RIZIKOVOSŤ MODELOV STROJOVÉHO UČENIA V ROZHODOVACEJ ČINNOSTI SÚDOV

THE RISKS OF MACHINE LEARNING MODELS IN JUDICIAL DECISION-MAKING

Romana Koneracká[1]

Abstrakt: Modely strojového učenia ako nástroje umelej inteligencie majú čoraz väčší potenciál stať sa integrálnou súčasťou rozhodovacej činnosti súdov. Technické limity systémov umelej inteligencie, ktoré sú častokrát právnou vedou opomínané, však vyvolávajú zásadné otázky predovšetkým z hľadiska zachovania základných princípov materiálneho právneho štátu a s tým spojenej nezávislosti súdnictva. Osobitná pozornosť sa v príspevku kladie na dve technicko-právne hrozby spojené s aplikáciou modelov strojového učenia, pričom referenčný rámec tvoria textové údaje. Jednou z hrozieb je pretrénovanie modelu (tzv. overfitting), keď model „príliš prispôsobí" svoje rozhodovanie konkrétnym údajom, na ktorých bol trénovaný. Druhou hrozbou sú adverzariálne útoky, teda úmyselné manipulácie vstupných údajov s cieľom ovplyvniť výstupy modelu. Na tomto základe autorka identifikuje vnútorný rozpor v kontexte Aktu o umelej inteligencii, ktorý akcentuje potrebu ľudského dohľadu pri používaní systémov umelej inteligencie vo vysokorizikových oblastiach ako napr. súdnictvo. Kontrola ľudských operátorov počas fázy tréningu modelov strojového učenia je však stále nedostatočne reflektovaná. Príspevok poukazuje na to, že ľudskí operátori podieľajúci sa na tréningu systémov umelej inteligencie disponujú vedomosťami o „slabých miestach" modelu, a teda ľudskí operátori predstavujú riziko realizácie strategicky cielených adverzariálnych útokov. Autorka sa následne zameriava na identifikáciu najoptimálnejšieho modelu strojového učenia v nadväznosti na nezávislosť súdnej moci.

Kľúčové slová: súdnictvo; modely strojového učenia; rizikovosť; pretrénovanie modelu; adverzariálne útoky.

Abstract: Machine learning models, as tools of artificial intelligence, have an increasingly strong potential to become an integral part of judicial decision-making. However, the technical limitations of AI systems—often overlooked by

[1] Katedra ústavného práva, Právnická fakulta Univerzity Komenského v Bratislave.

doi: 10.46282/bpf.2025.05

legal scholarship—raise fundamental questions, particularly regarding the preservation of the basic principles of the material rule of law and the associated independence of the judiciary. The contribution pays special attention to two technical-legal threats connected with the application of machine learning models, using textual data as the reference framework. One threat is model overfitting, where the model "over-adapts" its decision-making to the specific data on which it was trained. The second threat is adversarial attacks, meaning intentional manipulations of input data aimed at influencing the model's outputs. Based on this, the author identifies an internal contradiction within the AI Act, which emphasizes the need for human oversight when using AI systems in high-risk areas such as the judiciary. Yet human oversight during the training phase of machine learning models remains insufficiently addressed. The contribution points out that human operators involved in training AI systems possess knowledge of the model's "weak spots," and therefore represent a risk of carrying out strategically targeted adversarial attacks. The author then focuses on identifying the most optimal machine learning model in relation to the independence of the judiciary.

Key words: judiciary; machine learning models; risk assessment; overfitting; adversarial attacks.

Úvod

V súčasnosti je pojem umelá inteligencia natoľko prítomný v spoločenskom diskurze, že sa len ťažko nájde jednotlivec či už z radov odbornej verejnosti alebo laikov, ktorý by sa s týmto pojmom aspoň okrajovo nestretol. Čoraz naliehavejšia sa pritom javí potreba dôsledného rozlišovania medzi populárnou, často stereotypnou predstavou o umelej inteligencii a jej reálnym, technicky aj právne komplexným vymedzením, najmä v oblastiach vyžadujúcich vysoký stupeň integrity a nezávislosti.

Súdna prax nie je v tomto ohľade výnimkou. Práve naopak, predstavuje jednu z najcitlivejších a zároveň najambivalentnejších oblastí, čo sa týka systémov umelej inteligencie. Súdy sú konfrontované týmito „revolučnými" digitálnymi trendmi nielen externe, a to prostredníctvom prípadov, ktoré reflektujú digitálnu realitu, ale aj interne, nakoľko sa samé stávajú objektom digitalizácie ovplyvňujúcej ich fungovanie.

Hoci sa viaceré články a príspevky nepochybne zaoberajú využitím umelej inteligencie a modelov strojového učenia aj v medziach súdnej praxe, odpovede na mnohé zásadné otázky ostávajú zodpovedané implicitne, resp. vôbec alebo ide o otázky metodologicky nedostatočne rozpracované. Slovenská ale aj zahraničná právna veda, doposiaľ nevenovala dostatočnú pozornosť špecifickým rizikám, ktoré

doi: 10.46282/bpf.2025.05

plynú z vývoja systémov umelej inteligencie vo vlastnej réžii súdov, na rozdiel od komerčných riešení ako OpenAI, resp. GenAI (napr. ChatGPT, Gemini, DeepSeek atď.). Výsledkom sú redukcionistické tendencie vedúce k zjednodušovaniu samotnej predstavy o „AI v súdnictve", ktorá je verejnosťou často mylne stotožňovaná s plne automatizovaným rozhodovaním s absenciou ľudskej personálnej zložky alebo ľudského operátora. Na základe toho sú častokrát vyslovené absolútne závery typu „umelá inteligencia by mala byť na súdoch zakázaná" alebo naopak „umelá inteligencia by mala byť na súdoch povolená", a to bez riadnej identifikácie a hĺbkovej analýzy limitov podmieňujúcich a legitimujúcich takéto zakazovanie alebo povoľovanie. Je nevyhnutné porozumieť nielen tomu *aké* nástroje umelej inteligencie sa používajú ale aj *kedy* sa používajú a aké riziká sú následne spojené nielen s plne automatizovanými procesmi ale aj s procesmi, ktoré si vyžadujú prítomnosť ľudského operátora.

