Úloha a dôležitosť etických pravidiel v systémoch umelej inteligencie

Klučka, J.[*]

KLUČKA, J.: Úloha a dôležitosť etických pravidiel v systémoch umelej inteligencie.[**] Právny obzor, *108*, 2025, č. 3, s. 240 – 254. **https://doi.org/10.31577/pravnyobzor.2025.3.02**

The role and importance of ethical rules in artificial intelligence systems. In the field of artificial intelligence, its ethical rules represent a significant part, which is understood as a set of values, principles and techniques using generally accepted standards of „good and bad" and regulating ethical behaviour in the process of developing and using AI technologies. Ethical rules create a delicate balance between technological progress and the protection of human values such as privacy, justice and transparency. The real application of ethical values and principles is however possible only after their implementation in AI systems, in each phase of the „AI life cycle" from the initial design through its practical deployment to monitoring. Currently, there is neither general nor uniform practice about what method or what mechanisms should be used to implement ethical rules into AI systems. Common to the implementation procedures is however the fact that they focus on AI algorithms that immediately generate final decision of AI system containing an ethical rule used (applicable) in a specific situation. For this purpose are algorithms trained on a large amount of input data, which, however, may contain bias caused by humans in the preparation and the application of AI. If they are not properly addressed, the practical application of such „infected" algorithms can lead to discriminatory consequences and the strengthening of social inequality. These shortcomings should be identified by AI audit, the aim of which is to assess whether individual AI systems fulfil the expected functions while respecting the corresponding ethical and legal rules. However, there are currently no internationally accepted standards of procedure in the field of AI auditing. This is caused by e.g. insufficient adjustment of AI (since its very definition is still ambiguous), so the creation of common standards for its audit seems problematic. Although they can be identified on the basis of and through an audit, this alone is not capable of actually eliminating such algorithms or to block their application. System biases found their way into AI because, like all technologies, they were created by humans. As a result, the final results of audits reflect and preserve human biases and errors in society. Therefore, it is generally accepted that the only way how to "combat" AI bias is currently by rigorously testing and evaluating its data and algorithms, and using best practices for collecting and using data and creating AI algorithms.

Key words: *protection of human rights and moral values by ethical rules of artificial intelligence, implementation of ethical rules in artificial intelligence systems, role of audit in artificial intelligence, efforts to ensure impartial artificial intelligence free of bias*

Úvod

Článok je venovaný miestu a úlohe, ktorých význam značne narastá v dôsledku dynamického vývoja umelej inteligencie a rozširovania jej pôsobnosti do oblastí pôvodne

[*] Prof. JUDr. Ján Klučka, CSc., Ústav medzinárodného a európskeho práva, Právnická fakulta UPJŠ, Košice.

[**] Dielo bolo vytvorené bez finančnej podpory grantových schém.

vyhradených ľudským aktivitám. V jeho prvej časti autor prináša všeobecnejšiu charakteristiku významu a úlohy etických pravidiel v kontexte umelej inteligencie a venuje sa posudzovaniu všeobecných dokumentov právne nezáväznej povahy obsahujúcich etické pravidlá pre umelú inteligenciu, ako aj vzťahu etických pravidiel k relevantnej právnej úprave. Druhá časť je venovaná problémom implementácie etických pravidiel do systémov umelej inteligencie, auditu systémov umelej inteligencie a eliminácii skreslených algoritmov umelej inteligencie.

1. Význam a úloha etických pravidiel v kontexte umelej inteligencie

Pred časom sa autor na inom mieste pokúsil o analýzu vzťahu medzinárodného práva a umelej inteligencie a *vice versa.* Jeden zo záverov, ku ktorému dospel, bol ten, že: „Vzhľadom na možnú úlohu medzinárodného práva sa oprávnene poukazuje na skutočnosť, že „[r]*ealita je taká, že v najlepšom prípade bude až počas niekoľkých rokov možno zaregistrovať postupne vznikajúcu tradičnú úpravu AI, a to aj napriek jej rastúcemu rozširovaniu a uplatňovaniu. Aspoň dočasne bude takáto medzera právnej úpravy vypĺňovaná pravidlami soft law.“*[1] V tejto súvislosti možno uviesť, že formujúca sa právna úprava umelej inteligencie (ďalej len AI) nie je jediná, ktorá sa vzťahuje na AI. Vo všeobecnosti sa uznáva, že súčasťou komplexu pravidiel vzťahujúcich sa na AI sú aj *etické pravidlá,* ktoré sú s právnymi pravidlami vo vzájomnej interakcii. Etické pravidlá obsahujú právne nezáväzné pravidlá spoločné pre určité spoločenstvá ľudí, ktoré vytvárajú kultúru ich života, pričom ich rešpektovanie nie je právne vynútiteľné a sankcionovateľné. Zatiaľ čo etické pravidlá požadujú, *ako by sa mal* príslušný subjekt vo vzťahu k AI správať, právne pravidlá určujú, ako je príslušný subjekt *povinný sa* správať pod hrozbou sankcie. Napriek tomu, že etické pravidlá nie sú právne záväzné, obsahujú sociálne a morálne princípy a hodnoty, ktoré sú osobitne významné aj vo vzťahu k AI pre úpravu jej vzťahov k ľuďom a naopak. O reálnom etickom rozmere AI technológií možno hovoriť vtedy, ak by pristúpili (akceptovali) na systém etických pravidiel a používali (zohľadňovali) by ich vo svojej činnosti. Ak by sa uvažovaná akcia ocitla v rozpore s etickými pravidlami, AI by ju nemala vykonať. Pôjde teda o činnosti (akcie) spočívajúce v prijímaní rozhodnutí so značnými dôsledkami, ktoré ľudia budú považovať za etické alebo morálne. Existujú viaceré pokusy o definíciu etiky AI, ktoré majú spoločnú črtu v tom, že ju charakterizujú ako súbor hodnôt, princípov a techník používajúcich všeobecne akceptované štandardy „dobra a zla“ a upravujúce etické postupy v procese rozvoja a používania technológií AI. Účelom etického správania AI je ochrana morálnych hodnôt, ktoré sú všeobecne prijímané ako súčasť ľudskej povahy a prirodzenosti (ochrana ľudského života, sloboda, ľudská dôstojnosť a i.), ako aj viery a presvedčenia určitých spoločenstiev obyvateľov (náboženské, prípadne etnické obrady a pravidlá). Na rozdiel od

[1] KLUČKA, J. Medzinárodné právo, umelá inteligencia a viceversa. In *Časopis pro právní vědu a praxi.* Roč. XXIX, 2021, č.3, s. 553. ISSN 1210- 9126. Dostupné na: https://doi.org/10.5817/CPVP2021-3-4. Analýzu právnej úpravy umelej inteligencie priniesol v r. 2024 aj Právny obzor.

