Umělá inteligence a lidská práva: rizika, příležitosti a regulace

Výzkumný projekt TAČR č. TL05000484

Zpráva o rizicích porušení lidských práv AI technologiemi a možnosti nápravy

Dílčí zpráva

Dílčí zpráva je součástí výzkumného grantového projektu „Umělá inteligence a lidská práva: rizika, příležitosti a regulace“ Technologické agentury ČR (č. TL05000484), který je realizován interdisciplinárním týmem čtyřčlenného konsorcia složeného z AMBIS vysoká škola, a.s., Fakulty elektrotechnické ČVUT, Ústavu práva a technologií Masarykovy univerzity a prg.ai, z.s. v letech 2021–2023. Cílem projektu je identifikovat a zhodnotit rizika a příležitosti v oblasti vztahu umělé inteligence (AI) a lidských práv a navrhnout řešení, jak by AI technologie měly být vyvíjené, používané a regulované, aby neohrožovaly lidská práva a naopak pomohly jejich rozvoji a ochraně.

Dílčí zpráva představuje výsledky první části projektu, zpracované řešiteli z AMBIS vysoká škola, a.s., Fakulty elektrotechnické ČVUT a prg.ai, z.s., jejímž cílem bylo identifikovat zdroj porušení lidských práv v rámci všech fází životního cyklu AI a následně formulovat soubor doporučení k nápravě na úrovni technicky odborné. V následující fázi projektu budou týmem z Ústavu práva a technologií Masarykovy univerzity formulována doporučení k opatření na úrovni regulační. Druhá část projektu (10/2022 – 5/2023) se pak bude věnovat podpoře vývoje a nasazení AI technologií v nejrůznějších oblastech s cílem dále posílit a chránit lidská práva.

Dílčí zpráva nejprve seznamuje s riziky plynoucími z používání AI technologií pro lidská práva stejně jako s možnostmi nápravy na úrovni technické. Popisuje nejen rizika týkající se všech AI technologií, ale nabízí i systematizaci potenciálních rizik a jejich prevence v různorodých oblastech nasazení AI. Následně Dílčí zpráva představuje výsledky dotazníkového šetření mezi firmami, které dodávají AI řešení a produkty, provedeného pod záštitou prg.ai. Cílem bylo promítnout výzkumná zjištění do praxe a zmapovat současný stav reflexe vztahu AI a lidských práv na úrovni vývoje a provozu AI systémů. Závěrem Dílčí zpráva obsahuje soubor doporučení pro adaptaci vývoje, zavádění a používání AI technologií v souladu s mezinárodními normami ochrany lidských práv adresovaný subjektům životního cyklu AI. Pro jednotlivé AI technologie a jejich fáze životního cyklu tak výsledek poskytne vodítka pro přijetí potřebných opatření a pravidel.

Zpracovali: JUDr. Martina Šmuclerová, Ph.D., DEA, Ing. Luboš Král, Ph.D., Ing. Jan Drchal, Ph.D., Mgr. Jaroslav Šíp, Lenka Kučerová, MSc.

Praha, srpen 2022

Obsah

I. Lidská práva a rizika plynoucí z AI technologií

Úvod

Umělá inteligence představuje pro společnost velký přínos, ale i riziko do té míry, kterou společnost povolí. Zásadní úlohou je zajistit efektivní implementaci existujících lidskoprávních norem v oblasti AI[1], což vyžaduje úzkou interdisciplinární spolupráci mezi experty obou domén. Prvotním krokem je identifikace kořenové příčiny porušení lidského práva, která se může nalézat jak ve vývojové, tak v provozní fázi životního cyklu AI, což následně umožní vymezit adekvátní nápravu. Tato studie poukazuje na to, že přes rostoucí diverzitu AI technologií a multiplikaci sfér jejich užití v dnešní době jsou příčiny a potenciální riziko porušení lidských práv podobné a náprava na úrovni technické i regulační spočívá na stejných principech. Právo na soukromí, zákaz diskriminace a právo na spravedlivý proces mohou být ohroženy napříč všemi AI technologiemi (1.) a dále v jednotlivých oblastech nasazení AI může riziko porušení lidských práv vycházet např. z nevyvážených dat, nedostatečně definovaných požadavků a vlastností systému či změny kontextu nasazení, použití AI za nelegálním účelem či jejího zneužití (2.). Základem je zavést režim hodnocení rizik z pohledu lidských práv do celého životního cyklu AI a zahrnout jej do uživatelských požadavků a specifikací. Toto mj. zajistí, že AI systém bude vyvinut, testován a monitorován s ohledem na normy lidských práv.

1. Riziko porušení lidského práva společné všem AI technologiím a možnost nápravy

AI technologie jsou dnes nasazovány v nejrůznějších oblastech lidské činnosti a jejich diverzita a potenciál využití roste: biometrické AI technologie, vč. rozpoznávání obličejů, predikce v trestním řízení, policejní predikce, zdravotní diagnostika a robotická chirurgie, autonomní vozidla, automatizace systému sociálního zabezpečení, automatizace bankovních služeb (např. poskytnutí hypotéky), automatizace výběrových řízení a evaluace zaměstnanců, sociální kreditní systém, e-shopping a cenová diskriminace, chatboty, monitoring a revize online obsahu, autonomní zbraně a dále AI technologie využívané v oblasti migrace, zemědělství, marketingu, advokacii, turismu aj. Ačkoli má každá AI technologie jiný účel a může fungovat na různých principech, vnitřní procesy a základní technické prvky jsou identické a mohou obsahovat nedostatky vedoucí k porušení lidského práva.[2] Ve všech těchto

[1] Výzkumný rámec projektu je založený na mezinárodněprávním režimu ochrany lidských práv, který je právně závazný pro Českou republiku, tj. primárně Mezinárodním paktu o občanských a politických právech, 1966, 999 UNTS 171, Mezinárodním paktu o hospodářských, sociálních a kulturních právech, 1966, 993 UNTS 3, potažmo Evropské úmluvě o ochraně lidských práv, 1950, ETS 5.

[2] Životní cyklus AI systému je složen s následujících fází: 1. Analýza požadavků (*business understanding)* – sběr informací o aplikační doméně, datech, procesech a vhodných AI technologiích; 2. předzpracování dat (*data preparation*) – sběr surových vývojových dat (např. dat pro učení modelů strojového učení), čištění dat, jejich

sférách mohou být ohrožena tři základní lidská práva: právo na soukromí, zákaz diskriminace a právo na spravedlivý proces.

1.1. Právo na soukromí

Právo na soukromí je garantováno v čl. 17 Mezinárodního paktu o občanských a politických právech (*International Covenant on Civil and Political Rights* – ICCPR), který stanoví: „Nikdo nesmí být vystaven svévolnému zasahování do soukromého života, do rodiny, domova nebo korespondence ani útokům na svou čest a pověst.“[3] Podobně Evropská úmluva o lidských právech (EÚLP, *European Convention on Human Rights*) zakotvuje v čl. 8 „právo na respektování svého soukromého a rodinného života, obydlí a korespondence“.[4] Ostatní mezinárodní lidskoprávní instrumenty, stejně jako předpisy na regionální a národní úrovni, tyto normy reflektují a podle OSN právo na soukromí a nutnost zajistit jeho ochranu v právu i v praxi požívají „univerzálního uznání [jejich] zásadního významu a trvalé důležitosti“[5]. Bez ohledu na konkrétní systematizaci konceptů a termínů mezinárodněprávní úprava závazku chránit soukromý život zahrnuje širokou škálu aspektů a hodnot, které se týkají fyzické, psychologické a morální identity člověka, soukromí *stricto sensu* a lidské identity a autonomie. AI technologie jsou dotčeny především s ohledem na riziko v oblasti ochrany osobních údajů. Data jsou základem umělé inteligence a významná část může obsahovat osobní údaje. Osobní údaj nese informaci, která se týká identifikované či identifikovatelné fyzické osoby. Může např. obsahovat jméno a příjmení, adresu bydliště, číslo občanského průkazu, data o geolokaci na mobilním telefonu, fotografii a jiná biometrická data, stejně jako údaje o zdraví či náboženském vyznání. Evropský právní standard a definice osobních údajů jsou zakotveny v Obecném nařízení na ochranu osobních údajů (*General Data Protection Regulation* –

transformace, výběr atributů (podmnožiny vývojových dat určených k učení modelů), rozdělení dat na trénovací a testovací části (trénovací data jsou použita pro výběr a naučení modelu, např. neuronové sítě, testovací pro odhad kvality modelu v rámci následného nasazení do provozu) a anotace (definice očekávaných cílových výstupů systému vzhledem k jeho vstupům); 3. modelování (*modelling*) – výběr vhodného typu modelu a algoritmu pro jeho učení, samotné učení modelu, 4. evaluace (*evaluation*) – správnost systému je testována a ověřována z hlediska plánované funkčnosti, přičemž evaluace je založena na jedné či více metrikách, které číselně hodnotí jeho kvalitu, 5. nasazení (*deployment*) – systém je uveden do plného provozu. Tyto uvedené fáze odpovídají standardizovaným metodám, jako jsou Cross-Industry Standard Process for Data Mining (CRISP-DM), viz C. Shearer, ‘The CRISP-DM Model: The New Blueprint for Data Mining’, 5 *Journal of data warehousing* (2000) 13; nebo the Team Data Science Process, viz například Microsoft Azure, ‘What is the Team Data Science Process?’, dostupné na https://docs.microsoft.com/en-us/azure/architecture/data-science-process/overview. Životní cyklus AI je v této studii představován na příkladu systému založeném na strojovém učení, který je dostatečně reprezentativní. Ostatní paradigmata AI představovaná zejména metodami klasické AI, jako je plánování, agentní metody nebo herně-teoretické metody, budou typicky řešeny podmnožinou uvedených fází. Aktéři životního cyklu AI jsou uvedeni v pozn. č. 21. Ohledně pojmu AI a další terminologie viz ISO/IEC FDIS 22989 ‘Information technology — Artificial intelligence — Artificial intelligence concepts and terminology’ dostupné na https://www.iso.org.

[3] Mezinárodní pakt o občanských a politických právech, 1966, 999 UNTS 171, čl. 17 (1).

[4] Evropská úmluva o ochraně lidských práv, 1950, ETS 5.

[5] The right to privacy in the digital age. Report of the Office of the United Nations High Commissioner for Human Rights, UN Doc. A/HRC/27/37, 30 June 2014, para. 13 (překlad autorů).

GDPR)[6], účinném v EU od 25. května 2018. Osobní údaje mohou být obsaženy jak ve vstupních datech[7], tak v datech z nich získaných[8] a jsou používány v nejrůznějších technologiích, jako jsou rozpoznávání obličeje, zdravotní diagnostika, geografické lokalizace nebo nákupní zvyklosti. Právo na soukromí je potenciálně ohroženo během sběru dat a zpracování dat jak ve fázi vývoje, tak nasazení AI technologie.

V Evropské unii je tak právní ochrana zajištěna režimem GDPR, který se aplikuje i do sféry AI.[9] Tato otázka je tedy právně pokryta a výzva zůstává dvojí: za prvé, účinná a skutečná implementace pravidel GDPR v praxi během celého životního cyklu AI, vč. technických procesů vývoje AI, a za druhé, zavedení právního režimu ochrany osobních údajů celosvětově. Následně je třeba řešit v praxi tyto otázky: Pracuje AI vývojář[10] s anonymizovanými daty či používá pseudonymizaci? Je reálně proveditelné nahrazovat „skutečná" data syntetickými? V jakých oblastech a za jakých podmínek se použití syntetických dat doporučuje? Je možné omezit typ vstupních dat, a způsob jejich zpracování, která bude AI technologie sbírat a zpracovávat během nasazení? Např. vymezit Alexe, aby „neposlouchala", neukládala a nezpracovávala osobní a citlivé informace, které mohou být proneseny v domácnosti, kde je nasazena, a které jí nejsou určené?

Anonymizace se používá ve zdravotnictví, mobilních službách, při zpracování obrazu a videa, při sčítání lidu, zpoplatnění dat, reportování třetím stranám a v dalších oblastech. Spočívá v odstranění částí dat, které by mohly umožnit zpětné ztotožnění osob. Anonymizace však stále nedává plnohodnotnou záruku, že původní data nebudou rekonstruována. Pseudonymizace je naopak založena na tom, že informace, která přesně určuje identitu subjektu, je nahrazena „pseudonymy" či „identifikátory", což zabraňuje tomu, aby data odkazovala na konkrétní subjekt, avšak umožňuje jeho pozdější identifikaci ve specifických případech. Pseudonymizace se například využívá pro hodnocení zaměstnanců, statistické výzkumy (např. testování nových léčiv), sledování zákaznických preferencí, analýzu uživatelského chování pro cílenou reklamu. Pseudonymizace uchovává individualizaci analyzovaného datasetu, a tím i například souvislosti mezi událostmi a záznamy, vztahující se k jedné osobě, zatímco anonymizace tyto spojitosti ruší (např. ruší sledování pomocí cookies).

---

[6] Nařízení Evropského Parlamentu a Rady (EU) 2016/679 ze dne 27. dubna 2016 o ochraně fyzických osob v souvislosti se zpracováním osobních údajů a o volném pohybu těchto údajů a o zrušení Směrnice 95/46/ES (Obecné nařízení o ochraně osobních údajů). Čl. 4 (1) definuje osobní údaje jako „veškeré informace o identifikované nebo identifikovatelné fyzické osobě (dále jen ‚subjekt údajů')", přičemž identifikovatelnou fyzickou osobou je „fyzická osoba, kterou lze přímo či nepřímo identifikovat, zejména odkazem na určitý identifikátor, například jméno, identifikační číslo, lokační údaje, síťový identifikátor nebo na jeden či více zvláštních prvků fyzické, fyziologické, genetické, psychické, ekonomické, kulturní nebo společenské identity této fyzické osoby".

[7] Termín „vstupní data" v tomto dokumentu zahrnuje vývojová data i data, která AI systém zpracovává za provozu. Pokud se jedná pouze o jednu podskupinu, jsou užity termíny „vývojová data" či „operační data".

[8] Jedná se zejména o různé agreace původních dat, parametry modelů strojového učení, předpočítané indexy (například v AI systémech vyhledávajících v textech). V mnoha případech není zpětná transformace na původní data možná, není však obecně vyloučena.

[9] Oblast GDPR není předmětem této studie založené na mezinárodněprávním rámci ochrany lidských práv.

[10] Pojem zahrnuje v této studii všechny aktéry účastnící se vlastního vývoje AI, tj. datového analytika, systémového inženýra, specialitu přes strojové učení, vývojáře a testera.

Použití syntetických dat je možné, ale pouze v omezené míře. Jejich použití jde proti podstatě AI a strojového učení, neboť pokud by bylo možné generovat syntetická data, která by plně nahradila ta reálná, existovalo by algoritmické řešení problému a nasazení AI metod by nebylo nutné. Nahrazování reálných dat umělými vzorky může navíc závažně poškodit vypovídací hodnotu celých vstupních dat. V praxi jsou syntetická data používána k doplnění vývojových dat, kde není možné pokrýt celou množinu příkladů z časových či jiných praktických důvodů. Například automobilka Tesla využívá syntetických dat pro trénování situací, kde fyzická data z reálného dopravního provozu chybí či kde by jejich sběr byl časově náročný. Dalším příkladem je Amazon, který využívá syntetických dat pro trénování lokalizací služby Alexa pro jazykové mutace hindštinu, španělštinu používanou v USA či brazilskou portugalštinu.[11]

Nelehkou otázkou je omezení typu dat, která bude AI technologie „sbírat" za provozu, se současným zajištěním ochrany práva na soukromí. Toto téma bylo obzvlášť relevantní, zejm. v profesní sféře, s ohledem na fenomén home office během covidové éry, kdy některé společnosti instruovaly své zaměstnance, aby vypnuli Alexu a jiná domácí AI zařízení s cílem zabránit úniku obchodních informací a odposlechu.[12] Nejen že je v sázce kybernetická bezpečnost, ale i pouhé zpracování informací Alexou v sobě obsahuje prvek uchování dat v cloudu. Uživatel si ani nemusí být vědom takového toku dat či se AI zařízení může náhle „probudit". Výzvou je, jak *předem* nastavit AI technologii, již během fáze strojového učení, aby se vypnula, pokud je v sázce osobní či jinak citlivá informace, což je však těžko představitelné předtím, než ji uživatel vůbec vysloví. Vzhledem k tomu, že se zpracování dat – a tedy i jejich analýza – realizuje v cloudu, záznam dat, jinými slovy „absorpce dat", je nevyhnutelná. Jediná představitelná restrikce je okamžitá eliminace dat z cloudu po jejich zpracování. Nelze však zaručit, že datová stopa nebude zachována. Ochrana osobních či jinak citlivých informací v této konkrétní situaci tak neleží primárně v technické sféře, ale v lidském přístupu a preventivním chování.

Aby se zabránilo riziku porušení práva na soukromí při nakládání s osobními údaji, lze obecně zdůraznit tři základní přístupy ve vztahu k AI: za prvé, preference pro uzavřené systémy, tj. systémy, které neposkytují informace třetím stranám, ale používají je pouze pro deklarovaný účel[13]; za druhé, anonymizace či pseudonymizace vstupních dat; a za třetí, souhlas poskytnutý uživatelem pro zpracování jeho osobních údajů a jejich další použití.

---

[11] Viz J. Slifka, 'Tools for Generating Synthetic Data Helped Bootstrap Alexa's New-language Releases', *Amazon Science*, dostupné na https://www.amazon.science/blog/tools-for-generating-synthetic-data-helped-bootstrap-alexas-new-language-releases.

[12] Viz např. 'Protect Your Amazon Echo Privacy While Working From Home: 7 Simple Tricks', *CNET*, 26 February 2022, https://www.cnet.com/home/smart-home/protect-your-amazon-echo-privacy-while-working-from-home-7-simple-tricks/; 'Calif. Bar to Attorneys: Disable Alexa when Working from Home', 13 August 2021, *Reuters*, dostupné na https://www.reuters.com/legal/legalindustry/calif-bar-attorneys-disable-alexa-when-working-home-2021-08-13/; and 'Coronavirus: Employees Urged to Turn off Alexa-style Devices While Working from Home Due to Privacy Fears', 24 March 2020, *Yahoo news*, https://uk.news.yahoo.com/.

[13] Např. aby hlasoví asistenti typu Alexy používali data pouze pro své vlastní doučování a neposkytovali je třetím stranám či aby sociální média nesdílela data se třetí stranou – viz např. kauza Cambridge Analytica, kde osobní údaje milionů uživatelů Facebook byly sbírány britskou firmou Cambridge Analytica bez jejich souhlasu a užity především pro politickou reklamu, Amnesty International, ''The Great Hack': Cambridge Analytica is Just the Tip

1.2. Zákaz diskriminace

Diskriminace na jakémkoli základě, jako je rasa, barva, pohlaví, jazyk, náboženství, politické nebo jiné přesvědčení, národnostní nebo sociální původ, majetek, rod nebo jiné postavení (ICCPR, čl. 26)[14] stejně jako příslušnost k národnostní menšině (hodnota obsažená navíc v EÚLP, čl. 14)[15], je zakázána.[16] Diskriminace na základě těchto tzv. chráněných hodnot je protiprávní, pouze pokud došlo k „rozdílnému zacházení s osobami v analogické či relevantně podobné situaci“[17], či k opomenutí zacházet různě s osobami, které se nacházejí v relevantně odlišných situacích, bez „objektivního a důvodného ospravedlnění“[18]. To znamená, že takový rozdíl – či absence rozdílu – je protiprávní, ledaže by sledoval legitimní účel a použité prostředky by byly rozumně proporcionální danému účelu. Úmysl diskriminovat není určujícím faktorem, tím je pouze výsledek. V důsledku je tak jak přímá, tak nepřímá diskriminace zakázána. V případě přímé diskriminace je kritérium odlišného zacházení přímo chráněná hodnota, což je zakázáno. Naopak nepřímá diskriminace staví na kritériu, které je, *prima facie*, neutrální (jako např. geografická lokace), avšak její účinky dopadají disproporčně více na jednu skupinu osob.[19]

Jak se toto týká AI? Jaké mohou být potenciální zdroje neoprávněné diskriminace? Prvek způsobující v rámci AI neoprávněné odlišné zacházení (diskriminaci) na bázi chráněné hodnoty, tzv. algoritmická předpojatost či bias[20], může být vnesen do AI technologie na vícero úrovních: především skrze vstupní data, výběr vstupních atributů, rozdělení dat na trénovací a testovací množiny, anotace dat, výběr metrik pro měření kvality systému, přepoužívání existujících modelů (ve strojovém učení označováno jako transfer learning) či vlastnosti prostředí, do kterého je systém nasazen. Rovněž se může objevit v rámci provozu, pokud je systém modifikován na základě dat, která zpracovává (například model strojového učení může

of the Iceberg', 24 July 2019, dostupné na https://www.amnesty.org/en/latest/news/2019/07/the-great-hack-facebook-cambridge-analytica/, and 'Cambridge Analytica Scandal 'Highlights Need for AI Regulation'', *The Guardian*, 16 April 2018, dostupné na https:// www.theguardian.com/technology/2018/apr/16/cambridge-analytica-scandal-highlights-need-for-ai-regulation.