Príspevok poukazuje na dve technicko-právne hrozby, ktoré vyplývajú z využívania modelov strojového učenia, pričom referenčný rámec predstavujú textové údaje. Jednou z hrozieb je pretrénovanie modelu (tzv. overfitting), keď model „príliš prispôsobí" svoje rozhodovanie konkrétnym údajom, na ktorých bol trénovaný. Druhou hrozbou sú adverzariálne útoky (tzv. nepriateľské strojové učenie), teda úmyselné manipulácie vstupných údajov s cieľom ovplyvniť výstup modelu. Na tomto základe autorka, prostredníctvom stanovenia si hypotézy tvrdí, že pretrénovanie modelu zvyšuje úspešnosť adverzariálnych útokov a tým aj „zraniteľnosť" rozhodovacieho procesu súdov. Autorka následne identifikuje vnútorný rozpor v kontexte Aktu o umelej inteligencii (ďalej ako *„AIA"*), ktorý koncepčne vymedzuje systémy umelej inteligencie ako „výsledný produkt". AIA zároveň akcentuje potrebu ľudského dohľadu pri využívaní týchto systémov vo vysokorizikových oblastiach ako napr. súdnictvo. Kontrola samotných ľudských operátorov počas prvotnej fázy tréningu modelov strojového učenia je však stále nedostatočne reflektovaná. Autorka poukazuje na to, že ľudskí operátori podieľajúci sa na tréningu systémov umelej inteligencie disponujú vedomosťami o „slabých miestach" modelu, a teda ľudskí operátori predstavujú riziko realizácie strategicky cielených adverzariálnych útokov.

Výskumná otázka sa zameriava na identifikáciu najvhodnejšieho White-Box modelu, ktorý si aj napriek pretrénovaniu dokáže zachovať vysokú odolnosť voči adverzariálnym útokom a zároveň tým neohrozí nezávislosť súdnej moci.

Príspevok reaguje na otázky, ktoré právna veda doteraz riešila iba parciálne. Autorka sa usiluje o systematické posúdenie rizík pretrénovania (1) a adverzariálnych útokov (2) v súdnej praxi, a to naprieč jednotlivými typmi modelov strojového učenia (3). Príspevok sa zameriava na kritické zhodnotenie implementácie nástrojov umelej inteligencie v rámci „životného cyklu" rozhodovacej činnosti súdov, so zreteľom na súvisiacu *ex post* právnu zodpovednosť.

doi: 10.46282/bpf.2025.05

Pretrénovanie: „Achillova päta" modelov strojového učenia

Predtým ako vôbec môžeme uvažovať o nasadení systémov umelej inteligencie do prostredia súdov, resp. predtým ako vôbec môžeme vysloviť závery o bezpečnosti takého nasadenia, je nevyhnutné pochopiť, ako sa tento výsledný model učí. Napriek svojej výkonnosti vykazujú systémy umelej inteligencie určité nedostatky, ktoré sa najvýraznejšie prejavujú v samotnom procese učenia. Než pristúpime ku analýze rizikovosti jednotlivých modelov strojového učenia, je z hľadiska logickej štruktúry nevyhnutné v prvom rade objasniť „Achillovu pätu" *všetkých* modelov strojového učenia.

Strojové učenie je podmnožina umelej inteligencie, ktorá sa zameriava na rozpoznávanie a učenie sa vzorcov v datasetoch. Ide o aplikáciu umelej inteligencie umožňujúcu systémom automaticky sa zlepšovať a prispôsobovať na základe skúseností, bez toho, aby boli explicitne programované.[2] Výsledkom takého učenia je potom schopnosť systémov napr. analyzovať vzorce, vytvárať predikcie, sumarizácie či generalizácie atď. Strojové učenie teda spočíva v trénovaní systémov umelej inteligencie na sérií údajov alebo vstupov, z ktorých sa model na báze generalizácie učí. Úspešnosť jazykových modelov je, okrem iného, podmienená efektívnym využitím nástrojov na spracovanie prirodzeného jazyka (*„Natural Language Processing*" alebo *„NLP*"). *Per definitionem*, tieto nástroje „študujú" textové údaje, čím sa učia pochopiť ich úplný význam vrátane zámerov a emócií. V praxi si to môžeme predstaviť na príklade datasetu (údajov na vstupe) obsahujúceho 500 najreprezentatívnejších rozhodnutí súdu[3]. Model sa následne z týchto rozhodnutí učí rozpoznávať vzorce, analyzovať kontext alebo emócie a následne dokáže vygenerovať relevantný výstup. V tomto štádiu ale najčastejšie dochádza k pretrénovaniu (tzv. *overfitting*).

Pretrénovanie je notoricky známy pojem spájaný so všetkými modelmi strojového učenia. Podstata pretrénovania spočíva v tom, že s rastúcim počtom iterácií sa model veľmi dobre naučí trénovacie údaje a zvyšuje sa aj jeho generalizačná výkonnosť (t. j. klesá chybovosť). Existuje ale „bod zlomu" kedy ďalšie trénovanie zvyšuje presnosť trénovacích údajov, ale znižuje generalizačnú výkonnosť.[4] Inými slovami, umelá

[2] ABDOLRASOL, Maher G. M. – SUHAIL HUSSAIN, S. M. – USTUN, Taha Sehim *et al.*: Artificial Neural Networks Based Optimization Techniques: A Review. [online]. In: *Electronics Journal*, Vol. 10, 2021, s. 1. [cit. 10-09-2025]. Dostupné na: https://www.mdpi.com/2079-9292/10/21/2689.

[3] Pozn. autorky: reprezentatívnosť a kvalita rozhodnutí je zásadná čo sa týka využívania modelov strojového učenia v súdnej praxi aby sa predišlo tomu, že sa model bude učiť na rozhodnutiach, ktoré sú zo svojej podstaty diskriminačné alebo právne či ústavne neudržateľné. Z toho dôvodu by sa vo fáze predspracovania údajov mali uvedené kritéria odfiltrovať.

[4] ALIFERIS, Constantin – SIMON, Gyorgy: Overfitting, Underfitting and General Model Overconfidence and Under-Performance Pitfalls and Best Practices in Machine Learning and AI.

doi: 10.46282/bpf.2025.05

neurónová sieť (*artificial neural network* alebo *ANN*) sa ako typ matematického modelu napodobňujúceho kognitívne schopnosti človeka[5] (neurónové siete v ľudskom mozgu), príliš prispôsobí konkrétnym vstupným údajom. To má za príčinu, že model nedokáže generalizovať ani vytvárať správne výstupy založené na nových údajoch než na ktorých bol trénovaný. V jednoduchosti to znamená, že pretrénovaný model reaguje citlivo na malé zmeny vo vstupoch.