PINTEROVÁ, D. Právna regulácia umelej inteligencie (perspektívy a výzvy). In *Právny obzor.* Roč. 107, 2024, č. 4, s. 361 – 383.

právnych pravidiel, ktoré postupne vznikajú ako reagencia na rýchlo sa rozvíjajúci fenomén AI, etické pravidlá tvoria súčasť nadstavbovej úpravy jednotlivých spoločenstiev, ktoré sú v súčasnosti konfrontované s AI. Etické pravidlá v takomto kontexte vytvárajú „*krehkú rovnováhu medzi technologickým pokrokom a ochranou ľudských hodnôt, akými sú súkromie, spravodlivosť, transparentnosť*“.[2] Vzhľadom na svoje špecifiká v súčasnosti neexistuje všeobecne uznávaný systém etických pravidiel vzťahujúcich sa na umelú inteligenciu, ale systémy (regionálne či lokálne) aplikovateľné v špecifických národných, geografických, náboženských a etnických kontextoch. Vo vzťahu k AI sa preto nevynárajú otázky ich tvorby, ale skôr miesta, významu a možného výkladu a uplatňovania vo vzťahu k špecifikám vývoja, zavádzania a používania AI. Význam, úloha a rôznorodosť etických pravidiel vzrastá v dôsledku toho, ako sa AI dostáva bližšie k ľuďom a prostrediu, v ktorom vykonávajú svoje činnosti (nemocnice, domovy pre seniorov, autá bez vodičov a pod.),v ktorých dochádza k úzkym kontaktom (zdieľaniu) medzi ľuďmi a AI. Etické pravidlá sa tak uplatňujú v systémoch AI a v situáciách, keď by v rôznych prostrediach mohli úplne alebo čiastočne nahradiť človeka, resp. keď vykonávajú činnosti, ktoré by boli za „normálnych“ okolností vyhradené ľuďom, ktorým je etický aspekt ich činností vlastný. Nie je úplne jasné, akým spôsobom, prípadne v akom rozsahu môže takáto „kohabitácia“ v budúcnosti postihovať emočnú oblasť a spôsoby ľudského myslenia a vplývať na vzájomné pôsobenie a ovplyvňovanie ľudí v ich vzťahoch s AI. V širšom kontexte vývoj v oblasti takéhoto používania AI navodzuje aj ďalšie otázky, ako napríklad, čo je ľudský život, čo je ľudskosť, čo znamená ľudský život a dôstojnosť a aký je vzťah človeka k systémom AI. Vo vzťahu k rýchlo sa rozširujúcim (a pôvodne výlučne ľudským) oblastiam pôsobenia AI dokonca vyvstáva otázka, či by niektoré z nich, týkajúce sa etiky a morálnych hodnôt ľudí, nemali byť vylúčené, prípadne obmedzené z použitia systémov AI, prípadne stanoviť medze ich proporcionálneho využitia. Odborná spisba v tejto súvislosti pripomína, že: „*Čo sa v mladej, rýchlo rastúcej digitálnej civilizácii mení, je to, že rozhodnutia o našom individuálnom, rodinnom alebo spoločenskom živote delegujeme na technológie. V dôsledku toho možno napokon ľudskú existenciu „subdodávateľsky“ zabezpečiť AI softvérom.*“[3] Neobmedzená dôvera v pomoc a podporu AI môže viesť k postupným stratám poznávacích ľudských schopností a dokonca k celkovému úpadku ľudskej inteligencie, autonómie a integrity (*tzv. digitálna demencia*). Ako dôsledky nadmerného užívania progresívnej technológie možno uviesť internet a mobilné telefóny, ktoré u užívateľov vytvárajú silné závislosti, ktoré je potrebné liečiť. Ľudský rozmer etických pravidiel nemožno identifikovať v prostriedkoch AI pôsobiacich v rôznych „nezaľudnených“ zemských priestoroch, resp. v kozmickom

[2] KIROVA, V. D., KU, C. S., LARACY, J. R., MARLOWE, T. J. The Ethics of Artificial Intelligence in the Era of Generative AI. In *Journal of Systemics, Cybernetics and Informatics* [online]. 2023, č. 4 [cit. 22. 11. 2024]. Dostupné na: https://www.iiisci.org/DOIJSCI/SA445KR23/#/;https://doi.org/10.54808/JSCI.21.04.42.

[3] WALTZ, A., FIRTH-BUTTERFIELD, K. Implementing Ethics into Artificial Intelligence: A Contribution, from a Legal Perspective, to The Development of an AI Governance Regime. [online]. 2019 [cit. 22. 11. 2024]. Dostupné na: https://www.researchgate.net/publication/338083064_IMPLEMENTING_ETHICS_INTO_ARTIFICIAL_INTELLIGENCE_A_CONTRIBUTION_FROM_A_LEGAL_PERSPECTIVE_TO_THE_DEVELOPMENT_OF_AN_AI_GOVERNANCE_REGIME.

priestore a na kozmických telesách, pretože v nich absentuje interakcia s ľuďmi. Ide o dôsledok ľudsky orientovanej (*human oriented)* AI, ktorá vznikla a rozvíja sa s cieľom zlepšovať životné podmienky ľudí, ich slobody a tiež napomáhať rôznym potrebám medzinárodného spoločenstva. Cieľom etických projektov AI je napomôcť, aby sa AI systémy správali eticky alebo sa riadili etickými a morálnymi princípmi a pravidlami, a tým (spolu s právnymi pravidlami) znižovali riziká, ktoré pre ľudí a spoločnosť nesú so sebou systémy AI.

Konečným cieľom AI etiky je príprava a vytvorenie optimálneho modelu vzájomného vzťahu medzi ľuďmi a systémami AI. Odborná spisba pripomína, že: „*Nové technológie by sa mali využívať na ďalšiu podporu a zavádzanie etických hodnôt a princípov ako základných smerníc pre náš každodenný život a mali by sa použiť na rozvoj dobrej spoločnosti AI*."[4] Uplatňovanie etických hodnôt a princípov je však možné len pri eliminácii rizík, ktoré prináša AI, keďže: „*Napriek mnohým výhodám AI prináša značné riziká, ktoré je potrebné zvládnuť. Minimalizácia týchto rizík zdôrazní výhody a zároveň ochráni etické hodnoty definované základnými právami a základnými ústavnými princípmi, čím sa chráni ľudská spoločnosť*."[5]

1.1 Medzinárodné dokumenty právne nezáväznej povahy upravujúce etické pravidlá AI

Doterajšie úsilie medzinárodného spoločenstva sa vo vzťahu k etickým pravidlám AI o. i. sústreďovalo na navrhovanie a prípravu právne nezáväzných dokumentov (*guidelines*) identifikujúcich etické pravidlá, ktoré by mali byť rešpektované. Všeobecnou požiadavkou na akýkoľvek etický program systému AI je, aby pravidlá v ňom obsiahnuté boli zlučiteľné so zásadami ľudskej dôstojnosti, integrity, slobody, ochrany súkromia, kultúrnej či rodovej rozmanitosti, ako aj s ďalšími ľudskými právami. Táto úloha nepredstavuje vážnejší problém a samotná prax potvrdzuje rastúci počet takýchto právne nezáväzných usmernení.

V posledných rokoch bol prijatý celý súbor etických smerníc obsahujúcich princípy, ktoré by vývojári AI technológií mali v čo najväčšej miere dodržiavať.[6] Súčasná odborná spisba prináša celkom 146 etických *guidelines* prijatých rôznymi organizáciami, spoločnosťami a vládami v celosvetovom meradle.[7] Prax potvrdzuje, že tieto pravidlá môžu mať rôzne formy, počínajúc dobrovoľnými programami, štandardmi, kódexmi správa-

[4] CATHE, C., WACHTER, S., MITTELSTADT, B., TADDEO, M., FLORIDI, L. Artificial Intelligence and the 'Good Society': the US, EU, and UK approach. *Science and Engineering Ethics* [online]. 2018, č. 24 [cit. 22. 11. 2024]. Dostupné na: https://link.springer.com/article/10.1007/S11948-017-9901-7; https://doi.org/10.1007/s11948-017-9901-7.

[5] WALTZ, A., FIRTH-BUTTERFIELD, K., 2019.

[6] HAGENDORFF, T. The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds&Machines* [online]. 2020, č. 30 [cit. 22. 11. 2024]. Dostupné na: https://link.springer.com/article/10.1007/s11023-020-09517-8; https://doi.org/10.1007/s11023-020-09517-8.

