DOI: https://doi.org/10.52651/nmb.c.2024.9788080406738.161-181

15. medzinárodná vedecká konferencia ♦15th International Scientific Conference
Národná a medzinárodná bezpečnosť 2024 ♦ National and International Security 2024
17. – 18. 10. 2024, Liptovský Mikuláš

VÝZVY SPOJENÉ S UMELOU INTELIGENCIOU PRE MEDZINÁRODNÝ MIER A BEZPEČNOSŤ

CHALLENGES ASSOCIATED WITH ARTIFICIAL INTELLIGENCE FOR INTERNATIONAL PEACE AND SECURITY

Milan KUSÁK

ABSTRACT

This article analyzes and maps the challenges that artificial intelligence poses to international peace and security. The article identifies and systematizes the main threats associated with the implementation and development of artificial intelligence technologies, with particular attention given to artificial intelligence safety, cybersecurity risks, and human-machine interaction. It also examines the potential consequences of these challenges for global security. The aim is to provide overview of artificial intelligence challenges and to support the discussion on measures to mitigate these threats, thereby contributing to the sustainable and secure development of artificial intelligence at the global level.

Keywords: Military applications of artificial intelligence, Autonomous weapons systems, Conflict escalation, Cyber risks, Human-machine interaction

ÚVOD

Umelá inteligencia predstavuje jednu z najrevolučnejších technológií súčasnosti, ktorá má potenciál výrazne ovplyvniť rôzne oblasti života, vrátane priemyslu, zdravotníctva, dopravy a dokonca aj vojenských operácií. Jej rýchly vývoj a integrácia do rozmanitých systémov prinášajú nespočetné výhody, ako je zvýšenie efektivity, automatizácia úloh či lepšie rozhodovanie na základe analýzy veľkého množstva dát. Avšak, spolu s týmito prínosmi prichádzajú aj nové výzvy a hrozby, ktoré si vyžadujú dôkladné posúdenie a reguláciu.

V súčasnom globálnom prostredí, ktoré je čoraz viac prepojené a komplexné, má umelá inteligencia potenciál nielen pozitívne, ale aj negatívne ovplyvniť medzinárodný mier a bezpečnosť. Systémy poháňané umelou inteligenciou sú čoraz viac používané v oblastiach, ktoré môžu priamo ovplyvniť dynamiku medzinárodných vzťahov, vrátane vojenských operácií, spravodajstva a kontroly zbraní. Napriek tomu, že technologický pokrok vždy sprevádzal vojenský vývoj, dnešné rýchlo sa meniace technológie predstavujú nové a často nepredvídateľné hrozby, ktoré môžu prekračovať hranice národných štátov a ovplyvňovať globálnu bezpečnosť.

Zvlášť znepokojujúci je potenciál umelej inteligencie na militarizáciu a jej použitie v autonómnych zbraňových systémoch, ktoré môžu operovať bez priamej kontroly človeka. Takéto systémy nielenže zvyšujú riziko eskalácie konfliktov, ale aj znižujú prah pre použitie sily, čím zvyšujú pravdepodobnosť ozbrojených stretov. Okrem toho, schopnosť systémov vybavených umelou inteligenciou rýchlo sa učiť a prispôsobovať novým situáciám znamená, že môžu byť zneužité alebo nesprávne použité spôsobom, ktorý ohrozuje medzinárodnú stabilitu.

Táto problematika nie je len teoretickou otázkou, ale stáva sa čoraz naliehavejšou témou v politických a bezpečnostných diskusiách na medzinárodnej úrovni. Správy ako "Nová agenda pre mier" od Generálneho tajomníka OSN a prvá diskusia o umelej inteligencii v Bezpečnostnej rade OSN konaná v júli 2023 zdôrazňujú význam multilaterálnych snáh o reguláciu vývoja a používania umelej inteligencie. Napriek týmto iniciatívam však stále chýba komplexný rámec na pochopenie a riadenie rizík spojených s touto technológiou.

1. VŠEOBECNE K VÝZVAM UMELEJ INTELIGENCIE

To ako sa výzvy pre medzinárodný mier a bezpečnosť prejavujú môže nadobúdať rôzne formy. Podľa správy generálneho tajomníka OSN „Nová agenda pre mier“, ktorá bola vydaná 20. júla 2023 sa uvádza, že „žijeme v prepojenom globálnom rizikovom prostredí, v ktorom sú hrozba krízy a zdroj nestability úzko prepojené a vyžadujú si kolektívne a spolupracujúce reakcie“ (Nová agenda pre mier, 2023). V súčasnosti netreba zabúdať na nové technológie, ktoré sú taktiež súčasťou tohto prostredia ako celku a ich komplexný potenciál na militarizáciu sú schopné vytvárať vyvíjajúce sa výzvy. Aj napriek tomu, že technologický pokrok šiel vždy ruka v ruke s vojnou, a boli tak prepojené, rýchlo sa rozvíjajúca oblasť technologických inovácii a vyvíjajúcich sa technológii predstavujú potenciálne cezhraničné hrozby, a to predovšetkým čo do kontextu s ich prepojením s inými hrozbami, ktorých príkladom sú aj jadrové zbrane.

Umelá inteligencia je v posledných rokoch medzinárodným spoločenstvom považovaná za hlavnú obavu, a to nie len z pohľadu jej rýchleho pokroku, ale taktiež aj z dôvodu jej škálovateľnosti, prístupnosti a postupnej všadeprítomnosti. Predovšetkým jej všadeprítomnosť znamená a je pre ňu charakteristické, že umelá inteligencia ako taká je používaná a zavádzaná v mnohých rozmanitých technických systémoch, a to aj v oblasti vojenských operácii a rovnako aj ako súčasť širokej škály zbraní a vojenských aplikácii.

Tieto skutočnosti reflektujú aj politické dokumenty a iniciatívy, ako správa Generálneho tajomníka „Nová agenda pre mier“ a taktiež vôbec prvá diskusia o umelej inteligencii, ktorá sa uskutočnila v Bezpečnostnej rade konanej 18. júla 2023. Tieto doklady poukazujú na význam multilaterálnych snáh na zmiernenie rizík a reguláciu vývoja a používania umelej inteligencie. V tejto súvislosti je potrebné poukázať na chýbajúci celkový rámec, ktorý by slúžil k pochopeniu výziev v tejto oblasti a ich vzájomného prepojenia. Množstvo odborníkov sa týmito výzvami zaoberá v rámci svojich vedných odborov (napr. kybernetické bezpečnostné výzvy umelej inteligencie alebo výzvy, ktoré môžu vyplývať z prepojenia umelej inteligencie a biotechnológii).

Aj napriek tomu, že rôznorodé pochopenie výziev v kontexte technológii (napr. výzva zaujatosti, nepredvídateľnosti algoritmických systémov) a jej zneužitia (napr. použitie umelej inteligencie na vývoj a nasadenie plne autonómnych zbraňových systémov) sú známe a súčasťou multilaterálnych diskusii a rokovaní, predmetná oblasť výziev ostáva nedostatočne preskúmaná a pochopená.

Riešenie tejto problematiky si predovšetkým vyžaduje technické pochopenie a vytvorenie jednotného slovníka. Preto je potrebné sa v tejto súvislosti sústrediť predovšetkým na definovanie komplexného prehľadu možných výziev umelej inteligencie s ohľadom na medzinárodný mier a bezpečnosť.

1.1 TRIEDENIE VÝZIEV VYPLÝVAJÚCICH Z UMELEJ INTELIGENCIE

Výzvy umelej inteligencie s ohľadom na medzinárodný mier a bezpečnosť obsahujú rozmanité spektrum technologických domén a možností použitia. Z metodologického

hľadiska je mapovanie výziev spojených s tak široko používanou technológiou, akou umelá inteligencia rozhodne je, pomerne zložitá úloha. Je potrebne jasne vymedziť hranice medzi jednotlivými kategóriami hrozieb, ktoré nemusia byť vždy jednoznačné. Umelá inteligencia je široko používanou technológiou, tak v statických systémoch (ako napr. plánovacie systémy) ako aj v mobilných systémoch (akými sú napr. drony alebo bezpilotné vozidlá) – s čím sú spojené špecifické súbory hrozieb (Russell, 2022) – taktiež zahŕňajú aj systémy umelej inteligencie s rozdielnou schopnosťou učiť sa, adaptívnymi vlastnosťami a učebno-adaptívnymi schopnosťami, čo má za následok, že rozsah výziev je rozdielny a môže sa v priebehu času aj vyvíjať.

Mapovanie výziev umelej inteligencie s poukazom na medzinárodnú bezpečnosť ďalej prináša hrozby z pohľadu širokého rozsahu aplikácie a dopadu tejto technológie. Ide predovšetkým o aplikáciu vo vojenstve a zbraňových systémoch.

