov, ako aj použité algoritmy by sa mali dokumentovať podľa najlepšej možnej normy, aby sa umožnila vysledovateľnosť a zvýšenie transparentnosti. To sa týka aj rozhodnutí systému umelej inteligencie. Následkom toho bude možné určiť dôvody, prečo bolo rozhodnutie umelej inteligencie chybné, čo by zas mohlo pomôcť predchádzať budúcim omylom. Vysledovateľnosť teda uľahčuje kontrolovateľnosť a vysvetliteľnosť.

77. ***Vysvetliteľnosť.*** Vysvetliteľnosť sa týka schopnosti vysvetliť technické procesy systému umelej inteligencie, ako aj súvisiace ľudské rozhodnutia (napr. oblasti použitia systému umelej inteligencie). Podmienkou technickej vysvetliteľnosti je to, že rozhodnutia systému umelej inteligencie sú zrozumiteľné a že ich ľudia môžu sledovať. Okrem toho by mohlo byť potrebné robiť kompromisy medzi zvýšením vysvetliteľnosti systému (čo môže znížiť

[40] To sa týka súborov, v ktorých sa zopakuje každý krok procesu vývoja systému umelej inteligencie od výskumu a zberu pôvodných údajov až po výsledky.

[41] Ako príklad možno uviesť existujúce právne predpisy v oblasti ochrany súkromia, ako je všeobecné nariadenie o ochrane údajov alebo pripravované nariadenie o súkromí a elektronických komunikáciách.

jeho presnosť) alebo zvýšením jeho presností (na úkor vysvetliteľnosti). Vždy, keď má systém umelej inteligencie značný vplyv na životy ľudí, malo by byť možné dožadovať sa vhodného vysvetlenia rozhodovacieho procesu systému umelej inteligencie. Takéto vysvetlenie by sa malo poskytnúť v primeranej lehote a malo by byť prispôsobené odborným znalostiam dotknutej zainteresovanej strany (napr. laik, regulačný orgán alebo výskumný pracovník). Okrem toho by mali byť k dispozícii vysvetlenia o tom, do akej miery systém umelej inteligencie ovplyvňuje a formuje organizačný proces rozhodovania, možnosti výberu systému a dôvody jeho zavedenia (čím by sa zabezpečila transparentnosť obchodného modelu).

78. ***Komunikácia.*** Systémy umelej inteligencie by voči používateľom nemali vystupovať ako ľudské bytosti. Ľudia majú právo vedieť, že komunikujú so systémom umelej inteligencie. To znamená, že systémy umelej inteligencie musí byť možné rozpoznať. Navyše by mala existovať možnosť odmietnuť takúto komunikáciu v prospech komunikácie s človekom, ak je takáto interakcia potrebná v záujme zabezpečenia súladu so základnými právami. Okrem toho by sa schopnosti a obmedzenia systému umelej inteligencie mali oznámiť špecialistom na umelú inteligenciu alebo koncovým používateľom takým spôsobom, ktorý je vhodný pre daný prípad použitia. Toto oznámenie by mohlo obsahovať informácie o úrovni presnosti systému umelej inteligencie, ako aj o jeho obmedzeniach.

5. Rozmanitosť, nediskriminácia a spravodlivosť

79. V záujme dosiahnutia dôveryhodnej umelej inteligencie musíme v celom životnom cykle systému umelej inteligencie zabezpečiť inklúziu a rozmanitosť. To okrem zohľadnenia a zapojenia všetkých dotknutých zainteresovaných strán počas celého procesu znamená aj to, že je potrebné zabezpečiť rovnaký prístup prostredníctvom inkluzívnych procesov navrhovania, ako aj zabezpečiť rovnaké zaobchádzanie. Táto požiadavka úzko súvisí so *zásadou spravodlivosti*.

80. ***Zabránenie nespravodlivej zaujatosti***. Dátové súbory používané systémami umelej inteligencie (pri výcviku aj prevádzke) môžu byť negatívne ovplyvnené zahrnutím neúmyselnej historickej zaujatosti, neúplnosťou a zlými modelmi riadenia. Pretrvávanie takejto zaujatosti by mohlo viesť k neúmyselným (ne)priamym predsudkom a diskriminácii[42] proti určitým skupinám alebo ľuďom, čím by sa mohli zvýšiť predsudky a marginalizácia. Ujma môže vyplynúť aj z úmyselného zneužívania (spotrebiteľskej) zaujatosti alebo zo zapojenia sa do nekalej súťaže, ako je zjednotenie cien prostredníctvom kolúzie alebo netransparentný trh[43]. Rozpoznateľná a diskriminačná zaujatosť by sa mala odstrániť, pokiaľ možno už v štádiu zhromažďovania údajov. Nespravodlivá zaujatosť môže negatívne ovplyvniť aj spôsob, akým sa systémy umelej inteligencie vyvíjajú (napr. programovanie algoritmov). Tomu by sa mohlo zabrániť zavedením postupov v oblasti dohľadu, ktoré budú jasne a transparentne analyzovať účel, obmedzenia, požiadavky a rozhodnutia systému a zaoberať sa nimi. Rozmanitosť názorov navyše možno zabezpečiť prijímaním pracovníkov rozličného pôvodu, pochádzajúcich z rôznych kultúr a disciplín, čo by sa malo podporovať.

81. ***Prístupnosť a dizajn pre všetkých***. Najmä v oblastiach, v ktorých prevládajú vzťahy medzi podnikom a koncovým zákazníkom, by sa systémy mali zameriavať na používateľa a mali by sa navrhovať tak, aby všetkým ľuďom umožňovali využívať výrobky alebo služby umelej inteligencie bez ohľadu na ich vek, pohlavie, schopnosti alebo vlastnosti. Mimoriadny význam má prístupnosť k týmto technológiám pre osoby so zdravotným postihnutím, ktoré sa nachádzajú vo všetkých spoločenských skupinách. Systémy umelej inteligencie by nemali vychádzať z univerzálneho prístupu a mali by sa v ich prípade zvážiť zásady dizajnu pre

[42] Vymedzenie pojmu priamej a nepriamej diskriminácie sa nachádza napríklad v článku 2 smernice Rady 2000/78/ES z 27. novembra 2000, ktorá ustanovuje všeobecný rámec pre rovnaké zaobchádzanie v zamestnaní a povolaní. Pozri aj článok 21 Charty základných práv EÚ.

[43] Porovnaj štúdiu Agentúry Európskej únie pre základné práva: *BigData: Discrimination in data-supported decision making* (2018) (BigData: diskriminácia v rozhodovaní s podporou údajov): http://fra.europa.eu/en/publication/2018/big-data-discrimination.

všetkých[44], ktoré sa zaoberajú najširším možným okruhom používateľov so zreteľom na príslušné normy prístupnosti[45]. Tým sa zabezpečí nestranný prístup všetkých ľudí k existujúcim a novým ľudským činnostiam sprostredkovaným počítačmi a ich aktívna účasť na nich, a to s ohľadom na podporné technológie[46].

82. ***Účasť zainteresovaných strán***. V záujme vývoja systémov umelej inteligencie, ktoré sú dôveryhodné, sa odporúča viesť konzultácie so zainteresovanými stranami, ktoré môže systém priamo alebo nepriamo ovplyvniť počas svojho životného cyklu. Užitočné je vyžadovať pravidelnú spätnú väzbu aj po zavedení a vytvoriť dlhodobejšie mechanizmy pre účasť zainteresovaných strán, napríklad tým, že sa zabezpečí informovanie pracovníkov, konzultácie s nimi a ich účasť počas celého procesu realizácie systémov umelej inteligencie v organizáciách.

6. Spoločenský a environmentálny blahobyt

83. V súlade so *zásadami spravodlivosti a prevencie ujmy* by sa za zainteresované strany počas životného cyklu umelej inteligencie mala považovať širšia spoločnosť, ostatné cítiace bytosti a životné prostredie. Mala by sa podporovať udržateľnosť a ekologická zodpovednosť systémov umelej inteligencie a výskum do riešení umelej inteligencie, ktoré sa týkajú oblastí globálneho záujmu, ako sú napríklad ciele udržateľného rozvoja. V ideálnom prípade by sa umelá inteligencia mala používať v prospech všetkých ľudí vrátane budúcich generácií.

84. ***Udržateľná umelá inteligencia, ktorá je šetrná k životnému prostrediu***. Systémy umelej inteligencie predstavujú nádej na riešenie niektorých z najnaliehavejších spoločenských otázok, avšak musí sa zabezpečiť, že k nemu bude dochádzať spôsobom, ktorý bude čo najšetrnejší k životnému prostrediu. Proces vývoja systému, jeho zavádzania a používania, ako aj celý dodávateľský reťazec by sa mali posudzovať z tohto hľadiska, napríklad prostredníctvom kritického skúmania využívania zdrojov a spotreby energie v priebehu výcviku a výberom menej škodlivých možností. Podporovať by sa mali opatrenia zabezpečujúce, aby bol celý dodávateľský reťazec systému umelej inteligencie šetrný k životnému prostrediu.

85. ***Sociálny vplyv***. Všadeprítomné vystavenie sociálnym systémom umelej inteligencie[47] vo všetkých oblastiach našich životov (či už vo vzdelávaní, v práci, starostlivosti alebo pri zábave), môže zmeniť naše chápanie sociálneho činiteľa alebo ovplyvniť naše sociálne vzťahy a príslušnosť k skupine. Hoci sa systémy umelej inteligencie môžu použiť na zlepšenie sociálnych zručností[48], môžu rovnako prispieť k ich zhoršeniu. To by mohlo ovplyvniť aj telesnú a duševnú pohodu ľudí. Účinky týchto systémov sa preto musia pozorne sledovať a posudzovať.

86. ***Spoločnosť a demokracia.*** Okrem posudzovania vplyvu vývoja systému umelej inteligencie, jeho zavádzania a používania na úrovni jednotlivca by sa tento vplyv mal posudzovať aj zo spoločenského hľadiska, v ktorom sa zohľadní ich účinok na inštitúcie, demokraciu a spoločnosť vo všeobecnosti. Využívanie systémov umelej inteligencie by sa malo starostlivo zvážiť najmä v situáciách týkajúcich sa demokratického procesu, a to nielen vrátane politického rozhodovania, ale aj v súvislosti s voľbami.

---

[44] V článku 42 smernice o verejnom obstarávaní sa vyžaduje, aby sa v technických špecifikáciách zohľadnila prístupnosť a dizajn pre všetkých.

[45] Napríklad norma EN 301 549.

[46] Táto požiadavka súvisí s Dohovorom Organizácie Spojených národov o právach osôb so zdravotným postihnutím.

[47] Tento pojem označuje systémy umelej inteligencie, ktoré komunikujú a interagujú s ľuďmi tým, že simulujú spoločenské správanie buď v rámci interakcie s humanoidným robotom (stelesnená umelá inteligencia), alebo ako avatary vo virtuálnej realite. V dôsledku toho tieto systémy disponujú potenciálom meniť naše sociálno-kultúrne zvyky a charakter nášho sociálneho života.

[48] Pozri napríklad projekt financovaný z EÚ na vývoj softvéru založeného na technológii umelej inteligencie, ktorý robotom umožňuje účinnejšie komunikovať s autistickými deťmi na terapeutických sedeniach pod vedením človeka, čím pomáha zlepšovať ich sociálne a komunikačné zručnosti: http://ec.europa.eu/research/infocentre/article_en.cfm?id=/research/headlines/news/article_19_03_12_en.html?infocentre&item=Infocentre&artid=49968.

7. Zodpovednosť

87. Požiadavka zodpovednosti dopĺňa uvedené požiadavky úzko súvisiace so *zásadou spravodlivosti*. Vyžaduje si zavedenie mechanizmov na zabezpečenie zodpovednosti systémov umelej inteligencie a ich výsledkov, a to pred ich zavedením aj po ňom.

88. ***Kontrolovateľnosť.*** Požiadavka kontrolovateľnosti má za následok možnosť posudzovať algoritmy, údaje a procesy navrhovania. To však nevyhnutne neznamená, že informácie o obchodných modeloch a duševnom vlastníctve spojené so systémom umelej inteligencie musia byť vždy verejne dostupné. Hodnotenie internými a externými audítormi a dostupnosť takýchto hodnotiacich správ môže prispieť k dôveryhodnosti tejto technológie. Pri aplikáciách, ktoré majú vplyv na základné práva, vrátane aplikácií nevyhnutných z hľadiska bezpečnosti, by malo byť možné vykonávať nezávislú kontrolu systémov umelej inteligencie.

89. ***Minimalizácia negatívnych vplyvov a ich oznamovanie***. Musí sa zabezpečiť možnosť oznamovať opatrenia alebo rozhodnutia, ktoré prispievajú k určitým výsledkom systému, ako aj schopnosť reagovať na ich následky. Identifikácia, posúdenie, nahlásenie a minimalizovanie možných negatívnych účinkov systémov umelej inteligencie sú obzvlášť dôležité pre tých, ktorých sa tieto účinky (ne)priamo dotýkajú. Oznamovateľom, mimovládnym organizáciám, odborom či iným subjektom sa pri oznamovaní oprávnených obáv týkajúcich sa systému založeného na umelej inteligencii musí poskytnúť náležitá ochrana. Na minimalizáciu negatívneho vplyvu môžu byť užitočné posúdenia vplyvu (napr. vytváranie červených tímov alebo formuláre na posúdenie vplyvu algoritmov) pred vývojom systémov umelej inteligencie, ich zavedením a používaním aj počas nich. Tieto posúdenia musia zodpovedať riziku, ktoré systémy umelej inteligencie predstavujú.

90. ***Kompromisy.*** Pri vykonávaní uvedených požiadaviek môžu medzi nimi vznikať rozpory, čo môže viesť k nevyhnutným kompromisom. Tieto kompromisy by sa mali riešiť racionálne a metodicky v rámci aktuálneho stavu. To znamená, že by sa mali určiť príslušné záujmy a hodnoty dotknuté systémom umelej inteligencie a že v prípade vzniku konfliktu by sa kompromisy medzi nimi mali jasne vziať na vedomie a vyhodnotiť z hľadiska ich rizika pre etické zásady vrátane základných práv. V situáciách, v ktorých nie je možné určiť žiaden eticky prijateľný kompromis, by sa nemalo pokračovať vo vývoji systému umelej inteligencie, jeho zavádzaní a používaní v danej podobe. Každé rozhodnutie o tom, ktorý kompromis sa má prijať, by malo byť odôvodnené a riadne zdokumentované. Rozhodovací subjekt musí niesť zodpovednosť za spôsob, akým sa prijíma príslušný kompromis, a mal by nepretržite kontrolovať primeranosť výsledného rozhodnutia, aby sa zabezpečilo, že v prípade potreby sa v systéme môžu uskutočniť potrebné zmeny. [49]

91. ***Náprava***. Keď dôjde k nespravodlivému nepriaznivému vplyvu, mali by sa stanoviť dostupné mechanizmy, ktorými sa zabezpečí primeraná náprava[50]. Vedomie, že existuje možnosť nápravy, ak sa niečo pokazí, je zásadné na zabezpečenie dôvery. Osobitná pozornosť by sa mala venovať zraniteľným osobám alebo skupinám.