[14] ICCPR, *op. cit.* n. 1.

[15] EÚLP, *op. cit*. n. 1.

[16] Seznam chráněných hodnot není taxativní, je pouze příkladný (výčet v ICCPR i ECHR obsahuje obrat „z jakýchkoli důvodů, např.“ a „nebo jiné postavení“),

[17] Specifikace formulovaná Evropským soudem pro lidská práva (např. Biao v. Denmark [GC], 2016, para. 89; Carson and Others v. the United Kingdom [GC], 2010, para. 61; D.H. and Others v. the Czech Republic [GC], 2007, para. 175) (překlad autorů). Výbor OSN pro lidská práva hovoří jednoduše o „různém zacházení“ („differentiation of treatment“), UN Human Rights Committee, Thirty-seventh session (1989), General comment No. 18: Non-discrimination, para. 13, dostupné na https://www.ohchr.org/en/hrbodies/ccpr/pages/ccprindex.aspx.

[18] Specifikace formulovaná Evropským soudem pro lidská práva (např. Molla Sali v. Greece [GC], 2018, para. 135; Fabris v. France [GC], 2013, para. 56; D.H. and Others v. the Czech Republic [GC], 2007, para. 175) (překlad autorů). Výbor OSN pro lidská práva potvrzuje stejnou podmínku následovně: „Výbor podotýká, že ne každé odlišné zacházení představuje diskriminaci, pokud jsou kritéria takového rozlišování důvodná a objektivní a pokud je cílem dosáhnout účelu, který je podle Paktu legitimní“, *ibid*., para. 13 (překlad autorů).

[19] Viz např. judikatura Evropského soudu pro lidská práva na toto téma (Biao v. Denmark [GC], 2016, para. 103; D.H. and Others v. the Czech Republic [GC], 2007, para. 184; Sampanis and Others v. Greece, 2008, para. 67).

[20] Termín bias je zde třeba odlišit od standardně používaného pojmu „bias“ v AI sféře, kterým se myslí obecně „odchylka“/„chyba“ v datech, nikoli tedy nutně založená na tzv. chráněné hodnotě.

být kontinuálně doučován na základě aktuálních dat). Protiprávní používání AI technologie za diskriminačním účelem je rozebráno zvlášť v Sekci 2.3.

1.2.1. Nevyvážená data

Data jsou palivem pro AI technologie. Pokud nereflektují věrně realitu, obraz, který vytváří ve výsledném AI systému, je od samého začátku nevypovídající. Fragmentovaná, neúplná či zkreslená data, ať již úmyslně či nevědomky, představují primární a základní zdroj biasu. Náprava v tomto bodě by tak měla představovat prvořadý zájem všech aktérů životního cyklu AI[21].

Například pokud poskytovatel vývojových dat pro AI systém prediktivního soudnictví, který odhaduje míru pravděpodobnosti recidivy, dodá neúplná data z policejních a vyšetřovacích databází pouze vybraných lokalit, či zaměřených pouze na určité pohlaví, výstup daný AI technologií může obsahovat bias proti určité etnické nebo náboženské menšině či pohlaví. Vstupní data tak již vnesou neopodstatněnou váhu určitých vzorků (což může zahrnovat chráněnou hodnotu, jako je pohlaví, barva pleti, náboženské vyznání a příslušnost k národnostní menšině), která bude dále reprodukována v AI systému. Podobným případem je AI systém pro výběrové řízení do zaměstnání testující a predikující potenciální míru úspěšnosti kandidátů, který znevýhodňuje ženy. Bias může být způsoben použitím trénovacích dat, která obsahují vzorky týkající se především mužské populace, jejíž zastoupení v daném profesním sektoru bylo historicky převažující, a tedy data byla sebrána z této reálné situace.[22]

Tento problém se týká jakéhokoli datasetu, který není vyvážený – bias může být obsažen ve vývojových datech, ale může být vnesen následně i na úrovni zpracování dat, zejm. během výběru atributů či rozdělení na trénovací a testovací data. Může být rovněž způsoben nevhodným použitím syntetických dat.

Otázkou je, jak zabránit vnesení biasu či jak jej případně dodatečně odstranit. Nejprve je třeba zjistit, zda data nesoucí chráněnou hodnotu mohou vůbec být vzata v potaz či musejí být kompletně eliminována z AI systému. Kdy a za jakých podmínek mohou AI systémy pracovat s chráněnou hodnotou, to je technicky náročná úloha transferu lidského uvažování do počítačových modelů.

[21] Aktéry životního cyklu jsou výrobce (vyvíjí AI systém, vč. AI inženýrů), dodavatel (nechává AI systém vyvinout a uvádí jej na trh jako produkt či službu), zákazník (získává AI systém s cílem jej provozovat či používat), provozovatel (provozuje AI systém), uživatel (užívá AI systém jako koncový uživatel) a regulátor (definuje pravidla pro vývoj a provozování AI systému). Všichni aktéři se mohou účastnit stanovení uživatelských požadavků (tzv. user stories).

[22] Amazon čelil podobnému problému ve svém AI systému pro nábor zaměstnanců v r. 2018, viz https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

1.2.1.1. Chráněná hodnota

Eliminace chráněné hodnoty ze vstupních dat, či přesněji z vybraných atributů[23], není řešením: existují situace, které vyžadují přítomnost určité chráněné hodnoty, a navíc její odstranění není v praxi vždy technicky možné. Lze definovat následující principy:

Za prvé, chráněná hodnota *může být vzata v potaz*, pokud je odlišné zacházení na její bázi ospravedlnitelné a proporcionální vůči legitimnímu účelu, který sleduje. Příkladem může být užití AI v dermatologii, kde kritérium barvy pleti může být relevantní pro přesnou zdravotní diagnózu a určité fyzické rysy představují zásadní prvek. Podobně výběrové kritérium v rámci castingu do filmu založené na věku či pohlaví může být považováno za ospravedlnitelné a proporcionální vůči sledovanému účelu.[24] Chráněná hodnota tedy může právoplatně figurovat mezi atributy v AI systému a nemusí být odstraněna.

Za druhé, chráněná hodnota *nesmí být vzata v potaz*, pokud by odlišné zacházení z(ne)výhodňující jednu skupinu či jejího příslušníka nebylo ospravedlnitelné a proporcionální účelu. V tomto scénáři např. informace o etnickém původu či pohlaví nesmí figurovat mezi atributy, na kterých je založena AI technologie predikující pravděpodobnost recidivy. Stejně tak bankovní systém poskytování úvěrů využívající AI nesmí zohledňovat pohlaví či i nepřímé faktory jako mateřská dovolená.

Za třetí, chráněná hodnota *musí být vzata v potaz na rovnoměrné bázi*, aby bylo zabráněno z(ne)výhodňování jedné skupiny či jejího příslušníka. Cílem je zajistit objektivní reprezentaci bez diskriminace. Tento scénář připadá v úvahu, pokud chráněná hodnota představuje nezbytný a určující prvek pro samotné fungování AI technologie, např. pro rozpoznávání obličejů, kde identifikace obličejových rysů, vč. barvy, je potřeba právě, aby se zabránilo z(ne)výhodňování. Kromě toho tato potřeba může vyvstat, pokud je chráněná hodnota obsažena ve vstupních datech a odlišné zacházení na její bázi není ospravedlnitelné a proporcionální účelu, avšak není technicky možné ji odstranit. Jako příklad lze uvést rozpoznávání hlasu, kdy si rozpoznávaný hlas musí zachovat svou původní charakteristiku, lékařské expertní systémy, statistická analýza pro účely marketingu atd.

Zajištění rovnoměrné reprezentace, například prostřednictvím vyrovnání a vyvážení datasetu, vč. doplnění chybějících dat, je zásadní pro zabránění neoprávněné diskriminaci. Příkladem je doporučovací systém, tj. systém doporučující zboží v e-shopech, filmy, hudební playlisty atd. Pro podobný systém je typické, že bere v potaz kulturní prostředí, případně etnicitu uživatelů – je tedy nutné, aby vývojová data pokrývala kompletní reprezentaci daných chráněných hodnot.

[23] V praxi stačí, aby se AI inženýr vyvaroval chráněné hodnoty ve výběru atributů, neboť nepoužitá vstupní data nebudou vzata v potaz v algoritmech.

[24] I když viz. např. Aja Romano, 'The Debate over Bridgerton and Race' (7 January 2021) <https://www.vox.com/22215076/bridgerton-race-racism-historical-accuracy-alternate-history> accessed 17 June 2022. Toto zdůrazňuje potřebu kontinuální aktualizace AI systémů s ohledem na možný vývoj lidskoprávních hodnot, dále viz. Sekce 3.

1.2.2. Evaluační metriky a transfer learning

Dalším způsobem, jak se může při vývoji do AI systému zavléct bias, je volba evaluačních metrik a přepoužívání existujících AI řešení (transfer learning).

Evaluační metriky popisují kvalitu AI systému. V průběhu vývoje i v rámci nasazení je sledováno hned několik takových metrik. Metriky jsou typicky vybrány před fází modelování, kdy slouží k výběru vhodného typu modelu, dále pak v samotném průběhu učení modelů, ale i ve fázi evaluace, kdy se s jejich pomocí odhadují schopnosti AI systému po uvedení do provozu. Pokrývají široké spektrum od nejjednodušších, jakou je klasifikační přesnost, až po potenciálně velmi složité – například metriky odhadující profit automatizovaného systému pro obchodování na burze. Výběr metriky i detaily její implementace mohou negativně ovlivnit vlastnosti AI systému z hlediska biasu. Příkladem může být velmi častý stav, kdy jsou vývojová data nevyvážená (některé zastoupené skupiny jsou pokryty výrazně menším počtem vzorků než ty majoritní). V tomto případě může vést nevhodná volba evaluační metriky k tomu, že bude odhad kvality naučených modelů příliš optimistický, přičemž reálně budou predikce pro minoritně zastoupené skupiny víceméně náhodné.

Další problematickou metodologií, která může vést k zanesení biasu do vyvíjeného AI systému, je dnes velmi rozšířený transfer learning. Transfer learning je využíván zejména v kontextu strojového učení. Základní ideou je přepoužití existujících modelů, které byly historicky použity pro řešení podobného problému, který má řešit vyvíjený systém. Důvodem pro přepoužívání může být nedostatek dat pro specifickou cílovou úlohu, případně extrémní výpočetní náročnost. Jako příklad, kde je transfer learning masivně nasazován, mohou sloužit v podstatě všechny nejmodernější aplikace zpracování přirozeného jazyka, např. klasifikace textu, sumarizace textu, strojový překlad apod. Tyto úlohy jsou typicky postaveny na velkých neuronových jazykových modelech, které jsou předučeny na rozsáhlých korpusech (databázích vzorových textů), díky čemuž umí reprezentovat sémantiku obecných textů. AI vývojář pak takový jazykový model pouze dolaďuje pro specifickou úlohu, přičemž objem dat nutný pro doladění je řádově menší než objem dat použitý pro naučení přepoužitého jazykového modelu. Problémem z hlediska biasu je, že předučené modely jsou nejčastěji produkovány třetí stranou (vzhledem k výpočetní náročnosti jsou to často velcí komerční hráči na poli AI jako Google, Facebook, Amazon, případně velké univerzity). V mnoha případech pak nejsou dostupné detailní informace o tom, jak byl model předučen, a hlavně nebývají dostupná data, která byla k předučení použita. Potenciální problém pak vychází z toho, že vývojář cílového AI systému nemá kontrolu nad negativy (včetně biasu), která mohou být přítomna v převzatém modelu.

Další výzvy související se zákazem diskriminace, jako je nepřímá diskriminace či otevřený, pouze příkladný výčet chráněných hodnot, vyžadují další interdisciplinární výzkum. Tyto prvky, které jsou nejisté i v lidském světě, jsou o to více zvýrazněny, pokud je vyžadován jejich převod do počítačových modelů.

1.3. Právo na spravedlivý proces

Právo na spravedlivý proces představuje základní procesní lidské právo, které zajišťuje případné oběti porušení lidských práv přístup k nezávislému a nestrannému soudu včetně všech potřebných právních a procesních záruk. Článek 14 Mezinárodního paktu o občanských a politických právech, který zahrnuje celou škálu specifických oprávnění, stanoví v první větě odstavce 1 následující: „Všechny osoby jsou si před soudem rovny. Každý má úplně stejné právo, aby byl spravedlivě a veřejně vyslechnut nezávislým a nestranným soudem, který rozhoduje buď o jeho právech a povinnostech, nebo o jakémkoli trestním obvinění vzneseném proti němu.“[25] Výbor pro lidská práva OSN ve své obecné připomínce č. 32 zdůrazňuje: „Právo na rovnost před soudy a tribunály rovněž zajišťuje rovnost zbraní. […] . Princip rovnosti mezi stranami se vztahuje i na občanské řízení a požaduje, inter alia, aby měla každá strana příležitost popřít argumenty a důkazy předložené druhou stranou.“[26] Například Evropský soud pro lidská práva v příručce *Guide on Article 6 of the European Convention on Human Rights – Right to a Fair Trial* poznamenává: „neposkytnutí důkazů žalované straně může porušit rovnost zbraní stejně jako právo na kontradiktornost řízení“.[27]

Právo na spravedlivý proces je dnes ohroženo zejména s ohledem na nevysvětlitelnou (1.3.2.) a netransparentní (1.3.1.) AI. Pokud by osoba namítající porušení lidského práva neměla přístup k informacím a důkazům, na jejichž základě o ní AI rozhodla, její obhajoba by byla ohrožena, resp. znemožněna. Pokud by komplexní AI systém profilující jednotlivce určoval poskytování či přístup k veřejným službám, jako je sociální zabezpečení, policejní dohled či migrační kontrola apod., odepření přístupu k důkazům by zmařilo účinný, resp. reálný výkon práva dotčené osoby na spravedlivý proces. Stejně tak pokud je zaměstnanci odepřena rovná příležitost v kariérním postupu na základě vyhodnocení AI, nemožnost získat informace zakládající takové rozhodnutí podrývá dané osobě možnost své obhajoby. Zajištění, aby bylo možné odkrýt vnitřní logiku fungování AI systémů, tak představuje základ vlády práva a práva na spravedlivý proces. AI by měla odůvodnit svá rozhodnutí a poskytnout potřebnou evidenci, stejně jako jsou povinni lidé.

[25] ICCPR, *op. cit.* n. 1.

[26] Human Rights Committee, General Comment No. 32, Article 14: Right to Equality Before Courts and Tribunals and to a Fair Trial, UN Doc. CCPR/C/GC/32, 23 August 2007, para. 13 (překlad autorů).

[27] European Court of Human Rights, *Guide on Article 6 of the European Convention on Human Rights – Right to a Fair Trial (Criminal Limb)*, Updated on 31 December 2021, 2022, para. 174, dostupné na https://www.echr.coe.int/ documents/guide_art_6_criminal_eng.pdf (překlad autorů). Soud podobně potvrdil, že „právo na kontradiktornost řízení v principu znamená příležitost stran trestního či občanskoprávního řízení seznámit se se všemi předloženými důkazy a připomínkami a mít možnost se k nim vyjádřit“, European Court of Human Rights, *Guide on Article 6 of the European Convention on Human Rights – Right to a Fair Trial (Civil Limb)*, Updated on 31 August 2021, 2022, para. 377, dostupné na https://www.echr.coe.int/ documents/guide_art_6_eng.pdf (překlad autorů).

1.3.1. Operační transparentnost AI

Transparentnost AI systému zahrnuje vývojovou a operační transparentnost. Vývojová transparentnost slouží především aktérům schvalujícím systém do provozu. Týká se provádění vývojového procesu a umožňuje jeho kontrolu, vč. kvality. Operační transparentnost naopak vyjadřuje schopnost AI systému poskytnout informaci o tom, která vstupní data přispěla k výslednému výstupu systému a s jakou vahou. V provozní fázi tím i systém dodává informace pro svůj monitoring, čímž umožňuje zpětný přezkum a audit své činnosti. Z lidskoprávního hlediska je operační transparentnost podstatná pro požívání práva na spravedlivý proces. Jak uvedeno výše, právo na spravedlivý proces zahrnuje právo mít přístup a vyjádřit se ke všem argumentům a důkazům předloženým druhou stranou. Je tedy nezbytné, aby případný stěžovatel měl přístup k maximu dostupných informací o činnosti AI systému.

Vývojář AI technologie by měl být schopen poskytnout informace o všech fázích procesu vývoje, tj. analýzy požadavků, přípravy a předzpracování dat, modelování, testování a uvedení systému do provozu. Provozovatel systému by měl dodat provozní záznamy. To by mělo být základem transparentnosti AI systému, což podpoří právo na spravedlivý proces.

Hlavní problémy týkající se poskytování informací stěžovateli či soudci jsou následující: Za prvé, některá vstupní data mohou být důvěrná. Tato právní ochrana však může být zrušena soudním rozhodnutím v případě nutnosti. Za druhé, transfer learning dnes často používaný v aplikacích strojového učení ztěžuje přístup k informacím o přepoužitém modelu, ledaže by byla přijata právní norma stanovící, že takový model musí být doplněn o příslušné informace v souladu se standardizovaným popisem životního cyklu AI. Za třetí, převod výstupů modelu do čitelné a přístupné formy a její přizpůsobení konkrétnímu případu užití může být pracné i časově náročné. Informace z vývojového procesu jsou navíc srozumitelné pouze pro IT a AI experty. To se týká i vytvoření *ex post* metod na snížení dopadů nevysvětlitelnosti, které by byly k dispozici třetím stranám (viz níže). Měla by příslušná regulace povinnosti zajistit přístup k informacím ohledně AI systému vzít v potaz finanční aspekty a stejně tak i soudce nařizující takovou povinnost? Kromě toho by měly být zohledněny i otázky práva duševního vlastnictví a obchodního tajemství.

Zvláštní pozornost by měla být věnována zveřejnění informací o algoritmech AI systémů, které jsou používány pro výkon veřejné správy a služeb, především pokud se jedná o poskytování a přístup ke klíčovým veřejným službám či v případě donucovacích a soudních orgánů. „Algoritmus popisující slovně či graficky ‚krok za krokem' danou operaci" Systému náhodného přidělování případů (SLPS) založeného na AI, který v Polsku od roku 2018 rozhoduje o náhodném přidělování případů soudcům obecných soudů, byl v roce 2021 polským Nejvyšším správním soudem kvalifikován jako *veřejná informace*, a tedy podléhající povinnosti zveřejnění.[28] Dalším významným krokem vpřed je koncept tzv. AI rejstříků – „prostředek

[28] 'Algorithm of the System of Random Allocation of Cases Finally Disclosed!', *Foundation Moje Panstwo*, 22 September 2021, dostupné na https://mojepanstwo.pl/aktualnosci/773 (překlad autorů).

zajišťující transparentnost a veřejnou kontrolu AI [používaných státem]“[29]. Rejstřík algoritmů města Amsterdam (City of Amsterdam Algorithm Register)[30] a AI rejstřík města Helsinki (City of Helsinki AI Register)[31] patří mezi první veřejné rejstříky a nabízejí následující informace o AI systémech využívaných městem, jako jsou automatizované parkovací kontroly či chatboty zdravotnických center: účel a účinky, odpovědnost, datasety, zprocesování dat, nediskriminace, lidský dohled, rizika a opatření pro snižování rizik. Tyto prostředky posilují důvěru ve veřejné služby a přístup k informacím důležitým pro ochranu lidských práv, vč. práva na spravedlivý proces.

V každém případě však informace o stěžovateli vložené do AI systému musí uživatel systému stěžovateli poskytnout jako důkazní materiál ke kontrole – např. fotografie stěžovatele použitá při rozeznávání obličejů či údaj o bydlišti a jiné kontaktní údaje v rámci policejní predikce. I když nemusí představovat zdroj chyby vedoucí k porušení lidského práva, dostupnost takových dat musí být rovněž zajištěna v rámci výkonu práva na spravedlivý proces.

1.3.2. Nevysvětlitelná AI

Nevysvětlitelná AI má charakter černé skříňky (black box), čímž jsou míněny modely, které jsou příliš komplexní na to, aby byly interpretovatelné lidmi. Jedná se především o AI založenou na technice strojového učení, která staví na modelech s takto vysokou komplexitou. Příkladem tohoto typu modelu jsou umělé neuronové sítě. Jiné modely jako například rozhodovací stromy jsou více inspirovány sekvenčním způsobem lidského rozhodování, a proto jsou z podstaty snadněji vysvětlitelné. Modely založené na vyhledávání nejbližších sousedů, vycházejí z lidem blízké schopnosti hledat analogie (podobnosti k již viděným vzorkům dat). Grafické modely mohou názorným způsobem zobrazovat pravděpodobnosti pozorovaných událostí a vztahy mezi nimi. I když některé z uvedených typů modelů jsou z hlediska vysvětlitelnosti silnější, od jisté úrovně složitosti řešeného problému mají charakter black boxu všechny.