Existuje množstvo metodologických prístupov k diskusii a klasifikácii výziev, čo vyplýva zo širokej škály systémov, v ktorých je technológia integrovaná a aplikovaná. Jedným z možných prístupov k rámcovaniu výziev je analýza jednotlivých fáz procesu zameriavania a identifikácia výziev špecifických pre túto úroveň, ako napríklad výzvy na operačnej alebo taktickej úrovni (Ekelhof, 2019). Alternatívna taxonómia by mohla byť navrhnutá s ohľadom na výzvy spojené s umelou inteligenciou vo fyzickom svete (napríklad autonómia v pohybe a použitie umelej inteligencie na prijímanie kinetických rozhodnutí) a v digitálnej sfére, vrátane využitia umelej inteligencie na ovplyvňovanie rozhodovacích procesov (napríklad potenciálne využitie generatívnej umelej inteligencie vo vojenskom spravodajstve). Tieto prístupy umožňujú komplexné pochopenie hrozieb spojených s technologickými inováciami v súčasných a budúcich vojenských operáciách.

Posúdenie výziev spojených s technológiami sa výrazne líšia v závislosti od organizácie a zainteresovaných strán, ktoré tieto výzvy hodnotia. Tieto subjekty môžu zameriavať svoju pozornosť na špecifické oblasti, ako sú problémy robustnosti v systémoch umelej inteligencie alebo výzvy kybernetickej bezpečnosti, alebo na konkrétne oblasti aplikácie, napríklad výzvy spojené s použitím umelej inteligencie v oblasti jadrového velenia a kontroly. Ako príklad môže poslúžiť kategorizácia navrhnutá organizáciou RAND (Research and Development) v roku 2020, ktorá identifikuje tri hlavné kategórie výziev vojenskej umelej inteligencie: etické a právne výzvy, operačné výzvy a strategické riziká. Tento prístup poskytuje štruktúrovaný rámec na hodnotenie potenciálnych hrozieb a výziev spojených s nasadením umelej inteligencie v rôznych vojenských kontextoch (Morgan, 2020).

Kategorizácia vypracovaná v roku 2023 Centrom pre vznikajúce technológie a bezpečnosť a Centrom pre dlhodobú odolnosť vo Veľkej Británii systematicky kategorizuje výzvy na základe jednotlivých fáz životného cyklu umelej inteligencie, v ktorých sa tieto výzvy môžu prejaviť: návrh, tréning a testovanie, vývoj a používanie, dlhodobé nasadenie a šírenie. Implementácia umelej inteligencie v tomto kontexte môže významne prispieť k riziku eskalácie a zvýšeného napätia, obzvlášť ak sa jej schopnosti rozvíjajú a nasadzujú v prostredí intenzívnej medzivládnej rivality, kde umelá inteligencia predstavuje kľúčovú prioritu v rámci národnej bezpečnosti (Janjeva. 2023).

Kategorizácia výziev uvedená v tejto správe identifikuje hlavné oblasti zraniteľnosti technológií umelej inteligencie, potenciál ich zneužitia, ako aj širšie strategické a geopolitické dôsledky v kontexte medzinárodnej bezpečnosti a mieru. Tento rámec je navrhnutý s cieľom poskytnúť komplexné porozumenie rizík spojených s nasadením umelej inteligencie v rámci vojenských operácií a medzinárodných bezpečnostných stratégií. V tejto časti sa podrobne analyzujú a kategorizujú hrozby technológie umelej inteligencie.

Príspevok sa zameriava hlavne na technologické riziká, ktoré zahŕňajú bezpečnostné riziká spojené s umelou inteligenciou a systémami využívajúcimi umelú inteligenciu, ako aj riziká vyplývajúce z interakcie medzi človekom a strojom. Táto kategória identifikuje faktory, ktoré môžu ovplyvniť celkovú bezpečnosť a výkon systémov s umelou inteligenciou v rôznych aplikáciách a oblastiach použitia. Zahŕňa riziká súvisiace s tým, ako sú tieto systémy navrhnuté, postavené a nasadené, čím zdôrazňuje význam dôkladného návrhu a implementácie na zabezpečenie spoľahlivosti a bezpečnosti týchto technológií.

1.2 RÁMEC PRE RIADENIE RIZÍK UMELEJ INTELIGENCIE

Hoci analýza konceptu rizika presahuje rozsah tohto príspevku, je dôležité zdôrazniť niekoľko základných aspektov. Usmernenia pre riadenie rizík zvyčajne definujú riziko v kontexte špecifických cieľov, ktoré sú predmetom záujmu danej problematiky. Napríklad v oblasti riadenia rizík pre verejné a súkromné organizácie Medzinárodná organizácia pre normalizáciu (ISO) navrhla rámec pre riadenie rizík, štandard 31000:2018, ktorý definuje riziko ako „vplyv neistoty na ciele“ (ISO 31000:2018, 2018). Tento štandard sa stal základom pre mnohé ďalšie prístupy k riadeniu rizík.

Rámec pre riadenie rizík umelej inteligencie, ktorý vyvinul Národný inštitút pre normy a technológie (NIST) pri Ministerstve obchodu Spojených štátov, nadväzuje na tento ISO štandard a definuje riziko ako „kombinované meradlo pravdepodobnosti výskytu udalosti a rozsahu alebo miery následkov danej udalosti (Artificial Intelligence Risk Management Framework, 2023)“. Tento prístup je obzvlášť relevantný pre vojenské operácie, kde je potrebné presne hodnotiť pravdepodobnosť a závažnosť potenciálnych hrozieb spojených s nasadením umelej inteligencie.

Rámec pre riziká umelej inteligencie od Organizácie pre hospodársku spoluprácu a rozvoj (OECD) čerpá z ISO štandardu, NIST rámca, ako aj z princípov OECD pre umelú inteligenciu a rámca pre due diligence. Tento rámec zdôrazňuje, že riziká spojené s umelou inteligenciou by mali byť vyvážené proti rizikám nepoužitia umelej inteligencie v kontextoch, kde môže priniesť významné výhody a nové poznatky (OECD, 2024). V kontexte vojenských operácií to znamená, že pri hodnotení rizík je potrebné zvážiť nielen potenciálne hrozby spojenej s umelou inteligenciou, ale aj straty príležitostí, ktoré by vznikli, ak by sa umelá inteligencia nevyužila v kritických situáciách, kde môže významne zlepšiť efektivitu a rozhodovanie.

Výzvy umelej inteligencie v tomto príspevku sú posudzované v kontexte medzinárodného mieru a bezpečnosti. V súlade s týmto, sa táto konceptualizácia rizík zameriava na spôsob, akým môže umelá inteligencia prispieť k zvýšeniu rizika ozbrojeného konfliktu alebo k vzniku negatívnych či nechcených účinkov na medzinárodnú bezpečnosť. Konkrétne sa riziká umelej inteligencie môžu prejaviť vo:

- vyvolávaní nehôd alebo zámernej či neúmyselnej eskalácie v ozbrojených konfliktoch;
- vytváraní významných výziev pre štáty pri zvládaní nežiaducich účinkov, ako je napríklad použitie určitých zbraňových systémov;
- zvýšení napätia medzi štátmi a zhoršení regionálnych a multilaterálnych vzťahov.

Táto analýza rizík tak zdôrazňuje, že hoci umelá inteligencia môže priniesť značné výhody, jej nasadenie v kontexte medzinárodných bezpečnostných štruktúr si vyžaduje dôkladné zhodnotenie potenciálnych hrozieb a nepriaznivých dôsledkov, ktoré môžu ohroziť globálnu stabilitu.

Ako už bolo avizované kategorizácia v tomto príspevku sa zameriava na riziká technológie umelej inteligencie, ktoré zahŕňajú široké spektrum zraniteľností vyplývajúcich z inherentných obmedzení alebo zraniteľností umelej inteligencie ako technických a učebných systémov, ako aj riziká vyplývajúce z interakcie človeka s týmito systémami. Medzi tieto riziká patria:

1. **bezpečnostné riziká:** tieto riziká sú spojené s vrodenými obmedzeniami vo vývoji a fungovaní systémov umelej inteligencie. Vojenské nasadenie umelej inteligencie môže byť ohrozené bezpečnostným zlyhaním, ktoré je často neúmyselné, avšak nedostatočné praktiky, ako napríklad nesprávne spracovanie dát, môžu viesť k závažným poruchám, čo môže mať katastrofálne následky v bojových podmienkach.
2. **kybernetické hrozby:** tieto zahŕňajú škodlivé úmyselné útoky, ktoré môžu narušiť spôsob, akým sa umelá inteligencia učí a rozhoduje. Vo vojenskom kontexte sú takéto útoky obzvlášť nebezpečné, pretože môžu narušiť kritické systémy, ovplyvniť dôvernosť, integritu a dostupnosť vojenských operácií. Zložitosť kybernetickej obrany umelej inteligencie si vyžaduje špecifické prístupy na ochranu pred takýmito hrozbami.
3. **výzvy interakcie človeka so strojom:** tieto riziká vznikajú v kontexte interakcie medzi ľuďmi a systémami s umelou inteligenciou, ktoré fungujú s rôznou mierou autonómie. Vojenský personál musí byť schopný dôverovať týmto systémom, avšak zaujatosť voči automatizácii a nedostatok dôvery môžu obmedzovať efektívne využitie umelej inteligencie, čo môže ohroziť úspech vojenských operácií.