Pretrénovanie vieme veľmi jednoducho ilustrovať na vyššie uvedenom príklade. Model sa naučí s vysokou presnosťou analyzovať 500 rozhodnutí súdu, na ktorých bol trénovaný, a to vrátane konkrétnych detailov nazývaných aj „šum", ktorý pre posúdenie veci neobsahuje žiadne relevantné informácie. Akonáhle sa však objaví nové rozhodnutie obsahujúce mierne odlišné údaje než tie na ktorých bol model trénovaný, model ich nedokáže správne vyhodnotiť. V kontexte súdnych rozhodnutí, kde je každý prípad založený na konkrétnych skutkových okolnostiach, je schopnosť modelu správne generalizovať smerodajná. Pretrénovaný model, ktorý sa príliš zameria na detaily z predchádzajúcich rozhodnutí (t. j. „šum"), môže vyhodnotiť nové rozhodnutie nesprávne. Táto znížená schopnosť modelu generalizovať potom vedie k nepredvídateľným až arbitrárnym záverom, ktoré nereflektujú variabilitu spoločenských vzťahov a ktoré nemusia byť pre právne posúdenie veci relevantné.

Hoci je pretrénovanie považované za bežný jav,[6] je kľúčové pre pochopenie limitov umelej neurónovej siete a teda aj praktického využitia jazykových modelov v súdnej praxi.

Adverzariálne útoky: zraniteľnosť rozhodovacieho procesu súdov

Adverzariálne útoky alebo inak nazývané nepriateľské strojové učenie predstavuje externú manipuláciu vstupných údajov s cieľom ovplyvniť výstup modelu. Nedávne štúdie ukázali, že jazykové modely založené na spracovaní prirodzeného jazyka sú

[online]. In: *Artificial Intelligence and Machine Learning in Health Care and Medical Sciences. Best Practices and Pitfalls.* Berlin: Springer, 2024, s. 481. [cit. 10-09-2025]. Dostupné na: https://link.springer.com/chapter/10.1007/978-3-031-39355-6_10.

[5] WU, Caicong - CHENG, Xiuwan – YANG, Yinsheng: Decision-Making Modeling Method Based on Artificial Neural Network and Data Envelopment Analysis. [online]. In: *International Geoscience and Remote Sensing Symposium (IGARSS)*, 2004, s. 1. [cit. 10-09-2025]. Dostupné na: IEEE Xplore Full-Text PDF:.

[6] Pozn. autorky: pri trénovaní toho-ktorého modelu existujú opatrenia zamerané buď na reguláciu a minimalizáciu pretrénovania alebo na jeho úplnú prevenciu. Napriek týmto opatreniam však pretrénovanie predstavuje inherentnú vlastnosť každého modelu strojového učenia.

doi: 10.46282/bpf.2025.05

mimoriadne zraniteľné voči adverzariálnym útokom,[7] keďže v textových údajov sa dá ľahko upraviť pravopis, slová alebo štruktúra textu.[8]

Medzi najčastejšie uvádzané príklady adverzariálnych útokov patria tie, ktoré spočívajú v minimálnych perturbáciách vstupu či už na úrovni vektorov, jednotlivých slov, fráz alebo písmen. Tieto nepatrné zmeny, hoci pre človeka ostávajú väčšinou nepostrehnuteľné môžu viesť k tomu, že model vygeneruje nesprávny, zavádzajúci alebo nerelevantný výstup.

(i) *Útok prostredníctvom vkladania slov*. Vkladanie slov je typ reprezentácie slov vo forme numerických vektorov kedy slová s podobným významom a kontextom sú reprezentované podobnými vektormi.[9] Inými slovami, ide o metódu prevodu textu (ktorý systémy umelej inteligencie nedokážu priamo pochopiť) do číselnej podoby, ktorú je model schopný spracovať. Model dokáže identifikovať, ktoré slová sa používajú v podobnom kontexte a na základe týchto informácií ich usporiada do tesnej blízkosti vo vektorovom priestore.

Príklad: *sudca* → [0.12, 0.85, ...], *súd* → [0.11, 0.87, ...].

Adverzariálny útok tak tkvie v tom, že útočník mierne upraví vektory (číselné hodnoty), aby model predikoval nesprávne.[10]

(ii) *Útok prostredníctvom manipulácie so slovami a frázami.* Adverzariálny útok na úrovni slov môže byť založený na identifikácii tzv. „horúcich znakov" (*hot characters*), teda písmen s najväčším vplyvom na výstup modelu. Slová obsahujúce vyšší počet týchto znakov sa následne označujú ako „horúce slová" (*hot words*), ktoré je možné ďalej kombinovať do „horúcich fráz" (*hot phrases*). Manipulácia s týmito frázami môže ovplyvniť predikciu modelu. Útočník týmto postupom identifikuje „slabé miesta" v konkrétnej vete, ktoré následne umožňujú pridanie alebo odstránenie „horúcej frázy".[11]

---

[7] ZHANG, Yu. – YANG, Junan – LI, Xiaoshuai *et al.*: Textual Adversarial Attacking with Limited Queries. [online]. In: *Electronics Journal*, Vol. 10, 2021, s. 1. [cit. 10-09-2025]. Dostupné na: https://www.mdpi.com/2079-9292/10/21/2671.

[8] XU, H. – MA, Y. – LIU, H.-CH. *et al.*: Adversarial Attacks and Defenses in Images, Graphs and Text: A Review. [online]. In: *International Journal of Automation and Computing*. Berlin: Springer, 2020, s. 168. [cit. 10-09-2025]. Dostupné na: https://link.springer.com/article/10.1007/s11633-019-1211-x.

[9] NASSIF, Ali Bou – ELNAGAR, Ashraf – SHAHIN, Ismail – HENNO, Safaa: Deep learning for Arabic subjective sentiment analysis: Challenges and research opportunities. [online]. In: *Applied Soft Computing Journal.* 2021, s. 4. [cit. 10-09-2025]. Dostupné na: https://www.sciencedirect.com/science/article/abs/pii/S1568494620307742.