V praxi jde o to, že výrobce AI systému není s ohledem na black box schopen garantovat vysvětlení, jak systém došel k výběru informací, na nichž založil své rozhodnutí, což podkopává, resp. znemožňuje případnou obhajobu osoby, jíž se rozhodnutí týká.

Jak překonat tuto zdánlivě nepřekonatelnou výzvu? Jedna možnost je podpořit alternativní metody mitigace dopadů nevysvětlitelnosti (1.3.2.1.) a druhá je omezit použití black boxu (1.3.2.2.).

[29] M. Haataja, L. van de Fliert and P. Rautio, *Public AI Registers. Realising AI Transparency and Civic Participation in Government Use of AI*, Whitepaper (Saidot – Gemeente Amsterdam – Helsinki), September 2020, dostupné na https://algoritmeregister.amsterdam.nl/wp-content/uploads/White-Paper.pdf (překlad autorů).
[30] *City of Amsterdam Algorithm Register*, dostupné na https://algoritmeregister.amsterdam.nl/en/ai-register/.
[31] *City of Helsinki AI Register*, dostupné na https://ai.hel.fi/en/ai-register/.

1.3.2.1. Metody na snížení dopadů nevysvětlitelnosti

Snahy zmírnit dopady nevysvětlitelné AI se realizují ve dvou základních formách: vytvoření zjednodušeného zástupného modelu a *ex post* metody založené na dodatečném zkoumání vstupně-výstupních závislostí modelu. Podstatou z lidskoprávního hlediska je snížit riziko porušení práva na spravedlivý proces skrze posílený přístup k informacím.

Běžnou metodou přispívající k vysvětlení AI systému je natrénování zástupného modelu, který bude napodobovat model black boxu. Takový vysvětlující model představuje zjednodušenou verzi původního modelu. Doménový expert (např. lékař provádějící zdravotní diagnostiku či soudce v rámci predikce v soudním řízení) může použít zástupný model jako uživatelsky vstřícnou vizualizaci. Ačkoli umožní detekci a identifikaci hlavních příčin rozhodnutí AI a jeho zdrojových prvků, nevýhoda zástupného modelu spočívá v jeho nižší expresivitě a nutné ztrátě informací. Podpora probíhajících výzkumů vysvětlujících modelů (*XAI – eXplainable AI – surrogate models*) by měla být v centru pozornosti národních AI strategií s cílem posílit vysvětlitelnost AI. V oblasti XAI dnes ve světě probíhá rozsáhlý výzkum, přičemž velká část těchto aktivit se soustředí na vysvětlující metody pro modely neuronových sítí, které jsou aktuálně mezi nejmodernějšími aplikacemi strojového učení nejrozšířenější.[32]

Další způsob vysvětlení vnitřních procesů AI technologií spočívá v *ex post* srovnávací metodě založené na testování hypotetických situací. Vývojář AI systému může poskytnout testovací platformu umožňující třetí straně, či přesněji stěžovateli, prověřit vlastnosti AI. Systém může například sám poskytnout informaci o tom, jaká změna ve vstupních datech by vedla ke změně jeho výstupů (rozhodnutí, predikce atd.) v konkrétním případě. Například pokud finanční systém založený na AI odepře dotčené osobě půjčku, *ex post* srovnávací metoda může indikovat práh, od kterého by půjčka byla poskytnuta, např. požadovaný minimální plat. Jiným příkladem může být *ex post* srovnávací metoda, která poukáže na to, že chybný AI systém predikující pravděpodobnost recidivy v trestním soudnictví generuje rozhodnutí založené na atributech, které dotčená osoba nemůže ovlivnit (např. místo narození), čímž navíc odhalí nepřímou diskriminaci. Pokud takové testování provádí přímo výrobce systému, objektivita nemusí být zajištěna s ohledem na vysoce expertní a výlučnou technickou znalost daného AI systému.

1.3.2.2. Omezení použití black boxu

Další možností je omezit či přímo zakázat použití systémů založených na black boxu, a to alespoň v kritických AI aplikacích, a přesunout se k tzv. white box algoritmům. Jedná se o AI systémy, které jsou vysvětlitelné, avšak ve většině případů neschopné řešit komplexní úlohy jako black box systémy z důvodu neexistence algoritmického popisu úlohy. Především pro analýzu velkých dat, jako jsou např. databáze s fotografiemi ve vysokém rozlišení či textové

[32] XAI je značně pokryto významnými AI konferencemi jako Conference and Workshop on Neural Information Processing Systems (NeurIPS), International Conference on Learning Representations (ICLR) a AAAI Conference on Artificial Intelligence (AAAI).

databáze, bude black box systém pravděpodobně nezbytný. White box modely jsou převážně aplikovány ve zdravotnictví, letectví a dalších životně důležitých a kritických oblastech. Jedná se typicky o úlohy, jejichž rozhodovací modely lze například dobře vizualizovat.[33]

Podobně jako u existujících kategorií nasazení do životně důležitých („life-critical“) oblastí (jako je zdravotnická přístrojová technika, farmaceutické systému automatizace přípravy léčiv a další) a do kritických operací („mission-critical“) v oblastech jako letectví, které jsou podrobeny zvýšeným požadavkům v procesu schvalování před uvedením do provozu, může být představena myšlenka AI systémů kritických z hlediska lidských práv („human rights-critical“). „Human rights-critical“ AI technologie se vyznačují vysokým rizikem porušení lidského práva s vážným dopadem na člověka a musely by vždy být certifikované. Jejich množinu bude třeba náležitě vymezit, přičemž lze předpokládat, že zahrne např. predikci v soudnictví, která se týká zbavení svobody jednotlivce, či robotickou chirurgii a autonomní zbraně, jež by mohly mít závažný dopad na právo na život.[34]

Vysvětlitelností (*explainability*) je myšlena algoritmická vysvětlitelnost, tj. objasnění způsobu vedoucího k rozhodnutí AI formou pochopitelnou člověku. Toto je nezbytné např. v řadě „life-critical“ a „mission-critical“ systémů, které vyžadují plnou kontrolu nad všemi parametry a procesy vedoucími k výstupům systému během jeho vývoje. Patří sem například aplikace pro kardiostimulátory, systémy řízení letadel, zabezpečení provozu železnic a další. Vysvětlitelnost je třeba odlišit od transparentnosti (*transparency*), přesněji operační transparentnosti. Operační transparentnost poskytuje informaci, na které je rozhodnutí založeno, tj. která vstupní data a s jakou vahou byla vzata v potaz.

Tyto pojmy je třeba dále odlišit od věrohodnosti (*confidence*) a přesnosti (*accuracy*), které mají jiný účel. Obě hlediska vyjadřují odhad exaktnosti AI technologie: zatímco se přesnost týká kvality AI systému a určuje např. odhadovaný procentuální poměr správných predikcí[35], věrohodnost vyjadřuje subjektivní ohodnocení výsledku samotnou AI technologií. Příkladem věrohodnosti je oblast obrazového rozpoznávání, kdy AI technologie uvede, že prohlížený obrázek je ze 60 % pes a 40 % kočka, a poskytne tak vlastní hodnocení exaktnosti svého výstupu.

Je třeba poukázat na to, že nevysvětlitelnost nepředstavuje vždy překážku pro fungování AI systému z pohledu ochrany lidských práv. V praxi je relevantnost výše uvedených požadavků ovlivněna mírou autonomie systému, typem uživatele a jeho rozhodovací diskrecí. Jinými slovy, pokud je za provozu AI přítomen lidský dohled, požadavky budou jiné, než když je

[33] Příkladem jsou rozhodovací stromy, metoda nejbližších sousedů, lineární logistická regrese, naivní Bayes, případně problémově specifické parametrizované modely.

[34] Je evidentní, že některé „human rights-critical“ AI se mohou překrývat s existujícími systémy nasazenými do životně důležitých („life-critical“) oblastí, které mají především dopad na právo na život či právo na zdraví, nicméně každá kategorie je definována jiným kritériem a navíc současné „life-critical“ systémy ještě neoperují na bázi AI. Z toho vyplývá, že „life-critical“ systémy založené na AI budou spadat do “human rights-critical“ AI, neboť obsahují riziko porušení základních lidských práv.

[35] „Přesnost klasifikace“ říká, v kolika procentech případů predikoval model na testovacích datech správnou výstupní třídu, a je odhadem toho, jak se bude AI technologie chovat v reálném provozu.

systém plně autonomní. Na jedné straně škály AI systém ponechává finální rozhodnutí na uživateli – jedná se o tzv. asistivní systémy. Člověk primárně vykonává příslušnou činnost a AI systém mu předzpracovává vstupní informace, urychluje jeho rozhodování a doporučuje možná řešení. Člověk sám přijímá finální řešení a vykonává případnou kontrolu, revizi a opravu výstupů AI systému. Zvláštním typem asistivních technologií je pak rozšíření lidského pochopení situace prostřednictvím tzv. augmentované reality či inteligence (*augmented intelligence*). V tomto případě je podpořeno lidské rozhodování tím, že se běžně dostupné informace (např. obraz z videokamery) rozšíří o syntetické informace, které dodává AI systém, což může vést ke kvalitativně lepším nebo novým závěrům. Augmentovaná inteligence se tedy vyznačuje rozšířením vnímané situace, problému, úlohy tak, aby mohl člověk dospět ke správnému řešení. K řešení nedospívá AI systém, ale člověk, což je hlavní rozdíl od předchozích asistivních systémů, které se snaží automatizovat konkrétní proces a řešit úlohu autonomně.

Uživateli využívajícímu pomoc asistivní AI nebude nikterak ku prospěchu algoritmická vysvětlitelnost, nýbrž transparentnost rozhodnutí AI, která umožní prověřit správnost výstupů AI systému. Znalost vstupních informací a jejich váhy, se kterou přispěly k výstupům AI systému, poslouží uživateli k ověření spolehlivosti AI systému před přijetím finálního rozhodnutí. Takový postup bude zapotřebí především v expertní oblasti vyžadující profesní kvalifikaci jako např. ve zdravotnictví, v právní sféře či v systémech kybernetické bezpečnosti. Například soudci odsuzujícímu obviněného k trestu odnětí svobody, kterému asistuje AI systém ohledně stupně pravděpodobnosti recidivy, algoritmická vysvětlitelnost příliš neposlouží. Co ovšem bude třeba, je transparentnost AI rozhodnutí, tj. na základě jakých vstupních dat a kritérií je dané rozhodnutí založeno. Algoritmická vysvětlitelnost je důležitá na úrovni vývoje systému, ale nepřináší žádnou doplňkovou informaci pro uživatele pro přijetí vlastního rozhodnutí, a to ani v případě algoritmicky vysvětlitelných modelů. Algoritmická vysvětlitelnost dává jistotu, že systém bude fungovat v rámci předem určených provozních podmínek vždy správně. Při používání systému nebude uživatel studovat tyto software algoritmy, ale i když bude systém algoritmicky vysvětlitelný, bude v řadě případů vyžadovat podpůrné informace o tom, proč systém k danému rozhodnutí dospěl, na základě jakých vstupních informací a s jakou vahou, resp. jejich kombinaci. Naopak u aplikací používaných širokou veřejností jako prediktivní psaní textů na chytrém telefonu není potřeba ani vysvětlitelnost ani transparentnost, správnost výsledku je možné odvodit přímo z doporučovaného textu. Na praktické úrovni, jestliže je pro porozumění algoritmické vysvětlitelnosti potřeba vysoce expertní technická znalost, pro využití transparentnosti AI systému však bude rovněž nutná určitá odbornost, minimálně zaškolení uživatele, jak správně chápat a vyhodnocovat informace a výstupy systému.

Naopak na druhé straně spektra leží systémy, které jsou nasazovány jako plně autonomní a kde lidský prvek nemá roli finálního posuzovatele výstupů AI, tedy nedisponuje žádnou rozhodovací diskrecí. Tato nejpokročilejší forma AI umožňuje strojům, botům a systémům jednat samostatně a nezávisle na lidské intervenci. Ačkoli počet autonomních systémů

stoupá, předávání plné kontroly strojům zůstává prozatím otázkou. Kromě problematiky odpovědnosti se tento režim nehodí pro celou řadu použití a situací, zejména tam, kde je těžké zaručit minimální kvalitu výsledků za všech přípustných scénářů, tedy vstupních dat. S rostoucím stupněm autonomie systému uživatel pouze přejímá výsledky a není zapojen do rozhodovacího procesu, vč. případného přehodnocování výstupů AI systému. Tím se tedy snižuje požadavek na *a priori* operační transparentnost, přičemž klíčovým pro uživatele je zajištění vysoké míry přesnosti AI systému. Uživatel potřebuje autonomní AI technologii, která je přesná a spolehlivá, a požadavkem na systém by tedy měl být vysoký stupeň přesnosti blížící se 100 % a důkladné testování systému ve všech přípustných operačních scénářích. Operační transparentnost však bude potřebná *ex post* pro případný výkon práva na spravedlivý proces.

Závěrem lze shrnout, že výše uvedené informace o AI technologii a jejich zpřístupnění stěžovateli jsou nutným předpokladem pro skutečné požívání práva na spravedlivý proces. Vysvětlitelnost, transparentnost a přesnost představují v tomto ohledu zásadní faktory. Ačkoli algoritmická vysvětlitelnost nebude mít v praxi pro uživatele příliš informační hodnotu při používání systému, kritérium vysvětlitelnosti může posloužit jako požadavek na certifikaci AI systémů kritických z hlediska lidských práv („human rights-critical“). Tímto požadavkem se může podmínit jejich nasazení a v důsledku tím eliminovat black box AI. Kritérium transparentnosti může být naopak požadováno pro všechny AI technologie, a to v různých scénářích: odborníkovi, který používá technologii jako asistivní, poslouží transparentnost přímo v procesu rozhodování, kdy využije dodatečné informace k ověření validity výsledků AI systému; v případě autonomní AI technologie pak transparentnost pomůže pro případné *ex post* přezkoumání výsledků, pokud bude požadováno.

Transparentnost rozhodnutí AI však neřeší existenci black boxu. Transparentnost pouze částečně napomáhá odkrýt a objasnit informace o vnitřních procesech AI systému, a tím posílit právo na spravedlivý proces. S cílem plně zajistit právo na spravedlivý proces je naopak panaceou, všelékem, vysvětlitelnost jakožto podmínka certifikace všech „human rights-critical“ AI? A která AI technologie není „human rights-critical“? Tyto otázky zůstávají otevřené k diskusi.

2. Varieta rizik porušení jednotlivých lidských práv AI technologiemi a možnost nápravy

Kromě základních lidských práv a svobod, které jsou ohroženy napříč všemi AI technologiemi (Část 1.), jednotlivá lidská práva mohou být porušena s ohledem na specifickou oblast nasazení AI technologie. Vysokoškolské přijímací řízení založené na AI může mít dopad na právo na vzdělání, AI systém přidělující sociální dávky může porušit právo na sociální zabezpečení a AI používaná při moderování diskusí na internetu může narušit svobodu projevu. Tato různorodá rizika porušení lidských práv mají společné příčiny, jejichž identifikace je nutným předpokladem pro efektivní nápravu.

Dopad AI na lidská práva lze schematizovat do čtyř vektorů: za prvé, data obsahující bias způsobující diskriminaci navíc sekundárně poruší lidské právo v dané oblasti nasazení AI (2.1.); za druhé, pokud nasazení AI systému neodpovídá původně specifikovanému a plánovanému provoznímu prostředí nebo došlo ke změně prostředí či je systém provozován v okrajových podmínkách, může být ohroženo příslušné lidské právo v oblasti nasazení (2.2.); za třetí, AI technologie může porušit jednotlivá lidská práva, pokud je nasazena a používána pro ilegální účely (2.3.); a za čtvrté, AI technologie může porušit lidské právo, pokud je napadena a zneužita (2.4.).

2.1. Sekundární porušení lidských práv z důvodu biasu v datech

Pokud vstupní data, či vybrané atributy, obsahují bias, způsobené porušení zákazu diskriminace následně může vést i k porušení příslušného lidského práva v oblasti nasazení AI. Příkladem je vstupní rám na letišti s mechanismem rozpoznávání obličejů, který chybně nerozpozná obličejové rysy příslušníků určitého etnika a odepře jim tak vstup do letadla, což vede nejen k porušení zákazu diskriminace, ale i svobody pohybu.[36] Podobně použití dat obsahujících bias v rámci predikce v trestním soudnictví operující na bázi pravděpodobnosti recidivy[37] může způsobit nejen ilegální diskriminaci na základě rasy, pohlaví či náboženství, ale sekundárně i porušení práva na svobodu, práva na spravedlivý proces či práv dítěte[38]. Právo na přiměřenou životní úroveň může být porušeno AI systémem obsahujícím bias, který určuje úvěruschopnost jednotlivců, pokud je využit v oblasti služeb týkajících se bydlení – „majitelé bytů mohou odmítnout pronajmout byt, banky mohou zamítnout žádost o kreditní kartu a operátoři sítí odmítnou novou smlouvu“[39].[40] V tomto scénáři tedy dochází k

[36] Technologie rozpoznávání obličejů obsahující bias použitá např. policií v Argentině v r. 2019 vedla k opakovanému protiprávnímu zadržení téže osoby, a to důvodu váhání uživatelů AI systému jeho výstupy přezkoumat a neřídit se jimi, European Digital Rights (EDRi), ‘Dangerous by Design: A Cautionary Tale About Facial Recognition’, 12 February 2020, dostupné na https://edri.org/our-work/dangerous-by-design-a-cautionary-tale-about-facial-recognition/.

[37] Viz např. systém COMPAS používaný v některých státech USA za účelem predikce rizika recidivy, zhodnocený v ProPublica, ‘Machine Bias: There’s Software Used Across the Country to Predict Future Criminals’, 1 October 2020, dostupné na https://pace.coe.int/en/files/28723/html, section 2.3. Podobně Harm Assessment Risk Tool (HART), systém používaný policií v britském Durhamu a založený na řadě proměnných týkajících se záznamů v rejstříku trestů a společensko-demografického prostředí podezřelé osoby, jako jsou věk, pohlaví a geografická zóna. Používání nového datasetu zvaného „Mosaic“, který výslovně definoval kategorie jako „Asian heritage“ či „disconnected youth“, bylo zastaveno v r. 2018, viz M. Oswald *et als.*, ‘Algorithmic Risk Assessment Policing Models: Lessons from the Durham HART Model and ‘Experimental’ Proportionality’, 27 *Information & Communications Technology Law* (2018) 223; a Rada Evropy, *ibid.*, section 2.2.

[38] Viz podobně např. systém ProKid 12-SI používaný od r. 2009 policií v Nizozemí, který hodnotí riziko kriminality 12letých dětí, K. La Fors-Owczynik, ‘Profiling ‘Anomalies’ and the Anomalies of Profiling: Digitalized Risk Assessments of Dutch Youth and the New European Data Protection Regime’, in S. Adams, N. Purtova, and R. Leenes, *Under Observation: The Interplay Between eHealth and Surveillance* (2017) 107.

[39] Komentář k SCHUFA systému používanému v Německu v European Digital Rights (EDRi), ‘Use cases: Impermissible AI and Fundamental Rights Breaches’, August 2020, dostupné na https://edri.org/wp-content/uploads/2021/06/Case-studies-Impermissible-AI-biometrics-September-2020.pdf, s. 7.

[40] Dále např. Tina Cheuk, ‘Can AI be Racist? Color-evasiveness in the Application of Machine Learning to Science Assessments’ (2021) 105 Science Education 825; a David Leslie *et als.*, ‘Does “AI” Stand for Augmenting

multiplikaci porušení lidských práv. Primární porušení spočívá v porušení zákazu diskriminace a sekundární porušení se týká lidského práva v dané oblasti nasazení AI.

Výchozí bod pro řešení této multiplikace porušení lidských práv je otázka stupně autonomie AI systému, a tedy rozsah účasti lidského faktoru a výsledná právní odpovědnost za porušení lidských práv. Jinými slovy vztah mezi AI technologií a uživatelem. Pokud AI technologie chybně kvalifikuje jednotlivce jako osobu obviněnou ze spáchání trestného činu, na kterou byl vydán zatykač, a policie adekvátně zakročí a zatkne jej, daný uživatel se účastní sekundárního porušení lidského práva (práva na svobodu). Je účast uživatele (letiště) stejná v situaci, kdy vstupní rám na letišti fungující na bázi AI systému rozpoznávání obličejů chybně zablokuje pasažéra a neumožní mu nástup do letadla? Co plně autonomní vozidlo operující s daty obsahujícími bias a vykonávající chybné operační kroky, které způsobí nehodu s vážnými materiálními škodami, ublížením na zdraví či ztrátami na životech? Je prvek diskrece a lidského faktoru zde přítomen?