[7] HUANG, CH., ZHANG, Z., MAO, B., YAO, X. An Overview of Artificial Intelligence Ethics.*IEEE Transactions on ArtificialIntelligence* [online]. 2023, č. 4 [cit. 22. 11. 2024]. Dostupné na: https://scholars.ln.edu.hk/en/publications/an-overview-of-artificial-intelligence-ethics; https://doi.org/10.1109/TAI.2022.3194503.

nia, certifikačnými súbormi, modelovými zákonmi, usmerneniami, deklaráciami zásad a princípov, ktoré nevyžadujú žiadne spôsoby formalizovaného súhlasu. Autormi takýchto dokumentov sú rôzne nevládne organizácie, akademický sektor, vedeckovýskumné inštitúcie, ako aj subjekty z technologickej oblasti, resp. z oblasti priemyslu. Do prvej skupiny možno zaradiť Princípy AI z Asilomaru (2017), Montrealskú deklaráciu o zodpovednom rozvoji AI (2017), Harmonizované princípy AI (2018). Medzi pravidlami z priemyslového prostredia možno spomenúť Google a AI: Naše princípy (2018), AI princípy Microsoftu (2018), Skupina Sony a sprievodca etickými princípmi umelej inteligencie (2018), IBM princípy pre poznávaciu éru (2017), IEEE Globálna iniciatíva o Etike Autonómnych a inteligentných systémov (2016, 2019) a i. Z vyššej úrovne úpravy možno spomenúť Etické usmernenia pre dôveryhodnú AI Európskej komisie (2018), Pekinské princípy AI (2019), OECD Princípy AI (2019), Odporúčania UNESCO o Etike Umelej Inteligencie (2021) a pod. Tieto dokumenty definujú sociálne a etické hodnoty potrebné pre bezpečný rozvoj AI v budúcnom období (humanita, spolupráca, bezpečnosť, transparentnosť, ochrana súkromia, rešpektovanie ľudskej dôstojnosti a autonómie) a pod. Odlišujú sa v počte a obsahu etických pravidiel a tiež v definovaní technických postupov, ktorých použitím by malo byť dosiahnuté rešpektovanie etických princípov. Obsahovo sú formulované všeobecne a nevenujú pozornosť špecifickým etickým problémom, ktoré môžu vznikať v konkrétnej AI praxi (napríklad autá bez vodičov). Napriek rozdielnostiam, všetky zdôrazňujú potrebu rešpektu k základným ľudským právam, ľudskej dôstojnosti, rovnosti a zákazu diskriminácie. Keďže akceptovanie takýchto pravidiel spočíva na dobrovoľnom prístupe, vzniká riziko ich výberového (selektívneho) použitia založené na individuálnom záujme konkrétneho užívateľa namiesto všeobecnej akceptácie. Ďalší problém predstavuje možnosť výberu medzi existujúcimi etickými súbormi spracovanými rôznymi nevládnymi subjektmi. V odbornej spisbe sa preto oprávnene poukazuje na potrebu nejakého koordinačného orgánu, ktorý by sa mal pokúsiť tieto nedostatky riešiť.[8] Pokiaľ ide o význam práva v tejto oblasti, toto by mohlo zohrávať úlohu pri harmonizácii a vyvažovaní rozdielnych etických modelov. Žiaden z týchto dokumentov nemá univerzálne či prioritné postavenie, pretože nikto nemôže v globálnom meradle záväzne rozhodovať, čo je etické a čo neetické.

Odpoveď na otázku, či tieto princípy majú aj reálny vplyv na ľudskú činnosť v oblasti vývoja a výroby AI, je prevažne *negatívna,* pretože etické pravidlá obsiahnuté v týchto dokumentoch v praxi reálne netvoria základ na prijímanie konkrétnych etických rozhodnutí pre navrhovateľov, vývojárov a výrobcov prostriedkov AI, prípadne iných profesionálov v komunite AI. Rôzne etické kódexy tiež nemôžu fungovať ako náhrady za etické posudzovanie a uplatňovanie systémov AI, ktoré musia zohľadňovať skutočnosti vyplývajúce z konkrétneho kontextu a ktoré nie je možné reflektovať vo všeobecných formuláciách a usmerneniach etických kódexov. Je to aj z toho dôvodu, že pozornosť sa v AI výskume a využívaní sústreďuje prevažne na ich inovácie a technologické aspekty. Od-

[8] MARCHANT, G. „Soft Law" Governance of Artificial Intelligence [online]. 2019 [cit. 22. 11. 2024]. Dostupné na: https://escholarship.org/content/qt0jq252ks/qt0jq252ks_noSplash_1ff6445b4d4efd438fd6e06cc2df4775.pdf?t=po1uh8.

borná spisba v tejto súvislosti upozorňuje, že: „*Inžinieri a vývojoví pracovníci vo sfére AI nie sú systematicky školení v etických otázkach a nie sú ani zmocnení zaoberať sa vo svojej praxi etickými otázkami*.“[9] Potvrdzuje sa tiež, že softvéroví inžinieri venujú viac pozornosti tomu, ako a či AI systém alebo výrobok úspešne plní svoju technologickú funkciu, než sociálnym a etickým dôsledkom jeho praktickej aplikácie. Etická agenda býva posudzovaná ako nejaký „cudzí prvok“ prichádzajúci z externého prostredia a týkajúci sa technologickej AI agendy len ako právne nezáväzný prívesok. Vzhľadom na stále rastúci globálny a sociálny význam AI *narastá napätie* medzi sociálnymi a spoločenskými záujmami (vrátane etických) a potrebami a požiadavkami na bezpečnosť AI na jednej strane a obchodnými, ekonomickými a finančnými záujmami zainteresovaných subjektov na strane druhej. Sústavné zanedbávanie etických dôsledkov výskumu a uplatňovania AI nielenže môže viesť k porušovaniu etických pravidiel, pretože jeho dôsledkom môžu vznikať negatívne sociálne javy, nerešpektovanie sociálneho poriadku a iné neželané spoločenské dôsledky spôsobujúce užívateľom AI viac škody ako úžitku. V kontexte AI sa uvedené konštatovanie vzťahuje predovšetkým na tzv. autonómne AI systémy schopné konať a rozhodovať nezávisle a bez ľudskej kontroly. Tieto skutočnosti signalizujú nevyhnutnosť tvorby „ľudských“ monitorovacích a kontrolných mechanizmov, ktoré by umožňovali neetické konania a rozhodnutia AI systémovo vylúčiť, zastaviť alebo obmedziť. Výzva na ich tvorbu sa však nevzťahuje len na AI profesionálov, ale tiež na ďalšie skupiny zainteresovaných subjektov, zástupcov občianskej spoločnosti a ďalších sociálnych skupín. Ako príklady možno uviesť porušovanie práva na súkromie, diskrimináciu, zvyšovanie nezamestnanosti, rôzne bezpečnostné riziká a i. Postupne sa rozširuje tendencia využívať AI na nelegálne aktivity, v ktorých absentujú etické aspekty a ktoré sú na škodu iným, ako aj spoločnosti. Uvedené skutočnosti potvrdzujú *existenciu rizík*, ktoré so sebou prináša rozvoj a rastúce uplatňovanie AI pre výrobcov a vývojárov AI a celú spoločnosť, pokiaľ v nich absentuje etický rozmer, a podčiarkuje naliehavú potrebu venovať sa etickým otázkam AI v rastúcej miere a intenzite. Pretrvávajúce tolerovanie neetického správania rôznych technológií AI môže mať za následok postupnú stratu dôvery verejnosti ako aj verejnej podpory (akceptácie) umelej inteligencie a formovanie presvedčenia o akejsi cudzorodej a nie priateľskej technológii nerešpektujúcej tradičné a ustálené etické spoločenské a morálne hodnoty na ujmu ľudí. Odborná spisba tiež pripomína, že: „*Dôvera v oblasti rozvoja, rozmiestňovania a používania systémov AI sa nevzťahuje len na charakteristiku samotných systémov AI, ale súvisí aj s kvalitou sociálno-technického systému zahrňujúceho ich používanie.“*[10] Významnú súčasť budovania dôvery verejnosti k AI predstavuje jej transparentnosť, v ktorej zmysle: „*Užívatelia AI potrebujú vedieť, akým spôsobom sa bude správať za rôznych okolností, t. j. byť informovaní o dôvodoch a technologických parametroch takéhoto postupu*.“[11] V odbornej spisbe sa preto oprávnene poukazuje na potrebu nejakého koordinačného orgánu, ktorý