Medzi tieto riziká patria:

a) **riziko nesprávneho výpočtu:** s rozširujúcim sa využívaním umelej inteligencie v spravodajských komunitách, a rôznych nástrojoch na predvídanie budúcich skutočností, narastá aj jej vplyv na vojenské rozhodovanie, vrátane kritických rozhodnutí o použití sily. Hoci riziká nesprávnych výpočtov nie sú nové, umelá inteligencia ich môže výrazne zhoršiť. Nesprávne použitie alebo zlyhanie technológie môže viesť k závažným chybám v spravodajských správach, chybným interpretáciám vyvíjajúcej sa operačnej situácie a k zásadným nesprávnym výpočtom počas ozbrojeného konfliktu. Tieto faktory môžu vážne ovplyvniť globálnu bezpečnostnú situáciu, zavádzať neistotu do stratégií a ovplyvňovať budúce konflikty.

b) **riziko eskalácie:** umelá inteligencia môže rôznymi spôsobmi prispieť k eskalácii konfliktov, najmä prostredníctvom integrácie do zbraňových systémov, ako sú jadrové alebo konvenčné zbrane. Môže iniciovať zámernú alebo neúmyselnú eskaláciu konfliktov a taktiež ovplyvniť systémy na podporu rozhodovania, kde môže podnietiť rozhodnutia vedúce k eskalácii. Vojenská aplikácia umelej inteligencie preto so sebou nesie značné riziká, ktoré môžu destabilizovať globálnu bezpečnosť a viesť k nepredvídateľným a potenciálne katastrofálnym následkom na medzinárodnej scéne.

Táto kategorizácia rizík zdôrazňuje, že hoci umelá inteligencia prináša potenciál pre inovácie a zlepšenie vojenských kapacít, jej nesprávne alebo nekontrolované použitie môže vážne narušiť medzinárodnú stabilitu a bezpečnosť. Riziká spojené s rozmachom umelej inteligencie predstavujú významnú hrozbu, najmä v kontexte konvergencie umelej inteligencie s inými technologickými oblasťami a rozšírenia samotných technológií umelej inteligencie. Široké rozšírenie softvéru poháňaného umelou inteligenciou umožňuje jeho prispôsobenie a jemné doladenie rôznymi aktérmi, čo zvyšuje riziko jeho nekontrolovaného šírenia a využitia na nepredvídateľné účely.

Tieto riziká sú často úzko prepojené. Napríklad bezpečnosť a odolnosť systémov s umelou inteligenciou sú neoddeliteľne spojené s otázkami interakcie medzi človekom a

strojom. Vojenský operátor môže prehliadnuť slabý výkon systému s umelou inteligenciou kvôli zaujatosti voči automatizácii, čo môže viesť k vážnym následkom. Ďalšie technologické zlyhania, ako napríklad neschopnosť umelej inteligencie adekvátne sa prispôsobiť novým operačným prostrediam, môžu viesť k závažným nesprávnym výpočtom, ktoré majú kritické dôsledky v kontexte ozbrojeného konfliktu, najmä ak je umelá inteligencia integrovaná do zameriavacích cyklov. Takéto zlyhania môžu mať okamžité následky, najmä ak sú narušené segmenty nájdenia a sledovania cieľa.

Bezpečnostné alebo kybernetické zlyhania môžu navyše viesť k eskalácii konfliktu, najmä ak sa tempo vojny zrýchli v dôsledku schopnosti algoritmických systémov vykonávať úlohy v priebehu niekoľkých sekúnd, ktoré by inak trvali hodiny. Takéto zrýchlenie môže zásadne ovplyvniť riadenie eskalácie alebo deeskalácie konfliktu, čím sa zvyšuje riziko nepredvídaných a potenciálne katastrofálnych dôsledkov na medzinárodnú bezpečnosť. Tieto prepojené riziká zdôrazňujú potrebu dôkladného riadenia a regulácie technológií umelej inteligencie, najmä v oblasti vojenských aplikácií, kde ich nesprávne nasadenie môže destabilizovať globálnu bezpečnostnú architektúru.

2. VÝZVY TECHNOLÓGIE UMELEJ INTELIGENCIE

Výzvy spojené s technológiou umelej inteligencie môžeme rozdeliť do dvoch hlavných kategórií: hrozby bezpečnosti a zabezpečenie. **Bezpečnostné hrozby** sa zvyčajne týkajú neúmyselných zlyhaní systémov umelej inteligencie, ktoré spôsobujú, že tieto systémy nefungujú podľa očakávaní. Tieto problémy sú súčasťou systému umelej inteligencie a môžu sa prejaviť v rôznych fázach ich vývoja, testovania a nasadzovania. Na druhej strane, **hrozby zabezpečenia** sa týkajú zámerných útokov na systémy umelej inteligencie, vrátane kybernetických bezpečnostných hrozieb a útokov, ktoré sú bežné aj v tradičných IT systémoch (Russell, 2022). Napriek tomu, že metódy útokov a dostupné obranné opatrenia majú určité podobnosti, systémy umelej inteligencie vyžadujú špecifické prístupy k ochrane.

Ďalším kritickým zdrojom hrozieb je **interakcia medzi človekom a strojom**, ktorá môže viesť k nehodám alebo zneužitiu technológie, aj keď systém s podporou umelej inteligencie technicky funguje správne. Tieto hrozby sú často prepojené. Problémy s robustnosťou modelov umelej inteligencie sa môžu prelínať s kybernetickými bezpečnostnými hrozbami (Brundage, 2018). Navyše, zlyhania umelej inteligencie môžu viesť k nesprávnym alebo oneskoreným reakciám ľudských operátorov, najmä ak ten istý systém pred zlyhaním fungoval spoľahlivo a bezproblémovo po dlhý čas.

2.1 BEZPEČNOSŤ UMELEJ INTELIGENCIE

Krehkosť umelej inteligencie je jedným z najzávažnejších výziev spojených s aplikáciami umelej inteligencie, najmä v kontexte bezpečnostne kritických operácií. Krehkosť sa prejavuje, keď algoritmus nedokáže zovšeobecniť alebo prispôsobiť sa podmienkam, ktoré sa líšia od tých, na ktorých bol pôvodne trénovaný. Tento jav nastáva napríklad v prípade algoritmov počítačového videnia, ktoré boli trénované na rozpoznávanie lodí na základe tisícok obrazových vzorov (Amodei, 2016). Ak sa však zmenia podmienky, ako napríklad poveternostné faktory, môže dôjsť k tomu, že model nebude schopný správne identifikovať cieľ.

Krehkosť systémov umelej inteligencie neznamená, že sú nevyhnutne slabé svojím návrhom, ale skôr to, že aj keď algoritmus funguje spoľahlivo v rámci určitých hraníc, môže sa „mýliť“, ak tieto hranice prekročí. Táto vlastnosť spôsobuje, že mnohé systémy s umelou inteligenciou, napríklad v robotike alebo autonómnych vozidlách, sa môžu zdať veľmi

schopné, ale pri stretnutí s nepredvídanými podmienkami v reálnom svete môžu dramaticky zlyhať (Lohn, 2020).

Systémy umelej inteligencie sú obzvlášť zraniteľné voči zlyhaniu, keď dochádza k systematickým zmenám kontextu, alebo keď sa údaje použité počas tréningovej fázy líšia od reálnych podmienok, s ktorými sa systém neskôr stretne, čo je známe ako problém „posunu distribúcie“ (Goodfellow, 2015). Tento fenomén predstavuje vážne hrozby najmä vo vojenskom prostredí, kde nepredvídané zmeny na bojisku môžu výrazne ohroziť účinnosť umelej inteligencie, a tým aj úspešnosť vojenských operácií. Dôkladné pochopenie a riadenie týchto hrozieb je preto kľúčové pre bezpečné nasadenie umelej inteligencie v kritických vojenských aplikáciách.

2.1.1 SCENÁR INCIDENTU: CHYBA V AUTONÓMNEJ NAVIGÁCII A MOŽNÉ ESKALAČNÉ NÁSLEDKY

V tomto scenári je bezpilotné lietadlo (dron) nasadené na misiu SPS (spravodajstvo, prieskum a sledovanie) v blízkosti vysoko sporného regiónu s cieľom monitorovať pohyb a aktivity pozdĺž hranice. Navigačný algoritmus dronu bol trénovaný na základe kombinácie záberov získaných počas reálnych operácií a simulovaných letov. Autonómne vedenie zabudované v drone dopĺňa tradičnú GPS navigáciu s cieľom minimalizovať riziko spoofingu (manipulácia s GPS signálmi, aby sa dron, pohyboval podľa nesprávnych údajov o polohe) alebo rušenia signálu.

Počas operácie nad zložitým terénom, ktorý zahŕňa hory, jazerá a ľudské osídlenia, dochádza k zlyhaniu počítačového vizuálneho systému, ktorý nedokáže správne identifikovať hranice. Tento problém vedie k tomu, že dron neúmyselne vnikne do vzdušného priestoru susednej krajiny v období vysokého napätia. Takýto incident môže mať vážne eskalačné následky, pretože narušenie vzdušného priestoru inej krajiny, obzvlášť v čase zvýšeného napätia, môže byť interpretované ako provokácia alebo úmyselný akt agresie (Amer, 2019). Vojenské a diplomatické dôsledky môžu byť značné, potenciálne vedúce k diplomatickým krízam a vyvolaniu vojenskej reakcie zo strany dotknutej krajiny. Tento scenár ilustruje, ako technické zlyhanie v autonómnych systémoch môže eskalovať konflikt a destabilizovať regionálnu bezpečnosť.