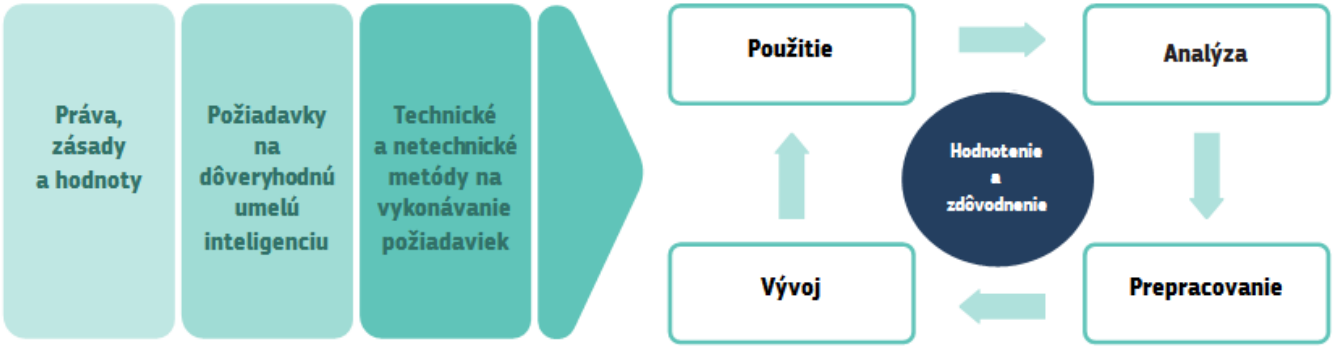
2. Technické a netechnické metódy na realizáciu dôveryhodnej umelej inteligencie

92. Na splnenie uvedených požiadaviek sa môžu využiť technické aj netechnické metódy. Tieto metódy sa týkajú všetkých štádií životného cyklu systému umelej inteligencie. Hodnotenie metód, ktoré sa uplatňujú s cieľom

[49] Tento cieľ môžu pomôcť dosiahnuť rôzne modely riadenia. Napríklad prítomnosť interného a/alebo externého odborníka na etiku (a na príslušný sektor) alebo etickej (či odvetvovej) rady by mohla byť užitočná, aby upozornili na oblasti možného konfliktu a aby navrhli spôsoby, ako daný konflikt čo najlepšie vyriešiť. Užitočné sú aj zmysluplné konzultácie a diskusie so zainteresovanými stranami vrátane subjektov, ktorým hrozí riziko, že sa ich systém umelej inteligencie nepriaznivo dotkne. Európske univerzity by mali prevziať vedúcu úlohu pri odbornej príprave potrebných odborníkov na etiku.

[50] Pozri aj stanovisko Agentúry Európskej únie pre základné práva s názvom *Improving access to remedy in the area of business and human rights at the EU level* (Zlepšenie dostupnosti nápravy v oblasti podnikania a ľudských práv na úrovni EÚ) (2017): https://fra.europa.eu/en/opinion/2017/business-human-rights.

plniť dané požiadavky, ako aj nahlasovanie a odôvodňovanie[51] zmien v procesoch vykonávania, by mali prebiehať priebežne. Keďže sa systémy umelej inteligencie neustále vyvíjajú a pôsobia v dynamickom prostredí, realizácia dôveryhodnej umelej inteligencie je nepretržitý proces, znázornený na obrázku 5.



Obrázok 3: Realizácia dôveryhodnej umelej inteligencie počas celého životného cyklu systému

93. Tieto metódy sa môžu chápať ako metódy, ktoré sa buď navzájom dopĺňajú, alebo sa navzájom nahrádzajú, keďže rôzne požiadavky (a rôzna citlivosť) môžu vyvolať potrebu rozdielnych metód vykonávania. Tento prehľad nemá byť komplexný, úplný ani záväzný. Jeho cieľom je skôr predložiť zoznam navrhovaných metód, ktoré môžu pomôcť pri vykonávaní dôveryhodnej umelej inteligencie.

1. Technické metódy

94. Tento oddiel obsahuje opis technických metód na zabezpečenie dôveryhodnej umelej inteligencie, ktoré je možné začleniť do fáz navrhovania systému umelej inteligencie, jeho vývoja a používania. Metódy uvedené ďalej sa od seba líšia v úrovni vyspelosti[52].

- *Architektúry pre dôveryhodnú umelú inteligenciu*

95. Požiadavky na dôveryhodnú umelú inteligenciu by sa mali premeniť na postupy a/alebo obmedzenia postupov, ktoré by mali byť zakotvené v architektúre systému umelej inteligencie. Tento cieľ by sa mohol dosiahnuť prostredníctvom súboru pravidiel tzv. bieleho zoznamu (správanie alebo stavy), ktorým by sa systém mal vždy riadiť, prostredníctvom obmedzení správania a stavov podľa tzv. čierneho zoznamu, ktoré by systém nikdy nemal porušiť, a kombinácie obidvoch súborov pravidiel alebo komplexnejších overiteľných záruk týkajúcich sa správania systému. Monitorovanie súladu systému s týmito obmedzeniami počas operácií sa môže uskutočniť podľa samostatného procesu.

96. Systémy umelej inteligencie so schopnosťou učenia sa, ktoré môžu dynamicky prispôsobiť svoje správanie, sa môžu chápať ako nedeterministický systém, ktorý sa môže prejavovať neočakávaným správaním. Často sa posudzujú teoretickou optikou cyklu „detekcie – plánovania – konania“. Prispôsobenie tejto architektúry na účely zabezpečenia dôveryhodnej umelej inteligencie si vyžaduje zapojenie požiadaviek vo všetkých troch krokoch tohto cyklu: i) v kroku týkajúcom sa „detekcie“ by systém mal byť vyvinutý tak, aby rozpoznal všetky prvky prostredia potrebné na zabezpečenie dodržania požiadaviek; ii) v kroku súvisiacom s „plánovaním“ by systém mal zvažovať iba plány, ktoré sú v súlade s požiadavkami; iii) v kroku „konanie“ by sa opatrenia systému

[51] Patrí sem napríklad vysvetlenie rozhodnutí prijatých pri navrhovaní, vývoji a zavádzaní systému s cieľom zapracovať do neho uvedené požiadavky.

[52] Hoci niektoré z týchto metód sú dostupné už teraz, v prípade ostatných je stále potrebné uskutočniť ďalší výskum. Oblasti, v ktorých je potrebný ďalší výskum, takisto poskytnú informácie pre druhý výstup expertnej skupiny na vysokej úrovni pre umelú inteligenciu, t. j. pre politiku v oblasti umelej inteligencie a investičné odporúčania.

mali obmedziť na správanie, ktorým sa tieto požiadavky realizujú.

97. Tento náčrt architektúry je všeobecný a poskytuje iba nedokonalý opis väčšiny systémov umelej inteligencie. Poskytuje však záchytné body týkajúce sa obmedzení a prístupov, ktoré by sa mali prejaviť v konkrétnych moduloch vedúcich k celkovému systému, ktorý bude dôveryhodný a bude sa za taký aj považovať.

- *Etika a právny štát už v štádiu návrhu (X v štádiu návrhu)*

98. Metódy na zabezpečenie hodnôt v štádiu návrhu tvoria presné a jasné prepojenie medzi abstraktnými zásadami, ktoré systém musí dodržiavať, a konkrétnymi rozhodnutiami súvisiacimi s vykonávaním. Rozhodujúcou pre túto metódu je myšlienka, že súlad s normami možno zapracovať do návrhu systému umelej inteligencie. Spoločnosti nesú zodpovednosť za určenie vplyvu svojich systémov umelej inteligencie od úplného začiatku, ako aj za identifikovanie noriem, ktoré by ich systém mal dodržiavať s cieľom zabrániť negatívnym následkom. V súčasnosti sa už široko používajú rôzne iné koncepcie „v štádiu návrhu“, napr. *ochrana súkromia v štádiu návrhu a bezpečnosť v štádiu návrhu.* Ako sme už uviedli, na to, aby si umelá inteligencia zaslúžila dôveru, musia byť jej procesy, údaje a výsledky bezpečné a mali by byť navrhnuté tak, aby boli odolné voči nepriateľským údajom a útokom. Mali by využívať mechanizmus na núdzové vypnutie a umožňovať obnovenie operácie po vynútenom vypnutí (napr. v dôsledku útoku).

- *Metódy vysvetľovania*

99. Na to, aby bol systém dôveryhodný, musíme byť schopní pochopiť, prečo sa správal určitým spôsobom a prečo poskytol daný výklad. Riešením tejto otázky v záujme lepšieho porozumenia príslušných mechanizmov systému a hľadaním riešení sa zaoberá celá oblasť výskumu, vysvetliteľná umelá inteligencia. Pre systémy umelej inteligencie založené na neurónových sieťach je to ešte aj v súčasnosti otvorená výzva. Procesy výcviku s neurónovými sieťami môžu viesť k parametrom siete nastaveným na číselné hodnoty, ktoré je ťažko uviesť do súladu s výsledkami. Navyše občas malé zmeny hodnôt údajov môžu viesť k dramatickým zmenám interpretácie, čo má za následok, že systém napríklad zamení školský autobus za pštrosa. Túto zraniteľnosť takisto možno zneužiť počas útokov na systém. Metódy týkajúce sa výskumu v oblasti vysvetliteľnej umelej inteligencie nie sú rozhodujúce iba na vysvetlenie správania systému voči používateľom, ale aj na zavedenie spoľahlivej technológie.

- *Testovanie a overovanie*

100. Z dôvodu nedeterministickej a kontextovej povahy systémov umelej inteligencie tradičné testovanie nestačí. Zlyhania koncepcií a tvrdení, ktoré systém využíva, sa môžu prejaviť iba vtedy, keď sa program použije na dostatočne realistické údaje. Následkom toho v záujme overenia a potvrdenia spracovania údajov sa príslušný model musí starostlivo monitorovať počas výcviku aj počas zavádzania z hľadiska jeho stability, odolnosti a prevádzky v rámci správne chápaných a predvídateľných ohraničení. Musí sa zabezpečiť, aby výsledok plánovacieho procesu zodpovedal vstupu a aby sa rozhodnutia prijímali spôsobom, ktorý umožní overenie príslušného procesu.

101. Testovanie a overovanie systému by sa malo uskutočniť čo najskôr, čím sa zabezpečí, že systém sa počas celého svojho životného cyklu, a najmä po zavedení bude správať zamýšľaným spôsobom. Malo by sa týkať všetkých zložiek systému umelej inteligencie vrátane údajov, nacvičených modelov, prostredí a správania systému ako celku. Testovanie a overovanie by mala navrhnúť a vykonať čo najrozmanitejšia skupina ľudí. Vyvinúť by sa mala rôznorodá metrika vzťahujúca sa na kategórie, ktoré sa testujú z rôznych uhlov pohľadu. Zvážiť možno testovanie nepriateľských útokov dôveryhodnými a rôznorodými tzv. červenými tímami, ktoré sa úmyselne pokúšajú „prelomiť“ systém s cieľom odhaľovať zraniteľné miesta, a „bug bounties“ (odmeny za objav chyby), ktoré motivujú outsiderov, aby odhaľovali a zodpovedne oznamovali systémové chyby a nedostatky. A napokon sa musí zabezpečiť, aby boli výstupy alebo opatrenia v súlade s výsledkami predchádzajúcich procesov, pričom sa porovnajú s politikami vymedzenými v minulosti s cieľom zabezpečiť, aby nedošlo ich porušeniu.

- *Ukazovatele kvality služby*

102. V prípade systémov umelej inteligencie možno definovať primerané ukazovatele kvality služby s cieľom získať základné poznatky o tom, či boli testované a či sa vyvíjali s prihliadnutím na bezpečnostnú ochranu a bezpečnosť. Tieto ukazovatele by mohli zahŕňať opatrenia na vyhodnotenie testovacích a výcvikových algoritmov, ako aj bežnú softvérovú metriku funkčnosti, výkonnosti, použiteľnosti, spoľahlivosti, bezpečnosti a udržiavateľnosti.

2. Netechnické metódy

103. Tento oddiel obsahuje opis rozmanitých netechnických metód, ktoré môžu zohrávať cennú úlohu pri zabezpečovaní a udržiavaní dôveryhodnej umelej inteligencie. Aj tieto metódy by sa mali hodnotiť **priebežne**.

- *Právny rámec*

104. Ako sme už uviedli, právny rámec na podporu dôveryhodnosti umelej inteligencie už v súčasnosti existuje – pripomeňme právne predpisy v oblasti bezpečnosti výrobkov a rámce zodpovednosti. Podľa miery, v akej právne predpisy podľa nás môžu potrebovať revíziu, úpravu alebo zavedenie (ako ochranné opatrenie a aj ako nástroj), sa táto otázka predloží v našom druhom výstupe, ktorý sa bude týkať politiky v oblasti umelej inteligencie a investičných odporúčaní.

- *Kódexy správania*

105. Organizácie a zainteresované strany sa môžu zaviazať dodržiavať usmernenia a môžu upraviť svoju chartu sociálnej zodpovednosti podnikov, kľúčové ukazovatele výkonnosti (KPI), svoje kódexy správania alebo dokumenty o vnútorných postupoch s cieľom doplniť úsilie o dosiahnutie dôveryhodnej umelej inteligencie. Organizácia, ktorá pracuje na systéme umelej inteligencie, môže vo všeobecnosti doložiť svoje úmysly a potvrdiť ich normami istých žiaducich hodnôt, ako sú základné práva, transparentnosť a predchádzanie ujme.

- *Normalizácia*

106. Normy, napr. týkajúce sa navrhovania, výroby a podnikateľskej praxe, môžu pre používateľov umelej inteligencie, spotrebiteľov, organizácie, výskumné inštitúcie a vlády plniť funkciu systému riadenia kvality tým, že ponúkajú možnosť uznávať a podporovať etické správanie prostredníctvom ich nákupných rozhodnutí. Okrem zvyčajných noriem existujú aj koregulačné prístupy: akreditačné systémy, profesijné etické kódexy alebo normy týkajúce sa návrhov v súlade so základnými právami. Súčasnými príkladmi sú napr. normy ISO alebo skupina noriem IEEE P7000, v budúcnosti by však mohlo byť vhodné označenie „Dôveryhodná umelá inteligencia“, ktorým sa prostredníctvom odkazu na konkrétne technické normy potvrdí, že systém vyhovuje napríklad požiadavkám na bezpečnosť, technickú odolnosť a vysvetliteľnosť.