Zahrnutí lidského prvku do sféry nasazení AI technologie může jistě snížit riziko vzniku sekundárního porušení lidského práva za podmínky, že lidská diskrece bude použita efektivně a rozhodnutí AI obsahující bias bude přezkoumáno a napraveno.[41] I když nelze reálně očekávat stagnaci vývoje AI na úrovni asistivních technologií, přítomnost lidského prvku lze definovat jako nutnou podmínku pro některé AI systémy kritické z hlediska lidských práv, jako je soudnictví, zdravotní péče apod. Přezkum výstupů asistivních technologií však vyžaduje určité záruky ve smyslu informovanosti a expertízy v doménové oblasti, aby mohl uživatel účinně a reálně zamezit sekundárnímu porušení lidských práv. Transparentnost rozhodnutí AI, která zajistí uživateli potřebné informace, tak představuje v tomto ohledu rovněž významný nástroj.

Klíč k zamezení porušení lidských práv v jakékoli oblasti používání AI z důvodu dat obsahujících bias, či nevyvážených dat obecně, spočívá v definici uživatelských požadavků a v obecných kontrolních mechanismech obsažených ve vývoji AI i v systému AI samotném.

2.2. Nasazení AI systému v jiném než v cílovém provozním prostředí

Lidská práva mohou být porušena, pokud je AI technologie nasazena v prostředí jiném, než pro které byla původně specifikována, nebo pokud se změní podmínky, za kterých je provozována, a ty již neodpovídají původním předpokladům. Případně se systém může dostat do stavu, kdy je provozován na hranici svých možností, například v okrajových situacích, ve kterých nebyl testován.

Inequality in the Era of Covid-19 Healthcare?' (2021) 372 (304) BMJ <https://www.bmj.com/content/bmj/372/bmj.n304.full.pdf> accessed 17 June 2022.

[41] Viz např. výzkum popsán v Zana Buçinca, Maja Malaya and Krzysztof Gajos, 'To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making' (2021) 5 Proceedings of the ACM on Human-Computer Interaction 1.

V prvním případě byla AI technologie naučena pro konkrétní geografický, kulturní, sociální či jiný kontext, ale je nasazena v jiném prostředí. Příkladem je AI systém hodnocení ve školách, který byl připraven na datech z většiny škol v zemi nepodporující inkluzivní vzdělávání a který byl však použit ve vzdělávacích institucích pro studenty se zvláštními potřebami. Výsledky mohou být diskriminující a právo na vzdělání může být dotčeno. Stejně tak, pokud bylo autonomní vozidlo vyvinuto pro automobilový trh v USA, bude autonomní řízení vykazovat významné nedostatky v případě, že bude nasazeno v zemi s arabským písmem či s odlišným dopravním značením, jako je v Asii či v Evropě.[42] Odlišné geografické nasazení může ohrozit řadu lidských práv, jako je právo na život či právo na zdraví. Naopak v druhém případě došlo ve stávajícím prostředí k jeho změně či rozšíření, které nebylo v původním návrhu systému zahrnuto. Příkladem může být vydání nové dopravní značky ve stávajícím regionu.

Co se týče problematiky okrajových vlastností, zde AI technologie funguje správně, avšak některá vývojová data chybí nebo nebyl systém na tyto vlastnosti testován, v důsledku čehož systém nepodává přesné výsledky. V tomto případě je autonomní vozidlo trénováno pro obvyklé počasí, avšak s extrémním větrem či se silnicemi pod neobvyklou extrémní sněhovou pokrývkou se nepočítalo. V těchto okrajových podmínkách, tj. hraničních, může nesprávné fungování AI technologie opět ohrozit lidský život či způsobit jinou újmu.

Jaké je řešení těchto nepřepokládaných situací vedoucích k porušení lidských práv? Podobně jako u prevence sekundárního porušení lidských práv plynoucího z nevyvážených dat (viz Sekce 2.1.) je klíčem dialog mezi výrobcem a zákazníkem. Kontext nasazení i okrajové vlastnosti AI systému by měly být jasně stanoveny v technických specifikacích, např. že autonomní vozidlo je určené pro dopravní nasazení v konkrétních zemích či regionech. Kromě toho je třeba kontext nasazení i okrajové vlastnosti verifikovat během procesu schvalování AI systému do provozu.

Pokud je kontext stanoven ve specifikacích, bude vzat v potaz a verifikován během fáze testování, kdy se AI systém prověřuje ve vztahu k výchozím požadavkům. Okrajové vlastnosti však zpravidla budou, jakožto nepředvídané skutečnosti, detekovány až během provozu AI technologie. Jinými slovy, až *ex post*, když problém mající dopad na lidské právo vznikne. Například AI systém pro stanovení dávek sociálního zabezpečení nepočítal s hraniční krizovou událostí, jako jsou přírodní katastrofy, a navrhuje udělování dávek běžných jako za normální situace. Ačkoli se může zdát, že AI technologie funguje správně, opomenutí správně ošetřit okrajové vlastnosti se projeví až v rozhodnutích za provozu systému. V takovém případě ani monitoring systému nedokáže předejít porušení lidského práva. Monitoring však zůstává zásadní pro detekování chyby, přijetí příslušné nápravy a pro zabránění opakování takového porušení lidského práva.

[42] Viz např. A. Sikander and H. Ali, 'Image Classification Using CNN for Traffic Signs in Pakistan', 2021, dostupné na http://arxiv.org/abs/2102.10130.

2.3. Použití AI technologie za nelegálním účelem

AI technologie může mít negativní dopad na lidská práva, pokud je vyvíjena a používána pro nelegální účely. Protiprávní monitorování obyvatel porušující jejich svobodu pohybu či právo na soukromí je již využíváno některými státy.[43] Tzv. zabíjející drony a jiné autonomní letální zbraně ohrožují lidské životy.[44] Prolomování hesel může být použito pro ilegální účely. Deep fakes produkované AI technologií mohou být nástrojem protiprávní manipulace vůči uživateli, např. skrze manipulaci s fotografiemi a obrázky či generování clickbaitových titulků.

Samozřejmě je možné přijmout příslušnou legislativu zakazující vývoj určitých AI technologií nebo jejich používání v konkrétní oblasti (např. manipulativní techniky, které podvědomě ovlivňují chování člověka a způsobují mu újmu, nebo sociální kreditní systém vedoucí k diskriminaci)[45] či stanovící pro vývoj a provoz AI právní restrikce, vč. povinné licence ve vybraných oblastech (na způsob zbrojní licence či licence pro provozování nebezpečných činností apod.). Výjimky by platily pro zvláštní subjekty, jako jsou státní bezpečnostní složky aj. (např. používání biometrických nástrojů na veřejných místech omezeno pouze pro potřeby donucovacích orgánů a za stanovených podmínek).

Na praktické úrovni vyvstává otázka, zda existuje technické opatření, které by účinně zabránilo protiprávnímu použití AI technologie, a v důsledku tak zamezilo porušení lidského práva. Jakési technické „tlačítko“, které by zastavilo či vypnulo AI technologii v případě jejího použití pro jiné účely, než pro které byl systém určen.

Podobně jako legislativní oblast i technická sféra nabízí standardní IT procesy, přičemž AI v tomto ohledu nevykazuje žádné zvláštnosti. Zajištění může být dvojí: za prvé, dodavatel prověří, komu je systém dodán, a za druhé jak dodavatel, tak zákazník mají možnost vypnout či omezit provoz AI technologie. První možnost je realizována běžným způsobem prostřednictvím licencí, které jsou udělovány při zahájení používání systému (vzdálené aktivace, bezpečnostní klíče, kontrola předplatitele a další). Například AI systém, který pracuje se zdravotnickými záznamy bude aktivován pouze pro ověřeného licencovaného uživatele,

[43] Přehled nasazení nástrojů AI dohledu ze strany států za účelem monitorování, sledování a kontroly obyvatel viz *AI Global Surveillance (AIGS) Index* amerického think tanku Carnegie Endowment for International Peace a S. Feldstein, ‘The Global Expansion of AI Surveillance’, Working Paper, Carnegie Endowment for International Peace, 2019, dostupné na https://carnegieendowment.org/2019/09/17/global-expansion-of-ai-surveillance-pub-79847.

[44] Viz např. Final report of the Panel of Experts on Libya established pursuant to Security Council resolution 1973 (2011), UN Doc. S/2021/229, 8 March 2021, para. 63. Pro právní zhodnocení používání autonomních letálních zbraní viz např. Elliot Winter, ‘The Compatibility of Autonomous Weapons with the Principles of International Humanitarian Law’ (2022) 27 Journal of Conflict and Security Law 1; Jai Galliott, Duncan MacIntosh and Jens David Ohlin (eds), *Lethal Autonomous Weapons : Re-examining the Law and Ethics of Robotic Warfare* (OUP 2021); a Andrea Spagnolo, ‘What do Human Rights Really Say About the Use of Autonomous Weapons Systems for Law Enforcement Purposes?’ in Elena Carpanelli and Nicole Lazzerini (eds), *Use and Misuse of New Technologies: Contemporary Challenges in International and European Law* (Springer 2019).

[45] Viz např. „Zakázané postupy v oblasti umělé inteligence“ stanovené v Čl. 5 Hlavy II Návrhu Nařízení Evropského parlamentu a Rady, kterým se stanoví harmonizovaná pravidla pro umělou inteligenci (Akt o umělé inteligenci) a mění určité legislativní akty Unie, COM/2021/206 final, Brussels, 21.4.2021, 2021/0106(COD). Ačkoli u předmětných postupů není výslovně uvedený link na lidská práva, všechny uvedené aktivity spočívají v porušení lidských práv.

kterým je konkrétní nemocnice či lékař, s cílem ochránit právo na soukromí. Jiným příkladem je monitorování IP adres. Druhou možností je paralyzující zásah do provozu AI systému, který může být vykonán buď na dálku v případě detekce problému či je mechanismus omezení provozu systému integrován přímo do AI systému. Například kamerový sledovací systém si musí pravidelně obnovovat svojí aktivaci na serveru provozovatele, aby mohl pokračovat v rozpoznávání lidí v rámci stanovených podmínek. Geolokační omezení mohou být zavedena výrobcem i na *ad hoc* bázi. Pokud se zjistí, že některý stát používá AI technologii protiprávně a porušuje tím lidská práva, např. za účelem sociálního kreditování, výrobce může danému státu zakázat přístup. AI nástroje s integrovaným provozním kontrolním mechanismem ovládaným na dálku, vč. aktivace a deaktivace, jsou typicky používané např. zpravodajskými službami.

Zavedení takovýchto kontrolních záruk je v zájmu různých aktérů životního cyklu AI i státu, přičemž dané omezení je stanoveno v požadavcích na systém.

V neposlední řadě je třeba zmínit obavy veřejnosti ze zlovolného použití AI vycházející často z „futuristického" a „sci-fi" pojetí AI, která překoná lidskou kontrolu a obrátí se proti svým lidským uživatelům.[46] Je opět důležité zdůraznit, že současný stav AI technologií, a podle všeho i AI blízké budoucnosti, nepředstavuje žádné zvláštní riziko zlovolného užití, které by se lišilo od rizik patrných v jakýchkoli jiných IT systémech. Žádná AI by nepoužila jiné mechanismy, než které by použil člověk ke zmanipulování AI operace, ačkoli AI může být rychlejší. Možnou ochrannou zárukou v IT systémech je dnes omezení či zastavení jejich provozu nezávislým systémem integrovaným do těchto mechanismů. Příkladem je kritická infrastruktura, konkrétně jaderné elektrárny, kde bezpečnostní systém funguje nezávisle a je zajištěn na vícero úrovních, což ztěžuje jeho případné zmanipulování. Stejné postupy lze aplikovat na AI technologie.

Zvláštní rizika týkající se AI vycházejí pouze ze skutečnosti, že jsou nasazovány do nových oblastí či automatizují komplexní úlohy, jako jsou sociální sítě, kamerové sledovací systémy, predikce v soudnictví, média a další. Rizika spočívají v nedostatečně ověřených scénářích používání, nedostatečném testování systému ve všech možných situacích, komplexitě interakcí systému s lidským uživatelem, neexistenci standardů či certifikačních postupů a převodu lidem srozumitelných definic těchto činností do technických specifikací, včetně například lidskoprávní terminologie. Ačkoli technické nástroje pro eliminaci rizik existují, jedná se o delší a iterativní proces.

[46] Viz např. L.B. Eliot, 'Using a Kill-Switch or Red Stop Button for AI is a Dicey Proposition, Including for Self-Driving Cars', *Forbes*, Oct 21, 2020, dostupné na https://www.forbes.com/sites/lanceeliot/2020/10/21/using-a-kill-switch-or-red-stop-button-for-ai-is-a-dicey-proposition-including-for-self-driving-cars/; a S. Russell, 'How to Stop Superhuman A.I. Before It Stops Us', *The New York Times*, 8 October 2019, dostupné na https://www.nytimes.com/2019/10/08/opinion/artificial-intelligence.html.

2.4. Zneužití AI technologie

Čtvrtý základní scénář, kdy může dojít k porušení lidského práva používáním AI technologie, spočívá v jejím zneužití. AI technologie, která je vyvinuta v souladu s právem a nasazena za legálním účelem, může být napadena a zneužita. Příkladem je IoT zařízení jako hlasem ovládané nákupní systémy a chytré domácnosti, které mohou být ovládnuty útočníkem a sloužit jako odposlouchávací zařízení porušující právo na respektování soukromého života. AI systém pro výběr uchazečů o zaměstnání může být digitálně zfalšován a znevýhodňovat tak určité skupiny obyvatelstva, což by zasáhlo do jejich práva na práci, či naopak zmanipulovaně prosazovat nevhodné kandidáty do státní služby. Takový potenciál zneužití se týká téměř všech oblastí provozu AI technologií.

Tyto případy spadají do standardní oblasti trestního práva a kybernetické bezpečnosti a nejsou zde dále rozpracovány.

3. Hodnocení rizik z pohledu lidských práv

Ačkoli jsou rizika porušení lidských práv AI technologiemi různorodá a mohou vyvstat v jakékoli fázi životního cyklu AI, společnou podmínkou jejich prevence a eliminace je zavedení mechanismu hodnocení rizik z pohledu lidských práv (tzv. *human rights risk assessment – HRRA*). Mechanismus by se měl stát integrální součástí životního cyklu AI s cílem zajistit soulad AI technologie s lidskými právy. Jedná se o zavedení lidskoprávního prvku do koncepčního rámce a obecných kontrolních mechanismů AI systému.

Na praktické úrovni jde o vymezení lidskoprávního rámce již v prvotní fázi životního cyklu AI, tj. v rámci analýzy požadavků. Zde se sbírají informace o aplikační sféře AI technologie, jinými slovy se mapuje, „co business potřebuje“. Všichni aktéři životního cyklu AI se mohou této úvodní části účastnit. Konkrétně se jedná o zavedení lidskoprávního prvku do třech základních nástrojů vymezujících koncept AI systému: analýza vstupů, která uvádí dostupná data a procesy používané v cílové doméně; uživatelské požadavky, které definují např. případy použití (služby, které bude AI systém poskytovat), detaily nasazení, specifikace běhového prostředí (např. hardware, další infrastruktura, personál nutný pro správu systému) a parametry dlouhodobé udržitelnosti (např. rozšiřitelnost); a specifikace popisující aplikační doménu, požadavky na vlastnosti AI technologie, architekturu, technické provedení a požadavky pro nasazení do provozu. Hodnocení rizik z pohledu lidských práv by mělo být představeno v rámci vstupní analýzy a dále rozpracováno v uživatelských požadavcích, což se následně promítne do specifikací AI systému. Toto zajistí jeho zahrnutí do celého cyklu AI, tedy že AI systém bude vyvinut, testován a monitorován s ohledem na lidskoprávní vymezení.

Požadavky definují různí aktéři životního cyklu AI v závislosti na stupni kritičnosti domény AI systému a jeho provozu. Čím kritičtější oblast, tím jsou stanoveny vyšší parametry. Stát jako regulátor by měl jistě vstoupit do sféry „human-rights critical“ AI, podobně jako tomu je dnes v případě existujících „life-critical“ a „mission-critical“ systémů. Naopak nejnižší úroveň je

zajištěna výrobcem, který pouze zapracovává požadavky uživatelů, vč. kontroly *ex post* díky zpětné vazbě. Toto se týká AI systémů, které nemají dopad na lidská práva či kde je riziko zanedbatelné, a tudíž preventivní požadavky nejsou specifikovány. Příkladem dopadu s minimální škodou může být zhoršení odezvy aplikace na mobilním telefonu, kdy skupina uživatelů dostane správný výsledek, ale s určitou prodlevou, která pouze zhorší uživatelský komfort při zachování použitelnosti systému. Naopak regulátor může pro „human rights-critical" AI stanovit nejpřísnější požadavky, kombinující prvky monitoringu, transparentnosti a vysvětlitelnosti.

Řada AI systémů bude provozována na základě podmínek stanovených pouze zákazníkem, což klade zvýšený důraz na tyto aktéry životního cyklu AI a jejich znalosti lidskoprávní perspektivy. Zákazník definuje potřeby s ohledem na efektivní fungování AI systému pro stanovený účel, přičemž musí vzít v potaz lidskoprávní aspekt. Dialog mezi výrobcem, zákazníkem a případným regulátorem je tak v tomto ohledu klíčový. Např. soudce bude potřebovat znát míru přesnosti AI systému pro spolehnutí se čistě na AI predikci a měl by mít k dispozici veškeré potřebné informace pro její přezkum. Podobně oddělení osobních půjček v bance využívající AI technologii pro hodnocení finanční solventnosti žadatele by mělo vzít v potaz přesnost a transparentnost AI systému v rámci svého celkového posouzení konkrétního případu, aby zamezilo porušení lidského práva. Je třeba definovat úroveň přesnosti AI technologie, aby uživatel získal reálnou představu o spolehlivosti rozhodnutí AI a případně nutnosti jeho přezkumu. Například v oblasti zdravotní přístrojové techniky je dnes standardně požadováno, aby systém ve všech definovaných situacích pracoval se 100% úspěšností. Naopak pro systémy rozpoznávání v dopravě, predikce trhu či doporučovací systémy je často požadován vysoký stupeň přesnosti, avšak ne nutně 100 %. Například používání softwaru pro rozpoznáváni obrazu k identifikaci poznávacích značek může být nastaveno na nižší stupeň, neboť neohrožuje žádný kritický zájem či chráněnou hodnotu. Výběr závisí na kritičnosti AI technologie, oblasti nasazení, ceně a času pro vývoj AI systému.

Dnes není prozatím běžné, aby aktéři vývojové a provozní fáze AI systému takové požadavky související s lidskoprávní problematikou definovali. Důraz by měl být kladen na zvýšení veřejného povědomí a informovanosti v této oblasti. Navíc ačkoli je očekávatelné, že soudní instituce či pracovní agentury budou schopny definovat uživatelské požadavky (např. požadovaný stupeň přesnosti či transparentnost), častým problémem bude jejich následný převod na měřitelné a testovatelné specifikace a jiné technické požadavky. Pokud nebudou takové požadavky existovat na úrovni regulátora, může být praktické zavést určitou formu AI poradenství či AI konzultačního orgánu, který by pomohl definovat příslušná kritéria pro hodnocení rizik z pohledu lidských práv. Vysoká technická komplexita AI technologií vyžaduje, aby byla expertní AI asistence k dispozici ve veřejné i business sféře, podobně jako tomu je dnes s již běžně dostupnými IT službami.

Dodržování požadavků a specifikací je kontrolováno během procesu schvalování AI systému do provozu, což slouží jako další pojistka proti porušení lidských práv. Jedná se o standardní proces schválení IT systému do provozu, přičemž se berou v potaz i požadavky specifické pro

AI. Schválení systému do provozu se může realizovat na vícero úrovních s rostoucím závazným účinkem: schválení výrobcem, schválení dodavatelem, schválení zákazníkem, schválení nezávislým soukromým subjektem a, na nejvyšší úrovni, právně závazné schválení státní regulační autoritou. Čím vyšší stupeň kritičnosti AI systémů z hlediska lidských práv, tím by byla vyšší úroveň autority požadované pro schválení. Požadavky specifické pro AI zahrnují zavedení průběžného monitoringu nasazeného systému, jeho transparentnost a vysvětlitelnost. Další požadavky se mohou týkat např. výběru vývojových dat a možných omezení na kontinuální učení systému[47]. V závislosti na konkrétní AI technologii a podmínkách jejího užití mohou být tyto požadavky různě kombinovány. Standardní požadavky by měly být doplněny o hodnocení rizik s ohledem na lidská práva. Toto by zajistilo verifikaci a otestování na absenci rizikových prvků pro porušení lidských práv (např. neoprávněné přítomnosti tzv. chráněné hodnoty v atributech či nevyvážená data).