[10] Útočník dokáže zanalyzovať, ako citlivý je model na malé zmeny v každom vektore slova. Potom vykoná nepatrnú zmenu vektora, aby vo výsledku narušil výstup napr. predikciu modelu.

[11] XU, H. – MA, Y. – LIU, H.-CH. *et al.*: Adversarial Attacks and Defenses in Images, Graphs and Text: A Review. [online]. In: *International Journal of Automation and Computing*. (premenovaný na *Machine Intelligence Research*). Berlin: Springer, 2020, Vol. 17, s. 168. [cit. 10-09-2025]. Dostupné na: https://link.springer.com/article/10.1007/s11633-019-1211-x.

doi: 10.46282/bpf.2025.05

Príklad:
Pôvodná veta v Náleze: *„… bolo porušené základné právo sťažovateľa.*"
Horúce znaky: p, r, á
Horúce slová: porušené, právo
Horúca fráza: porušené právo
Adverzariálny útok: *„… pravdepodobne bolo porušené základné právo sťažovateľa.*"
Uvedená modifikácia môže spôsobiť, že model vyhodnotí vstupné údaje nesprávne a dospeje k záveru, že základné právo sťažovateľa *„porušené nebolo*".
(iii) *Útok prostredníctvom manipulácie s písmenami, resp. znakmi.* Ďalším príkladom je adverzariálny útok nahradením písmena vo vete (*word embedding*) s cieľom oklamať model na úrovni znakov (kde je každé písmeno reprezentované vektormi). Zmena jedného písmena vo vete, mení predikciu modelu týkajúcu sa danej témy.[12]
Príklad:
Pôvodná veta v Náleze: *„… bolo porušené základné právo sťažovateľa.*"
Adverzariálny útok: *„… bolo porušené základné právo sťažovateľa.*"
Zmena na úrovni písmen sa ľudským okom dá chápať v zmysle chyby v písaní. Význam slova ostáva pre človeka ako taký nemenný, no model môže zmiasť do takej miery, že vygeneruje nesprávny výstup.
Uvedené stratégie dosahujú dobré výsledky z hľadiska efektívnosti útokov a s tým spojenej „kvality" výstupov.[13] Veľké jazykové modely (teda modely, ktoré by časom boli implementované do súdnej praxe) sú mimoriadne zraniteľné voči strategickým adverzariálnym útokom, ktoré využívajú slabé miesta modelu. Čím viac textových údajov, tým viac slabých miest. Ako vyplýva z predchádzajúcej časti príspevku, pretrénovaný model už vo všeobecnosti vykazuje neprimerane vysokú citlivosť na drobné zmeny vo vstupných údajoch. Vyššie uvedené tak priamo potvrdzuje, že pretrénovanie jazykových modelov výrazne zvyšuje ich zraniteľnosť voči adverzariálnym útokom.
Adverzariálne útoky teda predstavujú vážnu hrozbu pre robustnosť modelov a vyžadujú proaktívne opatrenia na zmiernenie ich rizík.[14] V súdnej praxi by to znamenalo, že akýkoľvek systém založený na takto pretrénovanom modeli môže byť o to ľahšie manipulovaný zo strany externých aktérov. Ak si daný scenár premietneme do prostredia súdov, tak sú prípustné úvahy do akej miery je vôbec opodstatnené nasadenie systémov umelej inteligencie na účely zefektívnenia

[12] Ibidem, s. 169.
[13] ZHANG, Yu. – YANG, Junan – LI, Xiaoshuai *et al.*: Textual Adversarial Attacking with Limited Queries. [online]. In: *Electronics Journal*, Vol. 10, 2021, s. 3. [cit. 11-09-2025]. Dostupné na: https://www.mdpi.com/2079-9292/10/21/2671.
[14] WANG, Yulong – SUN, Tong – LI, Shenghong et al.: Adversarial Attacks and Defenses in Machine Learning-Powered Networks: A Contemporary Survey. [online]. Cornell Tech: arXiv:2303.06302v1, 2023, s. 1. [cit. 11-09-2025]. Dostupné na: https://arxiv.org/pdf/2303.06302.

doi: 10.46282/bpf.2025.05

rozhodovacích procesov súdu. Vzhľadom na potenciálne ohrozenie nezávislosti súdnej moci *en bloc*, ktoré vyplýva zo zvýšenej zraniteľnosti jazykových modelov voči adverzariálnym útokom, je nevyhnutné analyzovať túto problematiku s dôrazom na špecifiká jednotlivých typov modelov strojového učenia.

Modely strojového učenia: ich robustnosť voči adverzariálnym útokom a dôsledky pre nezávislosť súdnej moci

Trénovanie modelov prostredníctvom strojového učenia sa realizuje rôznymi spôsobmi. Na základe štýlu učenia ich potom delíme do troch hlavných kategórií: učenie s učiteľom (*supervised learning*)[15]; učenie bez učiteľa (*unsupervised learning*); učenie posilňovaním (*reinforcement learning*).

1. Princíp učenia s učiteľom[16] je založený na trénovacích údajoch, ktoré obsahujú nielen vstupné, ale aj predpokladané výstupné údaje. Model má počas učenia prístup k správnym údajom alebo označeným výstupom. Tento typ učenia je analogický s ľudským učením sa zo skúseností s cieľom získať nové vedomosti za účelom zlepšiť našu schopnosť vykonávať určité úlohy. Keďže systémy umelej inteligencie nedisponujú vlastnými „skúsenosťami", proces strojového učenia im umožňuje získavať poznatky zo zhromaždených údajov.[17] Zhromažďovanie údajov je podmienené ich anotáciou, teda procesom označovania (v našom ilustračnom prípade ide o textové údaje - rozhodnutia súdu). Autonómne fungovanie systémov umelej inteligencie si vyžaduje rozsiahle množstvo takto označených údajov, ktoré pre model slúžia ako „zdroj poznania". To znamená, že ľudský operátor manuálne preskúma textové údaje (napr. nami vybraných 500 rozhodnutí súdu) a priradí im správne označenie podľa vopred definovaných pokynov. Pokyn môže byť formulovaný v zmysle: „identifikuj druh konania" (napr. konanie o ústavnej sťažnosti). Ľudský operátor si následne prečíta vybrané rozhodnutia súdu a k pokynom priradí správne označenia. Modely strojového učenia založené na učení s učiteľom sa týmto spôsobom môžu pokúsiť realizovať úlohy ako predikcie, vyhľadávanie relevantných rozhodnutí súdu alebo zhrnúť odôvodnenie rozhodnutia súdu. Ľudský operátor pri tomto type učenia hodnotí správnosť generovaných výstupov.