- *Certifikácia*

107. Keďže nemožno očakávať, že každý je schopný plne chápať fungovanie a účinky systémov umelej inteligencie, organizáciám, ktoré môžu širšej verejnosti dosvedčiť, že systém umelej inteligencie je transparentný, zodpovedný a spravodlivý, sa môže udeliť certifikácia[53]. Pri týchto certifikáciách sa budú uplatňovať normy vypracované pre rôzne oblasti použitia a technológie umelej inteligencie, ktoré sú primerane v súlade s priemyselnými a spoločenskými normami v rámci iného kontextu. Certifikácia však nikdy nemôže nahradiť zodpovednosť. Mali by ju teda dopĺňať rámce pre zodpovednosť vrátane vyhlásení o odmietnutí zodpovednosti, ako aj systémov preskúmania a mechanizmov nápravy[54].

[53] Tento prístup podporuje napr. iniciatíva IEEE pre etický dizajn: https://standards.ieee.org/industry-connections/ec/autonomous-systems.html.

[54] Viac informácií o obmedzeniach certifikácie sa nachádza na adrese: https://ainowinstitute.org/AI_Now_2018_Report.pdf.

- *Zodpovednosť prostredníctvom rámcov riadenia*

108. Organizácie by mali vytvárať interné aj externé rámce riadenia, ktoré budú zabezpečovať zodpovednosť za etickú dimenziu rozhodnutí spojovaných s vývojom umelej inteligencie, jej zavádzaním a používaním. To môže znamenať napríklad vymenovanie osoby, ktorá bude poverená etickými otázkami súvisiacimi s umelou inteligenciou, alebo interný/externý etický panel alebo etickú radu. Okrem možných úloh takejto osoby, panelu alebo rady bude ich úlohou zabezpečiť dohľad a poskytovať poradenstvo. Ako sme už uviedli, úlohu v tomto smere môžu zohrať aj certifikačné špecifikácie a/alebo orgány. V spolupráci s odvetvím a/alebo so skupinami pre verejný dohľad by sa mali zabezpečiť komunikačné kanály, prostredníctvom ktorých sa budú vymieňať najlepšie postupy, rozoberať dilemy alebo oznamovať vznikajúce problémy súvisiace s etickými obavami. Tieto mechanizmy môžu dopĺňať zákonný dohľad, nemôžu ho však nahrádzať (napr. v podobe vymenovania osoby zodpovednej za údaje alebo rovnocenných opatrení vyžadovaných zákonom v rámci právnych predpisov v oblasti ochrany údajov).

- *Vzdelanie a informovanosť na podporu etických postojov*

109. Dôveryhodná umelá inteligencia podporuje informovanú účasť všetkých zainteresovaných strán. Významnú úlohu zohráva komunikácia, vzdelávanie a odborná príprava, aby sa zabezpečilo rozšírenie poznatkov o možnom vplyve systémov umelej inteligencie, aj aby ľudia vedeli, že sa môžu zúčastniť na formovaní spoločenského vývoja. To sa týka všetkých zainteresovaných strán, napr. subjektov zapojených do tvorby produktov (projektanti a vývojári), používateľov (spoločnosti alebo jednotlivci) a iných dotknutých skupín (osoby, ktoré možno nekúpili ani nepoužívajú systém umelej inteligencie, o ktorých však systém umelej inteligencie rozhoduje, a spoločnosť vo všeobecnosti). V spoločnosti by sa mala podporovať základná gramotnosť v oblasti umelej inteligencie. Podmienkou vzdelávania verejnosti je zabezpečiť náležité zručnosti a odbornú prípravu etikov v tento oblasti.

- *Účasť zainteresovaných strán a sociálny dialóg*

110. Prínosov umelej inteligencie je veľa a Európa musí zabezpečiť, že budú k dispozícii všetkým. To si vyžaduje viesť otvorenú diskusiu a zapojiť sociálnych partnerov, zainteresované strany aj širokú verejnosť. Veľa organizácií už na diskusie o používaní systémov umelej inteligencie a analýzu údajov využíva panely zainteresovaných strán. V týchto paneloch zasadajú rôzni členovia, ako sú právni experti, technickí experti, etici, zástupcovia spotrebiteľov a pracovníci. Aktívne úsilie o účasť a dialóg o používaní a vplyve systémov umelej inteligencie pomáha pri hodnotení výsledkov a prístupov a môže byť obzvlášť užitočné v zložitých prípadoch.

- *Rozmanitosť a inkluzívne projektové tímy*

111. Rozmanitosť a inklúzia zohrávajú zásadnú úlohu vo vývoji systémov umelej inteligencie, ktoré sa budú využívať v skutočnom svete. Je kľúčové, že kým systémy umelej inteligencie plnia čoraz viac úloh samostatne, tímy, ktoré ich projektujú, vyvíjajú, testujú a udržiavajú, zavádzajú a/alebo obstarávajú, odzrkadľujú rozmanitosť používateľov a spoločnosti vo všeobecnosti. Prispieva to k objektivite a zohľadňovaniu rozličných stanovísk, potrieb a cieľov. V ideálnom prípade tieto tímy nie sú rozmanité iba z hľadiska pohlavia, kultúry a veku, ale aj z hľadiska odborného zázemia a súboru zručností.

Kľúčové usmernenia vyplývajúce z kapitoly II:

- ✓ Zabezpečiť, aby celý životný cyklus systému umelej inteligencie spĺňal požiadavky na dôveryhodnú umelú inteligenciu: 1. ľudský faktor a dohľad; 2. technická odolnosť a bezpečnosť; 3. správa súkromia a údajov; 4. transparentnosť; 5. rozmanitosť, nediskriminácia a spravodlivosť; 6. environmentálny a spoločenský blahobyt a 7. zodpovednosť.
- ✓ Zvážiť technické a netechnické metódy na zabezpečenie vykonávania týchto požiadaviek.

- ✓ Podporovať výskum a inováciu s cieľom pomôcť pri posudzovaní systémov umelej inteligencie a presadzovaní plnenia požiadaviek. Zverejňovať výsledky a otvorené otázky širšej verejnosti a systematicky pripravovať novú generáciu odborníkov na etiku v oblasti umelej inteligencie.
- ✓ Jasným a iniciatívnym spôsobom poskytovať zainteresovaným stranám informácie o schopnostiach a obmedzeniach systému umelej inteligencie, čo im umožní vytvoriť si realistické očakávania, a o spôsobe, akým dochádza k plneniu požiadaviek. Nezakrývať skutočnosť, že majú do činenia so systémom umelej inteligencie.
- ✓ Uľahčiť vysledovateľnosť a kontrolovateľnosť systémov umelej inteligencie, najmä v kritických kontextoch a situáciách.
- ✓ Zapájať zainteresované strany počas celého životného cyklu systému umelej inteligencie. Podporovať odbornú prípravu a vzdelávanie tak, aby boli všetky zainteresované strany informované a vyškolené v oblasti dôveryhodnej umelej inteligencie.
- ✓ Mať na pamäti, že môžu existovať zásadné rozpory medzi jednotlivými zásadami a požiadavkami. Nepretržite identifikovať, vyhodnocovať a dokumentovať tieto kompromisy a ich riešenia a informovať o nich.

III. Kapitola III: Posudzovanie dôveryhodnej umelej inteligencie

112. Na základe kľúčových požiadaviek v kapitole II sa v tejto kapitole uvádza neúplný **zoznam bodov na posudzovanie dôveryhodnej umelej inteligencie** (pilotná verzia) na účely **jej uvedenia do praxe**. Tento zoznam sa týka osobitne systémov umelej inteligencie, ktoré priamo komunikujú s používateľmi, a je určený v prvom rade vývojárom a prevádzkovateľom systémov umelej inteligencie (ktoré si buď vyvinuli sami, alebo ich získali od tretích strán). Tento zoznam bodov na posudzovanie sa nezaoberá prvou zložkou dôveryhodnej umelej inteligencie (zákonnou umelou inteligenciou) a jej uvedením do praxe. Súlad s týmto zoznamom bodov na posudzovanie nepredstavuje dôkaz o zhode s právnymi predpismi, ani nemá slúžiť ako pomôcka na zabezpečenie súladu s platným právom. Keďže každá aplikácia systémov umelej inteligencie je špecifická, zoznam bodov na posudzovanie sa bude musieť prispôsobovať podľa konkrétnych prípadov použitia a kontextov, v ktorých sa systémy prevádzkujú. Táto kapitola navyše obsahuje všeobecné odporúčanie o tom, ako zaviesť zoznam bodov na posudzovanie dôveryhodnej umelej inteligencie prostredníctvom štruktúry riadenia, ktorá zahŕňa prevádzkovú úroveň aj úroveň manažmentu.

113. Zoznam bodov na posudzovanie a štruktúra riadenia sa vypracujú v úzkej spolupráci so zainteresovanými stranami z verejného a súkromného sektora. Proces bude riadený ako „pilotný“ proces, ktorý umožní rozsiahlu spätnú väzbu z dvoch paralelných procesov:

    a) kvalitatívny proces, ktorým sa zabezpečí reprezentatívnosť v prípade, keď sa do zavádzania zoznamu bodov na posudzovanie a štruktúry riadenia do praxe prihlási a podrobnú spätnú väzbu poskytne úzky okruh spoločností, organizácií a inštitúcií (z rôznych sektorov a rozličnej veľkosti);

    b) kvantitatívny proces, pri ktorom sa do zavádzania zoznamu bodov na posudzovanie môžu prihlásiť a spätnú väzbu prostredníctvom otvorenej konzultácie môžu poskytovať všetky dotknuté zainteresované strany.

114. Po pilotnej fáze začleníme výsledky zo spracovania spätnej väzby do zoznamu bodov na posudzovanie a začiatkom roku 2020 pripravíme revidovanú verziu. Cieľom je vytvorenie rámca, ktorý sa môže používať horizontálne vo všetkých aplikáciách, a tak bude tvoriť základ na zabezpečenie dôveryhodnej umelej inteligencie vo všetkých oblastiach. Po ustanovení tohto základu sa bude môcť vytvoriť odvetvový rámec alebo rámec pre konkrétne aplikácie.

Riadenie

115. Spoločnosti, organizácie a inštitúcie by mohli zvážiť, ako by sa v ich organizácii mohol zaviesť zoznam bodov na posudzovanie dôveryhodnej umelej inteligencie. Môžu to spraviť tak, že proces posudzovania začlenia do existujúcich mechanizmov riadenia alebo zavedú nové procesy. Toto rozhodnutie bude závisieť od internej štruktúry organizácie, ako aj od jej veľkosti a dostupných zdrojov.

116. Z výskumu[55] vyplýva, že na dosiahnutie zmeny je nevyhnutná pozornosť manažmentu na najvyššej úrovni. Takisto sa ukazuje, že zapojenie všetkých zainteresovaných strán v spoločnosti, organizácii alebo inštitúcii pomáha prijatiu a relevantnosti zavedenia akéhokoľvek nového procesu (bez ohľadu na to, či je technologický alebo nie)[56]. Odporúčame preto zaviesť proces, ktorého súčasťou je zapojenie prevádzkovej úrovne, ako aj vrcholového manažmentu.

Úroveň	**Príslušné úlohy (závisí od organizácie)**
Vedenie a správna rada	Vrcholový manažment vedie diskusie o vývoji umelej inteligencie, jej zavádzaní alebo obstarávaní a vyhodnocuje ich a slúži ako rada pre riešenie sporov na hodnotenie všetkých inovácií a použití umelej inteligencie v prípade zistenia vážnych obáv. To sa týka osôb postihnutých možným zavedením systémov umelej inteligencie (napr. pracovníkov) a ich zástupcov počas celého procesu prostredníctvom postupov na informovanie, konzultácie a účasť.
Útvar funkcie dodržiavania súladu/právne oddelenie/ útvar pre sociálnu zodpovednosť podnikov	Útvar pre sociálnu zodpovednosť podnikov monitoruje používanie zoznamu bodov na posudzovanie a jeho nevyhnutný vývoj v záujme zvládnutia technologických alebo regulačných zmien. Aktualizuje normy alebo interné postupy týkajúce sa systémov umelej inteligencie a stará sa o to, aby bolo použitie takýchto systémov v súlade so súčasným právnym a regulačným rámcom a s hodnotami organizácie.
Vývoj výrobkov a služieb alebo rovnocenný útvar	Útvar pre vývoj výrobkov a služieb využíva zoznam bodov na posudzovanie na vyhodnotenie výrobkov a služieb založených na umelej inteligencii a zaznamenáva všetky výsledky. Tieto výsledky sú predmetom diskusií na úrovni manažmentu, na ktorej sa definitívne schvaľujú nové alebo prepracované aplikácie umelej inteligencie.
Zabezpečovanie kvality	Útvar na zabezpečovanie kvality (alebo rovnocenný útvar) zabezpečuje a overuje výsledky zoznamu bodov na posudzovanie a prijíma opatrenia na riešenie otázky na vyššej úrovni, ak výsledky nie sú uspokojivé alebo ak sa zistili nepredvídané výsledky.
Ľudské zdroje	Oddelenie ľudských zdrojov zabezpečuje správnu kombináciu spôsobilostí a rozmanitých profilov vývojárov systémov umelej inteligencie. Zabezpečuje, aby sa v rámci organizácie poskytovala primeraná úroveň odbornej prípravy v oblasti dôveryhodnej umelej inteligencie.
Obstarávanie	Útvar pre obstarávanie zabezpečuje, aby proces obstarávania výrobkov alebo služieb založených na umelej inteligencii obsahoval kontrolu dôveryhodnej umelej

[55] https://www.mckinsey.com/business-functions/operations/our-insights/secrets-of-successful-change-implementation.