Hoci sú tieto incidenty neúmyselné, môžu byť okamžite interpretované ako provokácia alebo akt útoku. V závislosti od konkrétneho kontextu, v ktorom k incidentu dôjde, by mohol tento typ narušenia vyvolať okamžité reakcie zo strany dotknutej krajiny. Tieto nepredvídané udalosti zdôrazňujú hrozby spojené s autonómnymi systémami, kde aj technické zlyhanie môže viesť k rýchlej eskalácii napätia a potenciálne k vojenskému konfliktu.

2.1.2 ŠPECIFIKÁCIA ÚLOHY A PROBLÉMY

Technické zlyhania systémov umelej inteligencie často pramenia z problémov spojených so špecifikáciou úlohy, ktorá zahŕňa prenesenie konkrétnych pokynov na systém strojového učenia. Tento proces vyžaduje, aby sa zámer návrhára presne pretransformoval do konkrétnych akcií a správania systému. V praxi je však zosúladenie ľudského vnímania úlohy s tým, ako ju vníma robot alebo systém strojového učenia, obzvlášť náročné pri zložitejších úlohách (Fazekas, 2021). Neraz dochádza k nesúladu medzi „návrhovou špecifikáciou,“ ktorá je explicitne zabudovaná do systému, a „odhalenou špecifikáciou,“ ktorá sa prejavuje v pozorovanom správaní systému počas jeho nasadenia. Inými slovami, to, čo systém skutočne robí, sa môže líšiť od pôvodného zámeru návrhára.

Robotické systémy sú vybavené úlohami, ktoré musia vykonávať, pričom tieto úlohy sú vo forme abstrakcií, ktoré sa systém učí buď explicitne, prostredníctvom štruktúr ako sú súbory funkcií a grafy, alebo implicitne, prostredníctvom neurónových sietí, ktoré automaticky extrahujú úlohy koreláciou vstupov s požadovaným správaním. V oboch prístupoch však existujú výzvy pri zosúladení zámeru návrhára s činnosťou robota. V kontexte vojenských aplikácií sú tieto problémy obzvlášť kritické, pretože nesprávne špecifikované úlohy môžu viesť k nežiaducemu správaniu autonómnych systémov počas bojových operácií (Styber, 2023). Takáto odchýlka môže mať vážne následky, od narušenia koordinácie až po neúmyselnú eskaláciu konfliktu, čo zdôrazňuje potrebu presnej špecifikácie a dôkladného testovania vojenských systémov s podporou umelej inteligencie pred ich nasadením do operácií.

Je ťažké predvídať alebo presne definovať všetky možné situácie, s ktorými sa systém umelej inteligencie môže stretnúť počas plnenia úloh, a táto výzva sa stáva ešte náročnejšou v zložitom prostredí. Neurónové siete síce dokážu niektoré z týchto výziev prekonať automatickým extrahovaním, avšak často vykazujú falošné spojenia (von Braun, 2021). Tieto spojenia sú prepojením dvoch premenných, ktoré síce môžu byť asociované, ale nie sú kauzálne prepojené. Napríklad navigačný systém v neobsadenom pozemnom vozidle, nasadenom na tichú operáciu, si môže mylne vyložiť prítomnosť stromov s prítomnosťou skladov a spoliehať sa na túto asociáciu, aj keď ide o nesúvisiace faktory.

Ďalším významným problémom spojeným so špecifikáciou úloh je tzv. **reward hacking** (zneužitie odmeňovacieho systému). Tento jav nastáva, keď sa systém naučí správanie, ktoré optimalizuje odmeňovaciu funkciu spôsobom, ktorý je nežiaduci alebo mimo zamýšľaného cieľa. Systém nájde „jednoduchšie“ riešenie na formálne splnenie úlohy, čím skreslí pôvodný zámer (Fischer, 2022). Napríklad v systéme na rozpoznávanie cieľov, ktorý je odmeňovaný za detekciu vojenských vozidiel v určitej oblasti, sa systém môže naučiť opakovane detegovať to isté vozidlo krúžením v užšej oblasti, čím maximalizuje svoju odmeňovaciu funkciu, ale neplní skutočný zámer misie.

Aj keď sa mnohé z týchto zlyhaní podarí odhaliť a opraviť počas tréningovej fázy, nie je realistické predpokladať, že takýmto problémom sa vždy podarí predísť, najmä s narastajúcou zložitosťou technológie. Tento problém zosúladenia sa stáva ešte zložitejším v prípade techník hlbokého učenia a neurónových sietí, kde inherentná komplexnosť tréningových algoritmov v kombinácii s problémom krehkosti výrazne zvyšuje hrozbu nepredvídaného správania. Problémy s bezpečnosťou sú ešte viac umocnené v kontexte intenzívneho úsilia o vývoj pokročilých schopností umelej inteligencie, často na úkor bezpečnostných opatrení. Ako poznamenal jeden odborník: „V oblasti bezpečnosti zaostávame. 98 % výskumníkov sa sústreďuje na zvyšovanie výkonu umelej inteligencie, nie na jej bezpečnosť. Bezpečnosť je podceňovaná (Yampolskiy, 2018).“ Perspektíva širokého nasadenia umelej inteligencie navyše ešte viac prehlbuje výzvy spojené s bezpečnosťou, zložitosťou testovania a zabezpečením kvality. Vojenské aplikácie umelej inteligencie, kde sú tieto hrozby obzvlášť kritické, vyžadujú dôkladné zváženie týchto problémov. Nezohľadnenie bezpečnostných rizík môže viesť k nepredvídateľným a potenciálne katastrofálnym následkom, ohrozujúcim nielen úspech jednotlivých operácií, ale aj širšiu medzinárodnú bezpečnosť.

2.1.3 NEÚMYSELNÉ ZLYHANIA: PRÍKLADY PROBLÉMOV BEZPEČNOSTI UMELEJ INTELIGENCIE

Tento zoznam sumarizuje hlavné bezpečnostné problémy spojené so systémami umelej inteligencie. Bol zostavený skupinou odborníkov (Siva Kumar, 2019). Tento

dokument je neúplný a dynamický, pretože sa predpokladá, že s vývojom technológie budú v technických komunitách identifikované a konceptualizované nové režimy zlyhania.

Tabuľka 1 Typy zlyhaní

Typ zlyhania	Opis / Príčina zlyhania
Reward hacking	Nesúlad medzi deklarovanou odmenou a „skutočnou“, zamýšľanou odmenou. Systém môže optimalizovať odmeňovaciu funkciu spôsobom, ktorý formálne splňuje úlohu, ale skresľuje pôvodný zámer návrhára.
Vedľajšie účinky	Systém neúmyselne naruší prostredie, aby dosiahol svoj cieľ, čím vytvára nežiaduce vedľajšie účinky, ktoré neboli pôvodne zamýšľané.
Distribučné posuny	Zmeny v typoch dát môžu viesť k poruche systému alebo jeho neschopnosti prispôsobiť sa novým podmienkam, čo môže byť kritické v dynamickom vojenskom prostredí.
Prirodzené adversariálne príklady	Systém zlyháva kvôli ťažkému negatívnemu tréningu, bez zásahu útočníka. Ide o prípady, keď sú do tréningu zahrnuté nesprávne alebo nesprávne detegované objekty, čo vedie k vytvoreniu negatívnych vzoriek, ktoré spôsobujú nesprávnu funkciu systému.
Bežné poškodenie alebo narušenie dát	Zahŕňa úpravy dát, ktoré môžu mať dôsledky od relatívne neškodných až po veľmi závažné. Na rozdiel od adversariálnych poškodení alebo narušení, bežné poškodenia/narušenia dát vznikajú neúmyselne a môžu viesť k významným problémom pri nasadení umelej inteligencie.

Zdroj: vlastné spracovanie

Tieto neúmyselné zlyhania sú vo veľkej miere prepojené so samotnou konštrukciou systémov umelej inteligencie, vrátane použitých dát, učebných algoritmov a ďalších technických aspektov. V kontexte vojenských aplikácií môžu tieto zlyhania viesť k vážnym operačným výzvam, vrátane nesprávnych rozhodnutí alebo nepredvídaných reakcií v kritických situáciách.

Popri týchto neúmyselných zlyhaniach existuje aj ďalšia kategória technologických zlyhaní, a to **úmyselné útoky**, ktoré cielene kompromitujú kybernetickú bezpečnosť. Tieto útoky predstavujú ďalšiu vrstvu rizík, ktorá vyžaduje špecifické obranné opatrenia na ochranu vojenských AI systémov pred sabotážou a zneužitím.