[15] Pozn. autorky: existujú aj modely ako semi-supervised a self-supervised. Ich technické parametre však s ohľadom na právnu orientáciu príspevku nie je potrebné detailne objasňovať.

[16] Pozn. autorky: osobitná pozornosť bude venovaná tréningu spočívajúcom v učení s učiteľom, vzhľadom na fakt, že je v súčasnosti najrozšírenejším typom učenia v oblastiach ako súdnictvo, zdravotníctvo a pod. Aj keď ide o najčastejšie využívaný model, jeho závislosť na ľudskom operátorovi prináša špecifické riziká.

[17] LIU, Bing: Web Data Mining. Exploring Hyperlinks, Contents, and Usage Data. [online]. Berlin: Springer, 2011, s. 63. [cit. 12-09-2025]. Dostupné na: https://link.springer.com/chapter/10.1007/978-3-642-19460-3_3.

doi: 10.46282/bpf.2025.05

Učenie s učiteľom založené predovšetkým na manuálnom označovaní údajov, sa na prvý pohľad javí ako najvhodnejší prístup z hľadiska požiadaviek kladených na ľudský dohľad nad automatizovanými procesmi systémov umelej inteligencie, ako sú tieto rámcovo definované v AIA. Na tomto mieste sa názory môžu začať rozchádzať, a to pri otázke či a do akej miery možno ľudský dohľad, tak ako je koncipovaný v AIA, považovať za „dôveryhodný"? Ak zohľadníme nami identifikovanú slabinu všetkých modelov strojového učenia v podobe pretrénovania a potenciál jej zneužitia prostredníctvom adverzariálnych útokov, možno dospieť k záveru, že so zvyšovaním požiadaviek na ľudský dohľad, a teda s rastúcim počtom ľudských operátorov, narastá aj riziko úspešnej realizácie niektorej z foriem útoku. Zároveň však nie je efektívnym riešením ani spoliehanie sa na dohľad jedným ľudským operátorom.
Ďalší aspekt vyžadujúci kritickú reflexiu je transparentnosť. Transparentnosť je „živou krvou" súdnictva, nakoľko sa od toho odvíja úroveň dôvery spoločnosti v súdnu moc a tým pádom aj jej legitimita. Požiadavka transparentnosti z toho dôvodu predstavuje inherentnú súčasť implementácie systémov umelej inteligencie do súdnej praxe. Jednoduché, malé jazykové modely (na účely súdnych procesov sa domnievame, že *White-Box* modely)[18] založené na učení s učiteľom sú preferované pre svoju transparentnosť nevyhnutnú vo vysokorizikových prostrediach. Paradoxne však otvorená architektúra takýchto modelov zároveň zvyšuje ich zraniteľnosť voči adverzariálnym útokom, pretože umožňuje útočníkovi identifikovať slabé miesta modelu na realizáciu potenciálne cielených adverzariálnych útokov. Útočník potrebuje **zacieliť adverzariálny útok**, a to často s využitím znalosti modelu (na ktorého učení aj potenciálne participoval).[19] V kontexte veľkých jazykových modelov využiteľných v súdnictve si tieto útoky vyžadujú istú úroveň stratégie, často s využitím znalosti modelu, čím sa zvyšuje riziko manipulácie vstupných údajov, najmä pri spracovaní rozsiahlych textových údajov akými sú rozhodnutia súdu, kde počet slabých miest narastá s veľkosťou a komplexnosťou textu.[20]

[18] Akýkoľvek model, ktorý je plne transparentný a spätne sa dá vysvetliť jeho myšlienková línia, sa nazýva „*White-Box*" model. To znamená, že dokážeme pochopiť, ako model funguje. NEUER, Marcus J.: Machine Learning for Engineers. Introduction to Physics-Informed, Explainable Learning Methods for AI in Engineering Application. [online]. Berlin: Springer, 2025, s. 3. [cit. 12-09-2025]. Dostupné na: https://www.scribd.com/document/865803419/Machine-Learning-for-Engineers-2025#page=150.

[19] Pri White-Box modeloch sa predpokladá, že útočníci majú úplnú znalosť cieľového modelu, vrátane jeho architektúry a parametrov. Preto môžu priamo vytvárať adverzariálne útoky akýmkoľvek spôsobom. Bližšie pozri: REN, Kui – ZHENG, Tianhang – QUIN, Zhan – LIU, Xue: Adversarial Attacks and Defenses in Deep Learning. [online]. In: *Engineering Journal,* 2020, s. 1. [cit. 12-09-2025]. Dostupné na: https://www.sciencedirect.com/science/article/pii/S209580991930503X.

[20] Pozn. autorky: Napriek narastajúcemu využívaniu modelov strojového učenia s učiteľom, väčšina primárnych alebo sekundárnych zdrojov (napr. aj AIA alebo vedecké články zaoberajúce sa problematikou umelej inteligencie) v súčasnosti analyzujú využitie Black-Box modelov alebo

doi: 10.46282/bpf.2025.05

2. Pri type učenia bez učiteľa model nemá k dispozícii predpokladané výstupné údaje alebo označené údaje, prostredníctvom ktorých ľudský operátor kontroluje správnosť výstupu. Namiesto toho model vo vstupných textových údajoch (t. j. naša vzorka 500 rozhodnutí súdu) hľadá podobnosti a vzory čím sa v údajoch snaží odhaliť skryté štruktúry,[21] ktoré človek môže prehliadnuť. Príkladom je situácia keď model hľadá skupiny súvisiacich údajov, „zhluky", ktoré vykazujú podobné správanie v procesoch z hľadiska ich charakteristík.[22] Svojou povahou je daný typ strojového učenia vhodný na analytické úlohy ako je rešerš, abstrakcia.

Modely založené na učení bez učiteľa sú vo všeobecnosti **odolnejšie voči pretrénovaniu** než modely založené na učení s učiteľom, nakoľko nepracujú s označenými vstupmi a výstupmi, ale hľadajú vzorce. Z toho dôvodu je menej časté, že by sa model príliš prispôsobil a „naučil naspamäť" konkrétne detaily, tzv. šum. Napriek tomu aj tieto modely môžu niekedy zachytávať šum alebo nepodstatné vzory, t. j. nie sú úplne bez rizika. Zároveň ale tieto modely nie sú priamo závislé od ľudského operátora v zmysle manuálneho označovania údajov.