[56] Pozri napríklad Bryson, A., Barth, E., a Dale-Olsen, H., *The Effects of Organisational change on worker well-being and the moderating role of trade unions* (Účinky organizačných zmien na pohodu pracovníkov a moderačná úloha odborov), *ILRReview*, roč. 66, č. 4, júl 2013; Jirjahn, U., a Smith, S.C., (2006), *What Factors Lead Management to Support or Oppose Employee Participation – With and Without Works Councils? Hypotheses and Evidence from Germany's Industrial Relations* (Ktoré faktory ovplyvňujú rozhodnutie vedenia podporovať účasť zamestnancov alebo sa proti nej postaviť – so zamestnaneckými radami a bez nich? Hypotézy a dôkazy z pracovnoprávnych vzťahov v Nemecku), roč. 45, č. 4, s. 650 – 680; Michie, J., a Sheehan, M., (2003), *Labour market deregulation, „flexibility" and innovation* (Deregulácia pracovného trhu, „flexibilita" a inovácia), *Cambridge Journal of Economics*, roč. 27, č. 1, s. 123 – 143.

inteligencie.

Každodenné operácie	Vývojári a projektoví manažéri používajú zoznam bodov na posudzovanie pri svojej každodennej práci a dokumentujú výsledky a závery posúdenia.

Používanie zoznamu bodov na posudzovanie dôveryhodnej umelej inteligencie

117. Pri používaní zoznamu bodov na posudzovanie v praxi odporúčame nevenovať pozornosť iba oblastiam záujmu, ale aj otázkam, na ktoré nie je možné dať (jednoduchú) odpoveď. Jedným z možných problémov by mohla byť nedostatočná rozmanitosť zručností a spôsobilostí členov tímu, ktorý vyvíja a testuje systém umelej inteligencie, a preto môže byť nevyhnutné zapojiť ďalšie zainteresované strany z prostredia organizácie alebo mimo nej. Dôrazne sa odporúča zaznamenávať všetky výsledky z technického hľadiska aj z hľadiska manažmentu, čím sa zabezpečí, že riešenie problémov bude pochopené na všetkých úrovniach štruktúry riadenia.

118. Tento zoznam bodov na posudzovanie je určený ako pomôcka pre špecialistov na umelú inteligenciu na účely vývoja dôveryhodnej umelej inteligencie, jej zavádzania a používania. Posúdenie by malo byť primerane prispôsobené konkrétnemu prípadu použitia. Počas pilotnej fázy môžu byť odhalené konkrétne citlivé oblasti a potreba ďalších špecifikácií v takýchto prípadoch sa vyhodnotí v nasledujúcom kroku. Hoci tento zoznam bodov na posudzovanie neposkytuje konkrétne odpovede na riešenie položených otázok, motivuje k premýšľaniu o krokoch, ktoré môžu pomôcť zabezpečiť dôveryhodnosť systémov umelej inteligencie, a o možných krokoch, ktoré by sa v tejto súvislosti mali prijať.

Súvislosť s existujúcim právom a procesmi

119. Pre subjekty zapojené do vývoja umelej inteligencie, jej zavádzania a používania je takisto dôležité uznať, že existujú rozličné platné právne predpisy, ktorými sa stanovujú konkrétne procesy a zakazujú určité výsledky, ktoré sa môžu prekrývať a kolidovať s niektorými z opatrení uvedených v zozname bodov na posudzovanie. Napríklad v právnych predpisoch v oblasti ochrany osobných údajov sa stanovuje niekoľko právnych požiadaviek, ktoré musia byť splnené zo strany účastníkov zberu a spracovávania osobných údajov. Keďže si však dôveryhodná umelá inteligencia takisto vyžaduje etické nakladanie s údajmi, takéto nakladanie s údajmi môžu pomôcť uľahčiť aj interné postupy a politiky zamerané na zabezpečenie súladu s právnymi predpismi v oblasti ochrany osobných údajov, a tak môže dopĺňať platné právne procesy. Súlad s týmto zoznamom bodov na posudzovanie však *nepredstavuje* dôkaz o zhode s právnymi predpismi, ani nemá slúžiť ako pomôcka na zabezpečenie súladu s platným právom. Cieľom tohto zoznamu je skôr poskytnúť súbor konkrétnych otázok adresátom, ktorých účelom je zabezpečiť, že ich prístup k vývoju a zavádzaniu umelej inteligencie bude zameraný na dôveryhodnú umelú inteligenciu, ktorú sa budú snažiť zaistiť.

120. Podobne veľa špecialistov na umelú inteligenciu už má nástroje na posúdenie a postupy na vývoj softvéru, ktorými zabezpečujú súlad aj s inými než právnymi normami. Posúdenie uvedené ďalej by sa nemalo nevyhnutne vykonávať samostatne, môže sa však začleniť do podobných existujúcich postupov.

ZOZNAM BODOV NA POSUDZOVANIE DÔVERYHODNEJ UMELEJ INTELIGENCIE (PILOTNÁ VERZIA)

1. Ľudský faktor a dohľad

Základné práva:

- ✓ V tých prípadoch použitia, v ktorých môže potenciálne dochádzať k negatívnemu vplyvu na základné práva, vykonali ste posúdenie vplyvu na základné práva? Identifikovali ste a zdokumentovali možné kompromisy medzi jednotlivými zásadami a právami?

- ✓ Ovplyvňuje systém umelej inteligencie rozhodovanie zo strany ľudských koncových používateľov (napr. odporúčané opatrenia alebo rozhodnutia, ktoré sa majú prijať, predloženie možností)?

  - Existuje v týchto prípadoch riziko, že sa systém umelej inteligencie dotkne ľudskej autonómie zasahovaním do rozhodovacieho procesu koncového používateľa spôsobom, ktorý nebol plánovaný?

  - Uvažovali ste o tom, či by mal systém umelej inteligencie používateľov informovať o tom, že rozhodnutie, obsah, rada alebo výsledok sú následkom algoritmického rozhodovania?

  - Ak je systém umelej inteligencie vybavený chatbotom alebo konverzačným systémom, sú ľudskí koncoví používatelia informovaní o tom, že nekomunikujú s človekom?

Ľudský faktor:

- ✓ Ak sa systém umelej inteligencie zavádza do pracovného a zamestnaneckého procesu, uvažovali ste o rozdelení úloh medzi systém umelej inteligencie a ľudských pracovníkov z hľadiska zmysluplných interakcií a primeraného ľudského dohľadu a kontroly?

  - Zvyšuje alebo posilňuje systém umelej inteligencie ľudské schopnosti?

  - Prijali ste bezpečnostné opatrenia s cieľom zabrániť prílišnej dôvere voči systému umelej inteligencie alebo nadmernej závislosti od neho v pracovných procesoch?

Ľudský dohľad:

- ✓ Uvažovali ste o tom, aká úroveň ľudskej kontroly by bola primeraná pre konkrétny systém umelej inteligencie a prípad použitia?

  - V náležitých prípadoch, môžete opísať úroveň ľudskej kontroly alebo zapojenia? Kto je „osobou poverenou vykonávať kontrolu“ a v akých situáciách dochádza k zásahu človeka alebo aké nástroje pri tom využíva?

  - Zaviedli ste mechanizmy a opatrenia, aby ste zabezpečili túto možnú ľudskú kontrolu či dohľad alebo aby ste zabezpečili, že rozhodnutia sa budú prijímať v rámci celkovej zodpovednosti ľudí?

  - Prijali ste nejaké opatrenia, aby ste umožnili audit a napravili problémy týkajúce sa správy autonómie umelej inteligencie?

- ✓ V prípade samoučiaceho sa alebo autonómneho systému umelej inteligencie alebo prípadu použitia takéhoto systému, zaviedli ste konkrétnejšie mechanizmy kontroly a dohľadu?

  - Aký druh mechanizmu odhaľovania a reakcií ste stanovili na posúdenie toho, či by mohli nastať problémy?

  - Zabezpečili ste vytvorenie tlačidla „stop“ alebo postupu na bezpečné zastavenie operácie v prípade potreby? Zastaví sa týmto postupom daný proces úplne, čiastočne alebo sa ním prevedie kontrola na človeka?

2. <u>Technická odolnosť a bezpečnosť</u>

Odolnosť voči útokom a bezpečnostná ochrana:

- ✓ Posúdili ste možné formy útoku, voči ktorým by systém umelej inteligencie mohol byť zraniteľný?
  - Uvažovali ste konkrétne o rôznych typoch a odlišnom charaktere zraniteľností, ako je znečistenie údajov, fyzická infraštruktúra, kybernetické útoky?
- ✓ Zaviedli ste opatrenia alebo systémy, aby sa zabezpečila integrita a odolnosť systému umelej inteligencie proti možným útokom?
- ✓ Posúdili ste, ako sa správa váš systém v nepredvídaných situáciách a prostrediach?
- ✓ Uvažovali ste o tom, či by váš systém mohol mať dvojaké použitie a do akej miery by ho mohol mať? Ak áno, prijali ste vhodné preventívne opatrenia proti tejto situácii (vrátane napríklad nezverejnenia výskumu alebo nezavedenia systému)?

Záložný plán a všeobecná bezpečnosť:

- ✓ Zabezpečili ste, aby mal váš systém dostatočný záložný plán v prípade, ak dôjde k nepriateľským útokom alebo ak nastanú iné neočakávané situácie (napr. technické postupy zmeny alebo vyžiadanie si ľudského operátora pred tým, než je možné pokračovať)?
- ✓ Uvažovali ste o úrovni rizika, ktoré vyvoláva systém umelej inteligencie v tomto konkrétnom prípade použitia?
  - Zaviedli ste nejaký proces na meranie a posúdenie rizík a bezpečnosti?
  - Poskytli ste potrebné informácie v prípade rizika pre telesnú nedotknuteľnosť ľudí?
  - Uvažovali ste o nejakej forme poistenia pre prípad možnej škody spôsobenej systémom umelej inteligencie?
  - Identifikovali ste možné bezpečnostné riziká (iných) predvídateľných použití technológií vrátane ich náhodného alebo zlomyseľného zneužitia? Máte plán na zmiernenie alebo na zvládanie týchto rizík?
- ✓ Posúdili ste, či existuje pravdepodobná možnosť, že by systém umelej inteligencie mohol spôsobiť škodu alebo ujmu používateľom alebo tretím stranám? Ak áno, posúdili ste jej pravdepodobnosť, potenciálnu škodu, postihnutú populáciu a závažnosť?
  - V prípade, ak existuje riziko, že by systém umelej inteligencie mohol spôsobiť škody, uvažovali ste o pravidlách týkajúcich sa zodpovednosti za škody a ochrany spotrebiteľov, a ako ste ich zohľadnili?
  - Zvážili ste možný vplyv alebo bezpečnostné riziko pre životné prostredie alebo zvieratá?
  - Zohľadnili ste vo svojej analýze rizika, či problémy s bezpečnosťou alebo sieťou (napríklad nebezpečenstvo v oblasti kybernetickej bezpečnosti) predstavujú bezpečnostné riziká alebo škody z dôvodu neúmyselného správania systému umelej inteligencie?
- ✓ Spravili ste odhad pravdepodobného dosahu zlyhania vášho systému umelej inteligencie, ktoré vedie

k nesprávnym výsledkom, ktoré vedie k nedostupnosti vášho systému alebo k tomu, že váš systém poskytuje spoločensky neprijateľné výsledky (napr. diskriminačné postupy)?

- Vymedzili ste prahové hodnoty a spôsob riadenia pre uvedené scenáre, na základe ktorých sa spustí vykonávanie alternatívnych či záložných plánov?
- Stanovili a vyskúšali ste záložné plány?

Presnosť

✓ Posúdili ste, aká úroveň a definícia presnosti bude potrebná v kontexte systému umelej inteligencie a prípadu použitia?

- Posúdili ste, akým spôsobom sa presnosť meria a zaisťuje?
- Zaviedli ste opatrenia na zabezpečenie úplnosti a aktuálnosti použitých údajov?
- Zaviedli ste opatrenia na posúdenie toho, či sú potrebné dodatočné údaje, napríklad na zvýšenie presnosti alebo na odstránenie zaujatosti?

✓ Posúdili ste ujmu, ktorá by vznikla v prípade, ak by systém umelej inteligencie robil nepresné predpovede?

✓ Zaviedli ste spôsoby na meranie toho, či váš systém vytvára neprijateľné množstvo nepresných predpovedí?

✓ Ak systém vytvára nepresné predpovede, zaviedli ste postupnosť krokov na vyriešenie tohto problému?

Spoľahlivosť a reprodukovateľnosť:

✓ Zaviedli ste stratégiu na monitorovanie a testovanie toho, či systém umelej inteligencie dosahuje ciele, plní svoj účel a zamýšľané použitia?

- Uskutočnili ste testovanie, aby ste zistili, či sa v záujme zabezpečenia reprodukovateľnosti musia zohľadniť nejaké osobitné situácie alebo konkrétne podmienky?
- Zaviedli ste procesy alebo metódy na overovanie s cieľom merať a zabezpečiť rozličné aspekty spoľahlivosti a reprodukovateľnosti?
- Zaviedli ste procesy na opis situácie, keď systém umelej inteligencie v určitých typoch prostredia zlyhá?
- Vypracovali ste zrozumiteľnú dokumentáciu k týmto procesom a zaviedli ste ich do praxe na účely testovania a overovania spoľahlivosti systémov umelej inteligencie?

  Zaviedli ste mechanizmy alebo nadviazali komunikáciu s cieľom ubezpečiť (koncových) používateľov o spoľahlivosti systému umelej inteligencie?

3. <u>Správa súkromia a údajov</u>

Rešpektovanie súkromia a ochrana osobných údajov:

✓ Stanovili ste mechanizmy v závislosti od prípadu použitia, ktoré ostatným umožňujú upozorňovať na

problémy súvisiace s otázkami súkromia alebo ochrany osobných údajov a ktoré sa týkajú procesov systémov umelej inteligencie na zber údajov (na účely výcviku aj prevádzky) a na spracovanie údajov?

- ✓ Posúdili ste typ a rozsah údajov vo vašich dátových súboroch (napr. či obsahujú osobné údaje)?
- ✓ Uvažovali ste o spôsoboch vývoja systému umelej inteligencie alebo výcviku modelu s minimálnym využitím potenciálne citlivých alebo osobných údajov alebo bez nich?
- ✓ Zabudovali ste mechanizmy na upozorňovanie na osobné údaje a na kontrolu nad nimi v závislosti od prípadu použitia (ako je platný súhlas a možnosť jeho odvolania v prípade potreby)?
- ✓ Prijali ste opatrenia na posilnenie súkromia, napríklad prostredníctvom šifrovania, anonymizácie a agregácie?
- ✓ Ak existuje zodpovedná osoba na ochranu údajov, zapojili ste ju do procesu už v počiatočnom štádiu?