Další prostředky pro zabránění porušení lidských práv jsou běžné kontrolní mechanismy v rámci životního cyklu AI, jako je testování a monitoring AI technologie. Testování AI je standardní proces vývoje softwaru a je prováděn podle specifikací AI systému. Může účinně podchytit chybu či slabé místo s rizikem pro lidská práva. Testování se provádí během vývojové fáze, před nasazením a může být dále rozšířeno na monitorování systému za provozu. Testování zahrnuje verifikaci (kontrola, že specifikace jsou správně implementovány), validaci (systém funguje dle požadavků uživatele) a kvalifikaci (systém je v souladu s regulacemi, standardy a certifikacemi). Monitoring je významnou součástí provozní fáze AI, neboť může detekovat nesrovnalosti, chyby a nedostatky, které se objeví během nasazení a používání, vč. rizika souvisejícího s lidskými právy. Monitoring AI může být automatizovaný či prováděný člověkem, a to v časových intervalech specifických pro konkrétní oblast nasazení. V rámci monitoringu jsou posuzovány výstupy systému ve vztahu k „předpokládaným“ výstupům, které mohou být definovány na základě prověřených historických dat, metodami detekce anomálií či přímým lidským dohledem nebo mohou být pouze zálohovány pro pozdější vyhodnocení. Auditem je pak formální evaluace celého systému, vč. monitoringu a souladu systému s prostředím, ve kterém je nasazen. Např. New York City Council přijal v prosinci 2021 první právně závazné nařízení v USA zakazující zaměstnavatelům ve městě používat AI systémy k přijímání a hodnocení zaměstnanců, ledaže by tyto systémy byly podrobeny ročnímu auditu prokazujícímu, že nediskriminují na základě pohlaví či rasy.[48] Audit rovněž slouží jako významný nástroj pro průběžnou aktualizaci AI systémů a hodnocení rizik z pohledu lidských práv s ohledem na vývoj lidskoprávních hodnot a jejich aplikaci.

Vzhledem k tomu, jaká rizika pro lidská práva AI technologie představují, zavedení výše uvedených procesních a technických omezení je pro ochranu lidských práv nezbytné. Je však

[47] Průběžné přiučování modelů AI systému pomocí dat získaných za běhu.

[48] New York City Council, Automated employment decision tools, 11 December 2021, 2021/144. Viz např. Nicol Turner Lee and Samantha Lai, ‘Why New York City is Cracking Down on AI in Hiring’ (*Brookings, Techtank*, 20 December 2021) <https://www.brookings.edu/blog/techtank/2021/12/20/why-new-york-city-is-cracking-down-on-ai-in-hiring/> accessed 17 June 2022.

nutné si uvědomit, že AI technologie není 100% perfektní, stejně jako není lidské rozhodování. Podstata nespočívá v zajištění jistoty, ale kontroly. Cílem je identifikovat a ošetřit potenciální rizika a příčiny a zavést pojistné mechanismy pro zabránění vzniku a omezení porušování lidských práv.

Závěr

Umělá inteligence nabízí společnosti velkou službu, avšak skýtá i rizika, kterým lze předcházet nebo je odstranit. Toto se ostatně týká všech technologií: téměř každá technologie může být zneužita a představovat nebezpečí pro člověka. Již stará řecká báje o letecké technologii Daidala nabádá k obezřetnosti před překročením lidských mezí: „Daidalos, vynálezce […] se svým synem Ikarem uprchl uvěznění tím, že odletěl na křídlech vyrobených z ptačích per a vosku […]. Daidalos se díky své letecké technologii zachránil, avšak Ikaros posouval své hranice příliš daleko, letě příliš vysoko a příliš blízko k silnému slunci, kde mu *nemesis* tavícím voskem a pádem do moře oplatil jeho mladickou *hybris*. Technologie jako taková zde není odsouzena, ale jde o varování před jejím neuváženým a nadměrným použitím.“[49] Podstatou je stanovit limity pro používání AI s cílem zabránit porušování lidských práv.

Předpokladem takové „human centric AI“ je identifikace rizik a jejich kořenových příčin, což umožní následně definovat nápravu. Tento proces stejně jako samotná nápravná opatření vyžadují holistický přístup slučující AI technickou sféru a lidskoprávní expertízu. Mechanismus hodnocení rizik z pohledu lidských práv by měl prostupovat všechny fáze životního cyklu AI a být integrován především do uživatelských požadavků a specifikací. To zajistí mj., že AI systém bude vyvíjen, testován a monitorován s ohledem na lidská práva. Požadavky týkající se transparentnosti, vysvětlitelnosti a certifikace, stejně jako výběru vývojových dat, kontextu nasazení či okrajových vlastností aj., jsou všechny vysoce relevantní pro hodnocení a ochranu lidských práv v rámci AI. Náprava dopadu AI na lidská práva by se měla realizovat současně na třech úrovních: regulační – stanovení pravidel, technické – implementace pravidel a osvětové – zvyšování povědomí o daných pravidlech a rizikách. V závislosti na stupni rizika a kritičnosti dopadu AI na lidská práva bude regulační kompetence v rukách různých aktérů a autorit, přičemž zahrne zákony a jiné právně závazné nástroje, soft law a praktická opatření (kodex chování, vodítka). Varieta a velké množství dotčených aktérů vyžaduje multidisciplinární spolupráci.

Proces převedení lidskoprávních norem do AI technologií bude čelit výzvě týkající se explicitního přenosu lidského uvažování do výpočetních modelů a kvantitativní a kvalitativní klasifikaci rizik pro lidská práva. Které AI systémy budou kvalifikovány jako „human rights-critical“? Jaké kritérium vedoucí k nepřímé diskriminaci by mělo být vymazáno z atributů? Tyto a další otázky provázejí dnešní společensko-technologický progres.

[49] Ferré, F., *Philosophy of Technology* (1998), s. 98 (překlad autorů).

II. Soubor doporučení pro subjekty životního cyklu AI

1. Hodnocení rizik pro lidská práva v životním cyklu AI systému: Soubor doporučení

Jedná se prozatím o prvotní návrh části souboru doporučení týkající se hodnocení rizik pro lidská práva v životním cyklu AI systému, která v konečné fázi bude obsahovat i konkrétní doporučení týkající se možností eliminace rizika při vývoji systému a za provozu.

Vstupní analýza AI systému				
Riziko	Příklad/Vysvětlení rizika	Dotčené fáze životního cyklu AI systému	Možnosti eliminace rizika při vývoji systému *bude doplněno v další fázi projektu*	Možnosti eliminace rizika za provozu *bude doplněno v další fázi projektu*
Systém bude provozován v doméně s vazbou na LP*	Při vstupní analýze je zjištěna vazba AI systému na LP problematiku. Systém bude například vyhodnocovat osobní údaje, počítat výši půjčky, identifikovat osoby na základě obrazu a jiných biometrických hodnot, hodnotit účastníky výběrového řízení či predikovat pravděpodobnost recidivy a další. Pokud se vazba na LP problematiku určí pouze indikativně, například osobou, která není odborníkem v oblasti LP, musí následovat expertní rozpracování odborníkem v rámci vstupní analýzy.	- Možný dopad do všech fází		
Systém bude při vývoji používat data s vazbou na LP, případně data citlivá na bias s dopadem do LP				
Byly identifikovány požadavky související s LP				

Vstupní data, strojové učení				
Riziko	Příklad/Vysvětlení rizika	Dotčené fáze životního cyklu AI systému	Možnosti eliminace rizika při vývoji systému *bude doplněno v další fázi projektu*	Možnosti eliminace rizika za provozu *bude doplněno v další fázi projektu*
Bias v datech	Bias z pohledu LP je neoprávněná či nevyvážená přítomnost tzv. chráněné hodnoty v datech, tj. nesoucí informaci o pohlaví, rase, barvě pleti, náboženství, politické příslušnosti apod. Došlo by tak k protiprávní diskriminaci. Příkladem biasu v datech v oblasti bankovnictví je nezahrnutí skupiny osob v určitém příjmovém rozsahu při přípravě systému pro ohodnocování možné výše poskytované půjčky, pokud je daná skupina kvalifikována právě jednou z chráněných hodnot. Mohlo by tak dojít k diskriminaci dané skupiny. Může se jednat jak o bias ve vývojových datech používaných například v procesu strojového učení, tak v datech používaných pro verifikaci, validaci a kvalifikaci systému. Při analýze rizik bereme v úvahu: - Je tzv. chráněná hodnota přítomna vývojových datech? - Je její přítomnost oprávněná? - Pokud je oprávněná, je zajištěna reprezentativnost a úplnost dat nesoucí tuto chráněnou hodnotu? - Analyzovali jsme takové riziko i u testovací sady?	- Možný dopad do všech fází		

Použití syntetických vývojových dat	Syntetická data se používají zejména pro doplnění vývojových dat, pokud není k dispozici dostatečné množství reálných dat nebo není časově možné všechna reálná data nasbírat. Příkladem je doplnění obrázků obličeje o uměle generované obličeje etnika, které v reálných snímcích chybí. Syntetická data se nepoužívají jako plnohodnotná náhrada reálných dat a jejich použití má své limity vycházející z komplexity řešené úlohy. V případě komplexní úlohy i syntetická data mohou obsahovat ostatní zde uvedená LP rizika. Syntetická data nejsou zpravidla vhodná pro pokrytí celé úlohy. Pokud by bylo možné celou úlohu pokrýt syntetickými daty, její algoritmické řešení je známé a není třeba používat metody strojového učení k jejímu řešení.	- Příprava vývojových dat - Verifikace, validace, kvalifikace - Provozní fáze
AI systém není vysvětlitelný	Jedná se o algoritmickou vysvětlitelnost AI systému. Nevysvětlitelná AI se chová jako černá skříňka: jde o AI modely, které jsou příliš složité na to, aby je člověk mohl jednoduše interpretovat. Tyto systémy jsou většinou založeny na technikách strojového učení – příkladem jsou neuronové sítě, které jsou v dnešní době široce používaným paradigmatem pro úlohy strojového učení a představují tak primární zdroj modelů černých skříněk. U nevysvětlitelné AI výrobce není schopen zaručit vysvětlení, jakým způsobem AI systém došel ke svému rozhodnutí/výstupu. Naopak u vysvětlitelných modelů lze čtením modelu vidět, na základě jakých informací a jakým způsobem systém dospěl k rozhodnutí/výstupu. Příkladem je vysvětlitelný doporučovací AI systém využívaný ve výběrovém řízení do zaměstnání, který doporučí kandidáta na pracovní místo, pokud je jeho předchozí praxe delší pěti let. U systému, který není vysvětlitelný a který je natrénovaný pro obdobnou funkcionalitu takové čtení modelu není možné, žádné takové lidsky srozumitelné rozhodovací postupy v něm nelze nalézt.	- Možný dopad do všech fází

Dotrénování systému s použitím provozních dat	Například systémy rozpoznávání řeči používají reálné promluvy nasbírané za provozu systému ke svému zlepšování, tedy k dotrénování svých modelů. Hlavní riziko spočívá ve volbě správné metody řízeného dotrénování a zajištění odpovídajících trénovacích dat sbíraných za provozu systému. Hrozí zavedení biasu a posun již natrénovaného systému tak, že nebude ve shodě s LP. Dalším rizikem je ovlivnění přesnosti systému a jeho věrohodnosti. Při dotrénování za provozu může být systém i manipulován, čemuž je nutné předcházet.	- Možný dopad do všech fází
Nejsou známé a definované okrajové/hraniční vlastnosti systému a systém je na nich provozován	Například u bankovního systému pro ohodnocování kredibility klienta se nepočítalo s tím, že bude nasazen v lokalitě s velkou mírou rizika ztráty zaměstnání nebo naopak v oblasti s velkým výskytem startup firem, které vytváří v krátkém časovém horizontu majetné klienty. I u pečlivě připravovaných vývojových dat nemusí být plně pokryty veškeré hraniční případy řešené úlohy nebo vzácné výskyty určitého jevu. Rizikem je neznalost hranic, za kterých systém ještě funguje správně a kdy naopak i s malou odchylkou ve vstupních provozních datech bude poskytovat nesprávné výsledky.	- Verifikace, validace, kvalifikace - Uvedení systému do provozu - Provozní fáze
Požadavky na data používaná při strojovém učení jsou stanovena natolik striktně, že neumožní vývoj systému bez biasu	Příkladem může být situace, kdy by u požadavků na data používaná pro strojové učení bylo striktně trváno na použití anonymizovaných dat, např. anonymizovaných registračních značek vozů v případě trénování rozpoznávání právě těchto značek. Pak nelze úlohu učení úspěšně realizovat a její funkce nebude odpovídat realitě. Pokud se tento problém nebude řešit již ve vstupní analýze, později způsobí daleko větší náklady při vývoji.	- Vstupní analýza AI systému - Příprava vývojových dat

Integrace AI systémů třetích stran, rozšíření/úprava/oprava stávajícího systému, certifikace				
Riziko	Příklad/ Vysvětlení rizika	Dotčené fáze životního cyklu AI systému	Možnosti eliminace rizika při vývoji systému *bude doplněno v další fázi projektu*	Možnosti eliminace rizika za provozu *bude doplněno v další fázi projektu*
Integrace AI systému třetí strany při vývoji (transfer learning)	Riziko se týká situací, kdy je přebírán již hotový AI systém od externího dodavatele a integrován do výsledného řešení. Rizika spočívají v neznalosti vývojové datové sady, způsobu ověřování přebíraného systému a požadavků, za kterých byl systém vyvíjen. Většinou není k dispozici relevantní analýza dopadů systému na LP spojená s očekávaným nasazením systému. Rizikem jsou i případné certifikace systému s ním dodávané, zejm. pokud nejsou vydány ověřenou autoritou či nejsou plně známé podmínky, za kterých k certifikaci došlo.	- Verifikace, validace, kvalifikace - Uvedení systému do provozu		
Certifikace systému	Riziko souvisí zejména s provozováním systému a dopadá na provozovatele. Způsobů certifikace z hlediska LP problematiky může být celá řada a provozovatel systému musí zvolit adekvátní k možným dopadům provozu na LP, případně se musí řídit povinnou certifikací, pokud existuje. Obecně může certifikovat každý z aktérů, který se účastní životního cyklu AI systému, a to s různou mírou autority certifikátu. Za nejnižší formu certifikace lze považovat certifikát vydaný přímo výrobcem systému, za nejvyšší pak certifikaci provedenou regulačním orgánem zřízeným státem.	- Uvedení systému do provozu		
Výrobce nezahrnul do vývoje požadavky regulátora, certifikační autority nebo doporučené postupy, pokud existují	Riziko nastává, pokud si výrobce nebo odběratel/provozovatel systému není vědom těchto specifických požadavků a nezahrnul je do analýzy systému. Může se jednat například o požadavek testování systému na konkrétní testovací sadě dat, kterou spravuje příslušný úřad. Riziko souvisí zejména s provozováním systému a dopadá na provozovatele.	- Uvedení systému do provozu		
Je provedeno rozšíření nebo oprava AI systému	Typickou situací je aktualizace systému, rozšíření funkcionality nebo odstranění chyby v systému. Při tomto zásahu může dojít ke změně parametrů systému, biasu kvůli rozšíření systému nebo vzniku nových rizik (ve smyslu rizik uvedených v této analýze) a to díky novým funkcionalitám.	- Možný dopad do všech fází		

Požadavky na AI systém				
Riziko	Příklad/Vysvětlení rizika	Dotčené fáze životního cyklu AI systému	Možnosti eliminace rizika při vývoji systému *bude doplněno v další fázi projektu*	Možnosti eliminace rizika za provozu *bude doplněno v další fázi projektu*
LP požadavky jsou zadány vágně a nejsou měřitelné a testovatelné	Riziko vágně zadaných požadavků v LP oblasti je vysoké, což je dáno vstupem AI do oblastí lidské činnosti, které zatím nebyly automatizovány. Vágnost požadavků může být způsobena nedostatečnou analýzou požadavků na systém či i obtížností převodu LP definic do technických parametrů. Navazujícím problémem je testovatelnost těchto vágních požadavků. Testy musí pokrývat celý testovaný případ a musí být konkrétní a měřitelné, což je u vágně definovaných požadavků těžké splnit. Příkladem je chráněná hodnota "barva pleti", kde není v technické oblasti jednoznačně zadefinováno, jak je tato hodnota reprezentována například v obrazových datech. Podobně by působil požadavek, že "systém nemá pracovat s chráněnou hodnotou", aniž by byla tato chráněná hodnota definována technickou specifikací.	- Vstupní analýza AI systému - Produktové a uživatelské požadavky - Technická specifikace a požadavky - Verifikace, validace, kvalifikace		
Nejsou identifikovány všechny LP uživatelské požadavky	Riziko nastává zejména v případě, kdy nejsou ke tvorbě požadavků na systém přizvány všechny klíčové osoby, kdy existuje malá zkušenost s vývojem a nasazením systému pro příslušnou doménu a také v případě neznalosti příslušných regulací, certifikačních postupů nebo doporučení dobré praxe. Příkladem je potřeba transparentnosti systému, který bude nasazen jako asistivní technologie v soudnictví. Pokud požadavek transparentnosti není zachycen, systém nebude poskytovat podpůrné informace uživateli (soudci), čímž oslabí či znemožní jeho schopnost v rámci svého rozhodování detekovat případnou chybu v predikci systému vedoucí k porušení LP.	- Vstupní analýza AI systému - Produktové a uživatelské požadavky.		

LP uživatelské požadavky jsou nereálné	Příkladem nereálných požadavků může být požadavek vytvoření plně autonomního systému tam, kde vzhledem ke složitosti rozhodovacího procesu má být nasazena asistivní technologie, podporující rozhodování lidského odborníka - typicky oblast soudnictví a zdravotnictví. Obecně, pokud je činnost řešena člověkem s určitou mírou chybovosti a tuto činnost hodláme automatizovat, je nutné nejprve provést analýzu, zda je tato chybovost automatizací odstranitelná nebo se jedná o hlubší problém, který bude přetrvávat i po automatizaci příslušné činnosti.	- Vstupní analýza AI systému		
		Provozování AI systému		
Riziko	Příklad/Vysvětlení rizika	Dotčené fáze životního cyklu AI systému	Možnosti eliminace rizika při vývoji systému *bude doplněno v další fázi projektu*	Možnosti eliminace rizika za provozu *bude doplněno v další fázi projektu*
Netransparentní AI systém	Transparentnost AI systému zahrnuje vývojovou a operační transparentnost. Vývojová transparentnost slouží především aktérům schvalujícím systém do provozu. Týká se provádění vývojového procesu a umožňuje jeho kontrolu, vč. kvality. Operační transparentnost naopak vyjadřuje schopnost AI systému poskytnout informaci o tom, která vstupní data přispěla k výslednému výstupu systému a s jakou vahou. Transparentní systém tak poskytuje uživateli za provozu důležité informace, tj. kromě výstupů systému má k dispozici i vstupní informace, které k výstupu přispěly a s jakou vahou. Z hlediska provozu je systém pro uživatele dostatečně jasný a zřetelný a jeho výstupy lze lehce ověřovat. Naopak u netransparentního systému uživatel tyto informace nemá a zná pouze výstup/výsledek. Pozn.: Transparentnost je třeba odlišit od algoritmické vysvětlitelnosti. Ta se týká způsobu algoritmické interpretace znalostí, tj. vysvětlení, jakým způsobem AI systém došel k výběru informací, na nichž postavil své rozhodnutí. Pro uživatele může být i algoritmicky vysvětlitelný systém netransparentní, protože neposkytuje informace o tom, na základě čeho se rozhodl.	- Produktové a uživatelské požadavky - Technická specifikace a požadavky - Verifikace, validace, kvalifikace - Provozní fáze		

AI systém není za provozu monitorovatelný	Nesoulad AI systému s LP lze detekovat za provozu pomocí jeho monitorování. Pokud není systém pro své monitorování připraven, nelze zpětně provádět ani jeho audit, analyzovat jeho výstupy, detekovat nesoulad a zjišťovat jeho příčinu. Monitorování AI systému za provozu přispívá k jeho transparentnosti.	- Produktové a uživatelské požadavky - Technická specifikace a požadavky - Verifikace, validace, kvalifikace - Provozní fáze
Systém je určen pro jiné než cílové provozní prostředí	Systém je nasazen v jiném cílovém prostředí, než bylo plánováno. Může se jednat např. o rozšíření okruhu uživatelů produktu, přičemž v daném prostředí nevyhovuje požadavkům na LP, či může dojít k instalaci systému v jiné zemi či regionu, kde existuje odlišný přístup nebo zákonné úpravy. Toto riziko může nastat i v rámci jedné firmy působící globálně. Další případem může být situace, kdy se v cílovém prostředí nasazení produktu změnil přístup k LP problematice a systém přestal být ve shodě s LP. Okrajové vlastnosti jsou řešeny samostatně v riziku uvedeném výše.	- Vstupní analýza AI systému - Produktové a uživatelské požadavky - Uvedení systému do provozu
Uživatel nebo provozovatel nepoužívá systém správně	Nesprávné používání systému uživatelem/provozovatelem může mít vícero variant a jejich kombinací: - systém je využíván i pro úlohy, pro které nebyl určen - asistivní systém je využíván jako plně autonomní - uživatel/provozovatel systému předkládá vstupy, pro které nebyl systém připraven - výstupy monitorování systému nejsou správně interpretovány. Příčinou je většinou neznalost a nesprávné proškolení uživatele a provozovatele.	- Uvedení systému do provozu - Provozní fáze