Aj keď je model bez učiteľa, spolieha sa na vektorové (číselné) reprezentácie slov. Adverzariálne útoky nie sú vylúčené a vo výsledku môžu skresliť naučené vzorce alebo vzťahy medzi slovami. To znamená, že model tieto vzorce alebo vzťahy nesprávne vyhodnotí a zle zatriedi údaje do skupín, resp. vygeneruje údaje, ktoré sú irelevantné (napr. rozhodnutia súdu, ktoré navzájom nijako nesúvisia). To samo osebe nepredstavuje až taký problém s ohľadom na nezávislosť súdnictva, keďže výstupy modelu učeného bez učiteľa predstavujú iba medzištádium a potenciálny analytický podklad pre ďalšie procesy, ktoré vedú k vyhotoveniu samotného rozhodnutia súdu. *Status quo* by tak ostal zachovaný. Rozhodovací proces by bol naďalej závislý na personálnej zložke súdu.

3. Učenie s posilňovaním funguje obdobne na princípe získavania skúseností modelu. Pri type učenia s posilňovaním však model (tzv. agent) získava skúsenosti interakciou v prostredí, v ktorom sa nachádza. Ako praktický príklad využitia modelov učenia posilňovaním možno uviesť autonómne vozidlá. Na začiatku učenia agent vykonáva náhodné akcie, pri ktorých získava odmeny (napr. v kontexte autonómnych vozidiel

---

výnimočne jednoduchých White-Box modelov. Doteraz však nebola systematicky preskúmaná kombinácia komplexných veľkých White-Box jazykových modelov založených na učení s učiteľom využívaných na účely súdneho rozhodovania, a to z hľadiska ich citlivosti na pretrénovanie a adverzariálne útoky. Situáciu ďalej komplikuje skutočnosť, že ľudskí operátori môžu disponovať vedomosťami o architektúre a procese trénovania modelu. Táto kombinácia so sebou opätovne prináša riziká, ktoré nie sú reflektované v súčasných reguláciách.

[21] NEUER, Marcus J.: Machine Learning for Engineers. Introduction to Physics-Informed, Explainable Learning Methods for AI in Engineering Application. [online]. Berlin: Springer, 2025, s. 141. [cit. 12-09-2025]. Dostupné na: https://www.scribd.com/document/865803419/Machine-Learning-for-Engineers-2025#page=150.

[22] Ibidem, s. 89.

doi: 10.46282/bpf.2025.05

sa počas tréningu model vyhne prekážke) alebo tresty (napr. model zlyhá a narazí do prekážky). Odmeny ho motivujú vykonávať akcie, ktoré mu priniesli pozitívnu spätnú väzbu. Tresty majú opačný efekt a agent sa ich snaží minimalizovať, tým že obmedzuje vykonávanie akcií, za ktoré bol v minulosti trestaný.

Modely založené na princípe učenia s učiteľom alebo bez učiteľa sú schopné generovať výsledky na základe tréningových údajov, avšak absentuje u nich mechanizmus, ktorý by im umožňoval *adaptívne* reagovať na dynamické prostredie v reálnom čase. Modely učené posilňovaním dokážu, naopak, využiť historické údaje a interakciou s prostredím v reálnom čase dokážu generovať výstupy na báze adaptívneho rozhodovania. Takýto prístup umožňuje vytvárať modely, ktoré sa samé prispôsobujú meniacim sa podmienkam.[23] Otvára sa tým možnosť automatizácie a efektívnejšej optimalizácie rôznych oblastí, súdnictvo nevynímajúc. Keďže sa však agent učí sám na základe vlastných skúseností, jeho rozhodovací proces je ťažko sledovateľný. To je v rozpore so základnými princípmi transparentnosti a legitimity súdnej moci. Absencia transparentnosti spôsobuje, že akékoľvek vyvodzovanie zodpovednosti je takmer nemožné.

V súvislosti s pretrénovaním majú modely učené na princípe posilňovania tendenciu si zapamätať a naučiť sa výlučne trénovacie údaje alebo vzory. To inými slovami znamená, že namiesto toho, aby sa naučili **generalizovať rôzne alternatívy riešenia úloh**, sú schopné opakovať iba konkrétne stratégie z tréningu. Hoci, počas tréningu dosahujú **optimálne výsledky**, ich správanie v nepredvídateľných situáciách (t. j. v situáciách, ktoré netvorili súčasť tréningu) môže byť výrazne odlišné a nekonzistentné.[24] Napriek tomu, že modely učené posilňovaním sú veľmi výkonné, považujú sa za potenciálne menej spoľahlivé a interpretovateľné pri praktickej implementácii do vysokorizikového prostredia.

Ako sporná sa opäť javí byť prítomnosť ľudského operátora. Už bolo uvedené, že model učený posilňovaním je ťažko interpretovateľný. To znamená, že adverzariálny útok zvyčajne nie je priamo detekovateľný a je možné ho identifikovať až s odstupom času na základe jeho reálnej interakcie s prostredím. Hoci sú White-Box modely učené posilňovaním menej rozšírené, ich použitie by v kontexte súdnictva pravdepodobne bolo nevyhnutné, pričom by zároveň prinášalo špecifické riziká vyplývajúce z prístupu ľudského operátora k internej architektúre modelu.

---

[23] SHAIK, Thanveer – TAO, Xiaohui – LI, Lin *et al.*: Predictive deep reinforcement learning with multi-agent systems for adaptive time series forecasting. [online]. In: *Knowledge-Based Systems*, Vol. 326, 2025, s. 2. [cit. 12-09-2025]. Dostupné na: https://www.sciencedirect.com/science/article/pii/S0950705125009864.