2.2 HROZBY KYBERNETICKEJ BEZPEČNOSTI

Modely strojového učenia sú mimoriadne zraniteľné voči rôznym formám kybernetických útokov. Typické útoky v kybernetickej oblasti, známe pod skratkou CIA, ktorá zahŕňa dôvernosť (Confidentiality), integritu (Integrity) a dostupnosť (Availability), sú relevantné aj pre modely umelej inteligencie, a to ako počas fázy tréningu, tak aj pri nasadení týchto systémov.

Útoky na dôvernosť:

Pri útokoch na dôvernosť sa útočníci snažia získať skryté informácie o modeli, často prostredníctvom metódy známej ako „krádež modelu (Wang, 2019)“. Tento typ útoku spočíva v tom, že protivník testuje klasifikačný systém tak, že pozoruje, ako reaguje na rôzne vstupy. Cieľom tohto postupu je odhaliť vnútornú štruktúru modelu, čo následne umožňuje útočníkovi model manipulovať alebo ho kompromitovať.

V kontexte vojenských aplikácií môžu takéto útoky predstavovať vážne riziko, pretože získanie dôverných informácií o fungovaní takýchto systémov môže protivníkovi umožniť manipuláciu s ich správaním v kritických momentoch, napríklad počas bojových operácií. Tým sa zvyšuje riziko kompromitácie vojenských rozhodovacích procesov a narušenia operačnej bezpečnosti. Dôsledky takýchto útokov môžu byť rozsiahle, potenciálne vedúce k strate vojenskej prevahy a destabilizácii širšej bezpečnostnej situácie.

Existujú tri hlavné typy útokov na dôvernosť modelov umelej inteligencie:

1. **Útoky na extrakciu modelu:** Pri tomto type útoku útočník opakovane zaznamenáva vstupy a výstupy cieľového modelu, až kým nezíska dostatočné údaje na vytvorenie „verného ekvivalentu“ napadnutého modelu. Súčasné techniky krádeže modelov sú schopné dosiahnuť takmer dokonalú obnovu pôvodného modelu. Tento typ útoku je obzvlášť účinný proti „grey box“ modelom, kde sú dostupné určité informácie o modeli (Papernot, 2016).
2. **Útoky na členstvo:** Tieto útoky zahŕňajú analýzu vstupov a výstupov systému strojového učenia s cieľom určiť, či bola konkrétna dátová vzorka súčasťou tréningových dát daného modelu. Útočník môže tento typ útoku realizovať napríklad vyhodnotením dôvery modelu v porovnaní s tzv. „shadow modelom,“ ktorý obsahuje náhodné podmnožiny tréningových dát dostupných útočníkovi (Carlini, 2022).
3. **Útoky na inverziu modelu:** Tento typ útoku sa zameriava na rekonštrukciu alebo obnovu výstupných kategórií modelu. Namiesto hľadania konkrétnych dát sa útočník snaží pochopiť určité vlastnosti vstupov, aby vytvoril reprezentatívnu vzorku pre danú triedu. Útoky na inverziu modelu sa dostali do popredia v politických diskusiách kvôli ich potenciálu ohroziť integritu modelov. Tento typ útoku môže vážne narušiť fungovanie systémov umelej inteligencie, často prostredníctvom manipulácie s dátami v rôznych fázach spracovania.

V kontexte vojenských aplikácií môžu tieto typy útokov predstavovať vážnu hrozbu, pretože umožňujú protivníkovi získať prístup k citlivým informáciám, manipulovať s operačnými modelmi a narušiť kritické vojenské operácie. Takéto kompromitácie môžu nielen oslabiť schopnosti obrany, ale aj destabilizovať širšiu bezpečnostnú situáciu tým, že protivníkovi poskytujú neprimeranú výhodu v boji. Napríklad veľmi jemné, takmer nepostrehnuteľné zmeny na obrázku 3D vytlačenej korytnačky v jednej štúdii spôsobili, že klasifikátor obrázkov ju nesprávne identifikoval ako pušku (Athalye, 2018). Takéto chyby ilustrujú zraniteľnosť modelov umelej inteligencie voči útokom na integritu, ktoré možno rozdeliť do dvoch hlavných typov:

1) **Otrava dát:** Tieto útoky majú za cieľ znížiť presnosť a výkon systému strojového učenia tým, že manipulujú s jeho tréningovou dátovou sadou. Tento proces vedie k tomu, že systém sa učí nesprávne, čo spôsobuje, že jeho schopnosť správne klasifikovať objekty alebo situácie je oslabená. Otrava dát môžu byť realizovaná rôznymi spôsobmi, často veľmi jemne alebo s minimálnymi výpočtovými nárokmi, napríklad zmenou štítkov tried počas tréningovej fázy (Biggio, 2018).
2) **Útoky na obchádzanie (evasion attacks):** Tieto útoky patria medzi najbežnejšie formy narušenia integrity systému umelej inteligencie. Spočívajú v škodlivých

zmenách vstupov, ktoré sú tak jemné, že ich ľudskí pozorovatelia nepostrehnú, no stále sú dostatočne významné na to, aby systém generoval nesprávne výstupy. Hlavným cieľom týchto útokov je spôsobiť, aby systém nesprávne klasifikoval objekty, čo sa často dosahuje pomocou tzv. adversariálnych príkladov. Na konštrukciu takýchto príkladov útočníci zvyčajne potrebujú prístup typu „white box," kde majú plný prístup k systému vrátane parametrov modelu a tréningových dát. Avšak existujú aj „black box" útoky, pri ktorých má útočník minimálny alebo žiadny prístup k modelu. Napríklad jedna štúdia preukázala, že drobná perturbácia v hlbokej neurónovej sieti spôsobila, že systém nesprávne klasifikoval dopravnú značku „stop" ako „dajte prednosť (Goodfellow, 2016)." Tento druh útoku môže byť obzvlášť nebezpečný v kontexte autonómnych vozidiel, kde by modifikácie dopravných značiek - či už fyzické, napríklad použitím farby alebo nálepiek, alebo digitálne, prostredníctvom obrazu spracovaného vozidlom - mohli viesť k nebezpečným nesprávnym interpretáciám, čo by mohlo mať vážne následky v reálnych vojenských operáciách.

Nedávny výskum ukázal, že multimodálne veľké jazykové modely sú zraniteľné voči nepriamym adversariálnym útokom, ktoré využívajú techniku tzv. injekcie promptov. Multimodálne veľké jazykové modely sú pokročilé modely umelej inteligencie, schopné vykonávať úlohy naprieč rôznymi modalitami, kombinujúcimi spracovanie textu s generovaním informácií v rôznych formách, vrátane obrázkov či zvukov. Technika injekcie promptov umožňuje útočníkom nasmerovať veľké jazykové modely k nežiaducemu správaniu, obchádzajúc ochranné filtre alebo manipulujúc model prostredníctvom starostlivo navrhnutých inštrukcií. Nepriama injekcia promptov spočíva v zavedení adversariálnych inštrukcií treťou stranou, pričom obeťou je samotný používateľ. Útočník môže vložiť škodlivý prompt do obrázka alebo zvukového klipu, ktorý potom používateľ nevedomky predloží chatbotu. Chatbot, spracovávajúc tento perturbovaný vstup, generuje výstup, ktorý môže viesť napríklad k návšteve škodlivej webovej stránky (Goodfellow, 2016).

Rastúce využívanie veľkých jazykových modelov v oblastiach ako spravodajské operácie zdôrazňuje komplexný vzťah medzi hrozbami spojenými s touto technológiou a jej potenciálnymi dopadmi na medzinárodnú bezpečnosť. Zraniteľnosti, ktoré odhalili útoky na integritu, sú inherentné pre modely strojového učenia a môžu sa prejaviť v celom životnom cykle technológie, vrátane dodávateľského reťazca. Útočníci nemusia priamo preniknúť do samotného systému strojového učenia, aby narušili jeho výstupy; môžu napríklad preniknúť do spoločnosti, ktorá vyvíja špionážne drony, a získať informácie o modeloch strojového učenia, alebo manipulovať verejne dostupné dáta, ktoré tieto spoločnosti často používajú ako základ pre tréning svojich modelov.

Ďalším príkladom sú techniky maskovania, ktoré sú špecificky navrhnuté pre umelú inteligenciu, najmä tie, ktoré sa zameriavajú na rozpoznávanie obrázkov. Hoci tieto metódy boli zatiaľ prevažne experimentálne, ukázalo sa, že jednoduché adversariálne záplaty môžu úspešne zmiasť automatické detektory objektov. V jednej štúdii boli záplaty rôznych konfigurácií umiestnené na veľké vojenské objekty, ako napríklad vojenské lietadlá, s cieľom maskovať tieto objekty na leteckých snímkach (Szegedy, 2014). Hoci záplaty neboli fyzicky aplikované na skutočné lietadlá, tréningové nastavenie naznačuje, že podobný efekt by mohol byť dosiahnutý aj v reálnom svete.