Oblast s vysokou rizikovostí porušení LP za provozu	Vzhledem ke komplexnosti LP problematiky, novým oblastem nasazování AI systémů, a tím i malé zkušenosti s jejich provozováním pro tyto nové činnosti, existuje vyšší riziko, že systém za provozu způsobí porušení LP. Při eliminaci tohoto rizika je i nutné zvážit míru škody. To by mělo být součástí vstupní analýzy systému.	- Možný dopad do všech fází
AI systém bude napaden a zneužit	Riziko se ošetřuje stejným postupem jako u všech IT a software systémů, tj. analýzou bezpečnosti systému, implementací bezpečnostních prvků, vzdálené správy za provozu, zajištěním provozní infrastruktury proti napadení apod. Toto riziko nemá přímo vztah k AI technologiím jako takovým, jedná se obecně o přístupy, které jsou obvyklé pro vývoj a nasazení jakékoli informační technologie, která je tomuto riziku vystavena. Nejsou zde tedy uváděné konkrétní příklady dopadu rizika do jednotlivých fází a postačuje odkaz na standardní metody implementace bezpečnostních opatření do IT technologií. Patří sem i situace, kdy je systém používán za nelegálním účelem.	- Možný dopad do všech fází

Systém bude provozován v plně autonomním režimu	Jde o situaci, kdy je AI systém používán jako plně autonomní systém a není tedy primárně určen jako asistivní ke zlepšení práce odborníka v dané oblasti nasazení, tj. řeší úlohy bez přispění lidského činitele. Na rozdíl od asistivních AI systémů tak jeho výstupy nebudou přehodnocovány doménovým expertem. S tím souvisí riziko nedostatečné, resp. nemožné zpětné vazby při provozování autonomních systémů. V procesu, kde je zapojen člověk, je obvyklé získávat zpětnou vazbu, klást otázky či vznášet námitky na provedenou činnost, rozhodnutí či výstupy, což zamezí případným nedostatkům a chybám, popř. umožní jejich nápravu při nasazení. Důležitá je vstupní analýza možných dopadů na LP a vyhodnocení provozních rizik. Pokud bude systém vyhodnocen jako "human-rights critical", pak platí maximální možná opatření pro eliminaci rizika. S nižší mírou dopadů na LP budou opatření úměrně redukována. Doporučené minimální požadavky by se měly týkat: - zajištění transparentnosti systému - zajištění monitorování systému - certifikace systému před uvedením do provozu	- Možný dopad do všech fází
Status AI systému jako "vyšší autority" pro uživatele	V očích lidského účastníka procesu, především uživatele, může být AI systém považován za prvek vyšší autority, jejíž závěry není nutné přehodnocovat či podrobovat kritickému přezkumu. Může to být dáno lidskou důvěrou v bezchybnost systému a v kompletní pokrytí znalosti automatem. Riziko pak spočívá v nekritickém přebírání výstupů systému i tam, kde je očekáváno jejich odborné posouzení před aplikací výsledků. Toto se týká především asistivních AI systémů, viz například "autonomní" řízení automobilů, které dnes zdaleka ještě není plně autonomní, je pouze asistivní, což některé řidiče neodrazuje od svěření plné kontroly nad řízením automatu.	- Produktové a uživatelské požadavky - Uvedení systému do provozu - Provozní fáze

* *LP = lidská práva*

2. Doprovodná zpráva

Při vývoji AI systému je vždy nejprve nutné zvážit provedení analýzy možných dopadů systému na lidská práva (dále ve zkratce „LP“), tedy posouzení relevantních rizik a zapracování jejich prevence a eliminace do celého životního cyklu (vývojové a provozní fáze). V případě existence certifikačních procesů a regulací je nutno tyto požadavky rovněž zahrnout do vstupní analýzy. V současnosti však takové postupy a doporučení z hlediska LP v podstatě neexistují. „Soubor doporučení“ pro subjekty životního cyklu AI tak nabízí seznam hlavních rizik pro lidská práva, která mohou vyvstat v jednotlivých fázích životního cyklu AI a jejich popis. V následující fázi projektu budou naformulována doporučení, jak jim předcházet, příp. jak je odstranit. Tato část navazuje na výzkum LP problematiky ve spojení s AI systémy (Část I Dílčí zprávy), kde byla provedena identifikace a základní posouzení rizik, stejně jako způsobů jejich zapracování do životního cyklu AI systému. Doprovodná zpráva k „Souboru doporučení“ představuje mechanismus hodnocení rizik v rámci standardních fází vývoje software produktů, jehož technické procesy podrobně vysvětluje.[50]

* * *

Východiskem pro zapracování požadavků na AI systém z hlediska LP by vždy měla být vstupní analýza rizik AI systému spojených s dopadem na lidská práva. Rizika by měla být hodnocena pro jednotlivé části životního cyklu a následně rozpracována do požadavků, specifikací a provozních opatření. Toto se promítne do případné potřeby certifikace systému a do jeho schvalování před uvedením do provozu. Pokud se při vstupní analýze dojde k závěru, že je produkt nasazován do oblasti, kde hrozí riziko porušení LP, je třeba eliminovat tato rizika od prvních vývojových fází systému.

V případě AI systémů mohou být rizika související s LP obtížně uchopitelná, zejména z následujících důvodů:

- AI systémy umožňují řešit úlohy, které před nástupem těchto technologií nebylo možné úspěšně automatizovat tak, aby došlo k nahrazení příslušné lidské činnosti a rozhodování. S tím souvisí, že se automatizují oblasti lidské činnosti, ve kterých zatím není s tímto typem automatizace dostatečná zkušenost a není tedy ani zkušenost s definicí uživatelských požadavků z hlediska LP problematiky a jejich následným převodem do technických požadavků, specifikací a testů.
- V oblasti provozování systémů, především uživatelem, může být AI systém považován za prvek „vyšší autority“, který je prost lidské chybovosti. Důsledkem může být, že výstupy již nebudou dále přezkoumávány a zpochybňovány, a riziko porušení LP se zvyšuje.

[50] Jedná se tedy o širší rámec zahrnující AI fáze představované z pohledu CRISP-DM, který pokrývá vývoj celého produktu.

- U AI systémů nasazených v oblastech s rizikem porušení LP lze očekávat, že koncovým uživatelem nebude odborník v oblasti informatiky a automatizace; bude docházet k běžné interakci s technicky neodbornými uživateli. Rizika spojená s LP se mohou projevit až za provozu, a je tedy nutné zvážit průběžné monitorování AI systémů za provozu s cílem ověřovat jejich shodu s LP.

2.1. Životní cyklus AI systému

U životního cyklu AI systému se v rámci hodnocení rizik primárně rozlišují následující části:

- Vývoj systému: zahrnuje kroky, jako je vstupní analýza, sběr uživatelských požadavků na systém, sestavení technické specifikace, návrh systému, vývoj a testování, předání do provozu.
- Provozování systému: spočívá v používání systému koncovým uživatelem a jeho provozování, včetně údržby a řešení provozních událostí.

2.1.1. Vývoj AI systému

Vývoj systému obsahuje fáze, které jsou z pohledu hodnocení rizik dopadů na LP významné, a je třeba provést hodnocení pro každou z nich zvlášť. Jednotlivé fáze vycházející ze standardních fází vývoje software produktů jsou v této zprávě stručně představeny, přičemž detailní popis je uvedený v externí literatuře[51].

Následující obrázek ukazuje vývojové fáze překreslené tak, že je vidět návaznost fází specifikačních (vlevo) a dále vývojových a testovacích (vpravo). Po ukončení vývoje, tedy po předání systému do provozu, následuje provozní fáze životního cyklu AI systému.

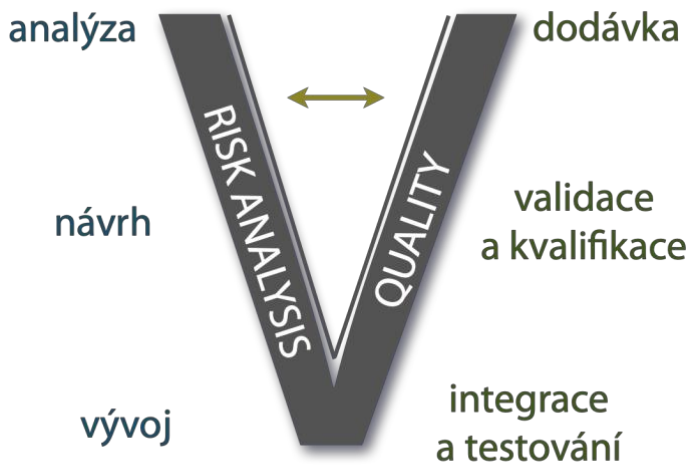
[51] Předmětný výzkum spadá do oblasti vývoje software produktů – software je implementací AI, hardware je pouze výkonná infrastruktura. Existuje velké množství literatury popisující vývojové procesy, přičemž předmětný výzkum se řídí tzv. agilními procesy a těmi, které se používají při vývoji „life-critical“ a „mission-critical“ aplikací. Odkazy na vývojové procesy:

- FDA Device Development Process.
- SCRUM guide
- LeSS - Large Scaling Scrum site
- Henrik Kniberg and Mattias Skarin - Kanban and Scrum
- Lisa Crispin and Janet Gregory - Agile Testing: A Practical Guide for Testers and Agile Teams

Odkazy na tématiku user stories a vytváření uživatelských požadavků:

- Product Backlog příklad
- User Story
- Splitting patterns
- Od vize k Product Backlogu

Jedná se o iterativní model, kde se fáze či celý cyklus několikrát opakuje dle potřeb; vývoj jednoho systému neprobíhá takto lineárně v čase. Obrázek tak spíše ukazuje logické řazení výstupů jednotlivých iterací. Z hlediska vstupní analýzy je důležité, že je obvyklé se vracet k jednotlivým fázím a provádět korekce dle nálezů, chyb, oprav, a tím tyto prvky eliminovat. V tomto směru bude provedeno hodnocení rizik, jejich předcházení a eliminace z pohledu LP.



2.1.1.1. Analýza systému (Vstupní analýza)

Jedná se o prvotní krok při realizaci systému a jejím účelem je identifikovat všechny oblasti, které budou dále rozpracovány pomocí požadavků na systém. Zahrnuje hodnocení rizik, všech účastníků vývoje a provozování systému, nasazení systému a podobně. Zjednodušeně řečeno, vymezuje celou oblast domény řešení.

Součástí vstupní analýzy by měla být analýza spojitostí a možných dopadů výsledného produktu na LP. V případě existence certifikačních procesů a event. regulací je pak v případech produktů, které spadají pod tyto regulace a certifikace, snazší tuto analýzu provést, protože budou existovat doporučení a závazné postupy třetích stran. V současnosti však takové postupy neexistují. Primárně by se tedy v rámci analýzy požadavků na software mělo provést hodnocení možného dopadu produktu na LP při jeho nasazení.

Pokud se při vstupní analýze dojde k závěru, že je produkt nasazován do oblasti, kde hrozí riziko porušení LP, měla by se tato rizika začít eliminovat od prvních vývojových fází produktu, a to v souladu s principy uvedenými v „Souboru doporučení“ u jednotlivých fází.

2.1.1.2. Produktové a uživatelské požadavky

Produktové a uživatelské požadavky se dále rozpadají na skupiny požadavků dle zacílení. Například požadavky na systém, uživatelské rozhraní, zabezpečení, provozování systému a

další. Záleží na kategorii produktu a jeho cílenému produkčnímu nasazení. Jiný okruh požadavků budou mít systémy pro nasazení do životně důležitých („life-critical“) oblastí a do kritických operací („mission-critical“), kde se například přidávají požadavky z hlediska bezpečnosti, certifikace systémů a jejich validace. Na druhém konci spektra pak budou ležet například volnočasové a zábavní aplikace.

Současné metody produktových a uživatelských požadavků pracují s pojmy, jako jsou „user stories“, které dávají do souvislosti technické a uživatelské aspekty požadavků a vysvětlují je v kontextu jejich používání. To je zásadní pro následné správné pochopení významu požadavků a jejich klíčových vlastností při implementaci.

Většinou se formulují dle schématu: „Jako <uživatel> chci <funkcionalitu>, abych získal <hodnotu>.“ Jsou tedy zaznamenávány informace o tom „Kdo <role uživatele>“, „Co <cíl>“, „Proč (důvod)“. Dále se rozpoznává vlastník požadavků, tedy ten, kdo odpovídá za jejich úplnost a správnost (vlastníků může být více za různé oblasti). Zdroje požadavků, tedy „Kdo“ – nositelé informací, mohou pocházet z různých domén, tj. z technologické, uživatelské či provozní části systému nebo od poskytovatele, a to dle komplexnosti vyvíjeného systému.

V případě, že se jedná o nasazení AI systému do oblasti s rizikem dopadu na LP, měly by produktové a uživatelské požadavky obsahovat vymezení s ohledem lidská práva.

2.1.1.3. Technické specifikace

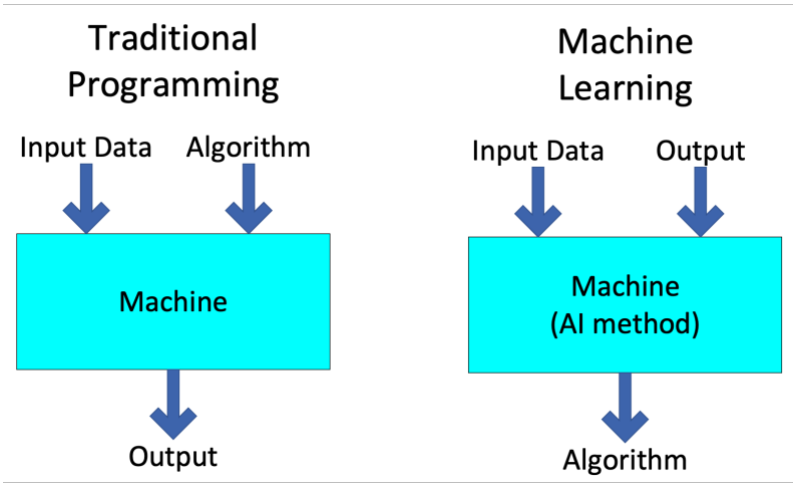
Technická specifikace se týká všech požadavků na technologii systému a způsobu implementace celého systému jako produktu. Zahrnuje technické specifikace, požadavky na systém, jeho interakci s prostředím, ve kterém bude provozován, a další technologické prvky. Oproti produktovým a uživatelským požadavkům se při popisu technické specifikace používají formalizované postupy návrhu softwaru.

Pokud je začleněna technologie, která může obsahovat implementační rizika související s LP, technická specifikace má obsahovat příslušný popis rizik a jejich vztahu k LP. Jedná se zejména o rizika, která vycházejí z vlastností technologií.

2.1.1.4. Strojové učení (ML)

Metody strojového učení, které se také nazývají metodami umělé inteligence, pracují jinak než algoritmické přístupy.

Následující obrázek znázorňuje hlavní rozdíl:



V případě algoritmického přístupu a programování vytváří algoritmus člověk, tj. člověk pochopí problém a převede ho do matematického popisu, algoritmu, naplní databázi znalostí a podobně. Naopak u strojového učení algoritmus vytváří stroj na základě učení s pomocí vstupních dat. Algoritmem je pak například natrénovaná neuronová síť, tedy soustava velkého množství nelineárních rovnic a jejich parametrů.

Zatímco u algoritmického přístupu, kdy člověk popisuje problém, je možné jeho popis číst a porozumět mu, u výsledku strojového učení člověk dostane černou skříňku, jejíž interní uspořádání nedokáže analyticky interpretovat a pochopit, a s výsledkem nelze případně dále pracovat na úrovni algoritmických oprav či změn. Změny se musí provádět novým učením či dotrénováním. Vnitřní interpretace není v tomto případě vysvětlitelná (explainable).

Pro eliminaci rizik spojených s LP je klíčové při trénování systému pomocí strojového učení správně vybrat trénovací a testovací data. Výběr trénovacích a testovacích dat tedy představuje v této fázi první krok pro možnou eliminaci těchto rizik.

2.1.1.5. Vývoj softwaru

Vývoj softwaru se dále člení do řady kroků, které odpovídají standardním vývojovým procesům. V rámci této fáze je především důležité, že se implementují testovací mechanismy a provádějí testy na úrovni technologie, softwaru a jednotlivých komponent a provádějí se revize kódu (více o procesu testování viz následující část). Jedná se souhrnně o testování technologického návrhu systému (funguje, jak byl navržen) a testování vzhledem k uživatelským požadavkům (jak byl požadován).

Při vývoji se vychází z požadavků na celý systém, tj. produktových, uživatelských a technických. Vlastní implementace požadavků nevnáší další možná rizika z hlediska LP. Pokud je vytvořen produkt, který obsahuje LP rizika, jsou tato rizika zanesena či popsána již v předchozích fázích. Vývojář pracuje s již připravenými vstupními daty v případě učení systémů založených na

strojovém učení, používá typ neuronových sítí specifikovaných v technické specifikaci a obecně technologie, které byly specifikovány. Vlastní vývoj tedy nevnáší další přidaná rizika.

2.1.1.6. Verifikace, validace a kvalifikace

Testování může probíhat v různých fázích vývoje a má více částí. Při vývoji jednotlivých bloků systému provádí vývojář přímo testy komponent či bloků, a to včetně nezávislé revize kódu jiným vývojářem, a dále probíhají integrační a systémové testy. Dělení a názvosloví se může lišit produkt od produktu. Tento typ testů má zajistit, že software pracuje tak, jak byl navržen, tedy jak byl specifikován.

Další množinou jsou testy, které mají zajistit, že systém bude fungovat tak, jak bylo požadováno v produktových a uživatelských požadavcích a v technických specifikacích. Jedná se o „kvalifikaci" software, jeho verifikaci a validaci. Například v případě regulovaného vývoje zdravotnických aplikací se ještě navíc k uživatelským a funkčním požadavkům validací ověřují požadavky regulátora a požadavky týkající se prostředí plánovaného nasazení.

Zjednodušeně řečeno, jedná se o testy typu „ověřuji, že produkt pracuje tak, jak byl navržen" a „ověřuji, že produkt pracuje tak, jak bylo požadováno", tedy o testování technologické a uživatelské.

Z hlediska vztahu k LP jsou verifikace i validace relevantní fáze ověřování softwaru. Pokud se jedná o produkt, který při nasazení může mít dopad na LP a obsahuje rizikové technologie, jeho funkcionalita vzhledem k rizikům spojeným s LP a technologiemi se otestuje ve fázi verifikace. Příslušné testovací požadavky vzejdou z technické specifikace. Pokud jsou na tento produkt navíc kladeny uživatelské a produktové požadavky z hlediska LP, příslušné testy jsou zahrnuty buď do fáze verifikace či validace. Záleží na konkrétním vývojovém procesu. V případě existence certifikací či regulací, které by se vztahovaly k softwaru a jeho nasazení do oblastí, kde by se mohla projevit rizika spojená s LP, tyto požadavky se ověří validací softwaru.

Hodnocení rizik uvádí dopad do fáze verifikace, validace a kvalifikace, nikoli však konkrétní testovací případy. Ty se upřesňují dle příslušné specifikace od úrovně uživatelské, přes architekturu až po systém.

2.1.1.7. Uvedení systému do provozu

Kromě technického uvedení systému do provozu, tedy zejm. přenosu systému z vývojové do provozní fáze a zajištění infrastruktury, se jedná o ověření shody systému s provozními podmínkami. Kromě provedení testů v předchozích krocích, které ověřují shodu s požadavky systému, se jedná o potvrzení této shody na úrovni certifikace systému. Ta může proběhnout na několika úrovních (od nejnižší úrovně certifikace po nejvyšší):

- výrobcem či dodavatelem,

- odběratelem či uživatelem,
- nezávislou stranou,
- nezávislou stranou, která je ověřenou certifikační autoritou,
- nezávislou stranou, která je státem ověřenou a pověřenou certifikační autoritou.

V současnosti neexistuje regulátor, certifikační autorita nebo doporučené postupy pro certifikaci systémů s ohledem na rizika spojená s LP. Lze tedy očekávat, že certifikace doporučené v překládaném hodnocení rizik se budou prozatím provádět na nižších úrovních. Pro ověření shody systému s provozem v příslušné doméně se očekává, že takové postupy pro certifikace budou vznikat, a to nejprve ze strany domén nejvíce rizikových, jako je soudnictví, zdravotnictví, veřejná správa, finančnictví a podobně.

2.1.2. Provozování systému

Provozním nasazením rozumíme nasazení produktu do provozu a používání systému v jeho předpokládaném cílovém prostředí, např. aplikace v telefonu, systém obrazového rozpoznávání vozidel v kameře, automatizace poskytování půjček. Do provozu vstupuje systém ve stavu po testování, který odpovídá produktovým a uživatelským požadavkům. První nasazení s sebou může nést rizika spojená se špatně identifikovanými či neúplnými požadavky a skrytými chybami.