[24] ZHANG, Chiyuan – VINYALS, Oriol – MUNOS, Remi – BENGIO, Samy: A Study on Overfitting in Deep Reinforcement Learning. [online]. Cornell Tech: arXiv:1804.06893v2, 2018, s. 2. [cit. 18-09-2025]. Dostupné na: https://arxiv.org/pdf/1804.06893.

doi: 10.46282/bpf.2025.05

Pri tomto type učenia je najčastejšou formou adverzariálneho útoku otrávenie odmien (*reward-poisoning*). Útočník má plnú znalosť reálneho prostredia (t. j. aj skutočných odmien), prípadne pozná aj učebný algoritmus agenta, resp. pozná oboje. Keďže všetky údaje sú pre útočníka dostupné naraz, v takom prípade dochádza k adverzariálnym útokom, ktoré pri minimálnych zásahoch do systému odmeňovania dokážu agenta naviesť k osvojeniu nežiaducich zmanipulovaných výstupov.[25] V praxi majú prístup k agentovi najčastejšie ľudskí operátori zodpovední za jeho tréning a to tak, že upravujú rozhodovaciu politiku agenta aby robil „správne" alebo očakávané rozhodnutia. Adverzariálne útoky zo strany externých aktérov sú reálnou hrozbou, no najväčšie riziko predstavujú najmä ľudskí operátori, pretože disponujú priamym (White-Box) prístupom k údajom, politike a mechanizmom odmeňovania, čo im umožňuje realizovať adverzariálne útoky (úmyselné aj neúmyselné) typu *reward-poisoning*, *data-poisoning* alebo iné. Tie sú pre externého aktéra ťažko realizovateľné v dôsledku absencie prístupu k architektúre modelu. Navyše adverzariálne útoky pri type učenia posilňovaním sú ťažko detekovateľné a k ich identifikácii dochádza v praxi až s odstupom času.

Záver

Je jedna vec špekulovať či nástroje umelej inteligencie dokážu zefektívniť a optimalizovať súdne procesy alebo či naopak ohrozia nezávislosť súdnej moci.
Úroveň potenciálneho rizika pri nasadzovaní nástrojov umelej inteligencie do súdnictva je však závislá nielen od typu modelu strojového učenia ale aj od spoľahlivosti personálnej zložky zodpovednej za tréning modelov a za výber údajov, na ktorých sa model učí.
Príspevok podrobne analyzoval dve technicko-právne hrozby spojené s nasadením modelov strojového učenia do rozhodovacej činnosti súdov, pričom sa zameral na ich zraniteľnosť voči pretrénovaniu (overfitting) a adverzariálnym útokom. Ukázalo sa, že zraniteľnosť jazykových modelov sa zvyšuje v dôsledku ich pretrénovania, čo ich robí neprimerane citlivými na drobné, strategicky cielené manipulácie vstupných údajov. To priamo ohrozuje základné princípy materiálneho právneho štátu a nezávislosť súdnictva, nakoľko by to mohlo viesť k nepredvídateľným, arbitrárnym alebo dokonca úmyselne zmanipulovaným výstupom.
Analýza rizikovosti realizovaná naprieč jednotlivými typmi modelov strojového učenia (s učiteľom, bez učiteľa a učenie posilňovaním) v kontexte ich odolnosti voči adverzariálnym útokom viedla k nasledujúcim záverom.

[25] RAKHSHA, Amin – ZHANG, Xueshou – ZHU, Xiaojin – SINGLA, Adish: Reward Poisoning in Reinforcement Learning: Attacks Against Unknown Learners in Unknown Environments. [online]. Cornell Tech: arXiv:2102.08492, 2021, s. 8. [cit. 18-09-2025]. Dostupné na: https://arxiv.org/pdf/2102.08492v1.

doi: 10.46282/bpf.2025.05

(i) Učenie s učiteľom založené predovšetkým na manuálnom označovaní údajov, sa na prvý pohľad javí ako najvhodnejší prístup z hľadiska požiadaviek kladených na ľudský dohľad nad automatizovanými procesmi systémov umelej inteligencie, ako sú tieto rámcovo definované v Akte o umelej inteligencii. Ak zohľadníme nami identifikovanú slabinu všetkých modelov strojového učenia v podobe pretrénovania a potenciál jej zneužitia prostredníctvom adverzariálnych útokov, možno dospieť k záveru, že so zvyšovaním požiadaviek na ľudský dohľad, a teda s rastúcim počtom ľudských operátorov, narastá aj riziko úspešnej realizácie niektorej z foriem útoku. V prípade generovania výstupu v podobe sumarizácie alebo predikcie na základe právneho posúdenia veci následne môže dôjsť k ovplyvneniu výsledného rozhodnutia súdu.
(ii) Učenie bez učiteľa (vhodné pre analytické úlohy) vykazuje vyššiu odolnosť voči pretrénovaniu. Tieto modely nie sú priamo závislé od ľudského operátora v zmysle manuálneho označovania údajov. Adverzariálne útoky nie sú vylúčené a vo výsledku môžu skresliť naučené vzorce alebo vzťahy medzi slovami. To znamená, že model tieto vzorce alebo vzťahy nesprávne vyhodnotí a zle zatriedi údaje do skupín, resp. vygeneruje údaje, ktoré sú irelevantné (napr. rozhodnutia súdu, ktoré navzájom nijako nesúvisia). Inými slovami, výstupy tohto modelu slúžia ako podklady pre rozhodnutie, čím je minimalizované priame ohrozenie nezávislosti rozhodovacieho procesu, a to aj napriek tomu, že dôjde k realizácii adverzariálneho útoku.
(iii) Pri učení s posilňovaním sa model učí sám na základe vlastných skúseností a jeho rozhodovací proces je ťažko sledovateľný. Keďže všetky údaje sú pre útočníka dostupné naraz, v takom prípade dochádza k adverzariálnym útokom, ktoré pri minimálnych zásahoch do systému odmeňovania dokážu agenta naviesť k osvojeniu nežiaducich zmanipulovaných výstupov. Napriek tomu, že modely učené posilňovaním sú veľmi výkonné a autonómne, považujú sa za potenciálne menej spoľahlivé a interpretovateľné pri praktickej implementácii do vysokorizikového prostredia. Absencia transparentnosti spôsobuje, že adverzariálne útoky sú ťažko detekovateľné a k ich identifikácii dochádza v praxi až s odstupom času. Akékoľvek vyvodzovanie zodpovednosti je v tomto smere takmer nemožné.
(iv) Prostredníctvom komplexnej analýzy sme identifikovali vnútorný rozpor vo filozofii, na ktorej je vystavaný Akt o umelej inteligencii. Kým Akt o umelej inteligencii akcentuje ľudský dohľad pri využívaní systémov umelej inteligencie vo vysokorizikových oblastiach (napr. súdnictvo), nedostatočne reflektuje a reguluje dohľad a kontrolu nad samotnými ľudskými operátormi a ich činnosťou vo fáze tréningu jednotlivých modelov. Práve ľudskí operátori disponujú vedomosťami o „slabých miestach" modelu čím predstavujú najväčšie riziko realizácie niektorej z foriem strategicky cielených adverzariálnych útokov.
Nasadzovanie systémov umelej inteligencie do súdnictva by sa malo obmedziť na nástroje vhodné pre analytické úlohy, ktoré priamo nenahrádzajú rozhodovaciu