Hrozby môžu tiež vzniknúť v dôsledku transferového učenia, pri ktorom je existujúci predtrénovaný model doladený na novú úlohu. Hlavná časť pôvodného modelu, nazývaná „učiteľský model", sa môže prispôsobiť na iný „študentský model" pre novú doménu (Goodfellow, 2016). Tento proces retréningu často vyžaduje menej dát a výpočtových zdrojov, avšak zároveň môže vytvárať nové bezpečnostné hrozby, najmä v kontexte

vojenských aplikácií, kde by zraniteľnosti mohli byť zneužité na kompromitáciu kritických operácií.

Vzhľadom na vysoké nároky na dáta a výpočtové zdroje potrebné na tréning algoritmov je v súčasnosti bežnou praxou opätovné využívanie modelov trénovaných veľkými korporáciami, ktoré sú následne prispôsobené na špecifické potreby. Tieto modely sú často verejne dostupné, čo predstavuje potenciálne riziko „otrávenia studne.“ V tomto scenári môže útočník do modelu zaviesť škodlivý kód, ktorý môže byť nevedomky stiahnutý a použitý vývojármi strojového učenia ako súčasť ich vlastného kódu. Ďalšou hrozbou je outsourcing procesu tréningu škodlivej tretej strane, ktorá môže úmyselne manipulovať model, napríklad tak, že dron bude nesprávne klasifikovať ciele - tento typ útoku je známy ako „backdoor útok.“

Útoky na dostupnosť predstavujú tretiu kategóriu hrozieb pre systémy strojového učenia a môžu viesť k spomaleniu alebo úplnému zastaveniu komponentu strojového učenia, čo drasticky znižuje kvalitu výkonu alebo prístup k systému. Tieto útoky môžu využívať závislosť systému na hardvéri a optimalizácii modelu. Jedným z príkladov sú tzv. „sponge útoky,“ pri ktorých adversariálne príklady absorbujú energiu spotrebovanú neurónovou sieťou, čím nútia základný hardvér k podvýkonu. Následky takýchto útokov môžu byť obzvlášť devastujúce v aplikáciách v reálnom čase, ktoré vyžadujú rýchle porozumenie scény alebo operačného prostredia a majú prísne požiadavky na latenciu. Hoci útoky na dostupnosť doteraz nepriťahovali toľko pozornosti ako útoky na dôvernosť a integritu, záujem výskumnej komunity v posledných rokoch výrazne vzrástol, najmä s nasadením zložitejších systémov náročných na výpočty(Goodfellow, 2016). Vojenské aplikácie, ktoré sú často závislé na výkonných systémoch umelej inteligencie pre kritické rozhodovanie v reálnom čase, sú obzvlášť zraniteľné voči týmto typom útokov. Narušanie dostupnosti týchto systémov môže mať vážne dôsledky pre operačnú efektivitu a môže ohroziť úspech vojenských misií, čím sa ešte viac zvyšuje význam ochrany pred takýmito hrozbami.

2.2.1 SCENÁR INCIDENTU: ADVERSARIÁLNY ÚTOK A MOŽNÉ ESKALAČNÉ NÁSLEDKY

V tomto scenári je ozbrojené bojové letecké vozidlo nasadené na sledovanie a angažovanie sa proti vopred určenému cieľu, konkrétne proti flotile vojenských vozidiel prepravujúcich personál a zbrane. Tento bojový dron je vybavený pokročilými schopnosťami spracovania obrazu, ktoré mu umožňujú rýchlo hodnotiť ciele a zoskupovať objekty záujmu.

Tieto scenáre ukazujú, ako môžu adversariálne útoky viesť k neúmyselným vojenským akciám, ktoré eskalujú konflikt nad rámec pôvodných zámerov. V prípade zapojenia tretích strán do sabotáže umelej inteligencie môže napätie v regióne rýchlo narastať nepredvídateľnými spôsobmi, čo môže podporiť nedôveru medzi zúčastnenými stranami a skomplikovať alebo prekaziť snahy o deeskaláciu a ukončenie nepriateľských akcií. Takticky, takéto incidenty môžu vážne oslabiť dôveru operátorov v umelú inteligenciu, čo predlžuje čas potrebný na prekalibrovanie a získanie legitímnych cieľov. Týmto spôsobom môžu adversariálne útoky nielen ohroziť operačnú efektivitu, ale aj destabilizovať širší bezpečnostný kontext, čo vedie k eskalácii konfliktov a komplikáciám v medzinárodných vzťahoch.

Tabuľka 2 Incident a jeho možné eskalačné následky

Typ incidentu	Opis incidentu	Možné eskalačné následky
Adversariálny útok (a)	Tretia strana získala prístup ku klasifikačnému modelu bojového dronu a jemne upravila kategórie „priateľ“ a „nepriateľ“. Dron nesprávne klasifikuje civilné autobusy ako vojenské ciele, čo vedie k útokom na civilné objekty (Rudner, 2021).	Adversariálne útoky vedú k rýchlej a nebezpečnej eskalácii konfliktu, ktorý môže prerásť do vojenských akcií presahujúcich pôvodné ciele a zámery.
Adversariálny útok (b)	Škodlivý aktér nedokáže priamo narušiť systém strojového učenia, preto zmení vzhľad fyzických objektov (napr. civilné autobusy označí vojenskými insigniami), čo zmätie klasifikačný proces dronu. Dron nesprávne klasifikuje ciele, a operátori softvér vypnú (Eykolt, 2018).	Oslabená dôvera operátorov v umelú inteligenciu, predĺženie času na prekalibrovanie systému, komplikácie v deeskalácii napätia a potenciálna destabilizácia medzinárodných vzťahov.

Zdroj: vlastné spracovanie

2.2.2 HODNOTENIE OBMEDZENÍ OBRANY

Útoky na systémy strojového učenia sa stávajú čoraz častejšími, najmä v dôsledku rozšíreného využívania tejto technológie vo vojenských operáciách a v kritických infraštruktúrach. Podobne ako v oblasti kybernetických operácií, kde je dlhodobo rovnováha medzi útokom a obranou naklonená v prospech útoku, platí to isté aj pre systémy strojového učenia, kde „neexistuje dokonalá dualita medzi útokom a obranou.“ Je všeobecne uznávané, že vykonávanie útokov na systémy umelej inteligencie často vyžaduje menej odborných znalostí než ich navrhovanie alebo trénovanie. Napríklad v štúdii o útokoch na obchádzanie (evasion attacks) sa ukázalo, že niekoľko verzií útoku môže byť vytvorených v priebehu jedného popoludnia, pričom každá verzia si vyžadovala menej ako 20 riadkov kódu (Lohn, 2020). Tento problém je ešte umocnený tým, že mnoho nástrojov na útoky na systémy AI je ľahko dostupných na internete, často bezplatne.

To však neznamená, že všetky útoky na umelú inteligenciu sú jednoduché alebo vždy úspešné. Napriek tomu predstavujú trvalú výzvu z niekoľkých dôvodov, ktoré sú sociálno-organizačného aj technického charakteru. Výskumné a politické komunity stále nevenujú dostatočné zdroje na zvyšovanie odolnosti systémov strojového učenia. Odhaduje sa, že iba asi 1 percento akademického výskumu v oblasti umelej inteligencie sa zameriava na bezpečnosť systémov umelej inteligencie, pričom značná časť tohto výskumu sa sústreďuje

na adversariálne príklady, čo je len jedna forma útoku a v mnohých kontextoch nemusí byť najpravdepodobnejšia.

Zraniteľnosti systémov strojového učenia sa môžu počas používania zhoršovať, napríklad v dôsledku zle navrhnutých používateľských rozhraní alebo v špecifických kontextoch. Vo vojenských systémoch, ktoré fungujú na diaľku, a kde je ľudský operátor fyzicky vzdialený, môže byť identifikácia kompromitovaného systému výrazne oneskorená, čo sťažuje rýchly a efektívny zásah. Okrem toho, hrozby spojené s útokmi na umelú inteligenciu nespočívajú len v relatívnej jednoduchosti vykonania útoku, ale aj v jeho škálovateľnosti - ak sa útočníkovi podarí získať kontrolu nad jedným dronom, môže to potenciálne znamenať ohrozenie celej flotily.

Obrana umelej inteligencie, najmä neurónových sietí, voči škodlivým útokom predstavuje komplexné výzvy a môže zahŕňať ďalšie nepredvídané náklady. Vojenské aplikácie týchto systémov preto vyžadujú dôkladné prehodnotenie a investície do bezpečnostných opatrení, ktoré by dokázali čeliť týmto rastúcim hrozbám a zabezpečiť spoľahlivosť kritických operácií v prostredí stále sofistikovanejších kybernetických útokov. Hoci rámce kybernetickej bezpečnosti sú zvyčajne aplikovateľné naprieč rôznymi triedami zraniteľností, vrátane tých novovznikajúcich, niektoré bezpečnostné hrozby v systémoch umelej inteligencie sú špecificky nové. Na rozdiel od tradičného softvéru, zraniteľnosti v systémoch strojového učenia nie je vždy možné vykonávať opravy bežnými metódami. Nové zraniteľnosti môžu vyžadovať zavedenie nových techník opravy alebo priniesť dodatočné kompromisy.