Kvalita a integrita údajov:

- ✓ Uviedli ste svoj systém do súladu s potenciálne relevantnými normami (napr. ISO, IEEE) alebo so široko zavedenými protokolmi na účely každodennej správy a riadenia údajov?
- ✓ Stanovili ste mechanizmy dohľadu na účely zberu, uchovávania, spracúvania a používania údajov?
- ✓ Posúdili ste mieru, do akej máte kontrolu nad kvalitou používaných externých zdrojov údajov?
- ✓ Zaviedli ste procesy na zabezpečenie kvality a integrity svojich údajov? Uvažovali ste o iných procesoch? Ako overujete, že vaše súbory údajov neboli ohrozené alebo že sa nestali terčom útoku hekerov?

Prístup k údajom:

- ✓ Aké protokoly, procesy a postupy boli dodržané na účely riadenia a zabezpečenia riadnej správy údajov?
  - Posúdili ste, kto môže mať prístup k údajom používateľov a za akých okolností?
  - Zabezpečili ste, aby boli tieto osoby kvalifikované a mali prístup k údajom a aby mali potrebné spôsobilosti na porozumenie podrobností politiky ochrany údajov?
  - Zabezpečili ste, aby mechanizmus dohľadu zaznamenával to, kedy došlo k prístupu k údajom, kde, ako, kto k nim pristupoval a na aký účel?

4. <u>Transparentnosť</u>

Vysledovateľnosť:

- ✓ Zaviedli ste opatrenia, ktoré môžu zabezpečiť vysledovateľnosť? Tieto opatrenia si môžu vyžadovať zdokumentovanie:
  - ➢ metód použitých na navrhnutie a vývoj algoritmického systému:
    - v prípade systému umelej inteligencie založeného na pravidlách by sa mala zdokumentovať metóda programovania alebo spôsob, akým bol vytvorený model,
    - v prípade systému umelej inteligencie založeného na učení sa by sa mala zdokumentovať

metóda výcviku algoritmu vrátane toho, ktoré vstupné údaje boli zhromaždené a vybraté, a toho, ako došlo k ich získaniu a výberu,

- metód použitých na testovanie a validáciu algoritmického systému:
  - v prípade systému umelej inteligencie založeného na pravidlách by sa mali zdokumentovať scenáre alebo prípady použité na testovanie a validáciu,
  - v prípade systému umelej inteligencie založeného na učení sa by sa mali zdokumentovať informácie o údajoch použitých na testovanie a validáciu,
- výsledkov algoritmického systému:
  - zdokumentovať by sa mali výsledky algoritmu alebo rozhodnutia prijaté na jeho základe, ako aj prípadné ďalšie rozhodnutia, ktoré vyplývajú z iných prípadov (napr. v prípade iných podskupín používateľov).

Vysvetliteľnosť:

- ✓ Posúdili ste mieru, do akej sú rozhodnutia systému umelej inteligencie a v dôsledku nich aj jeho výsledky zrozumiteľné?
- ✓ Presvedčili ste sa o tom, že vysvetlenie, prečo systém uskutočnil istú voľbu vedúcu k určitému výsledku, bude zrozumiteľné pre všetkých používateľov, ktorí môžu požadovať vysvetlenie?
- ✓ Posúdili ste, do akej miery rozhodnutie systému ovplyvňuje rozhodovacie procesy organizácie?
- ✓ Posúdili ste, prečo bol v tejto konkrétnej oblasti zavedený daný systém?
- ✓ Posúdili ste obchodný model súvisiaci s týmto systémom (napr. ako vytvára hodnotu pre organizáciu)?
- ✓ Projektovali ste systém umelej inteligencie už od začiatku so zreteľom na možnosť výkladu?
  - Uskutočnili ste výskum a snažili ste sa pre danú aplikáciu používať najjednoduchší model, ktorý je možné najľahšie vykladať?
  - Posúdili ste, či môžete analyzovať svoje údaje z výcviku a testovania? Môžete tieto údaje neskôr zmeniť a aktualizovať?
  - Posúdili ste, či po výcviku a vývoji modelu máte nejaké voľby na preskúmanie možnosti výkladu alebo či máte prístup k vnútornému postupu práce modelu?

Komunikácia:

- ✓ Oznámili ste (koncovým) používateľom, prostredníctvom vyhlásenia o odmietnutí zodpovednosti alebo inými prostriedkami, že nekomunikujú s iným človekom, ale so systémom umelej inteligencie? Označili ste svoj systém umelej inteligencie ako takýto systém?
- ✓ Zaviedli ste mechanizmy na informovanie používateľov o dôvodoch a kritériách, ktoré tvoria základ výsledkov systému umelej inteligencie?
  - Oznámili ste túto skutočnosť zamýšľaným používateľom jasne a zrozumiteľne?
  - Stanovili ste procesy, v ktorých sa zohľadňuje spätná väzba používateľov, a použili ste ich na pozmenenie systému?
  - Viedli ste aj komunikáciu o možných alebo vnímaných rizikách, ako je zaujatosť?

- V závislosti od prípadu použitia, uvažovali ste o komunikácii s inými adresátmi, tretími stranami alebo širokou verejnosťou a o transparentnosti voči nim?

✓ Vysvetlili ste, aký je účel systému umelej inteligencie a kto alebo čo môže mať z tohto produktu alebo služby prospech?

- Boli v prípade produktu stanovené scenáre použitia a boli zrozumiteľne oznámené, pričom sa zohľadnili aj alternatívne formy komunikácie, aby sa zabezpečilo, že scenáre budú zrozumiteľné a primerané pre daného používateľa?

- V závislosti od prípadu použitia, uvažovali ste o ľudskej psychológii a možných obmedzeniach, ako je riziko zámeny, konfirmačné skreslenie alebo kognitívna únava?

✓ Viedli ste zrozumiteľnú komunikáciu o vlastnostiach, obmedzeniach a možných nedostatkoch systému umelej inteligencie:

- v prípade vývoja: s kýmkoľvek, kto tento systém zavádzal do produktu alebo služby?

- v prípade zavádzania: s koncovým používateľom alebo spotrebiteľom?

5. <u>Rozmanitosť, nediskriminácia a spravodlivosť</u>

Zabránenie nespravodlivej zaujatosti:

✓ Zabezpečili ste stratégiu alebo súbor postupov s cieľom zabrániť vzniku alebo posilneniu nespravodlivej zaujatosti v systéme umelej inteligencie, pokiaľ ide o použitie vstupných údajov, ako aj o návrh algoritmu?

- Posúdili a uznali ste možné obmedzenia vyplývajúce zo zloženia použitých dátových súborov?

- Uvažovali ste o rozmanitosti a reprezentatívnosti používateľov v rámci údajov? Uskutočnili ste testovanie konkrétnych populácií alebo prípadov problémového použitia?

- Uskutočnili ste výskum dostupných technických nástrojov a používali ste ich na zlepšenie svojho porozumenia údajom, modelu a výkonnosti?

- Zaviedli ste procesy na testovanie a monitorovanie možných prípadov zaujatosti počas fázy vývoja, zavádzania a používania systému?

✓ V závislosti od prípadu použitia, zabezpečili ste mechanizmus, ktorý umožňuje ostatným upozorňovať na problémy súvisiace so zaujatosťou, diskrimináciou alebo slabou výkonnosťou systému umelej inteligencie?

- Uvažovali ste o jasných krokoch a spôsoboch komunikácie týkajúcej sa toho, ako na tieto problémy upozorniť a komu sa majú predložiť?

- Zohľadnili ste pri tom nie len (koncových) používateľov, ale aj ostatných potenciálne nepriamo postihnutých systémom umelej inteligencie?

✓ Posúdili ste, či existuje akákoľvek možná variabilita rozhodnutí, ku ktorej môže dochádzať za tých istých podmienok?

- Ak áno, uvažovali ste o tom, čo by mohlo byť jej pravdepodobnou príčinou?
- V prípade variability, stanovili ste mechanizmus na meranie alebo posúdenie možného vplyvu tejto variability na základné práva?

✓ Zabezpečili ste primerané pracovné vymedzenie pojmu „spravodlivosť“, ktoré uplatňujete pri navrhovaní systémov umelej inteligencie?

- Používa sa vaše vymedzenie bežne? Uvažovali ste o iných vymedzeniach pred tým, než ste vybrali toto?
- Zabezpečili ste kvantitatívnu analýzu alebo metriku na meranie a testovanie použitého vymedzenia spravodlivosti?
- Stanovili ste mechanizmy na zabezpečenie spravodlivosti vo svojich systémoch umelej inteligencie? Uvažovali ste aj o iných možných mechanizmoch?

Prístupnosť a dizajn pre všetkých:

✓ Zabezpečili ste, aby systém umelej inteligencie vyhovoval širokému okruhu individuálnych preferencií a schopností?

- Posúdili ste, či systém umelej inteligencie môžu používať osoby so špeciálnymi potrebami, so zdravotným postihnutím alebo osoby, ktorým hrozí riziko vylúčenia? Ako bola táto skutočnosť začlenená do návrhu systému a ako sa overuje?
- Zabezpečili ste, aby informácie o systéme umelej inteligencie boli prístupné aj pre používateľov podporných technológií?
- Zapojili ste túto komunitu do vývoja systému umelej inteligencie alebo radili ste sa s ňou?

✓ Zohľadnili ste vplyv svojho systému umelej inteligencie na skupinu potenciálnych používateľov?

- Je tím zapojený do vytvárania systému umelej inteligencie reprezentatívny, pokiaľ ide o cieľovú skupinu používateľov? Je reprezentatívny z hľadiska širšej populácie a aj vzhľadom na iné skupiny, ktoré by systémom mohli byť dotknuté okrajovo?
- Posúdili ste, či by mohli existovať osoby alebo skupiny, ktoré by mohli byť neprimerane postihnuté negatívnymi následkami?
- Poskytli vám spätnú väzbu iné tímy alebo skupiny, ktoré majú odlišné zázemie a skúsenosti?

Účasť zainteresovaných strán:

✓ Uvažovali ste o mechanizme, ako zahrnúť účasť rôznych zainteresovaných strán do vývoja a používania systému umelej inteligencie?

✓ Pripravili ste cestu pre zavedenie systému umelej inteligencie vo vašej organizácii tým, že ste vopred informovali a zapojili dotknutých pracovníkov a ich zástupcov?

6. <u>Spoločenský a environmentálny blahobyt</u>

Udržateľná umelá inteligencia, ktorá je šetrná k životnému prostrediu:

- ✓ Zaviedli ste mechanizmy na meranie vplyvu vývoja, zavádzania a používania systému umelej inteligencie na životné prostredie (napr. energia, ktorú spotrebúva dátové stredisko, typ energie, ktorú používajú dátové strediská, atď.)?

- ✓ Zabezpečili ste opatrenia na zníženie vplyvu životného cyklu vášho systému umelej inteligencie na životné prostredie?

Sociálny vplyv:

- ✓ Ak systém umelej inteligencie priamo komunikuje s ľuďmi:

  - Posúdili ste, či systém umelej inteligencie motivuje ľudí, aby si vytvorili vzťah a sympatie k systému?

  - Zabezpečili ste, aby systém umelej inteligencie jasne naznačoval, že jeho sociálna interakcia je simulovaná a že nemá schopnosti „porozumenia“ a „cítenia“?

- ✓ Zabezpečili ste, že sociálnemu vplyvu systému umelej inteligencie sa náležite rozumie? Posúdili ste, či napríklad existuje riziko straty pracovných miest alebo zručností pracovníkov? Aké kroky ste prijali, aby ste zabránili týmto rizikám?

Spoločnosť a demokracia:

- ✓ Posúdili ste širší spoločenský vplyv používania systému umelej inteligencie nad rámec individuálneho (koncového) používateľa, napríklad jeho vplyv na možné nepriamo dotknuté zainteresované strany?

7. <u>Zodpovednosť</u>

Kontrolovateľnosť:

- ✓ Zaviedli ste mechanizmy, ktoré uľahčujú kontrolovateľnosť systému internými a/alebo nezávislými subjektmi, ako je zabezpečenie vysledovateľnosti a vedenie záznamov o procesoch a výsledkoch systému umelej inteligencie?

Minimalizácia negatívneho vplyvu a jeho oznamovanie:

- ✓ Vykonali ste posúdenie rizika alebo vplyvu systému umelej inteligencie, v ktorých sa zohľadňujú rôzne zainteresované strany dotknuté systémom priamo a nepriamo?

- ✓ Zaviedli ste rámce na vzdelávanie a odbornú prípravu s cieľom vypracovať postupy na vyvodenie zodpovednosti?

  - Ktorých pracovníkov alebo zložiek tímu sa to týka? Uplatňujú sa tieto rámce aj po skončení fázy vývoja?

  - Je súčasťou tejto odbornej prípravy aj výučba možného právneho rámca platného pre systém umelej inteligencie?

  - Uvažovali ste o vytvorení „rady pre preskúmanie etickej umelej inteligencie“ alebo podobného mechanizmu na diskusiu o celkovej zodpovednosti a etických postupoch vrátane potenciálne nejasných sivých zón?

- ✓ Okrem interných iniciatív alebo rámcov na dohľad nad oblasťou etiky a zodpovednosti, existuje nejaký druh externej pomoci alebo boli vytvorené aj audítorské postupy?

- ✓ Sú zavedené nejaké postupy pre tretie strany (napr. dodávateľov, spotrebiteľov, distribútorov/predajcov) alebo pracovníkov, aby mohli nahlasovať prípadné zraniteľnosti, riziká alebo prípady zaujatosti v systéme/aplikácii umelej inteligencie?

Dokumentácia kompromisov:

- ✓ Stanovili ste mechanizmus na identifikovanie príslušných záujmov a hodnôt dotknutých systémom umelej inteligencie a možných kompromisov medzi nimi?
- ✓ Aký postup používate na rozhodovanie o týchto kompromisoch? Postarali ste sa o zdokumentovanie tohto rozhodnutia o kompromisoch?

Schopnosť nápravy:

- ✓ Stanovili ste primeraný súbor mechanizmov, ktoré umožňujú nápravu v prípade výskytu nejakej ujmy alebo nepriaznivého vplyvu?
- ✓ Zaviedli ste mechanizmy na poskytovanie informácií (koncovým) používateľom/tretím stranám o možnostiach nápravy?

Vyzývame všetky zainteresované strany, aby vyskúšali tento zoznam bodov na posudzovanie v praxi a aby poskytli spätnú väzbu o jeho realizovateľnosti, úplnosti, relevantnosti pre konkrétnu aplikáciu alebo oblasť umelej inteligencie, ako aj o prekrývaní či doplnkovosti s existujúcimi postupmi zabezpečenia súladu alebo posudzovania. Na základe tejto spätnej väzby sa začiatkom roku 2020 Komisii predloží návrh revidovanej verzie zoznamu bodov na posudzovanie dôveryhodnej umelej inteligencie.