[9] HAGENDORFF, T., 2020, s. 108.
[10] HUANG, CH., ZHANG, Z., MAO, B., YAO, X., 2023, s. 809.
[11] WALTZ, A., FIRTH-BUTTERFIELD, K., 2019, s. 189.

by sa mal pokúsiť tieto nedostatky riešiť.[12] Takýto orgán by sa mal pokúsiť o formulovanie spoločných (základných) etických princípov, ktoré by sa jednotne uplatňovali v jednotlivých oblastiach pôsobenia AI. Uvedené by však nemalo byť na ujmu uplatňovaniu špecifických etických princípov uplatňovaných v určitých geografických oblastiach pôvodu náboženského, kultúrneho a pod. Vzrastajúci význam etických aspektov AI vyvoláva nutnosť aktívneho zapojenia širšej verejnosti do procesu ich rozširovania, a to prostredníctvom ich výučby, školení, tréningov, verejných prednášok a vystúpení medzi finálnymi užívateľmi AI, vývojovými pracovníkmi a podobne.

1.2 Vzťah etických pravidiel k právnej úprave

Ako bolo spomenuté, etické pravidlá predstavujú len jeden spôsob úpravy AI. Ďalším prostriedkom je právna úprava, ktorá má vzťah k AI. Hoci právna úprava a etické pravidlá predstavujú dve odlišné oblasti,[13] pre potreby úpravy AI ich treba posudzovať spoločne berúc do úvahy ich konvergentný charakter. Súčasná prax potvrdzuje, že vo vzťahu právnych a etických pravidiel možno identifikovať viacero situácií. Aj keď je určité konanie v súlade s právom, môže byť súčasne neetické, prípadne, ak je určité správanie považované za etické, nemusí byť v súlade s právom. Ideálna situácia nastáva vtedy, keď je právo založené (vychádza) a je v súlade etickými pravidlami. Je preto významné, aby právna úprava AI akceptovala etické pravidlá platné v konkrétnom spoločenstve (komunite) a aby napríklad hospodárske subjekty v snahe dosiahnuť väčšie ekonomické výhody neporušovali etické pravidlá. Etické pravidlá týkajúce sa AI by mali byť založené predovšetkým na základných právach upravených v medzinárodných dohovoroch o ľudských právach, pretože*: „Dodržiavanie základných práv v demokratickom rámci a v rámci právneho štátu predstavuje najsľubnejší základ na určenie abstraktných etických zásad a hodnôt, ktoré sa môžu zaviesť do praxe v kontexte umelej inteligencie.“*[14] Viaceré z týchto práv sú právne vymožiteľné tak, že súlad s nimi je právne garantovaný (právo na súkromie, sloboda prejavu, sloboda zhromažďovania, zákaz diskriminácie, rovnosť pred zákonom a pod.). Aj právne nezáväzné *guidelines* čerpajú inšpiráciu z oblasti základných ľudských práv a viaceré z nich boli transponované do právnych pravidiel v oblasti ľudských práv. Umelý (teda nie ľudský) pôvod AI preto nie je a ani nemôže byť dôvodom na nerešpektovanie základných ľudských práv napríklad poukazom na špecifiká prostriedkov ich porušovania (autonómne AI systémy). Odborná spisba pripomína, že: „*Zároveň by na práva občanov mohli mať aplikácie AI aj negatívny vplyv a tieto práva by sa mali chrániť.*“[15] V tejto súvislosti sa AI charakterizuje ako *zákonná,* pričom v základných ľudských právach sú obsiahnuté *štyri* etické zásady, ktoré by sa

[12] MARCHANT, G., 2019, s. 4.

[13] Právo nemožno ignorovať, ani ho stotožňovať s etikou, a naopak.

[14] Expertná skupina na vysokej úrovni pre umelú inteligenciu zriadená Európskou komisiou v júni 2018: Etické usmernenia pre dôveryhodnú umelú inteligenciu, s. 12. [online]. 2019 [cit. 22. 11. 2024]. Dostupné na: https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/JURI/DV/2019/11-06/Ethics--guidelines-AI_SK.pdf.

[15] Tamže, s. 13.

mali dodržiavať. Ide o *rešpektovanie ľudskej autonómie, predchádzanie ujmy ‚spravodlivosť a vysvetliteľnosť*.[16] Etická zásada *rešpektovania ľudskej autonómie* vylučuje obmedzenie slobody a autonómie ľudí ich podriadením, donútením alebo manipulovaním prostredníctvom a na základe systémov AI. V dôsledku toho môže dochádzať k narušovaniu ľudskej dôstojnosti, ktorá prislúcha a tvorí základ každého ľudského práva. Súčasťou tejto zásady je aj zabezpečenie účinného ľudského dohľadu a kontroly nad činnosťami systémov AI počas ich celého „životného cyklu". Zásada prevencie ujmy znamená, že systémy AI by nemali spôsobovať ujmu alebo inak nepriaznivo pôsobiť na ľudí. Takéto systémy by mali byť technicky odolné (robustné) a bezpečné, aby sa zaistilo ich zneužitie. Vývoj a používanie AI by mal byť tiež spravodlivý, t. j. vylučujúci, aby ľudia boli vystavení nespravodlivej a jednostrannej zaujatosti, diskriminácii, nerovnakému prístupu k AI, ako aj k tovaru, vzdelávaniu a technológiám a i. Dôležitú súčasť dôvery k systémom AI predstavuje jej vysvetliteľnosť, čo znamená, že jej procesy musia byť transparentné, o ich účeloch je potrebné otvorene informovať a musia byť vysvetliteľné osobám, ktorých sa to týka. Pre úplnosť možno spomenúť, že nie všetky etické pravidlá sú pravidelne transponované do právnych predpisov, napríklad pre nedostatok konsenzu príslušných orgánov pri rozhodovaní o tom, či by mali nadobudnúť právnu povahu, prípadne z iných dôvodov. Ďalšou skutočnosťou je, že právne predpisy nie vždy držia krok s technologickým vývojom AI, takže sa môžu ocitnúť v rozpore s etickými pravidlami, prípadne určité problémy späté s AI právna úprava vôbec nerieši. V tomto smere by mal mať systém AI zabezpečený súlad s etickými normami, pretože *„etické princípy môžu napomáhať pozitívnemu využívaniu systémov AI aj keď nie sú spôsobilé predchádzať, postihovať alebo sankcionovať negatívne spôsoby ich využívania"*.[17] Vo vzťahu k právnym pravidlám môžu etické pravidlá dopĺňať a posilňovať právnu úpravu, prípadne napomáhať, kompletizovať a rozširovať jej výklad, a tým napomáhať jednotlivcom, spoločnostiam a iným inštitúciám v jej správnom uplatňovaní. Nezanedbateľný je prínos etických pravidiel aj v prípadoch relevantnej právnej úpravy, pretože do procesu aplikácie systémov AI vnášajú usmerňujúce etické a morálne hodnoty pre ich účastníkov na inej ako legislatívnej úrovni, a tým napomáhajú dôveryhodnosti systémov AI a zvýšeniu ochrany príslušných subjektov.