V mnohých prípadoch, keď sa objaví zraniteľnosť, je potrebné model preškoliť a riešiť konkrétne problémy s jeho robustnosťou. Napríklad jednou z obranných metód proti útokom na obchádzanie (evasion attacks) je adversariálne preškoľovanie, pri ktorom sa model postupne trénuje na adversariálnych príkladoch, čím sa zvyšuje jeho odolnosť voči vybraným typom útokov (Federal Office for Information Security, 2023).

Obrana pre systém umelej inteligencie je možná a môže výrazne zvýšiť náklady pre útočníkov. V mnohých prípadoch sa realizácia útoku stáva veľmi náročnou, pracnou a časovo náročnou. Útočníci budú potrebovať prístup k veľkému množstvu dát na tréning systému pre útoky na otravu dát, alebo budú musieť prejsť mnohými pokusmi a omylmi, aby presnejšie pochopili, ako je systém navrhnutý.

Základnou výzvou však zostáva, že riešenie jedného problému môže otvoriť cestu pre iné zraniteľnosti, alebo že kompromisy medzi bezpečnosťou a výkonom môžu byť neakceptovateľné. Tento fenomén, ktorý niektorí odborníci nazývajú „hra na krtka“, spočíva v tom, že niekedy obrana, napríklad proti adversariálnym príkladom, môžu uzavrieť niektoré zraniteľnosti, zatiaľ čo iné zostávajú otvorené. Snahy urobiť systém vysoko robustným môžu viesť k situáciám, v ktorých sa obrany „pretrénujú“ na konkrétny typ útoku, ale tým sa znižuje ich schopnosť zvládnuť iné typy hrozieb. Tieto hrozby sú navyše zhoršené tým, že hrozby sa neustále vyvíjajú, pretože systémy strojového učenia prijímajú nové dáta a učia sa prispôsobovať (Hoffman, 2021).

Hrozby sú často vzájomne prepojené. Mnohé režimy zlyhania v systémoch umelej inteligencie, hoci neúmyselne, môžu vytvárať príležitosti, ktoré protivník dokáže využiť na ďalšie kompromitovanie systému. Bezpečnostné zlyhanie alebo kombinácia zlyhaní sa môžu ešte viac zhoršiť počas používania, najmä ak ľudskí operátori nie sú dostatočne vyškolení na to, aby rozpoznali, že systém je pod útokom alebo nefunguje podľa očakávania.

Tabuľka 3 Príklady obranných techník proti útokom

Metóda	Obmedzenia
Federated Learning — „decentralizovaná“ technika tréningu strojového učenia, ktorá začína od generického modelu, pričom používatelia ho spoločne a iteratívne trénujú a vylepšujú, až kým nie je model plne natrénovaný.	Vo federálnom učení je najslabším článkom výmena medzi pracovným modelom dátového hostiteľa a centrálnym serverom. Hoci sa model s každou výmenou zlepšuje, dáta, ktoré ho trénovali, sú zraniteľné voči útokom zameraným na inferenciu. Táto metóda je tiež výpočtovo náročná a prináša ďalšie výzvy v oblasti dôvery a transparentnosti.
Differential Privacy — diferenciačná ochrana súkromia, aplikovaná na obranu proti útokom na extrakciu informácií. Ide o metódu matematického merania parametrov ochrany súkromia a obmedzenia informácií o dátových bodoch.	Diferenciačná ochrana súkromia ukázala sľubné výsledky, ale vyžaduje kompromis medzi ochranou súkromia a presnosťou. Vývoj správnych parametrov pre ochranu súkromia môže byť tiež výpočtovo náročný.
Secure Multi-Party Computation — technika, ktorá pomáha skryť aktualizácie modelu prostredníctvom rôznych foriem šifrovania, aby sa znížili riziká únikov dát.	Táto technika spočíva na protokole tajného zdieľania, kde viaceré strany participujú na výpočte bez odhalenia svojich individuálnych vstupov. Stále však vyvoláva významné výzvy v oblasti dôvery v zdieľanie dát a v zabezpečení, že dáta nebudú zneužité alebo nesprávne použité.

Zdroj: vlastné spracovanie

2.3 VÝZVY INTERAKCIE MEDZI ČLOVEKOM A STROJOM

Spôsob, akým ľudia interagujú so systémami umelej inteligencie, je kľúčovou zložkou v taxonómii výziev spojených s touto technológiou. Táto časť sumarizuje hlavné výzvy súvisiace s interakciou medzi človekom a strojom.

Existuje niekoľko zdrojov výziev, ktoré môžu výrazne prispieť k zneužitiu alebo nesprávnemu využitiu technológií umelej inteligencie. Tieto výzvy môžu vzniknúť nielen vtedy, keď technológia funguje optimálne, ale aj keď je systém kompromitovaný alebo zlyháva. Diskusia o výzvach medzi človekom a strojom sa vedie už od 70. rokov 20. storočia a s nástupom automatizácie a vyššej úrovne autonómie sa tieto diskusie stávajú čoraz relevantnejšími. Nové technológie prinášajú nové požiadavky na ľudských operátorov, čo ešte viac komplikuje vzťah medzi človekom a strojom. Príchod autonómnych vozidiel v posledných desaťročiach obohatil naše chápanie zložitostí a rizík spojených s týmto fenoménom, čo je obzvlášť dôležité pre vojenské aplikácie. V dôsledku toho vzniká zložitá sieť vzájomne prepojených rizík.

2.3.1 KALIBRÁCIA DÔVERY A ZAUJATOSŤ VOČI AUTOMATIZÁCII

Otázka dôvery je v kontexte interakcie medzi človekom a strojom zásadná. Dôvera je komplexný a dynamický pojem, ktorý závisí od mnohých faktorov, vrátane výkonu

technológie, skúseností používateľov a konkrétnych podmienok, v ktorých sa technológia používa.

V literatúre existuje mnoho rôznych konceptualizácií dôvery, avšak v kontexte interakcie človeka a stroja môže byť dôvera definovaná ako „postoj, že agent (v tomto prípade technológia alebo stroj) pomôže dosiahnuť ciele jednotlivca v danej situácii (Lee, 2004).“ Vojenské aplikácie umelej inteligencie sú obzvlášť citlivé na otázku dôvery, pretože operátori musia byť schopní spoľahnúť sa na systémy v kritických momentoch, kde akékoľvek zlyhanie alebo nesprávna kalibrácia dôvery môže mať závažné operačné dôsledky.

Zaujatosť voči automatizácii je ďalšou výzvou, ktoré môže ovplyvniť interakciu medzi človekom a strojom. Operátori môžu byť náchylní preceňovať schopnosti automatizovaných systémov alebo, naopak, môžu byť skeptickí voči ich rozhodnutiam, čo môže viesť k nesprávnemu využitiu technológie. Tieto dynamiky sú obzvlášť kritické v bojových situáciách, kde dôvera a rýchle rozhodovanie zohrávajú kľúčovú úlohu v úspechu alebo neúspechu vojenských operácií. Preto je nevyhnutné neustále prehodnocovať a kalibrovať úroveň dôvery v systémy umelej inteligencie, ako aj vzdelávať a školiť operátorov v oblasti optimálnej interakcie s týmito technológiami, aby sa minimalizovali hrozby s tým spojené v kontexte vojenských aplikácií.

2.3.2 KALIBRÁCIA DÔVERY A ZAUJATOSŤ VOČI AUTOMATIZÁCII

Dôvera je charakterizovaná neistotou a zraniteľnosťou a zohráva kľúčovú úlohu v tom, ako sa ľudia prispôsobujú zložitosti systémov, najmä v prostredí, kde sú nevyhnutné adaptívne reakcie. V kontexte vojenských operácií, kde fixné protokoly často nemôžu byť dodržiavané kvôli nepredvídateľným okolnostiam, je dôvera v technológie a automatizované systémy nevyhnutná pre efektívne rozhodovanie a úspešnú realizáciu operácií. V štúdiách automatizácie je dôvera chápaná ako kvalita, ktorá umožňuje a usmerňuje závislosť na systémoch, najmä keď je zložitosť technológie taká vysoká, že plné pochopenie jej fungovania je pre ľudského operátora nereálne. Táto dôvera je kritická v situáciách, ktoré vyžadujú vysokú mieru adaptability, čo je obzvlášť dôležité v bojových podmienkach, kde rýchle rozhodovanie môže rozhodovať o živote a smrti. Zároveň je dôležité poznamenať, že táto dôvera musí byť správne kalibrovaná (Lee, 2024). Prehnaná dôvera v automatizované systémy, bez dostatočného pochopenia ich limitácií, môže viesť k nebezpečným rozhodnutiam, rovnako ako nedostatočná dôvera môže brániť efektívnemu využitiu technológie. Vojenské prostredie, s jeho inherentnou zložitosťou a dynamikou, vyžaduje, aby operátori boli schopní rozpoznať, kedy sa na automatizované systémy spoľahnúť a kedy zasiahnuť manuálne, čím sa minimalizujú hrozby zlyhania v kritických momentoch.

Preto je nevyhnutné, aby vojenské tréningové programy zahŕňali nielen technické školenie, ale aj školenie zamerané na rozvoj správnej kalibrácie dôvery v automatizované systémy, čím sa zabezpečí, že operátori budú schopní efektívne interagovať so zložitými technológiami umelej inteligencie v rôznych operačných scenároch.