Kľúčové usmernenia vyplývajúce z kapitoly III:

- ✓ Prijať **zoznam bodov na posudzovanie** dôveryhodnej umelej inteligencie, ktorý sa využije pri vývoji umelej inteligencie, jej zavádzaní alebo používaní, a prispôsobiť ho konkrétnemu prípadu použitia, v ktorom sa systém používa.
- ✓ Pamätať si, že tento zoznam bodov na posudzovanie nikdy **nebude úplný**. Pri zabezpečení dôveryhodnej umelej inteligencie nejde o odškrtávanie políčok, ale o nepretržité identifikovanie požiadaviek, vyhodnocovanie riešení a zabezpečovanie lepších výsledkov počas celého životného cyklu systému umelej inteligencie a o zapojenie zainteresovaných strán do týchto činností.

C) PRÍKLADY PRÍLEŽITOSTÍ A VÁŽNYCH OBÁV VYVOLANÝCH UMELOU INTELIGENCIOU

121. V tomto oddiele uvádzame príklady vývoja a používania umelej inteligencie, ktoré by sa mali podporovať, ako aj príklady, keď vývoj, zavádzanie a používanie umelej inteligencie môžu byť v rozpore s našimi hodnotami a môžu vzbudzovať osobitné obavy. Medzi tým, čo by sa malo a čo sa môže dosiahnuť s pomocou umelej inteligencie, musí byť rovnováha a náležitá pozornosť sa musí venovať tomu, čo by sa s pomocou umelej inteligencie robiť nemalo.

1. Príklady príležitostí dôveryhodnej umelej inteligencie

122. Dôveryhodná umelá inteligencia môže predstavovať veľkú príležitosť na podporu zmierňovania naliehavých výziev, ktorým čelí spoločnosť, ako je starnutie obyvateľstva, rastúca sociálna nerovnosť a znečisťovanie životného prostredia. Tento potenciál je takisto vnímaný na celosvetovej úrovni, napríklad sa odráža v cieľoch

OSN v oblasti udržateľného rozvoja[57]. Tento oddiel sa zameriava na otázku, ako podporiť európsku stratégiu v oblasti umelej inteligencie, ktorá rieši niektoré z týchto výziev.

a) Opatrenia v oblasti klímy a udržateľná infraštruktúra

123. Hoci by boj proti zmene klímy mal byť najvyššou prioritou tvorcov politík z celého sveta, digitálna transformácia a dôveryhodná umelá inteligencia majú veľký potenciál znížiť vplyv ľudstva na životné prostredie a umožniť účinné a efektívne využívanie energie a prírodných zdrojov[58]. Dôveryhodná umelá inteligencia sa môže napríklad spojiť s technológiou veľkých dát (big data) s cieľom presnejšie odhaľovať energetické potreby, čo povedie k efektívnejšej energetickej infraštruktúre a nižšej spotrebe[59].

124. V súvislosti s odvetviami, ako je verejná doprava, sa systémy umelej inteligencie pre inteligentné dopravné systémy[60] môžu použiť na minimalizovanie tvorby kolón, optimalizáciu plánovania trás, umožnenie väčšej nezávislosti pre osoby so zrakovým postihnutím[61], optimalizáciu energeticky účinných motorov, a teda posilnenie úsilia o dekarbonizáciu, a zníženie environmentálnej stopy v záujme ekologickejšej spoločnosti. V súčasnosti na celom svete každých 23 sekúnd zomrie na následky automobilovej nehody jeden človek[62]. Systémy umelej inteligencie by mohli pomôcť podstatne znížiť počet úmrtí, napríklad prostredníctvom zlepšenia reakčného času a lepšieho dodržiavania pravidiel[63].

b) Zdravie a pohoda

125. Technológie dôveryhodnej umelej inteligencie sa môžu použiť (a už sa aj používajú) na zabezpečenie inteligentnejšej a cielenejšej liečby a môžu pomáhať pri predchádzaní život ohrozujúcim chorobám[64]. Lekári a zdravotnícki pracovníci môžu teoreticky vykonávať presnejšiu a podrobnejšiu analýzu komplexných údajov o zdravotnom stave pacienta, a to dokonca ešte pred tým, než ochorie, vďaka čomu môžu poskytovať na mieru prispôsobenú preventívnu liečbu[65]. Umelá inteligencia a robotika môžu byť v kontexte starnúcej populácie Európy cennými nástrojmi, ktoré pomáhajú opatrovateľom, podporujú starostlivosť o staršie osoby[66] a sledujú

57 https://sustainabledevelopment.un.org/?menu=1300.

58 Niekoľko projektov EÚ je zameraných na vývoj inteligentných sietí a na uskladňovanie energie, ktoré majú potenciál prispieť k úspešnému digitálnemu prechodu v oblasti energetiky vrátane prostredníctvom riešení založených na umelej inteligencii a ďalších digitálnych riešení. Na doplnenie úsilia týchto jednotlivých projektov, Komisia spustila iniciatívu BRIDGE, ktorá prebiehajúcim projektom pre inteligentnú sieť a uskladňovanie energie v rámci programu Horizont 2020 umožňuje vytvoriť spoločný pohľad na prierezové otázky: https://www.h2020-bridge.eu/.

59 Pozri napríklad projekt Encompass: http://www.encompass-project.eu/.

60 Nové riešenia založené na umelej inteligencii pomáhajú pripravovať mestá na budúcnosť mobility. Pozri napríklad projekt financovaný z prostriedkov EÚ s názvom Fabulos: https://fabulos.eu/.

61 Pozri napríklad projekt PRO4VIP, ktorý je súčasťou európskej stratégie Vision 2020 na boj proti slepote, ktorej vzniku možno predchádzať, najmä slepote v dôsledku staroby. Mobilita a orientácia patrili medzi prioritné oblasti projektu.

62 https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries.

63 Európsky projekt UP-Drive je napríklad zameraný na riešenie opísaných dopravných problémov poskytnutím príspevkov umožňujúcich postupnú automatizáciu a spoluprácu medzi vozidlami, čo umožňuje bezpečnejší, inkluzívnejší a cenovo dostupnejší dopravný systém: https://up-drive.eu/.

64 Pozri napríklad projekt REVOLVER (Repeated Evolution of Cancer – Opakovaný vznik rakoviny): https://www.healtheuropa.eu/personalised-cancer-treatment/87958/, alebo projekt Murab, ktorý umožňuje vykonávať presnejšie biopsie a ktorý je zameraný na rýchlejšie diagnostikovanie rakoviny a ďalších ochorení: https://ec.europa.eu/digital-single-market/en/news/murab-eu-funded-project-success-story.

65 Pozri napríklad projekt Live INCITE: www.karolinska.se/en/live-incite. Toto konzorcium obstarávateľov zdravotnej starostlivosti vyzýva odvetvie, aby vyvinulo inteligentné riešenia umelej inteligencie a ďalšie riešenia v oblasti IKT, ktoré umožnia opatrenia v oblasti životosprávy v peroperatívnom procese. Jeho cieľ sa týka nových inovatívnych riešení v oblasti elektronických zdravotníckych služieb, ktoré môžu osobne ovplyvniť pacientov, aby prijali nevyhnutné opatrenia týkajúce sa ich životosprávy pred operáciou aj po nej s cieľom optimalizovať výsledky zdravotnej starostlivosti.

66 Projekt CARESSES financovaný z EÚ sa zaoberá robotmi určenými pre starostlivosť o staršie osoby a je zameraný na kultúrnu citlivosť robotov: roboti prispôsobujú svoje konanie a verbálny prejav tak, aby zodpovedali kultúre a zvykom staršej osoby, ktorej pomáhajú: http://caressesrobot.org/en/project/. Pozri aj aplikáciu umelej inteligencie s názvom Alfred, čo je

stav pacientov v reálnom čase, čím zachraňujú životy[67].

126. Dôveryhodná umelá inteligencia môže pomáhať aj v širšej miere. Napríklad môže skúmať a určovať všeobecné trendy v odvetví zdravotnej starostlivosti a liečby[68], čím prispieva k skoršiemu odhaleniu ochorení, efektívnejšiemu vývoju liekov, cielenejším formám liečby[69] a v konečnom dôsledku k väčšiemu počtu zachránených životov.

c) Kvalitné vzdelávanie a digitálna transformácia

127. Nové technologické, hospodárske a environmentálne zmeny znamenajú, že spoločnosť sa musí stať iniciatívnejšou. Vlády, popredné subjekty odvetvia, vzdelávacie inštitúcie a odbory stoja pred povinnosťou preniesť občanov do nového digitálneho veku tým, že zabezpečia, aby mali správne zručnosti na obsadenie pracovných miest budúcnosti. Technológie dôveryhodnej umelej inteligencie by mohli pomôcť poskytovať presnejšiu prognózu toho, ktoré zamestnania a povolania sa narušia v dôsledku technológie, ktoré nové úlohy vzniknú a ktoré zručnosti budú potrebné. Tým by mohli pomôcť vládam, odborom a priemyslu s plánovaním (re)kvalifikácie pracovníkov. Mohli by tiež ukázať občanom, ktorí môžu mať strach z prepúšťania, akou cestou sa rozvíjať pre novú úlohu.

128. Umelá inteligencia okrem toho môže byť významným nástrojom na boj proti nerovnosti vo vzdelávaní a pomôcť vytvoriť individualizované a adaptabilné vzdelávacie programy, ktoré by všetkým mohli pomôcť so získaním nových kvalifikácií, zručností a spôsobilostí v súlade s ich vlastnou schopnosťou učiť sa[70]. Vďaka nej by sa mohla zvýšiť rýchlosť učenia aj kvalita vzdelávania, a to od základnej školy až po univerzitu.

2. Príklady vážnych obáv, ktoré vzbudzuje umelá inteligencia

129. Vážne obavy z umelej inteligencie vzbudzuje porušenie niektorej zo zložiek dôveryhodnej umelej inteligencie. Mnohé z obáv uvedených ďalej už patria do rozsahu pôsobnosti existujúcich právnych požiadaviek, ktoré sú záväzné a musia sa teda dodržať. Jednako však aj v situácii, keď sa preukázal súlad s právnymi požiadavkami, nemusí tento súlad predstavovať riešenie celého spektra etických obáv, ktoré mohli vzniknúť. Keďže sa naše porozumenie primeranosti pravidiel a etických zásad neustále vyvíja a môže sa časom meniť, tento neúplný zoznam obáv sa v budúcnosti môže zúžiť, rozšíriť, upraviť alebo aktualizovať.

a) Identifikácia a sledovanie osôb pomocou umelej inteligencie

130. Umelá inteligencia umožňuje čoraz účinnejšiu identifikáciu osôb zo strany verejných aj súkromných subjektov. Pozoruhodnými príkladmi škálovateľnej technológie umelej inteligencie na identifikáciu je systém na

---

virtuálny asistent, ktorý starším osobám pomáha ostať aktívnymi: https://ec.europa.eu/digital-single-market/en/news/alfred-virtual-assistant-helping-older-people-stay-active. Navyše v rámci projektu EMPATTICS (EMpowering PAtients for a BeTTer Information and improvement of the Communication Systems – Posilnenie postavenia pacientov pre lepšiu informovanosť a zlepšenie komunikačných systémov) sa uskutoční výskum a vymedzí spôsob, akým zdravotnícki pracovníci a pacienti využívajú technológie IKT vrátane systémov umelej inteligencie na plánovanie zásahov s pacientmi a na monitorovanie pokroku ich telesného a duševného stavu: www.empattics.eu.

[67] Pozri napríklad projekt MyHealth Avatar (www.myhealthavatar.eu), ktorý ponúka zobrazenie zdravotného stavu pacienta v digitálnej podobe. Výskumný projekt spustil aplikáciu a online platformu, ktoré zbierajú digitálne dlhodobé informácie o vašom zdravotnom stave a poskytujú k nim prístup. Majú podobu celoživotného zdravotného spoločníka (tzv. avatara). Projekt MyHealthAvatar takisto predpovedá, aká je vaše riziko mozgovej príhody, cukrovky, srdcovocievnej choroby a hypertenzie.

[68] Pozri napríklad projekt ENRICHME (www.enrichme.eu), ktorý bojuje proti postupnému znižovaniu kognitívnych schopností v starnúcej populácii. Integrovaná platforma pre aktívny a asistovaný život (AAL) a pohyblivý služobný robot pre dlhodobé monitorovanie a interakciu budú pomáhať starším osobám, aby si uchovali dlhšie nezávislosť a ostali dlhšie aktívni.

[69] Pozri napríklad použitie umelej inteligencie spoločnosťou Sophia Genetics, ktorá využíva štatistickú inferenciu, rozpoznávanie vzorcov a strojové učenie na maximalizáciu hodnoty genomických a rádiomických údajov: https://www.sophiagenetics.com/home.html.