2. Implementácia etických pravidiel do systémov AI

Je zjavné, že iba formálne prihlásenie či ochota prijímať etické princípy negarantuje aj ich reálne uplatňovanie v AI. V dôsledku toho musia byť *implementované* do technických systémov a prevádzkových procesov umelej inteligencie. Odborná spisba poukazuje na skutočnosť, že: *„Transformácia etických princípov do vykonateľných usmernení je kľúčom k realizácii etickej AI. Ide o integrovanie etického hľadiska do každej fázy život-*

[16] Tamže, s. 14.

[17] CARILLO, M. R. Artificial intelligence From ethics to Law. *Telecommunications Policy* [online]. 2020, č. 6 [cit. 22. 11. 2024]. Dostupné na: https://www.sciencedirect.com/science/article/abs/pii/S030859612030029X; https://doi.org/10.1016/j.telpol.2020.101937.

ného cyklu AI, od počiatočného návrhu cez jej praktické nasadenie až po monitorovanie.“[18] Hlavný problém spočíva v tom, ako zabezpečiť, aby etické pravidlá boli účinne implementované do systémov AI (softvér) preto, aby sa AI systém skutočne eticky správal, a to aspoň v špecifických či kritických situáciách. V súčasnosti nepanuje všeobecná predstava a už vôbec nie zhoda v tom, aký spôsob, resp. aké mechanizmy by mali byť použité na implementáciu etických pravidiel do AI, prípadne, či by takéto mechanizmy mali obsahovať určitý *mix* spôsobov predstavujúcich výsledok koordinácie a diskusie medzi zainteresovanými subjektmi.[19] Tiež sa poukazuje na skutočnosť, že: „*Oblasť etiky AI je stále v plienkach a je potrebná jej konceptualizácia na to, aby sa dosiahol rozvoj AI obsahujúci jej etický rozmer.“*[20] Odborná spisba v tejto súvislosti uvádza dva možné spôsoby implementácie etických pravidiel do AI, a to technický a regulačný. Zatiaľ čo: *„…technické riešenia sú priamo implementované do AI v jej jednotlivých prostriedkoch, regulačné prístupy zaväzujú výrobcov a/alebo užívateľov takýchto prostriedkov, aby zabezpečili súlad ich užívania s konkrétnymi normatívnymi štandardmi.“*[21] V takomto kontexte implementované etické pravidlá predstavujú akési „inštrukcie“ správania sa AI v rámci určitého spoločenstva (mesto, región, krajina), ktoré by mali reflektovať ich morálne a etické hodnoty. Ide o proces, v ktorého rámci treba identifikovať etické pravidlá, ktoré by mali byť rešpektované pri používaní AI, ďalej implementovať tieto pravidlá do AI a napokon vyhodnotiť, či implementácia takýchto pravidiel do AI zodpovedá komunitárnym hodnotám a vyhodnoteniu stupňa a úrovne ich rešpektovania v procese interakcie medzi človekom a AI. Vzhľadom na skutočnosť, že etické princípy sú pravidelne formulované všeobecne, vyvstáva problém, ako ich implementovať do značne diverzifikovaného systému vedeckej, technickej a ekonomickej praxe uplatňovania AI ďalej voči rôznym geograficky rozptýleným skupinám výskumníkov, vývojárov s rôznymi prioritami a rôznou úrovňou zodpovednosti. Samotné etické dokumenty zvyčajne neobsahujú technické postupy, metodológie, technológie či iné detaily, ktorých použitím by sa dosiahol výskum, vývoj a uplatňovanie prostriedkov AI etickým spôsobom. V procese uplatňovania sú AI systémy v zásade konfrontované s dvomi druhmi situácií zahrňujúcimi etické aspekty (pravidlá). Jednoduchší je ten, keď ide o konkrétnu situáciu vyžadujúcu si uplatnenie etického pravidla. Komplikovanejšia je tá, keď ide o dve alebo viaceré situácie vyžadujúce uplatnenie etických princípov, pričom je potrebné zvážiť (rozhodnúť),

[18] SULLIVAN, M. Key principles for ethical AI development [online]. 2023 [cit. 22. 11. 2024]. Dostupné na: https://transcend.io/blog/ai-ethics.

[19] Príklady viacerých veľkých spoločností poskytujú prehľad, akým spôsobom si vytvárajú vlastné etické systémy a ako ich princípy reálne aplikujú vo svojej činnosti. Spoločnosť GOOGLE zabezpečuje implementáciu etických princípov kombináciou výukových programov, posudzovania etických pravidiel a technických prostriedkov, Microsoft pracuje s internými usmerneniami, ktoré sa vzťahujú na návrhy, výrobu a testovanie modelov svojich produktov s cieľom znižovať riziko AI a maximalizovať výhody, IBM sa zameriava na dôveryhodnú AI, pričom jej dôveryhodnosť garantuje sústavný monitoring a časté kontroly produktov s cieľom zabezpečenia dôvery rôznymi zainteresovanými subjektmi.

[20] LO PIANO, S. Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward. *Humanities and SocialSciencesCommunications* [online]. 2020, č. 9 [cit. 22. 11. 2024]. Dostupné na: https://www.nature.com/articles/s41599-020-0501-9; https://doi.org/10.1057/s41599-020-0501-9.

[21] WALTZ, A., FIRTH-BUTTERFIELD, K., 2019, s. 198.

ktoré z pravidiel etickej povahy je v danom momente potrebné preferovať, resp. dať mu prednosť, a z akých dôvodov *(ethical dilemmas*).[22] Odborná spisba v tejto súvislosti uvádza, že: *„... pokiaľ sú takéto princípy výslovne formulované ... poskytujú pádne a logické vysvetlenie, prečo bola vybraná situácia s jedným etickým pravidlom pred druhou situáciou"*.[23] V konkrétnostiach procesu implementácie etických princípov možno uviesť, že sa sústreďuje predovšetkým na algoritmy AI, ktoré bezprostredne generujú jej rozhodnutie obsahujúce etické pravidlo použité (použiteľné) v konkrétnej situácii. Na to, aby sa tento etický aspekt naplnil, sú algoritmy *eticky trénované* na veľkom množstve vstupných údajov, ktoré však môžu obsahovať skreslené údaje. Ak sa náležite nerieši, môže praktická aplikácia takto „infikovaných" algoritmov viesť k diskriminačným dôsledkom a posilneniu spoločenskej nerovnosti. V systéme AI, v ktorom sú vstupné údaje skreslené, je veľká pravdepodobnosť, že aj výstupné algoritmické rozhodnutia údaje budú skreslené. V tejto súvislosti možno spresniť, že samotné algoritmické pravidlá AI nie sú eticky neutrálne, pretože sa do nich vkladajú prostredníctvom a na základe výsledkov etického tréningového procesu vstupných údajov vstupujúcich do systému AI. Aj tréningové údaje však pravidelne obsahujú rôzne odchýlky, prípadne skreslenia *(bias),* ktoré sa negatívne prejavujú v rozhodnutiach a predpovediach založených na algoritmoch. Je preto potrebné pozorne skúmať údaje použité na trénovanie a identifikovať akékoľvek potenciálne zdroje skreslenia, ktoré v nich môžu byť obsiahnuté. Problémom je, ako nájsť vhodné spôsoby a zadávať vstupné údaje do AI bez rizika ich skreslenia. Pôvod skreslenia údajov umelej inteligencie možno nájsť v raných štádiách strojového učenia AI, keď boli súbory údajov obmedzené a vyžadovali si manuálne spracovanie. Spočiatku sa systémy AI považovali za nestranné, objektívne nástroje, ktoré objektívne rozhodujú na základe bezchybných algoritmov. Čoskoro sa však ukázalo, že algoritmy obsahujú skreslenia nadobudnuté počas tréningového procesu vstupných údajov. Výber vhodných údajov do aplikácií AI je preto *rozhodujúci* pre budovanie a tréning algoritmov. Pri výbere etických údajov je nevyhnutná relevantnosť údajov, ktoré by mali priamo súvisieť s problémom, ktorý sa má riešiť, kvalita údajov s minimom chýb, množstvo údajov atď. Následné algoritmické skreslenie môže pochádzať z viacerých zdrojov, napríklad zo skreslených tréningových údajov, z návrhov algoritmov, ako aj z ľudských skreslení a odchýlok obsiahnutých v AI systémoch. Všeobecne sa uznáva, že tréningový proces algoritmov (čo ako dokonalý) nedokáže eliminovať všetky prípadne skreslenia a odchýlky, ktoré sa negatívne prejavia po praktickom použití algoritmov ako nerešpektovanie etických princípov a hodnôt. V dôsledku toho postupne silnelo presvedčenie o potrebe osobitného *externého mechanizmu testovania* a vyhodnocovania etických a iných aspektov výsledkov dosiahnutých aplikáciou AI algoritmov. Takýto mechanizmus postupne