doi: 10.46282/bpf.2025.05

autonómiu sudcu. Akékoľvek nasadenie systémov umelej inteligencie spočívajúcich v generovaní výstupov ako sú predikcie alebo samotné odôvodnenia rozhodnutí spočívajúce v právnom posúdení veci, by malo byť založené na regulácii obsahovo zameranej na tréningovú fázu modelov strojového učenia vrátane otázky o zodpovednosti ľudských operátorov. Taktiež by sa mala akcentovať potreba metodiky zameranej na minimalizáciu a prevenciu pretrénovania a adverzariálnych útokov, a to prostredníctvom testovacích údajov a adverzariálnych príkladov (*adversarial examples*), najmä pri učení posilňovaním.
Efektívnosť a optimalizácia súdnictva by sa nemali dosahovať na úkor transparentnosti, najmä v krajinách kontinentálneho práva, kde súdy čelia nadmernej administratívnej záťaži. Hoci transparentnosť predstavuje jedno z rizík zavádzania nástrojov umelej inteligencie do prostredia súdov, mala by byť dôsledne reflektovaná nielen transparentnosť životného cyklu od vstupu po výstup modelu ale najmä transparentnosť tréningového procesu vybraných modelov vo vzťahu k ľudským operátorom. Koncept dôveryhodnej umelej inteligencie, ako je akcentovaný Aktom o umelej inteligencii, je udržateľný do takej miery do akej bude dôveryhodne nastavený výber ľudských operátorov, ktorí dohliadajú na tréning modelov, ako aj výber údajov, na ktorých sa model trénuje.

Zoznam použitej literatúry

1. ABDOLRASOL, Maher G. M. – SUHAIL HUSSAIN, S. M. – USTUN, Taha Sehim *et al.*: Artificial Neural Networks Based Optimization Techniques: A Review. In: *Electronics Journal*, Vol. 10, 2021, 43 s. Dostupné na: https://www.mdpi.com/2079-9292/10/21/2689.
2. ALIFERIS, Constantin – SIMON, Gyorgy: Overfitting, Underfitting and General Model Overconfidence and Under-Performance Pitfalls and Best Practices in Machine Learning and AI. In: *Artificial Intelligence and Machine Learning in Health Care and Medical Sciences. Best Practices and Pitfalls.* Berlin: Springer, 2024, s. 477 - 524. Dostupné na: https://link.springer.com/chapter/10.1007/978-3-031-39355-6_10.
3. LIU, Bing: Web Data Mining. Exploring Hyperlinks, Contents, and Usage Data. Berlin: Springer, 2011, 624 s. Dostupné na: https://link.springer.com/chapter/10.1007/978-3-642-19460-3_3.
4. NASSIF, Ali Bou – ELNAGAR, Ashraf – SHAHIN, Ismail – HENNO, Safaa: Deep learning for Arabic subjective sentiment analysis: Challenges and research opportunities. In: *Applied Soft Computing Journal.* 2021, 27 s. Dostupné na: https://www.sciencedirect.com/science/article/abs/pii/S1568494620307742.
5. NEUER, Marcus J.: Machine Learning for Engineers. Introduction to Physics-Informed, Explainable Learning Methods for AI in Engineering Application.

Berlin: Springer, 2025, 258 s. Dostupné na: https://www.scribd.com/document/865803419/Machine-Learning-for-Engineers-2025#page=150.
6. RAKHSHA, Amin – ZHANG, Xueshou – ZHU, Xiaojin – SINGLA, Adish: Reward Poisoning in Reinforcement Learning: Attacks Against Unknown Learners in Unknown Environments. Cornell Tech: arXiv:2102.08492, 2021, 22 s. Dostupné na: https://arxiv.org/pdf/2102.08492v1.
7. REN, Kui – ZHENG, Tianhang – QUIN, Zhan – LIU, Xue: Adversarial Attacks and Defenses in Deep Learning. In: *Engineering Journal,* 2020, 15 s. Dostupné na: https://www.sciencedirect.com/science/article/pii/S209580991930503X.
8. SHAIK, Thanveer – TAO, Xiaohui – LI, Lin *et al.*: Predictive deep reinforcement learning with multi-agent systems for adaptive time series forecasting. In: *Knowledge-Based Systems*, Vol. 326, 2025, 16 s. Dostupné na: https://www.sciencedirect.com/science/article/pii/S0950705125009864.
9. WANG, Yulong – SUN, Tong – LI, Shenghong *et al.*: Adversarial Attacks and Defenses in Machine Learning-Powered Networks: A Contemporary Survey. Cornell Tech: arXiv:2303.06302v1, 2023, 46 s. Dostupné na: https://arxiv.org/pdf/2303.06302.
10. WU, Caicong - CHENG, Xiuwan – YANG, Yinsheng: Decision-Making Modeling Method Based on Artificial Neural Network and Data Envelopment Analysis. In: *International Geoscience and Remote Sensing Symposium (IGARSS)*, 2004, s. 2435 – 2438. Dostupné na: IEEE Xplore Full-Text PDF:.
11. ZHANG, Yu. – YANG, Junan – LI, Xiaoshuai *et al.*: Textual Adversarial Attacking with Limited Queries. In: *Electronics Journal*, Vol. 10, 2021, 12 s. Dostupné na: https://www.mdpi.com/2079-9292/10/21/2671.
12. ZHANG, Chiyuan – VINYALS, Oriol – MUNOS, Remi – BENGIO, Samy: A Study on Overfitting in Deep Reinforcement Learning. Cornell Tech: arXiv:1804.06893v2, 2018, 25 s. Dostupné na: https://arxiv.org/pdf/1804.06893.

Kontaktné údaje

Mgr. Romana Koneracká
romana.koneracka@flaw.uniba.sk
Katedra ústavného práva
Právnická fakulta Univerzity Komenského v Bratislave

doi: 10.46282/bpf.2025.05