2.2. Předcházení a eliminace rizik v rámci jednotlivých fází životního cyklu

Úvahy nad možnostmi předcházení, detekce a eliminace rizik v návaznosti na LP problematiku a provozování systému automatizace lidské činnosti by se primárně měly odvíjet od rizik, které analogicky mohou vyvstat při příslušné lidské činnosti, s rozšířením o postupy pro AI a pro informační technologie.

Prvky předcházení a eliminace rizik budou v „Souboru doporučení“ uvedeny pro celý životní cyklus AI systému, rozdělený na dvě primární části: vývojovou a provozní.

U jednotlivých kroků vývojové fáze je dotčena zejména vstupní analýza AI systému, příprava specifikací se zahrnutím rizik spojených s LP, testování systému před nasazením a jeho předání do provozu, včetně případné certifikace. Hodnocení rizik v „Souboru doporučení“ uvádí rizika specifická pro LP v rámci AI systémů, následně budou vypracovány možné návrhy na jejich eliminaci.

Dalším důležitým prvkem bude identifikace návazných specifických rizik a jejich závažnosti v konkrétní doméně. Účelem této studie není vyjmenovat technické detaily každé takové domény a uvést veškeré možné dopady AI systému na příslušná LP. Již známá, doménově

specifická rizika dopadů lidské činnosti na LP pak bude rovněž vhodné zohlednit a adaptovat je pro AI automatizační systém.

Identifikace a eliminace provozních rizik systému s dopadem na LP je také novou oblastí. Za provozu AI systému může dojít z různých důvodů k odchylce od předpokládaného fungování systému, a to i v případě, kdy systém po uvedení do provozu veškeré předpoklady a požadavky splňuje. Pro eliminaci rizik souvisejících provozováním AI systému je žádoucí nasadit systém transparentní, certifikovaný a umožňující monitoring.

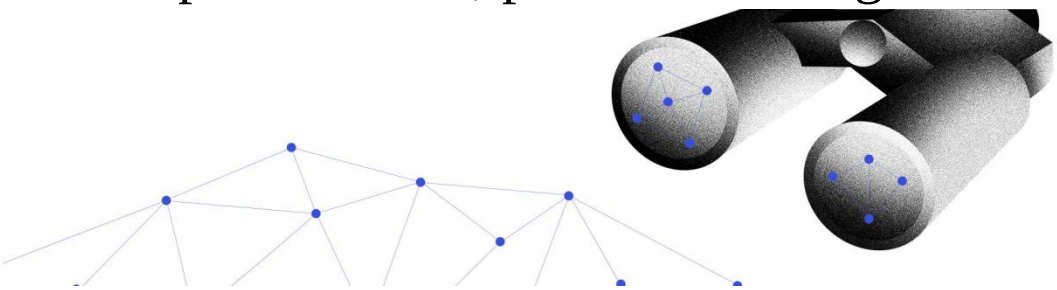
Úkolem monitoringu je právě vyhodnocení a upozornění na možné odchylky výstupů systému z hlediska rizik pro LP. Monitoring může být realizován automatem či člověkem nebo formou auditu, popř. kombinací. Vyhodnocování monitoringu může být založeno na historicky předchozích výstupech pozorované činnosti, jejíž výstupy jsou považovány z hlediska LP za bezrizikové. V případě automatického monitoringu je obtížnou částí převedení logiky lidského pozorovatele do softwaru. Odstranění detekovaného rizika může být, dle jeho povahy, provedeno přímo za provozu systému, nebo v případě složitějšího problému může být nutný návrat do vývojové fáze.

Rizika jsou rozdělena pro přehlednost do následujících skupin:

- **Vstupní analýza AI systému**: V této fázi by mělo být zejména identifikováno, zda systém může mít dopad na LP a v jakém rozsahu a jaká konkrétní rizika s tím souvisí. Rovněž by měly být analyzovány návaznosti na možné certifikace a regulace a globálnost či lokálnost nasazovaného systému.
- **Vstupní data, strojové učení**: Práce se vstupními daty je pro přípravu a provozování AI systémů klíčová. Případný bias systému je závislý právě na vstupních datech, a to jak použitých pro přípravu systému a jeho testování, tak provozních datech pro průběžné dotrénování systému a zlepšování jeho vlastností apod.
- **Integrace AI systémů třetích stran, rozšíření/úprava/oprava stávajícího systému, certifikace**: Při vývoji nového AI systému se většinou nelze vyhnout používání systémů třetích stran, včetně AI systémů. Ve většině případů je tak k dispozici málo informací o datech použitých pro přípravu AI systému, jeho testování a fungování za hraničních podmínek. Pokud se přihlédne k tomu, že v současnosti neexistují regulérní certifikace systémů v souvislosti s LP problematikou, je přebírání a integrace systémů od třetích stran velice riziková a je nutné tomu přizpůsobit jak vývoj, tak provozování nového systému.
- **Požadavky na AI systém**: Funkce AI systému je postavena na základě požadavků všech aktérů působících v celém životním cyklu AI systému. Kromě zahrnutí požadavků spojených s LP problematikou je klíčové, aby byly požadavky technicky realizovatelné, měřitelné a testovatelné. Pokud například nedojde k dialogu mezi uživatelem systému a AI odborníky, mohou požadavky srozumitelné člověku být příliš vágní pro automatizaci a tedy neimplementovatelné. V řadě případů bude nutné požadavky technicky konkretizovat a zúžit pro příslušnou doménu nasazení AI systému.

- **Provozování AI systému**: Bylo by chybné přepokládat, že správně navržený a otestovaný systém, který v okamžiku svého nasazení funguje v souladu s LP, tak bude fungovat i nadále. Díky změně vnějších podmínek, legislativy, nasazení systému na jiné pobočce, změnou systému, který se vylepšuje za provozu, a řadou dalších okolností, tak tomu nemusí být. Je tedy často nutné systém dále monitorovat a kontrolovat soulad s LP.

Dotazník pro projekt „Umělá inteligence a lidská práva: rizika, příležitosti a regulace“



Dotazník byl realizován v rámci projektu „Umělá inteligence a lidská práva: rizika, příležitosti a regulace“, č. TL05000484, financovaného Technologickou agenturou České republiky, formou online sběru v období 26.4.-9.5.2022. Předem vytipovaní potenciální respondenti z řad firem, kterých bylo 55, byli s žádostí o vyplnění osloveni e-mailem a následně i telefonicky. Celkem bylo získáno 23 odpovědí, tedy necelých 42% oslovených.

Zpracovali:

Luboš Král, ČVUT
Lenka Kučerová, prg.ai
Jaroslav Šíp, prg.ai
Martina Šmuclerová, AMBIS vysoká škola

V Praze 23.5.2022

Manažerské shrnutí

Dotazníkového šetření se zúčastnilo 23 firem působících ve všech oblastech životního cyklu AI, které zahrnují vytváření uživatelských požadavků s pochopením cílové domény, přípravu dat strojového učení, trénování a přípravu AI systémů, vývoj produktu/řešení, nasazení výsledného systému a jeho provozování. Většinou se jednalo o malé podniky (60.9%) a střední podniky (26.1%), dále se zúčastnilo 8.7% velkých firem a jeden mikropodnik.

Více než polovina firem (60.9%) se domnívá, že nenasazuje AI systémy v oblastech, kde při jejich provozování existují rizika spojená s lidskoprávní problematikou, tj. nejsou si ve své činnosti vědomy rizik týkajících se práva na soukromí, zákazu diskriminace, práva na spravedlivý proces (zejm. přístup k důkazům) a sekundárně dalších jednotlivých práv jako právo na vzdělání, právo na práci, právo na sociální zabezpečení, atd. Ostatní firmy vidí rizika zejm. v případném biasu, např. u rekomendačních systémů, při detekci osob v bezpečnostních systémech autonomního řízení nebo ve zdravotnictví při selekci pacientů pro operační zákroky, či v ohrožení práva na soukromí při sdílení polohy uživatelem prostřednictvím aplikace. Dále může docházet k dopadům na specifická lidská práva v jednotlivých oblastech nasazení AI jako jsou zpracování dat v oblasti zemědělství, zejm. v rozvojových zemích (v rámci potravinové bezpečnosti jde o právo na zdraví, na důstojný život), bankovnictví, mobilní služby, finanční systémy, developerské a konstrukční činnosti a zdravotnické databáze.

Většina dotázaných firem (87%) pracuje s osobními údaji a bere v potaz GDPR. Ve valné většině vyžadují souhlas se zpracováním dat (82%), v závěsu je i využití anonymizace (63,6%). Pseudonymizaci provádí pouze 36.4%. Jsou i firmy, které žádný z těchto postupů neaplikují.

Řešení v oblasti rozpoznávání obličejů dodávají dvě společnosti, přičemž jedna z nich charakteristiky typu rasa, barva pleti a etnikum neidentifikuje. Druhá řeší problémy s kvalitou rozpoznávání obecně u skupin lidí s nějakým charakteristickým prvkem, jako je nezvyklý typ účesu. Pokud se jedná o společensky citlivé charakteristiky, jako je rasa, barva pleti, etnikum, gender apod., nelze riziko špatné klasifikace úplně eliminovat, lze ho pouze snížit, a to co nejlepším vyvážením trénovací a testovací sady, pokud jsou dané charakteristiky předem dány.

Více jak tři čtvrtiny firem používají transfer learning – velké množství firem tedy navazuje na již existující natrénované systémy a rozšiřuje je.

Většina firem (69.6%) nezpracovává v uživatelských požadavcích požadavky související s LP od třetích stran. Pokud takové požadavky zapracovávají, pak se jedná o rovnoměrné rozložení mezi právem na soukromí a zákazem diskriminace. Konkrétně se požadavky od externích subjektů týkají např. odstranění dat, otázek práva na soukromí v souladu s doporučeními ÚOOÚ či vysvětlitelnosti verdiktu a poskytnutí důkazů pro soud v případě odhaleného podvodu.

Obdobná situace je i v případě zapracování vlastních požadavků souvisejících s LP problematikou. Zde je zdůrazněn především požadavek na odstranění biasu a vysvětlitelnost rozhodnutí a vnitřní etický kodex.

Pokud již k zapracování požadavků na LP problematiku dojde, 26.1% firem využívá existujících norem, 21.7% doporučení od etablovaných organizací a nejméně firem (13%) používá certifikované postupy. Zdrojem takových externích požadavků jsou například normy z různých krajů anebo požadavky od velkých firem, pro které je systém vyvíjen. V odpovědi jsou uvedeny odkazy na normy, ty ale až na GDPR nejsou pro tuto oblast relevantní (vývoj zdravotnických zařízení a obecné normy na bezpečnost informací); obdobné technické normy ani v mezinárodním měřítku neexistují.

Skoro polovina firem (43.5%) bere LP prvek na úrovni vývoje SW v potaz, a to na různé úrovni zpracování. Například nepracují s atributy uživatelů, nesbírají informace s chráněnou hodnotou a věnují se zajištění dostatečné variability dat, genderově neutrálního designu, filtrování dat před jejich použitím a detekci pouze neutrálních atributů v rozpoznávání popisu člověka.

Pouze 3 společnosti řešily provozní problémy spojené s narušením LP. To se týkalo zejména problémů, které vznikly použitím volně dostupných datasetů, řešení pro call centra či výzkumu a zakázek, které nesplňují etické normy firmy.

30.4% společností zahrnuje do testů požadavky související s LP, což odpovídá počtu společností, které při vývoji s LP pracují. Lze tedy předpokládat, že pokud společnost s LP pracuje, zahrnuje i tyto požadavky do testů. Jedná se zejména o testy na bias a o anonymizaci dat.

Obdobný počet společností (34.8 %) používá nějakou formu monitorování systému za provozu. Většinou se jedná o logování provozních informací a následné offline vyhodnocení, případně používají systémy třetích stran. Většinou se nejedná o monitoring, který by byl zaveden s cílem předcházet rizikům a zachycovat porušení související s LP, i když jeho výstupy mohou takové analýze sloužit. Pouze 16.7% společností provádí monitorování s takovým cílem.

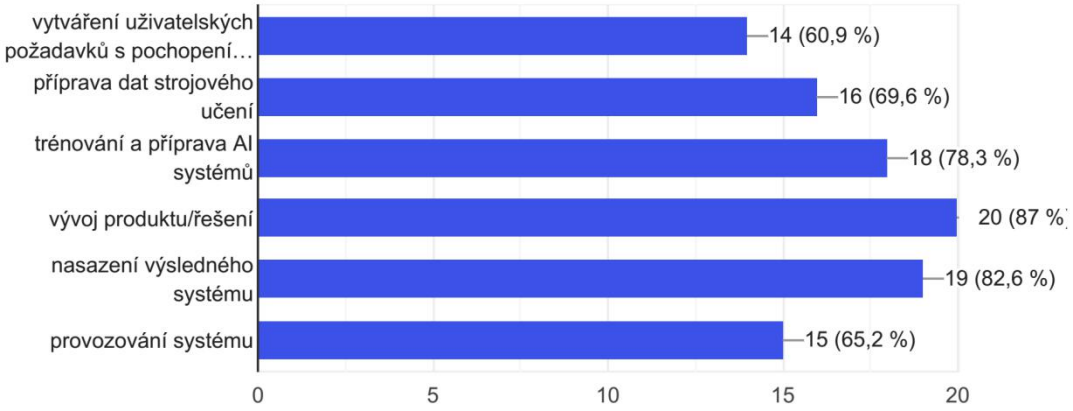
Výsledky dotazníkového šetření

1. V jaké fázi životního cyklu AI Vaše společnost působí?

23 odpovědí

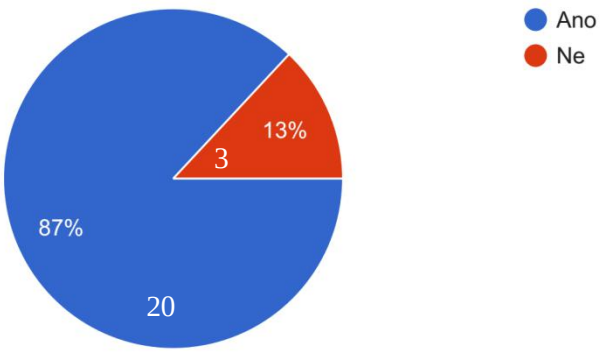
Odpovědi:

1. vytváření uživatelských požadavků s pochopením cílové domény
2. příprava dat strojového učení
3. trénování a příprava AI systémů
4. vývoj produktu/řešení
5. nasazení výsledného systému
6. provozování systému



2. Při práci s daty v rámci výše uvedených fází berete v potaz GDPR? Pracujete s osobními údaji?

23 odpovědí

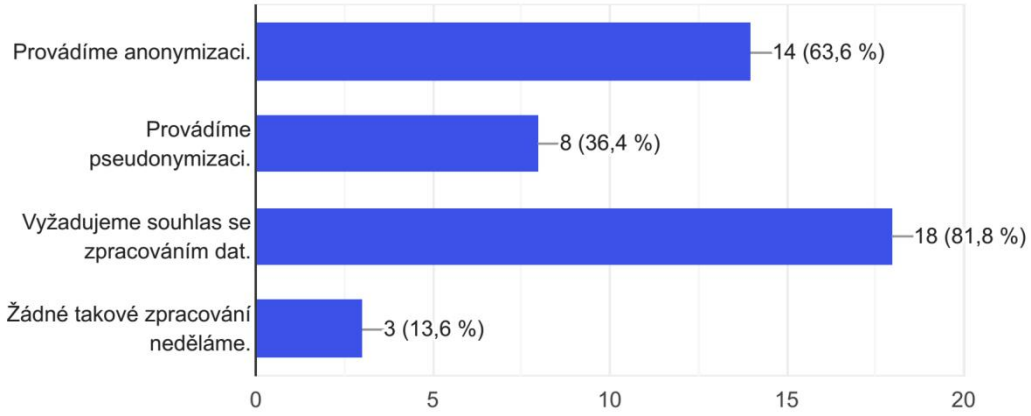


3. Pokud ano, osobní údaje podrobujete anonymizaci či pseudonymizaci anebo vyžadujete souhlas se zpracováním dat?

22 odpovědí

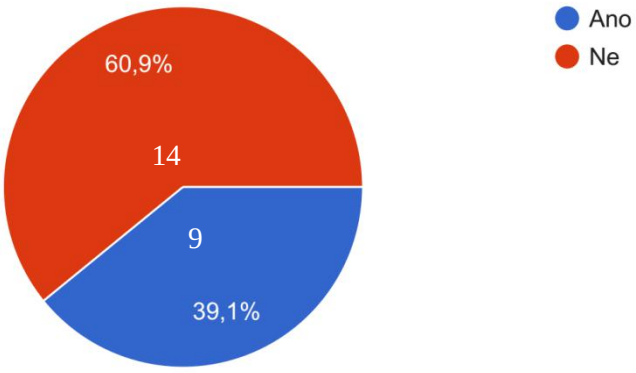
Odpovědi:

1. Provádíme anonymizaci.
2. Provádíme pseudonymizaci.
3. Vyžadujeme souhlas se zpracováním dat.
4. Žádné takové zpracování neděláme.



4. Jsou vaše systémy nasazovány v oblastech, kde při jejich provozování existují rizika spojená s lidskoprávní problematikou? Primárně se jedná o právo na soukromí, zákaz diskriminace a právo na spravedlivý proces (zejména přístup k důkazům) a sekundárně např. o právo na vzdělání, právo na práci, právo na sociální zabezpečení, právo na zdraví apod.

23 odpovědí



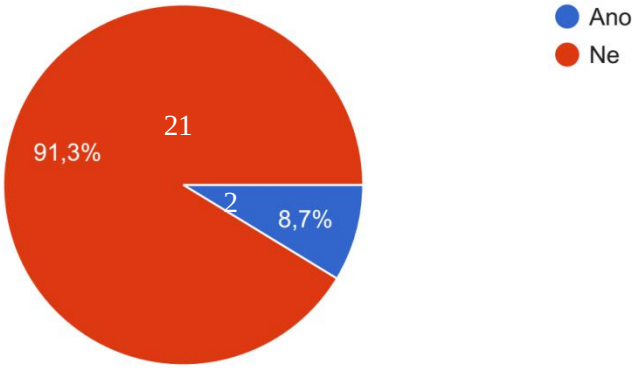
Pokud jste v předchozí otázce odpověděli „ano“, prosím upřesněte.

9 odpovědí

- Bias rekomendačních systémů není asi tak závažný jako uvedené příklady, ale může mít negativní vliv třeba na polarizaci společnosti.
- Uživatelé sdílí se systémem svou polohu skrze aplikaci (soukromí).
- Věnujeme se datům ze zemědělství, často v rozvojových zemích, kde je ohrožena potravinová bezpečnost. Naše výstupy mohou být relevantní pro některá základní lidská práva − na důstojný život, na zdraví.
- Sbíráme data pro hodnocení rizika konkrétních entit ve finančních institucích.
- Bias napříč populací. Výskyt specifických nálezů u různé populace.
- Zabezpečení finančních systémů.
- Neřešíme to vůbec přímo, ale pracujeme se zdravotnickými daty a potenciálně se to někdy těchto věcí může dotýkat (např. algoritmus hledající optimální párování pacientů, tzn. de facto pomáhající rozhodnout, kdo dostane ledvinu a kdo ne). I když tento algoritmus zrovna není AI, ale optimalizační. U těch AI věcí asi přímo ne.
- Spolupracujeme s klinikami, firmami v developerské a konstrukční činnosti, mobilními operátory, bankami.
- Detekce osob v bezpečnostních systémech autonomního řízení musí fungovat na všech rasách a nediskriminovat podle toho, co máme v trénovací sadě a co ne.

5. Pracujete s technologií rozpoznávání obličejů?

23 odpovědí



Pokud jste v předchozí otázce odpověděli „ano“, prosím upřesněte, jak zajišťujete, aby nedošlo k riziku různorodé kvality rozpoznání obličejových charakteristik s ohledem na rasu, barvu pleti a etnikum.

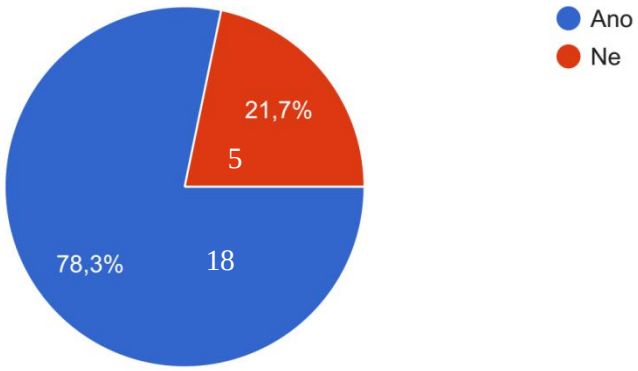
2 odpovědi

- V našich řešeních charakteristiky typu rasa, barva pleti a etnikum neidentifikujeme.