[22] Praktický príklad možno uviesť z prevádzky vozidiel bez vodiča, ktoré môžu byť konfrontované so situáciou neodvrátiteľnej zrážky, pričom je potrebné rozhodnúť buď o prioritách záchrany životov cestujúcich v samotnom vozidle, cestujúcich v ďalšom vozidle, prípadne chodcov, medzi životom detí a dospelých a pod.

[23] ANDERSON, M., ANDERSON, S. L. Toward Ensuring Ethical Behavior from Autonomous Systems: A Case- Supported Principle-Based Paradigm.*Industrial Robot* [online]. 2015, č. 4 [cit. 22. 11. 2024]. Dostupné na: https://www.researchgate.net/publication/268522350_Toward_Ensuring_Ethical_Behavior_from_Autonomous_Systems_A_Case-Supported_Principle-Based_Paradigm; https://doi.org/10.1108/IR-12-2014-0434.

vstupuje do oblasti AI a v súčasnosti je známy pod názvom *audit AI* algoritmov. Z faktických okolností podnecujúcich jeho nástup možno spomenúť viaceré verejné škandály týkajúce sa neobjektívnych výsledkov, nedostatku transparentnosti a zneužívania údajov pri uplatňovaní AI algoritmov. To viedlo k rastúcej nedôvere voči AI a zvýšeným požiadavkám na povinné etické audity algoritmov. Súčasné návrhy na etické hodnotenie algoritmov sú však buď príliš vysoké na to, aby sa dali uviesť do praxe bez ďalších usmernení, alebo sa zameriavajú na špecifické a technické termíny spravodlivosti alebo transparentnosti, ktoré nezohľadňujú konkrétne záujmy zainteresovaných strán alebo širší spoločenský kontext.

2.1 Audit AI a rešpektovanie etických princípov AI

Audit systémov AI predstavuje perspektívny a sľubný spôsob, ako pochopiť a zvládnuť etické problémy a spoločenské riziká spojené so súčasnými systémami AI. V praxi vykonávajú audit nezávislé audítorské spoločnosti a jeho cieľom je posúdiť, či jednotlivé systémy AI spĺňajú predpokladané funkcie pri rešpektovaní etických a právnych pravidiel. Ako bolo spomenuté, audit AI algoritmov predstavuje externý mechanizmus, pričom odborná spisba ho charakterizuje ako: *„posudzovanie negatívneho vplyvu algoritmu na práva a záujmy zainteresovaných strán so zodpovedajúcou identifikáciou situácií a/alebo vlastností algoritmov, ktoré spôsobujú tieto negatívne vplyvy“*.[24] V praxi audit predstavuje *„zbieranie dát o správaní sa algoritmu, ktorý bol použitý v konkrétnom kontexte, a následne použitie týchto dát na zistenie či správanie sa algoritmu nemalo negatívny vplyv na záujmy alebo práva osôb dotknutých jeho použitím“*.[25] Algoritmy umelej inteligencie by mali byť transparentné, spravodlivé a neskreslené, aby sa zabezpečilo, že jej rozhodnutia nebudú ovplyvnené skreslenými zdrojmi údajov. Vzhľadom na komplexnú povahu sa audity zvyčajne neobmedzujú len na posudzovanie etických algoritmov, ale zaoberajú sa aj ich vzťahom k vstupným údajom a k výsledkom dosiahnutým použitím algoritmov. V rámci auditu údajov použitých v systéme AI sa hodnotí ich správnosť, komplexnosť a prípadná existencia skreslení a omylov. Pri posudzovaní algoritmov sa zisťuje, či fungujú stanoveným spôsobom a neobsahujú skreslenia alebo omyly. Napokon audit výsledkov vyhodnocuje finálny výstup dosiahnutý použitím algoritmu, pričom sa hodnotí jeho správnosť, primeranosť a súlad s predpokladaným a očakávaným výsledkom. Keďže samotný audit a jeho výsledok napomáha objasneniu procesu tvorby a rozhodnutí systémov AI, prispieva k dôvere užívateľov, zákazníkov, ako aj regulátorov, pretože identifikuje riziká, ktoré by mohli vzniknúť (skreslenie a algoritmické omyly) a viesť k finančným a reputačným škodám. Formálny výsledok auditu predstavujú správy, resp. stanoviská (*opinions*) audítorskej spoločnosti. Správa audítora je formálnym stanoviskom buď interného, alebo nezávislého externého audítora, ktorá poskytuje „uží-

[24] BROWN, S., DAVIDOVIC, J., HASAN, A. The algorithm audit: Scoring the algorithms that score us. *Big Data& Society* [online]. 2021, č. 1 [cit. 22. 11. 2024]. Dostupné na: https://journals.sagepub.com/doi/10.1177/2053951720983865; https://doi.org/10.1177/2053951720983865.

[25] Tamže, s. 24.

vateľovi“ (ako sú organizácia alebo vláda) podklad na ich ďalší postup a prípadne rozhodovanie na základe svojich zistení. Ak správa signalizuje určité skreslenia alebo iné pochybenia algoritmu, audítor v nej *odporučí* zmeny, a to aj v údajoch použitých na trénovanie algoritmu za účelom zmiernenia identifikovaných problémov. V oblasti auditu však v súčasnosti neexistujú medzinárodne akceptované štandardy a kritériá audítorského procesu. Absencia štandardov je o. i. spôsobená nedostatočnou úpravou AI (pretože jej definícia je stále nejednoznačná), a preto niet divu, že za týchto okolností je vytvorenie spoločných štandardov auditu problematické. Proces auditu AI ďalej komplikuje vážny nedostatok pracovníkov s požadovanými zručnosťami. Vznikajúca technológia auditu AI je stále relatívne nová a kvalifikovaných odborníkov je málo. Pre úplnosť možno spomenúť, že okrem systémov externého auditu pôsobia aj *interné audity,* ktoré sa využívajú na kontrolu systémov vnútornej kontroly organizácie, procesov riadenia rizík a celkového riadenia (finančné audity). Poskytujú nezávislé a objektívne hodnotenia, ktoré pomáhajú identifikovať slabé stránky auditovaného subjektu, zlepšovať dodržiavanie relevantných predpisov a zvyšovať prevádzkovú efektivitu. Prax potvrdzuje niekoľko stupňov systémového prístupu audítorov k algoritmom. Najvyšším stupňom je tzv. biela schránka (*white box),* v ktorej sú všetky komponenty algoritmov vrátane vstupných údajov a tréningových parametrov k dispozícii pre potreby auditu. Najnižším stupňom je čierna skrinka (*blackbox),* kde nie je pre audit možný prístup k algoritmom alebo tréningovým dátam.