[70] Pozri napríklad projekt MaTHiSiS zameraný na poskytovanie riešenia pre afektívne učenie v príjemnom vzdelávacom prostredí tvorenom exkluzívnymi technologickými zariadeniami a algoritmami: (http://mathisis-project.eu/). Pozri aj projekt Watson Classroom spoločnosti IBM alebo platformu spoločnosti Century Tech.

rozpoznávanie tváre a ďalšie nedobrovoľné metódy identifikácie využívajúce biometrické údaje (t. j. detektory lži, posudzovanie osobnosti na základe mikrovýrazov a automatická detekcia hlasu). Identifikácia jednotlivcov niekedy môže byť žiaduci výsledok, ktorý je v súlade s etickými zásadami (napríklad pri odhaľovaní podvodov, prania špinavých peňazí alebo financovania terorizmu). Automatická identifikácia však vzbudzuje vážne obavy právnej aj etickej povahy, keďže môže mať neočakávaný dosah na mnohých psychologických a sociálno-kultúrnych úrovniach. Na dodržanie autonómie občanov Únie je potrebné primerané používanie kontrolných postupov v rámci umelej inteligencie. Jasné vymedzenie toho, či sa umelá inteligencia môže použiť na automatickú identifikáciu osôb, kedy sa môže použiť a ako, a rozlišovanie medzi identifikáciou osôb a ich vyhľadávaním a sledovaním a medzi cieleným a masovým sledovaním budú kľúčové na vytvorenie dôveryhodnej umelej inteligencie. Uplatňovanie týchto technológií musí byť jasne zaručené v existujúcom práve[71]. Ak právnym základom takejto činnosti je „súhlas“, musia sa vytvoriť praktické prostriedky[72], ktoré budú umožňovať vydanie účelného a overeného súhlasu v súvislosti s osobou, ktorú automaticky identifikoval systém umelej inteligencie alebo rovnocenné technológie. To sa týka aj použitia „anonymných“ osobných údajov, ktoré sa môžu odanonymizovať.

b) Tajné systémy umelej inteligencie

131. Ľudia by vždy mali vedieť, či komunikujú priamo s iným človekom alebo so strojom, a spoľahlivé dosiahnutie tohto cieľa je povinnosťou špecialistov na umelú inteligenciu. Špecialisti na umelú inteligenciu by preto mali zabezpečiť, aby ľudia boli na to, že komunikujú so systémom umelej inteligencie, upozornení (napríklad prostredníctvom vydania zrozumiteľných a transparentných vyhlásení o odmietnutí zodpovednosti) alebo aby sa mohli o tejto skutočnosti informovať a potvrdiť súhlas s touto komunikáciou. Treba upozorniť, že existujú sporné prípady, ktoré celú vec komplikujú (napr. hlas nahovorený človekom filtrovaný umelou inteligenciou). Treba mať na pamäti, že zamieňanie ľudí a strojov by mohlo mať mnoho následkov, ako je vytvorenie si vzťahu, ovplyvňovanie alebo zníženie ľudskej hodnoty[73]. Vývoj ľuďom podobných robotov[74] by preto mal byť predmetom dôkladného etického posudzovania.

c) Hodnotenie občanov s podporou umelej inteligencie v rozpore so základnými právami

132. Spoločnosti by sa mali usilovať o ochranu slobody a autonómie všetkých občanov. Akákoľvek forma hodnotenia občanov môže viesť k strate tejto autonómie a môže ohrozovať zásadu nediskriminácie. Hodnotenie by sa malo používať iba vtedy, ak pre to existuje jasný dôvod a ak sú dané opatrenia primerané a spravodlivé. Normatívne hodnotenie občanov (všeobecné posudzovanie „morálnej osobnosti“ alebo „etickej integrity“) vo *všetkých* aspektoch a vo veľkom rozsahu zo strany verejných orgánov alebo súkromných subjektov tieto hodnoty ohrozuje, najmä ak sa používajú v rozpore so základnými právami a ak sa používajú neprimerane a bez vytýčeného a oznámeného legitímneho účelu.

133. Hodnotenie občanov (vo väčšej či menšej miere) sa už v súčasnosti často používa v čisto opisných hodnoteniach špecifických pre konkrétnu oblasť (napr. školské systémy, elektronické učenie sa a vodičské preukazy). Aj v týchto užších použitiach by sa občanom mali sprístupniť celkom transparentné postupy, vrátane informácií o procese, účele a metodike hodnotenia. Treba upozorniť, že transparentnosť nemôže zabrániť nediskriminácii ani zabezpečiť spravodlivosť a nie je všeliekom na problém hodnotenia. *V* ideálnom prípade, ak je to možné, by občania mali mať možnosť rozhodnúť sa bez trestu pre neúčasť v systéme hodnotenia. V opačnom prípade sa musia poskytnúť mechanizmy námietok a nápravy hodnotenia. To je

---

71 V tejto súvislosti možno pripomenúť článok 6 všeobecného nariadenia o ochrane údajov, v ktorom sa okrem iného stanovuje, že spracovanie údajov je zákonné iba vtedy, keď má platný právny základ.

72 Ako sa ukazuje v prípade aktuálnych mechanizmov na udeľovanie informovaného súhlasu na internete, spotrebitelia súhlas zvyčajne udeľujú bez zmysluplného zvažovania. Z tohto hľadiska sa len ťažko môžu označiť za praktické.

73 Madary a Metzinger (2016), *Real Virtuality: A Code of Ethical Conduct.* Recommendations for Good Scientific Practice and the Consumers of VR-Technology. (Reálna virtualita: kódex etického správania. Odporúčania pre osvedčenú vedeckú prax a spotrebiteľov technológie virtuálnej reality) *Frontiers in Robotics and AI*, roč. 3, č. 3.

74 To sa týka aj avatarov riadených umelou inteligenciou.

obzvlášť dôležité v situáciách, keď medzi stranami existuje asymetria moci. Takéto možnosti vystúpenia by mali byť zabezpečené v návrhu technológie v prípadoch, keď to je potrebné v záujme dodržania súladu so základnými právami a keď to je v demokratickej spoločnosti nevyhnutné.

d) Smrtiace autonómne zbraňové systémy (LAWS)

134. V súčasnosti sa neznámy počet krajín a odvetví zaoberá výskumom a vývojom smrtiacich autonómnych zbraňových systémov, a to od riadených striel so schopnosťou selektívneho výberu cieľov po učiace sa stroje s kognitívnymi zručnosťami, ktoré im umožňujú rozhodovať s kým, kedy a kde bojovať, a to bez intervencie človeka. To vzbudzuje základné etické obavy, ako je skutočnosť, že by existencia týchto systémov mohla viesť k nekontrolovaným pretekom v zbrojení na historicky bezprecedentnej úrovni a mohli by vznikať vojenské kontexty, v ktorých bude dochádzať k takmer úplnému zrieknutiu sa ľudskej kontroly a pri ktorých sa neriešia riziká zlyhania. Európsky parlament vyzval na naliehavé vypracovanie spoločného, právne záväzného stanoviska venovaného etickým a právnym otázkam ľudskej kontroly, dohľadu, zodpovednosti a vykonávania medzinárodného práva v oblasti ľudských práv, medzinárodného humanitárneho práva a vojenských stratégií[75]. S odvolaním sa na cieľ Európskej únie presadzovať mier zakotvený v článku 3 Zmluvy o Európskej únii stojíme za uznesením Parlamentu z 12. septembra 2018 a všetkými súvisiacimi snahami v oblasti smrtiacich autonómnych zbraňových systémov a vyjadrujeme im podporu.

e) Možné dlhodobejšie obavy

135. Vývoj umelej inteligencie sa stále týka konkrétnych oblastí a vyžaduje, aby dobre pripravení ľudskí vedci a inžinieri presne stanovovali jej ciele. Pri odhade trendov do budúcnosti z dlhodobého hľadiska však možno predpokladať určité závažné dlhodobé obavy[76]. Z prístupu založeného na riziku vyplýva, že na tieto obavy by sa mal brať stály zreteľ, pokiaľ ide o možné nepoznané neznáme skutočnosti a tzv. čierne labute[77]. Mimoriadny dosah týchto obáv v spojení s aktuálnou neistotou v rámci príslušných trendov si vyžadujú pravidelné posudzovanie týchto tém.

D) ZÁVER

136. Tento dokument predstavuje etické usmernenia pre umelú inteligenciu, ktoré vypracovala expertná skupina na vysokej úrovni pre umelú inteligenciu (AI HLEG).

137. Uznávame pozitívny vplyv, ktorý už systémy umelej inteligencie majú a budú aj naďalej mať z komerčného aj spoločenského hľadiska. Rovnako sa však usilujeme zabezpečiť, aby sa riziká a ďalšie nepriaznivé účinky, s ktorými sa tieto technológie spájajú, riešili náležite a primerane vzhľadom na aplikáciu umelej inteligencie. Umelá inteligencia je technológia, ktorá je zároveň transformatívna a rušivá, a jej vývoj v posledných niekoľkých rokoch uľahčila dostupnosť obrovského množstva digitálnych údajov, veľkého technologického pokroku v oblasti výpočtovej a úložiskovej kapacity, ako aj značných vedeckých a technických inovácií, pokiaľ ide o metódy a nástroje umelej inteligencie. Systémy umelej inteligencie budú naďalej ovplyvňovať spoločnosť a občanov spôsobmi, ktoré si zatiaľ nevieme predstaviť.

138. V tejto súvislosti je dôležité budovať systémy umelej inteligencie, ktoré sú hodné dôvery, keďže ľudia budú schopní s istotou a plne využívať ich prínos len vtedy, keď bude táto technológia vrátane procesov a ľudí, ktorí ju vytvárali, dôveryhodná. Pri vypracúvaní týchto usmernení preto naším hlavným cieľom bola dôveryhodná umelá inteligencia.

---

75 Uznesenie Európskeho parlamentu 2018/2752(RSP).

76 Hoci sa niektorí ľudia domnievajú, že všeobecná umelá inteligencia, umelé vedomie, umelé morálne agensy, superinteligencia alebo transformatívna umelá inteligencia môžu byť príkladmi takýchto dlhodobých obáv (ktoré v súčasnosti neexistujú), mnohí ďalší sú presvedčení, že sú nerealistické.

77 Pojmom čierna labuť označujeme veľmi zriedkavú udalosť, ale s mimoriadnym dosahom. Ide o taký vzácny jav, že nemusel byť pozorovaný. Pravdepodobnosť jeho výskytu sa teda zvyčajne môže odhadnúť iba s veľkou neistotou.

139. Dôveryhodná umelá inteligencia má tri zložky: 1. mala by byť zákonná a zabezpečovať dodržiavanie všetkých platných zákonov a právnych predpisov; 2. mala by byť etická a zabezpečovať súlad s etickými zásadami a hodnotami a 3. mala by byť odolná, a to z technického aj sociálneho hľadiska, keďže jej cieľom je zabezpečovať, aby systémy umelej inteligencie aj pri dobrých úmysloch nespôsobili neúmyselnú ujmu. Každá zložka je potrebná, ale nestačí na dosiahnutie dôveryhodnej umelej inteligencie. V ideálnom prípade pôsobia všetky tri zložky vo vzájomnom súlade a prekrývajú sa vo svojom fungovaní. Ak vzniknú rozpory, mali by sme sa snažiť o ich zosúladenie.

140. V kapitole I sme formulovali základné práva a zodpovedajúci súbor etických zásad, ktoré sú v kontexte umelej inteligencie rozhodujúce. V kapitole II sme uviedli sedem kľúčových požiadaviek, ktoré by systémy umelej inteligencie mali spĺňať s cieľom realizovať dôveryhodnú umelú inteligenciu. Navrhli sme technické a netechnické metódy, ktoré môžu pomôcť pri realizácii týchto požiadaviek. A napokon v kapitole III sme uviedli zoznam bodov na posudzovanie dôveryhodnej umelej inteligencie, ktorý môže prispieť pri zavádzaní týchto siedmich požiadaviek do praxe. V poslednom oddiele sme poskytli príklady prospešných príležitostí a vážnych obáv, ktoré vzbudzujú systémy umelej inteligencie. Dúfame, že vyvolajú ďalšiu diskusiu.

141. Európa má jedinečné výhodné postavenie založené na jej sústredenom úsilí stavať občanov do centra svojho snaženia. Táto pozornosť je zapísaná v samotnej DNA Európskej únie prostredníctvom zmlúv, na ktorých je postavená. Tento dokument tvorí súčasť vízie, ktorá podporuje dôveryhodnú umelú inteligenciu, o ktorej sme presvedčení, že by mala byť základom, na ktorom Európa môže vybudovať svoje vedúce postavenie v oblasti inovatívnych a špičkových systémov umelej inteligencie. Táto ambiciózna víziapomôže zabezpečiť individuálnu aj spoločnú ľudskú prosperitu občanov Únie. Naším cieľom je vytvoriť kultúru „dôveryhodnej umelej inteligencie pre Európu“, prostredníctvom ktorej môžu všetci využívať výhody umelej inteligencie spôsobom, ktorým sa zabezpečí úcta k našim základným hodnotám: k základným právam, demokracii a právnemu štátu.

GLOSÁR

142. Tento glosár sa týka usmernení a jeho účelom je pomôcť porozumieť pojmom použitým v tomto dokumente.

Umelá inteligencia alebo systémy umelej inteligencie

143. Systémy umelej inteligencie sú softvérové (a prípadne aj hardvérové) systémy navrhnuté ľuďmi[78], ktoré vzhľadom na komplexný cieľ konajú vo fyzickom alebo digitálnom rozmere tak, že vnímajú svoje prostredie prostredníctvom získavania údajov, interpretácie zhromaždených štruktúrovaných alebo neštruktúrovaných údajov, odvodzovania poznatkov alebo spracúvania informácií odvodených z týchto údajov a že rozhodujú o najlepších krokoch, ktoré sa majú vykonať na dosiahnutie daného cieľa. Systémy umelej inteligencie môžu buď používať symbolické pravidlá, alebo sa naučiť numerický model, a takisto môžu upraviť svoje správanie na základe analýzy vplyvu, aký malo ich predchádzajúce konanie na ich prostredie.

144. Umelá inteligencia ako vedecká disciplína obsahuje niekoľko prístupov a techník, ako je strojové učenie (ktorého konkrétnymi príkladmi sú hĺbkové učenie a učenie posilňovaním), strojové odvodzovanie (ktorého súčasťou je plánovanie, programovanie, reprezentácia a odvodzovanie poznatkov, vyhľadávanie a optimalizácia) a robotika (do ktorej patrí kontrola, percepcia, senzory a ovládače, ako aj začlenenie všetkých ostatných techník do kyberneticko-fyzických systémov).

145. Súbežne s týmto dokumentom sa zverejňuje samostatný dokument, ktorý pripravila expertná skupina na vysokej úrovni pre umelú inteligenciu, v ktorom rozpracovala vymedzenie pojmu *systémy umelej inteligencie* používaného na účely tohto dokumentu a ktorý sa volá „Vymedzenie pojmu umelej inteligencie: hlavné schopnosti a vedecké disciplíny“.

Špecialisti na umelú inteligenciu

146. Pojmom špecialisti na umelú inteligenciu označujeme všetky osoby alebo organizácie, ktoré vyvíjajú (vrátane výskumu, návrhu alebo poskytovania údajov), zavádzajú (vrátane vykonávania) alebo používajú systémy umelej inteligencie, s výnimkou osôb a organizácií, ktoré systémy umelej inteligencie používajú vo funkcii koncového používateľa alebo spotrebiteľa.

Životný cyklus systému umelej inteligencie

147. Životný cyklus systému umelej inteligencie tvorí jeho fáza vývoja (vrátane výskumu, návrhu, poskytovania údajov a obmedzeného skúšania), zavádzania (vrátane vykonávania) a používania.

Kontrolovateľnosť

148. Kontrolovateľnosť označuje možnosť, aby systém umelej inteligencie bol predmetom posudzovania jeho algoritmov, údajov a procesov navrhovania. Ide o jednu zo siedmich požiadaviek, ktoré by systém dôveryhodnej umelej inteligencie mal spĺňať. To však nevyhnutne neznamená, že informácie o obchodných modeloch a duševnom vlastníctve spojené so systémom umelej inteligencie musia byť vždy verejne dostupné. Zabezpečenie mechanizmov vysledovateľnosti a vedenia záznamov od včasnej fázy návrhu systému umelej inteligencie môže pomôcť kontrolovateľnosti systému.