- Riziko různé kvality rozpoznání obličejových charakteristik na předem nespecifikované podmnožině osob nelze nikdy vyloučit. V principu lze vždy pro danou rozpoznávací metodu (rozpoznávací klasifikátor) najít nějakou podmnožinu osob, na které bude kvalita rozpoznání horší než na jiných. Například se může ukázat, že daná verze klasifikátoru má horší úspěšnost rozpoznání na osobách s novým nezvyklým typem účesu. Daný účes může být navíc specifický pro určitou úzkou skupinu lidí (např. mladí fanoušci fotbalového klubu XX). Rázem tedy máme „diskriminující“ klasifikátor. (Navíc není zřejmé, koho tato vlastnost vlastně diskriminuje, ty s tímto účesem nebo všechny ostatní?) Vzhledem k tomu, že nelze předem specifikovat všechny možné skupiny lidí, nelze ani vytvořit nediskriminující metodu. Pouze lze snížit riziko rozdílné kvality rozpoznávání u společností aktuálně citlivě vnímaných podmnožin/skupin, které jsou předem známy. Obecně se toto riziko snižuje co nejlepším vyvážením trénovací a testovací sady s ohledem na společensky citlivé charakteristiky, jako je rasa, barva pleti, etnikum, gender apod. Hlavní riziko vnímáme spíše ve způsobu užití daného produktu.

6. Používáte transfer learning?

23 odpovědí

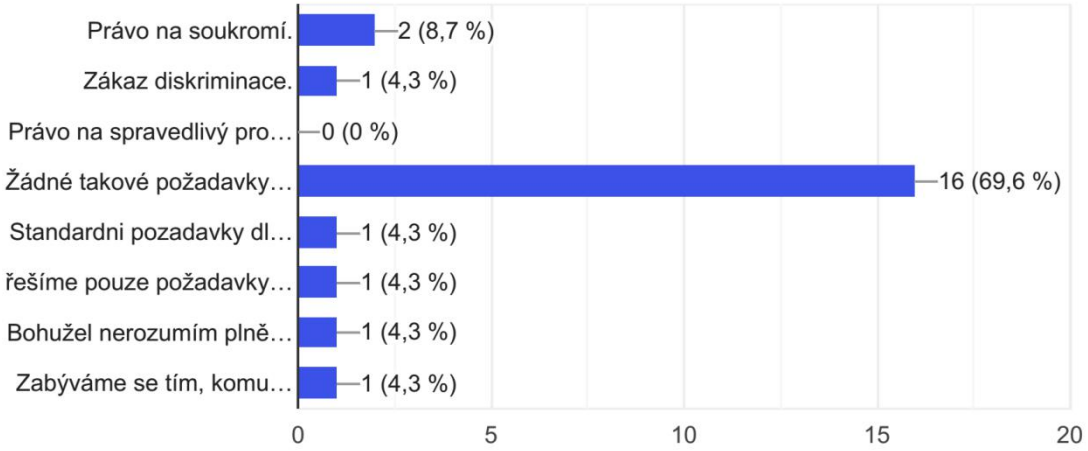


7. Zpracováváte v uživatelských požadavcích požadavky na zapracování a ošetření rizik souvisejících s lidskoprávní problematikou přicházející od externích subjektů? Specifikujte prosím oblasti lidského práva:

23 odpovědí

Odpovědi:

1. Právo na soukromí.
2. Zákaz diskriminace.
3. Právo na spravedlivý proces (zejm. přístup k důkazům).
4. Žádné takové požadavky v uživatelské specifikaci nezpracováváme.
5. Jiné:
    1. Standardní požadavky dle GDPR.
    2. Řešíme pouze požadavky, které splňují naše přísná etická kritéria.
    3. Bohužel nerozumím plně otázce. V rámci GDPR a dalších certifikátů se snažíme splňovat všechny požadavky.
    4. Zabýváme se tím, komu a do jakých regionů naše produkty dodáváme. Neprodáváme naše produkty do zemí, které jsou dle Democracy Index označeny za autoritářské.



Pokud takové požadavky zpracováváte, můžete nám uvést konkrétnější příklady, čeho se týkají a proč je uživatelé vyžadují?

5 odpovědí

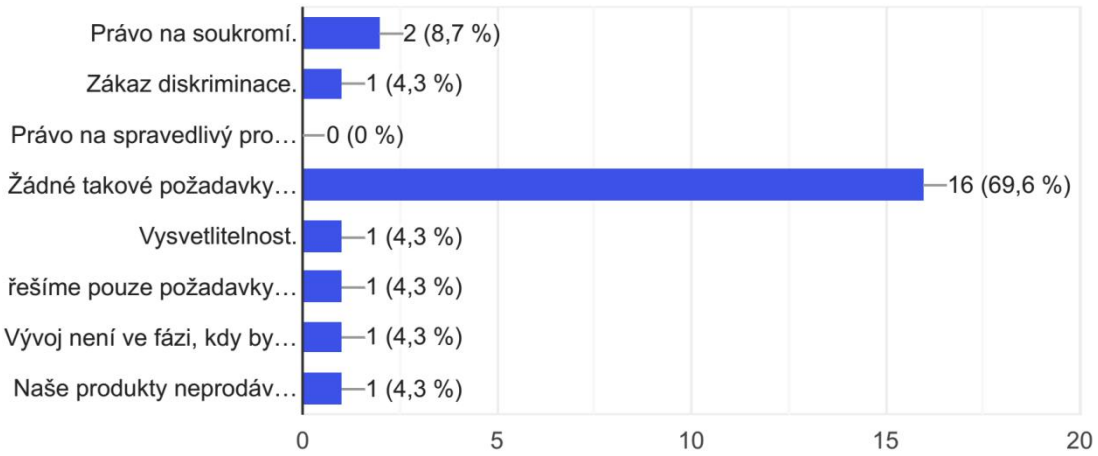
- Odstranění dat, vysvětlitelnost verdiktu, poskytnutí důkazů pro soud v případě odhaleného podvodu.
- Odchozí telefonní kampaně.
- Většinou předpokládáme, že při využití externích datasetů/pre-trained modelů byly data/modely z tohoto pohledu ošetřeny.
- Právo na soukromí se snažíme implementovat dle doporučení klientů, která vychází z doporučení a směrnice ÚOOÚ a EU. My zasahujeme ve chvíli, když dle našich znalostí vidíme rozpor mezi požadovaným řešením a těmito nařízeními.
- Bezpečnost systémů nasazovaných v autonomním řízení vyžaduje správnou funkcionalitu nehledě na rasu člověka.

8. Vytváříte a zapracováváte vlastní požadavky související s lidskoprávní problematikou při vývoji AI systému?

23 odpovědí

Odpovědi:

1. Právo na soukromí.
2. Zákaz diskriminace.
3. Právo na spravedlivý proces (zejm. přístup k důkazům).
4. Žádné takové požadavky v uživatelské specifikaci nezpracováváme.
5. Jiné:
    1. Vysvětlitelnost.
    2. Řešíme pouze požadavky, které splňují naše přísná etická kritéria.
    3. Vývoj není ve fázi, kdy by bylo toto na pořadu dne. Etika AI je pravidelný bod „all-hands“ agendy alespoň jednou měsíčně.
    4. Naše produkty neprodáváme do zemí, které jsou dle Democracy Index označeny za autoritářské.



Pokud takové požadavky zapracováváte, můžete nám uvést konkrétnější příklady, čeho se týkají a proč jste je zahrnuli do specifikací?

4 odpovědi

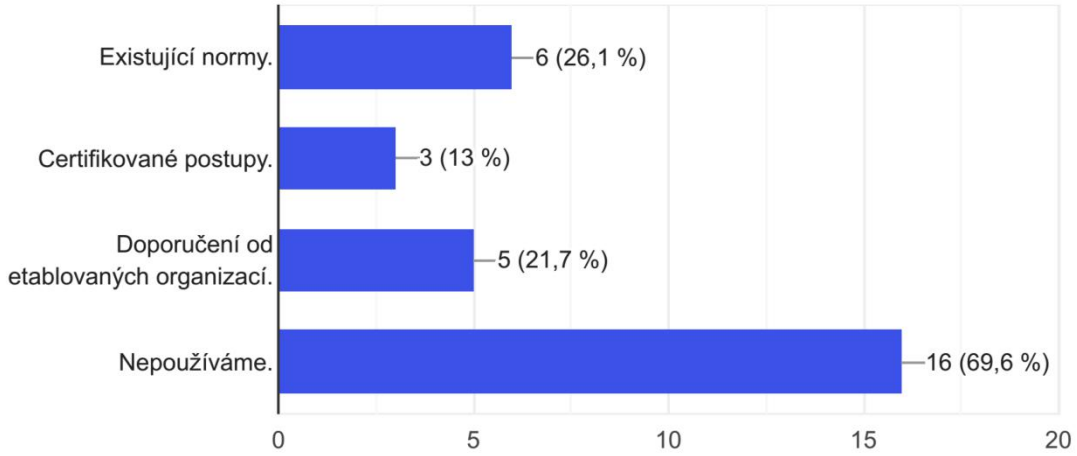
- Snažíme se modely „debiasovat“, ale často je to obtížné, protože biasy jsou v trénovacích datech a vycházejí z chování uživatelů.
- Vzhledem k doméně klademe důraz na vysvětlitelnost rozhodnutí. Data jsou vlastnictvím (tzn. pod správou) našich zákazníků a přístup k důkazům a další aspekty řeší přímo oni.
- Odchozí telefonní kampaně.
- Požadavky na vymazání dat uživatele.

9. Pokud ve vašem systému zapracováváte požadavky související s lidskoprávní problematikou, používáte k tomu:

23 odpovědí

Odpovědi:

1. Existující normy.
2. Certifikované postupy.
3. Doporučení od etablovaných organizací.
4. Nepoužíváme.



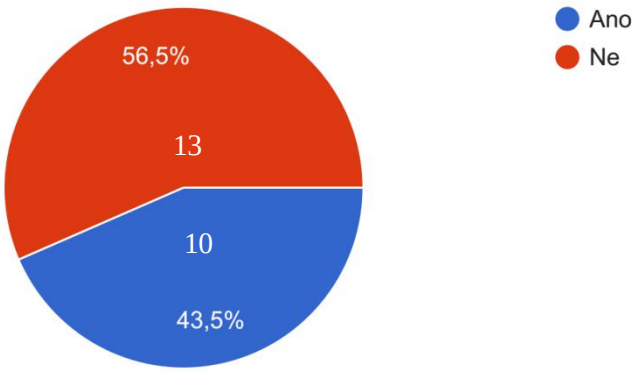
Pokud ano, můžete uvést konkrétní příklady?

6 odpovědí

- Máme vlastní rozsáhlý etický výzkum; systematicky mapujeme normy z různých krajů.
- Neexistuje ani v mezinárodním měřítku, děláme v tomto směru výzkum.
- GCR, ISO 13485
- Řešíme jen technické otázky a dáváme doporučení zákazníkům, kteří přijímají finální rozhodnutí.
- GDPR, ISO 27XYZ
- Našimi klienty jsou často velké korporace, které do řešení těchto otázek investují poměrně dost prostředků včetně doporučení od etablovaných organizací. Získané informace využíváme i pro diskuse s jinými klienty.

10. Berete v potaz lidskoprávní prvek na úrovni softwarového vývoje – např. selekce vstupních dat či atributů, aby neobsahovaly tzv. chráněnou hodnotu (informaci o etniku, pohlaví, náboženství apod.), či zajištění rovného početního zastoupení dat každé takové kategorie s cílem eliminovat riziko diskriminace apod.?

23 odpovědí



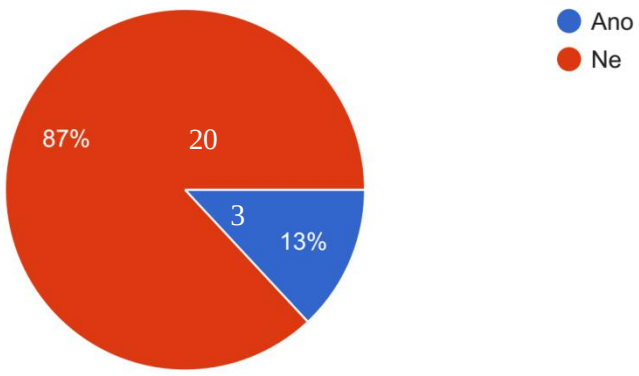
Pokud ano, prosím upřesněte:

11 odpovědí

- Podle typu výzkumu.
- Nepracujeme s atributy uživatelů.
- Abychom zajistili správnou funkci systému, musíme trénovat na datech s co největším množstvím a rozmanitostí variant.
- Nesbíráme informace s chráněnou hodnotou.
- Sestavení datasetů z různých pracovišť napříč trhy, kde plánujeme systém dodávat.
- Nepracujeme s obsahem dokumentů a transakcí, ale s technickými parametry a vztahy.
- Genderově neutrální design hlasového rozhraní, práce s daty.
- Zajištění rovného početního zastoupení dat vzhledem k předem zvoleným kategoriím, pokud je to dosažitelné, s cílem eliminovat riziko diskriminace.
- Dal jsem ne, ale v našem případě například pohlaví někdy může být dobrý prediktor nějakých zdravotnických věcí a je to podle nás naprosto relevantní používat. V žádném produkčním systému to ale zatím nemáme.
- Většinou jde o vyfiltrování informací, které nelze použít.
- Např. v detekci lidí v kamerových datech se naše metody soustředí na rozpoznání atributů popisujících člověka obecně a nikoliv takových, které jsou validní jen pro určitou rasu. Potřebujeme spolehlivě detekovat všechny.

11. Řešili jste provozní problémy spojené s narušením lidských práv a případně požadavky na úpravu systému s tím související? A to i v případě, že jste při vývoji AI systému a uvedení do praxe nezapracovávali požadavky související s lidskoprávní problematikou.

23 odpovědí



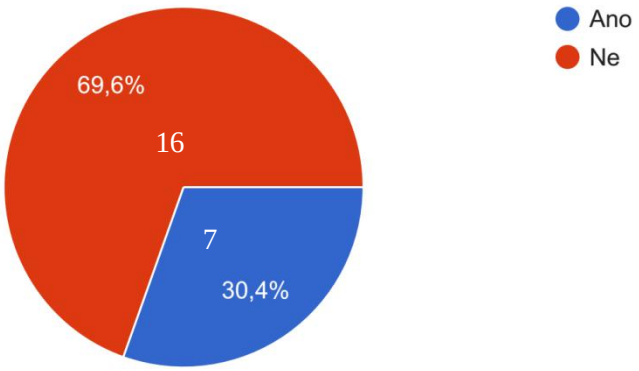
Pokud ano, prosím upřesněte:

3 odpovědi

- Odmítáme výzkum a zakázky, které nesplňují naše etické normy.
- Ano, přestali jsme užívat volně dostupné datasety a navrhli vlastní postupy.
- Úpravy architektury a designu telefonních řešení.

12. Máte v při vývoji AI systému v testech a testovacích specifikacích zahrnuty požadavky související s lidskoprávní problematikou?

23 odpovědí



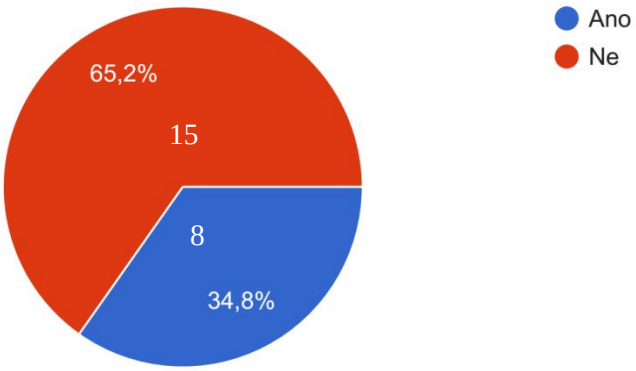
Pokud ano, můžete uvést příklady?

7 odpovědí

- Různé.
- Podporujeme vývoj open source repsys pro odhalování biasu v rekomendačních systémech.
- Testy na omezení diskriminace.
- Anonymizace dat. Vysvětlitelnost výstupů systému.
- Pouze na úrovni zajištění rovného početního zastoupení dat v testovacích sadách.
- Tak záleží to na oblasti, ale příkladem může být potřeba smazání dat na požadavek klienta. Existenci vstupních dat můžeme smazat hned, ale nelze to jednoduše vyjmout z natrénovaného modelu.
- Anonymizace dat. Navíc musíme být schopni vyhledat a smazat nahrávku konkrétních lidí, pokud by si to přáli. Naše nahrávací auta jsou označena a poskytují kontaktní informace.

13. Používáte nějakou formu monitorování systému za provozu, která by umožňovala předcházet rizikům a zachycovat porušení související s lidskoprávní problematikou? Může se jednat o jakoukoliv formu od logování, přes online vyhodnocování výstupů systému až po auditování systému.

23 odpovědí



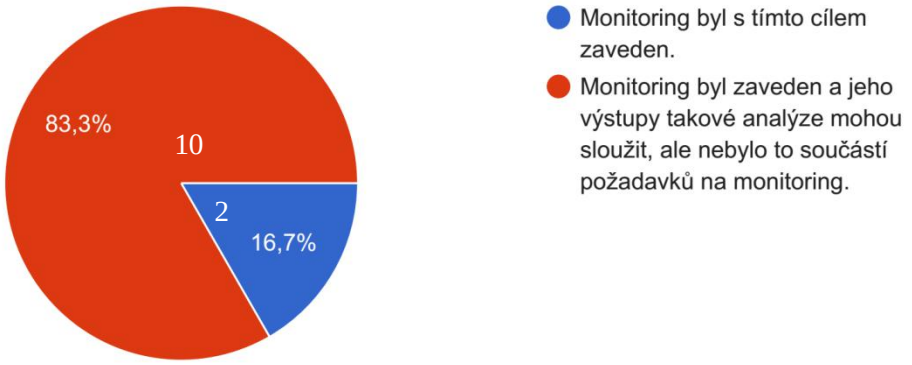
Pokud ano, můžete uvést příklady?

9 odpovědí

- Dnes převážně řešení třetích stran + ruční monitoring přes QC.
- Online je to moc těžké, jsme rádi za offline zpětné vyhodnocení a předcházení při trénování.
- Neprovozujeme tyto systémy.
- Sledujeme logování a provádíme ad-hoc testy.
- Logs. Response loop.
- Logging, monitoring, assisted continuous learning.
- Máme normální logging a monitoring serverů a přicházejících dat na AWS, ale neslouží to přímo k monitorování za účelem lidskoprávní problematiky.
- Monitorujeme, ale často nelze zabránit dané situaci. Většinou se zjistí problém až po vložení vstupních dat, někde pouze ze zpětné vazby osoby, které se to týká.
- Anonymizace obličejů a registračních značek.

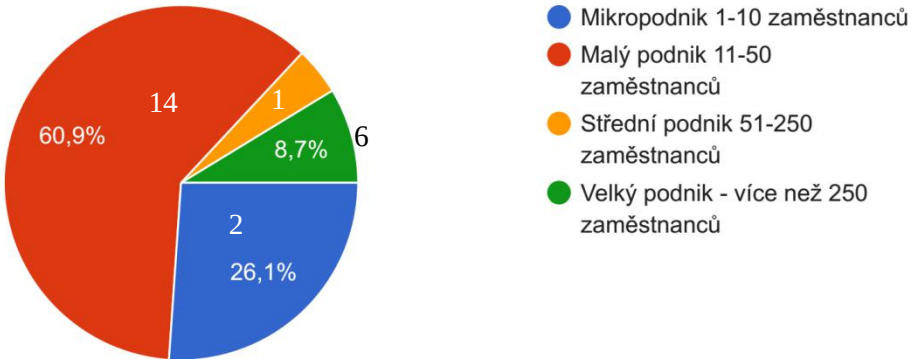
14. Pokud nějaký systém monitoringu existuje, byl zaveden, kromě jiného, také s cílem předcházet rizikům a zachycovat porušení související s lidskoprávní problematikou?

12 odpovědí



15. Velikost Vaší pobočky (v případě zahraničního subjektu) nebo firmy (v případě českého subjektu) v České republice:

23 odpovědí



Prostor pro Vaše případné komentáře a/nebo podněty k dotazníku a/nebo celkovému projektu.

4 odpovědi

- Nepracujeme s daty, která jsou obvykle spojována s AI a lidskoprávní problematikou (tj. hlavně data o lidech, obličeje atp.), ale přesto mohou mít naše výstupy lidskoprávní dopad (např. v oblasti potravinové bezpečnosti nebo důstojného života).
- Naše systémy jsou ve vývoji, nejsou v produkci. Uživatelská data dosud nezpracováváme.
- Myslím, že my se pohybujeme tak nějak okolo tohoto tématu, ale aktuální to teď pro nás úplně není.
- Ve 3 jsem vybral „Vyžadujeme souhlas se zpracováním dat“ a zároveň „Žádné takové zpracování neděláme“, neboť souhlas zajišťuje Správce, nikoli my jakožto Zpracovatel.