2.1.1 Postupy na eliminovanie skreslených algoritmov

Prax potvrdzuje, že mnohé algoritmy obsahujú skreslenia nadobudnuté počas tréningového procesu vstupných údajov. Hoci môžu byť identifikované na základe a prostredníctvom auditu, tento nie je spôsobilý takéto algoritmy reálne eliminovať, resp. zablokovať. Pokiaľ ide o zdroje skreslení, odborná spisba uvádza, že: „*Systémové skreslenia si našli cestu do AI, pretože ako všetky technológie, boli vytvorené ľuďmi.“*[26] V dôsledku toho aj finálne výsledky auditov odrážajú a uchovávajú ľudské skreslenia a pochybenia v systémoch AI. Keďže vstupné dáta, ich tréningový proces a ani samotný audit nie sú schopné reálne eliminovať neželané algoritmické odchýlky, začala sa venovať pozornosť ďalšiemu konaniu, ktorého cieľom je eliminácia, prípadne zablokovanie potenciálne škodlivých algoritmov (*tzv. debiasing procedure).* Takéto konanie sa stáva mimoriadne potrebnou a neoceniteľnou súčasťou tréningového procesu, pretože použitie nesprávnych údajov môže viesť ku škodlivým výsledkom na ujmu zainteresovaných strán. Jeho nevyhnutným predpokladom je dobré poznanie funkcií a prijímania rozhodnutí systému AI, pretože: „*inak nemôžeme prísť na to, ako opraviť údaje, na ktorých bol AI systém trénovaný“.*[27] Odborná spisba uvádza a analyzuje viacero spôsobov a techník, ktoré by mohli zmierniť alebo odstrániť algoritmické skreslenia. Ich základ predstavuje systematický a trvalý systém analýz vstupných dát, algoritmov, ľudského prvku, ako aj kontextu,

[26] CANNINGS, N. How to Tackle AI bias [online]. 2022 [cit. 22. 11. 2024]. Dostupné na: https://www.spiceworks.com/tech/artificial-intelligence/guest-article/how-to-tackle-ai-bias/.

[27] Tamže.

v ktorého rámci dochádza k uplatneniu algoritmov. Identifikovanie rôznych príčin skreslenia môže napomôcť vytvoreniu najlepšieho modelu strojového učenia, ktorý je správny a bez skreslenia.

Ďalšia navrhovaná technika určená na zníženie počtu skreslení sa sústreďuje na také používanie dát, aby sa zvýšila ich kvalita, zaviedla diverzifikácia tréningových dát, ako aj osobitné testy algoritmov na prítomnosť skreslení, a prijatie zodpovedajúceho regulačného rámca na úpravu týchto opatrení. Ďalší návrh sa snaží o odstránenie negatívnych dôsledkov skreslení prostredníctvom tzv. ochrancov AI (*guardians AI*). Títo by mali pôsobiť v rámci osobitného monitorovacieho systému AI, ktorý by kontroloval dodržiavanie právnych a etických pravidiel, pričom by mohli: *„technicky zasahovať do základného systému umelej inteligencie a priamo napravovať nezákonné alebo neetické rozhodnutia.“*[28] Nevýhoda tohto návrhu spočíva v jeho *ad hoc* povahe, pretože aj keď by jednorazovo mohol napravovať nezákonné alebo neetické rozhodnutie, nedotýka sa jeho skreslenia, ktoré je jeho príčinou. Ďalší návrh sa zameriava na ľudský aspekt kontroly nad AI. Keďže AI nemôže byť výlučne pod vlastnou kontrolou, vyžaduje si nevyhnutný trvalý ľudský dohľad. „*Človek by mal mať vždy možnosť nerešpektovať rozhodnutia AI. V konečnom dôsledku by vývojári AI mali zvážiť typ technického opatrenia, ktoré by zabezpečovalo ľudský dohľad, napríklad tlačidlo zastavenia (stop button) alebo postupu na prerušenie operácie s cieľom zabezpečiť (obnoviť)ľudskú kontrolu.“*[29] Nevýhoda posledne menovaných spôsobov spočíva v ich *ad hoc* povahe a účinkoch, pretože aj keď by jednorazovo mohli napravovať nezákonné alebo neetické rozhodnutia, nedotýkajú sa existencie skreslení, ktoré sú ich príčinou. Postupne ako sa s rozvojom prostriedkov AI budú zvyšovať ich schopnosti a komplexnosť, bude nevyhnutné vyvíjať sofistikovanejšie kontrolné a bezpečnostné systémy, aby sa predišlo vzniku nebezpečných situácií, ako aj škôd. V dôsledku uvedeného sa objavila otázka, či v takejto situácii je možné (hoci nepravdepodobné), aby vznikla absolútne nestranná AI. Dôvodom nepravdepodobnosti je absencia nestrannosti ľudskej mysle. Odborná spisba v tejto súvislosti uvádza, že: *„V skutočnom svete vieme, že je to nepravdepodobné. AI je určovaná údajmi, ktoré dostáva a z ktorých sa učí. Sú to však ľudia, ktorí generujú údaje, používané AI. V dôsledku toho je možné, že sa nikdy nedosiahne úplne nestranná ľudská myseľ, a preto ani nestranný systém AI. Koniec koncov sú to ľudia, ktorí generujú skreslené údaje, zhromažďujú ich, používajú a vytvárajú algoritmy AI, a sú to rovnako ľudia, ktorí overujú údaje, aby odhalili a opravili ich skreslenia.“*[30] Všeobecne sa uznáva, že „bojovať“ proti skresľovaniu AI je v súčasnosti možné dôsledným testovaním a vyhodnocovaním jej údajov a algoritmov a používaním osvedčených postupov na zhromažďovanie a používanie údajov a vytváranie algoritmov AI.

[28] WALTZ, A., FIRTH-BUTTERFIELD, K., 2019, s. 201.

[29] European Parliament: EU guidelines on ethics in artificial intelligence: Context and implemetation [online]. 2019 [cit. 22. 11. 2024]. Dostupné na: https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/640163/EPRS_BRI(2019)640163_EN.pdf.

[30] LARKIN, Z. AI Bias- What is it and how to avoid? [online]. 2024 [cit. 22. 11. 2024]. Dostupné na: https://levity.ai/blog/ai-bias-how-to-avoid.