Zaujatosť

149. Zaujatosť je sklon k predsudkom v prospech alebo neprospech nejakej osoby, predmetu alebo postavenia. Existuje veľa spôsobov, akými zaujatosť môže v systémoch umelej inteligencie vznikať. Napríklad v dátových systémoch umelej inteligencie, ako sú systémy vytvorené prostredníctvom strojového učenia, môže zaujatosť pri zbere údajov a výcviku systému mať za následok systém umelej inteligencie, v ktorom sa zaujatosť prejavuje. V prípade logickej umelej inteligencie, ako sú systémy založené na pravidlách, môže zaujatosť

[78] Ľudia navrhujú systémy umelej inteligencie priamo, môžu však používať aj techniky umelej inteligencie na optimalizáciu návrhu týchto systémov.

vzniknúť v dôsledku toho, ako môže znalostný inžinier chápať pravidlá uplatňované v konkrétnom prostredí. Zaujatosť môže vzniknúť aj v dôsledku online učenia sa a úpravy prostredníctvom interakcie. Vzniknúť môže aj prostredníctvom personalizácie, pri ktorej sa používateľom predložia odporúčania alebo informačné kanály, ktoré sú prispôsobené vkusu používateľa. Nesúvisí nevyhnutne s ľudskou zaujatosťou alebo zberom údajov riadeným ľuďmi. Vzniknúť môže napríklad prostredníctvom úzkeho kontextu, v ktorom sa systém používa, pričom v tomto prípade neexistuje príležitosť na jeho generalizáciu voči ostatným kontextom. Zaujatosť môže byť dobrá alebo zlá, úmyselná alebo neúmyselná. V istých prípadoch môže zaujatosť viesť k diskriminačným a/alebo nespravodlivým výsledkom, ktoré sa v tomto dokumente označujú ako nespravodlivá zaujatosť.

Etika

150. Etika je akademická disciplína, ktorá je pododborom filozofie. Vo všeobecnosti sa zaoberá otázkami ako „Čo je dobré konanie?“, „Aká je hodnota ľudského života?“, „Čo je spravodlivosť?“ alebo „Čo je dobrý život?“ V akademickej etike existujú štyri hlavné oblasti výskumu: i) metaetika, ktorá sa týka väčšinou zmyslu a referencie normatívneho výroku a otázky toho, ako možno určiť ich pravdivostné hodnoty (ak existujú); ii) normatívna etika, čiže praktické prostriedky určovania morálneho postupu prostredníctvom skúmania noriem, pokiaľ ide o správne a nesprávne konanie, a pridelenia hodnoty konkrétnemu konaniu; iii) deskriptívna etika, ktorá je zameraná na empirický prieskum morálneho správania a presvedčení ľudí a iv) aplikovaná etika, ktorá sa týka toho, čo máme povinnosť (alebo povolenie) robiť v konkrétnej (často historicky novej) situácii alebo v konkrétnej oblasti (často historicky bezprecedentnej) možností konania. Aplikovaná etika sa zaoberá situáciami zo skutočného života, pri ktorých sa rozhodnutia musia prijímať pod časovým tlakom a často so zníženou racionálnosťou. Etika umelej inteligencie sa vo všeobecnosti chápe ako príklad aplikovanej etiky a zameriava sa na normatívne otázky, ktoré vznikajú pri navrhovaní, vývoji, vykonávaní a používaní umelej inteligencie.

151. V diskusiách o etike sa často používajú výrazy „morálny“ a „etický“. Výraz „morálny“ sa vzťahuje na konkrétne, vecné vzorce správania, zvyky a obyčaje, ktoré možno nájsť v určitom čase v konkrétnych kultúrach, skupinách alebo u jednotlivcov. Výraz „etický“ sa vzťahuje na hodnotiace posúdenie tohto konkrétneho konania a správania zo systematického, akademického hľadiska.

Etická umelá inteligencia

152. V tomto dokumente sa pojem etická umelá inteligencia používa na označenie vývoja, zavádzania a používania umelej inteligencie, ktorou sa zabezpečuje súlad s etickými normami vrátane základných práv, ako osobitných morálnych nárokov, etických zásad a súvisiacich základných hodnôt. Ide o druhú z troch hlavných zložiek nevyhnutných na dosiahnutie dôveryhodnej umelej inteligencie.

Umelá inteligencia zameraná na človeka

153. Cieľom prístupu k umelej inteligencii zameraného na človeka je zabezpečiť, aby ľudské hodnoty boli rozhodujúce pre spôsob, akým prebieha vývoj, zavádzanie, používanie a monitorovanie systémov umelej inteligencie, a to prostredníctvom zabezpečenia dodržiavania základných práv vrátane práv stanovených v zmluvách Európskej únie a Charte základných práv Európskej únie, ktoré všetky spája odkaz na spoločný základ zakotvený v úcte k ľudskej dôstojnosti, v rámci ktorej má ľudská bytosť jedinečné a nescudziteľné morálne postavenie. To znamená aj zohľadnenie prírodného prostredia a ďalších živých bytostí, ktoré sú súčasťou ľudského ekosystému, ako aj udržateľný prístup umožňujúci prosperitu budúcich generácií.

Vytváranie červených tímov

154. Vytváranie červených tímov je činnosť, pri ktorej „červený tím“ alebo nezávislá skupina vyzve organizáciu na zlepšenie svojej účinnosti tým, že prijme nepriateľskú úlohu alebo stanovisko. Táto prax sa používa najmä s cieľom pomôcť identifikovať a riešiť možné zraniteľné miesta v oblasti bezpečnostnej ochrany.

Reprodukovateľnosť

155. Pojmom reprodukovateľnosť sa opisuje to, či sa pokus v oblasti umelej inteligencie prejavil rovnakým správaním pri jeho opakovaní za rovnakých podmienok.

Odolná umelá inteligencia

156. Odolnosť systému umelej inteligencie zahŕňa technickú odolnosť (podľa potreby v danom kontexte, ako je oblasť použitia alebo fáza životného cyklu), ako aj odolnosť zo sociálneho hľadiska (zabezpečenie náležitého zohľadnenia kontextu a prostredia, v ktorých sa systém prevádzkuje). Je to zásadné na zabezpečenie toho, aby sa napriek dobrým úmyslom nemohla spôsobiť neúmyselná ujma. Odolnosť je tretia z trojice zložiek potrebných na dosiahnutie dôveryhodnej umelej inteligencie.

Zainteresované strany

157. Pojem zainteresované strany označuje všetky subjekty, ktoré vykonávajú výskum umelej inteligencie, vyvíjajú ju, navrhujú, zavádzajú alebo používajú, ako aj subjekty, ktoré sú (priamo alebo nepriamo) dotknuté umelou inteligenciou – okrem iného vrátane spoločností, organizácií, výskumných pracovníkov, verejných služieb, inštitúcií, organizácií občianskej spoločnosti, vlád, regulačných orgánov, sociálnych partnerov, jednotlivcov, občanov, pracovníkov a spotrebiteľov.

Vysledovateľnosť

158. Vysledovateľnosť systému umelej inteligencie znamená schopnosť sledovať údaje systému, procesy vývoja a zavádzania, obvykle prostredníctvom zdokumentovaných zaznamenaných podkladov.

Dôvera

159. Toto vymedzenie sme prevzali z literatúry: „Dôvera sa chápe ako: 1. súbor konkrétnych presvedčení týkajúcich sa zhovievavosti, spôsobilosti, integrity a predvídateľnosti (dôverujúce presvedčenia); 2. ochota jednej strany spoľahnúť sa na druhý subjekt v rizikovej situácii (dôverujúci zámer) alebo 3. kombinácia týchto prvkov.“[79] Hoci „dôvera“ zvyčajne nie je vlastnosť pripisovaná strojom, cieľom tohto dokumentu je upozorniť na význam schopnosti dôverovať nie len skutočnosti, že systémy umelej inteligencie sú v súlade so zákonom, eticky vyhovujúce a odolné, ale že túto dôveru možno prisudzovať všetkým ľuďom a procesom, ktoré sú súčasťou životného cyklu systému umelej inteligencie.

Dôveryhodná umelá inteligencia

160. Dôveryhodná umelá inteligencia má tri zložky: 1. mala by byť zákonná a zabezpečovať dodržiavanie celého platného práva a právnych predpisov; 2. mala by byť etická a preukazovať úctu k etickým zásadám a hodnotám a zabezpečovať súlad s etickými zásadami a hodnotami a 3. mala by byť odolná, a to z technického aj sociálneho hľadiska, keďže systémy umelej inteligencie môžu aj pri dobrých úmysloch spôsobiť neúmyselnú ujmu. Dôveryhodná umelá inteligencia sa netýka iba dôveryhodnosti samotného systému umelej inteligencie, ale zahŕňa dôveryhodnosť všetkých procesov a subjektov, ktoré sú súčasťou životného cyklu systému.

Zraniteľné osoby a skupiny

161. Z dôvodu ich rôznorodosti neexistuje žiadna všeobecne akceptovaná alebo všeobecne dohodnutá právna definícia zraniteľných osôb. To, čo určuje zraniteľnú osobu alebo skupinu, často závisí od kontextu. Úlohu môžu zohrávať dočasné životné udalosti (ako je detstvo alebo choroba), trhové činitele (ako je informačná asymetria alebo trhová sila), ekonomické činitele (ako je chudoba), činitele súvisiace s identitou jednotlivca (ako je pohlavie, náboženstvo alebo kultúra) alebo iné faktory. V Charte základných práv EÚ sa v článku 21 zakazuje diskriminácia z týchto dôvodov, ktoré môžu okrem iného slúžiť ako referencia: najmä z dôvodu pohlavia, rasy, farby pleti, etnického alebo sociálneho pôvodu, genetických vlastností, jazyka, náboženstva alebo viery, politického alebo iného zmýšľania, príslušnosti k národnostnej menšine, majetku, narodenia,

[79] Siau, K., Wang, W., *Building Trust in Artificial Intelligence, Machine Learning, and Robotics* (Budovanie dôvery k umelej inteligencii, strojovému učeniu a robotike), *CUTTER BUSINESS TECHNOLOGY JOURNAL*, roč. 31, s. 47 – 53, 2018.

zdravotného postihnutia, veku alebo sexuálnej orientácie. Ďalšie ustanovenia práva sa zaoberajú právami osobitných skupín okrem skupín uvedených vyššie. Žiaden takýto zoznam nie je úplný a časom sa môže meniť. Zraniteľná skupina je skupina osôb, ktoré majú spoločnú jednu alebo viacero vlastností typických pre zraniteľnosť.

Tento dokument vypracovali členovia expertnej skupiny na vysokej úrovni pre umelú inteligenciu

uvedení ďalej v abecednom poradí

Pekka Ala-Pietilä, predseda expertnej skupiny na vysokej úrovni pre umelú inteligenciu
AI Finland, Huhtamaki, Sanoma
Wilhelm Bauer
Fraunhofer
Urs Bergmann – spoluspravodajca
Zalando
Mária Bieliková
Slovenská technická univerzita v Bratislave
Cecilia Bonefeld-Dahl – spoluspravodajkyňa
DigitalEurope
Yann Bonnet
ANSSI
Loubna Bouarfa
OKRA
Stéphan Brunessaux
Airbus
Raja Chatila
Iniciatíva IEEE pre etiku inteligentných/autonómnych systémov a univerzita Sorbonna (Sorbonne Université)
Mark Coeckelbergh
Viedenská univerzita (Universität Wien)
Virginia Dignum – spoluspravodajkyňa
Univerzita v Umeå (Umeå universitet)
Luciano Floridi
Univerzita v Oxforde (University of Oxford)
Jean-Francois Gagné – spoluspravodajca
Element AI
Chiara Giovannini
ANEC
Joanna Goodey
Agentúra pre základné práva
Sami Haddadin
Mníchovská škola robotiky a strojovej inteligencie (Munich School of Robotics and MI)
Gry Hasselbalch
The thinkdotank DataEthics a Univerzita v Kodani (Københavns Universitet)
Fredrik Heintz
Univerzita v Linköpingu (Linköpings universitet)
Fanny Hidvegi
Access Now
Eric Hilgendorf
Univerzita vo Würzburgu (Universität Würzburg)
Klaus Höckner
Hilfsgemeinschaft der Blinden und Sehschwachen
Mari-Noëlle Jégo-Laveissière
Orange
Leo Kärkkäinen
Nokia Bell Labs
Sabine Theresia Köszegi
TU Wien
Robert Kroplewski
Právny zástupca a poradca poľskej vlády
Elisabeth Ling
RELX
Pierre Lucas
Orgalim – Európske technologické odvetvia
Ieva Martinkenaite
Telenor
Thomas Metzinger – spoluspravodajca
JGU Mainz a Európske združenie univerzít
Catelijne Muller
ALLAI Holandsko a EHSV
Markus Noga
SAP
Barry O'Sullivan, podpredseda AI HLEG
University College Cork
Ursula Pachl
BEUC
Nicolas Petit – spoluspravodajca
Univerzita v Liège (L'Université de Liège)
Christoph Peylo
Bosch
Iris Plöger
BDI
Stefano Quintarelli
Garden Ventures
Andrea Renda
College of Europe a CEPS
Francesca Rossi
IBM
Cristina San José
Európska banková federácia
George Sharkov
Digital SME Alliance
Philipp Slusallek
Nemecké výskumné centrum pre umelú inteligenciu (DFKI)
Françoise Soulié Fogelman
Konzultantka pre oblasť umelej inteligencie
Saskia Steinacker – spoluspravodajkyňa
Bayer
Jaan Tallinn
Ambient Sound Investment
Thierry Tingaud
STMicroelectronics
Jakob Uszkoreit
Google
Aimee Van Wynsberghe – spoluspravodajkyňa
TU Delft
Thiébaut Weber
ETUC
Cecile Wendling
AXA
Karen Yeung – spoluspravodajkyňa
Univerzita v Birminghame (University of Birmingham)

Urs Bergmann, Cecilia Bonefeld-Dahl, Virginia Dignum, Jean-François Gagné, Thomas Metzinger, Nicolas Petit, Saskia Steinacker, Aimee Van Wynsberghe a Karen Yeung vystupovali ako spravodajcovia k tomuto dokumentu.

Pekka Ala-Pietilä predsedá expertnej skupine na vysokej úrovni pre umelú inteligenciu. Barry O'Sullivan je jej podpredseda, ktorý koordinuje druhý výstup AI HLEG. K obsahu tohto dokumentu prispela aj Nozha Boujemaaová, podpredsedníčka skupiny do 1. februára 2019, ktorá koordinovala prvý výstup.

Nathalie Smuha poskytla redakčnú podporu.