Záver

V oblasti úpravy umelej inteligencie predstavujú jej etické pravidlá významnú súčasť. Možno ich charakterizovať ako súbor hodnôt, princípov a techník používajúcich všeobecne akceptované štandardy „dobra a zla“ a upravujúce morálne správanie v procese rozvoja a používania technológií AI. Účelom etického správania AI je ochrana morálnych hodnôt, ktoré sú všeobecne prijímané ako súčasť ľudskej povahy a prirodzenosti (ochrana ľudského života, sloboda, ľudská dôstojnosť a i.), ako aj viery a presvedčenia určitých skupín obyvateľov (náboženské obrady a pravidlá). Etické pravidlá tak vytvárajú krehkú rovnováhu medzi technologickým pokrokom a ochranou ľudských hodnôt, akými sú súkromie, spravodlivosť, transparentnosť. Reálne uplatňovanie etických hodnôt a princípov je však možné len po ich implementácii do systémov AI, a to do každej fázy „životného cyklu AI“ od počiatočného návrhu cez jej praktické nasadenie až po monitorovanie. V súčasnosti nepanuje všeobecná predstava o tom, aký spôsob, resp. aké mechanizmy by mali byť použité na implementáciu etických pravidiel do AI, prípadne či by takéto mechanizmy mali obsahovať určitý *mix* spôsobov predstavujúcich výsledok koordinácie a diskusie medzi zainteresovanými subjektmi. Spoločné pre procedúry implementácie je fakt, že sa sústreďujú predovšetkým na algoritmy AI, ktoré bezprostredne generujú jej finálne rozhodnutie obsahujúce etické pravidlo použité (použiteľné) v konkrétnej situácii. Algoritmy sú preto eticky trénované na veľkom množstve vstupných údajov, ktoré však môžu obsahovať skreslené údaje. Ak sa náležite neriešia, môže praktická aplikácia takto „infikovaných“ algoritmov viesť k diskriminačným dôsledkom a posilneniu spoločenskej nerovnosti. Tieto nedostatky by mal posúdiť audit umelej inteligencie, ktorého cieľom je posúdiť, či systémy AI spĺňajú predpokladané funkcie pri rešpektovaní zodpovedajúcich etických a právnych pravidiel. V oblasti auditu však v súčasnosti neexistujú žiadne medzinárodne akceptované štandardy postupu. Absencia štandardov je o. i. spôsobená nedostatočnou úpravou AI (keďže jej samotná definícia je stále nejednoznačná), a preto niet divu, že za týchto okolností je vytvorenie spoločných štandardov pre jej audit problematické. Proces auditu AI ďalej komplikuje vážny nedostatok pracovníkov s požadovanými zručnosťami. Prax potvrdila, že mnohé algoritmy obsahujú skreslenia nadobudnuté počas tréningového procesu vstupných údajov. Hoci môžu byť identifikované na základe a prostredníctvom auditu, tento samotný nie je spôsobilý takéto algoritmy reálne eliminovať, resp. zablokovať. Systémové skreslenia si totiž našli cestu do AI, pretože – ako všetky technológie – boli vytvorené ľuďmi.

V dôsledku toho aj finálne výsledky auditov odrážajú a uchovávajú ľudské skreslenia a pochybenia. Všeobecne sa uznáva, že „bojovať“ proti skresľovaniu AI je v súčasnosti možné dôsledným testovaním a vyhodnocovaním jej údajov a algoritmov, a používaním osvedčených postupov na zhromažďovanie a používanie údajov a vytváranie algoritmov AI.

Literatúra

ANDERSON, M., ANDERSON, S. L. Toward Ensuring Ethical Behavior from Autonomous Systems: A Case-Supported Principle-Based Paradigm. *Industrial Robot* [online]. 2015, č. 4 [cit. 22. 11. 2024]. Dostupné na:

https://www.researchgate.net/publication/268522350_Toward_Ensuring_Ethical_Behavior_from_Autonomous_Systems_A_Case-Supported_Principle-Based_Paradigm; https://doi.org/10.1108/IR-12-2014-0434

BROWN, S., DAVIDOVIC, J., HASAN, A. The algorithm audit: Scoring the algorithms that score us. *Big Data & Society* [online]. 2021, č. 1 [cit. 22. 11. 2024]. Dostupné na: https://journals.sagepub.com/doi/10.1177/2053951720983865; https://doi.org/10.1177/2053951720983865

CANNINGS, N. How to Tackle AI bias [online]. 2022 [cit. 22. 11. 2024]. Dostupné na: https://www.spiceworks.com/tech/artificial-intelligence/guest-article/how-to-tackle-ai-bias/

CARILLO, M. R. Artificial intelligence: From ethics to law. *Telecommunications Policy* [online]. 2020, č. 6 [cit. 22. 11. 2024]. Dostupné na: https://www.sciencedirect.com/science/article/abs/pii/S030859612030029X; https://doi.org/10.1016/j.telpol.2020.101937

CATHE, C., WACHTER, S., MITTELSTADT, B., TADDEO, M., FLORIDI, L. Artificial Intelligence and the 'Good Society': the US, EU, and UK approach. *Science and Engineering Ethics* [online]. 2018, č. 24 [cit. 22. 11. 2024]. Dostupné na: https://link.springer.com/article/10.1007/S11948-017-9901-7; https://doi.org/10.1007/s11948-017-9901-7

HAGENDORFF, T. The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds & Machines* [online]. 2020, č. 30 [cit. 22. 11. 2024]. Dostupné na: https://link.springer.com/article/10.1007/s11023-020-09517-8; https://doi.org/10.1007/s11023-020-09517-8

HUANG, CH., ZHANG, Z., MAO, B., YAO, X. An Overview of Artificial Intelligence Ethics.*IEEE Transactions on Artificial Intelligence* [online]. 2023, č. 4 [cit. 22. 11. 2024]. Dostupné na: https://scholars.ln.edu.hk/en/publications/an-overview-of-artificial-intelligence-ethics; https://doi.org/10.1109/TAI.2022.3194503

KIROVA, V. D., KU, C. S., LARACY, J. R., MARLOWE, T. J. The Ethics of Artificial Intelligence in the Era of Generative AI. In *Journal of Systemics, Cybernetics and Informatics* [online]. 2023, č. 4 [cit. 22. 11. 2024]. Dostupné na: https://www.iiisci.org/DOIJSCI/SA445KR23/#/;https://doi.org/10.54808/JSCI.21.04.42

KLUČKA, J. Medzinárodné právo, umelá inteligencia a vice versa. In *Časopis pro právní vědu a praxi.* Roč. XXIX, 2021, č. 3, s. 553. ISSN 1210-9126; https://doi.org/10.5817/CPVP2021-3-4

LARKIN, Z. AI Bias- What is it and how to avoid it? [online]. 2024 [cit. 22. 11. 2024]. Dostupné na: https://levity.ai/blog/ai-bias-how-to-avoid

LO PIANO, S. Ethical principles in machine learning and artificial intelligence: cases fromthe field and possible ways forward. *Humanities and Social SciencesCommunications* [online]. 2020, č. 9 [cit. 22. 11. 2024]. Dostupné na: https://www.nature.com/articles/s41599-020-0501-9; https://doi.org/10.1057/s41599-020-0501-9

MARCHANT, G. „Soft Law" Governance of Artificial Intelligence [online]. 2019 [cit. 22. 11. 2024]. Dostupné na: https://escholarship.org/content/qt0jq252ks/qt0jq252ks_noSplash_1ff6445b4d4efd438fd6e06cc2df4775.pdf?t=po1uh8

PINTEROVÁ, D. Právna regulácia umelej inteligencie.(perspektívy a výzvy). In *Právny obzor.* Roč. 107, 2024, č. 4, s. 361 – 383

SULLIVAN, M. Key principles for ethical AI development [online]. 2023 [cit. 22. 11. 2024]. Dostupné na: https://transcend.io/blog/ai-ethics

WALTZ, A., FIRTH-BUTTERFIELD, K. Implementing Ethics into Artificial Intelligence: A Contribution, from a Legal Perspective, to The Development of an AI Governance Regime. [online]. 2019 [cit. 22. 11. 2024]. Dostupné na: https://www.researchgate.net/publication/338083064_IMPLEMENTING_ETHICS_INTO_ARTIFICIAL_INTELLIGENCE_A_CONTRIBUTION_FROM_A_LEGAL_PERSPECTIVE_TO_THE_DEVELOPMENT_OF_AN_AI_GOVERNANCE_REGIME

Dokumenty

Expertná skupina na vysokej úrovni pre umelú inteligenciu zriadená Európskou komisiou v júni 2018: Etické usmernenia pre dôveryhodnú umelú inteligenciu,s.12. [online]. 2019 [cit. 22. 11. 2024]. Dostupné na: https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/JURI/DV/2019/11-06/Ethics-guidelines-AI_SK.pdf