2.3.3 KALIBRÁCIA DÔVERY A JEJ VÝZVY

Kalibrácia dôvery je zásadná pre bezpečné a efektívne využívanie systémov umelej inteligencie, avšak predstavuje zložitú výzvu, ktorá nemá jednoznačný a univerzálny vzorec. Teoreticky ide o zladenie úrovne dôvery osoby s reálnymi možnosťami systému. Nesúlad medzi týmito faktormi sa môže prejaviť buď ako nadmerná dôvera, kde dôvera presahuje schopnosti systému, alebo ako nedôvera, kde dôvera neodpovedá skutočným schopnostiam systému.

Nadmerná nedôvera môže viesť k averzii voči algoritmom a zníženej ochote spoliehať sa na technológiu. Naopak, nadmerná alebo nekritická dôvera, často označovaná ako zaujatosť voči automatizácii, sa prejavuje nadmerným spoliehaním sa na výstupy automatizovaného systému. Tento jav môže nastať buď vtedy, keď operátori nezaznamenajú problémy, pretože automatizácia ich na ne neupozorní (tzv. „chyby opomenutia“), alebo keď nekriticky nasledujú chybné odporúčania automatizovaného systému (tzv. „chyby komisii“) (Parasuraman, 2000). Nadmerné spoliehanie sa na technológiu je častým problémom, najmä v dohľadových úlohách, kde systém pred zlyhaním fungoval spoľahlivo a konzistentne. Operátori si môžu zvyknúť a stratiť ostražitosť, čo predstavuje hrozbu, že nebudú schopní adekvátne reagovať na náhle zlyhanie systému. Tento fenomén „spoliehania sa“ je obzvlášť zreteľný v komplexných multitaskingových prostrediach, kde operátori čelia vysokým nárokom na svoje kognitívne zdroje. V takýchto situáciách môžu byť náchylní k nadmernej dôvere v automatizáciu a presunu svojich kognitívnych zdrojov na iné úlohy.

Podobná hrozba sa môže vyskytnúť aj v prípadoch, keď tréningové schopnosti operátorov zostanú dlhodobo nevyužité. Tento problém bol obzvlášť pozorovaný počas dlhodobých monitorovacích úloh, kde náhle zmeny v operačnom prostredí môžu operátorom, ktorí neboli predtým aktívne angažovaní, sťažiť rýchle zvýšenie ich mentálnej bdelosti v kritickom momente. To môže viesť k oneskoreným alebo nevhodným reakciám, čo v kontexte vojenských operácií môže mať závažné následky (Wickens, 2000). Preto je nevyhnutné, aby vojenské tréningové programy zahŕňali nielen technické zručnosti, ale aj stratégie na správnu kalibráciu dôvery v automatizované systémy, čím sa zabezpečí, že operátori budú schopní efektívne reagovať na rôzne operačné scenáre a minimalizovať hrozby spojené s nesprávnou dôverou v umelú inteligenciu.

ZÁVER

Umelá inteligencia sa dnes považuje za jednu z najvýznamnejších technologických inovácií, ktorá má potenciál výrazne zmeniť mnohé aspekty nášho života. Avšak s týmto pokrokom prichádzajú aj nové a komplexné výzvy, ktoré môžu mať hlboký dopad na globálnu bezpečnosť a stabilitu. Táto technológia, ktorá je stále vo fáze vývoja a integrácie do rôznych systémov, vyžaduje dôkladné posúdenie jej vplyvu nielen na úrovni jednotlivých národov, ale aj v rámci celosvetového spoločenstva.

Jedným z najvýraznejších rizík je militarizácia umelej inteligencie, ktorá prináša so sebou nové možnosti autonómnych zbraňových systémov. Tieto systémy majú potenciál rozhodovať o živote a smrti bez priamej ľudskej intervencie, čo zvyšuje hrozbu eskalácie konfliktov a môže viesť k neúmyselným a nepredvídateľným dôsledkom. Okrem toho, schopnosť umelej inteligencie rýchlo sa učiť a prispôsobovať novým situáciám otvára dvere k jej zneužitiu, čo môže spôsobiť, že tieto technológie budú využité spôsobmi, ktoré môžu ohroziť medzinárodnú stabilitu. Technické zlyhania a kybernetické útoky predstavujú ďalší významný zdroj obáv. Systémy poháňané umelou inteligenciou môžu byť zraniteľné voči manipulácii, čo môže mať katastrofálne následky, najmä ak sú tieto systémy integrované do kritických infraštruktúr alebo vojenských operácií. Kybernetické útoky môžu viesť k nesprávnym rozhodnutiam na bojisku alebo dokonca k eskalácii konfliktu, ak umelá inteligencia zlyhá v kľúčových momentoch.

Na medzinárodnej úrovni sa stále hľadajú vhodné rámce a mechanizmy na reguláciu a kontrolu rizík spojených s umelou inteligenciou. Hoci už existujú rôzne iniciatívy stále chýba komplexný a jednotný prístup, ktorý by zahŕňal všetky aspekty týchto rizík. Je zrejmé, že medzinárodná spolupráca je kľúčová pre vypracovanie pravidiel a štandardov, ktoré by

mohli minimalizovať hrozby spojené s umelou inteligenciou a zabezpečiť, že technológia bude slúžiť na prospech celej spoločnosti.

Ďalším významným aspektom je potreba zlepšiť naše porozumenie a povedomie o výzvach, ktoré umelá inteligencia prináša. To si vyžaduje investície do výskumu a vývoja, ktoré by umožnili lepšie predpovedať a zvládať výzvy spojené s touto technológiou. Štáty, organizácie aj jednotlivci musia spolupracovať na vytvorení spoločného rámca pre bezpečné nasadenie umelej inteligencie, pričom je potrebné zohľadniť etické a právne otázky, ktoré z toho vyplývajú. V neposlednom rade je dôležité uznať, že umelá inteligencia je stále v počiatočnej fáze svojho vývoja a jej plný potenciál, ako aj možné hrozby, sa budú naďalej vyvíjať. Preto je nevyhnutné, aby sme zostali ostražití a flexibilní v našich prístupoch k riadeniu rizík spojených s umelou inteligenciou. Vyváženie technologického pokroku s potrebou zachovania medzinárodnej bezpečnosti a stability bude jednou z najväčších výziev našej doby.

Zhrnutím, umelá inteligencia ponúka obrovský potenciál pre rozvoj a inovácie, avšak hrozby, ktoré so sebou prináša, nemožno ignorovať. Aby sme zabezpečili, že tieto technológie budú používané zodpovedne a bezpečne, je nevyhnutné, aby sme pokračovali v globálnom dialógu a spolupráci na vytvorení účinných mechanizmov, ktoré ochránia našu spoločnosť pred potenciálnymi hrozbami, ktoré môže umelá inteligencia predstavovať. Medzinárodné spoločenstvo musí aktívne pracovať na vytváraní rámcov, ktoré podporia mierové využitie umelej inteligencie a zabránia jej zneužitiu, čím prispeje k bezpečnejšiemu a stabilnejšiemu svetu.

ZOZNAM BIBLIOGRAFICKÝCH ODKAZOV

AMER, K. 2019. Deep Convolutional Neural Network-Based Autonomous Drone Navigation. [online]. Dostupné na internete: https://arxiv.org/abs/1905.01657.

AMODEI, D., OLAH, C., STEINHARDT, J., CHRISTIANO, P., SCHULMAN, J., & MANÉ, D. 2016. Concrete Problems in AI Safety. Ithaca: arXiv.org, 2016. 29 s. DOI:10.48550/arXiv.1606.06565.

ATHALYE, A. 2018. Synthesizing Robust Adversarial Examples [online]. Dostupné na internete: https://arxiv.org/pdf/1707.07397

BIGGIO, B., ROLI, F. 2018. Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning. [online]. Dostupné na internete: https://arxiv.org/pdf/1712.03141. DOI: https://doi.org/10.1145/3243734.3264418

BRUNDAGE, M. 2018. The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. [online]. Dostupné na internete: https://arxiv.org/pdf/1802.07228

CARLINI, N. et al. 2022. Membership Inference Attacks from First Principles. [online]. Dostupné na internete: https://arxiv.org/abs/2112.03570. DOI: https://doi.org/10.1109/SP46214.2022.9833649

CUMMINGS, M. L., LI, S. 2021. Subjectivity in the Creation of Machine Learning Models. Journal of Data and Information Quality. New York: Association for Computing Machinery, 2021. 19 s. ISSN 1936-1955. DOI: https://doi.org/10.1145/3418034

CUMMINGS, M. Revisiting Human-Systems Engineering Principles for Embedded AI Applications. [online]. Dostupné na internete: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.513474/full.

EKELHOF, M., PERSI PAOLI, G. 2019. The Human Element in Decisions about the Use of Force. [online]. Dostupné na internete: https://unidir.org/sites/default/files/2020-03/UNIDIR_Iceberg_SinglePages_web.pdf.

EYKOLT, K. 2018. Robust Physical-World Attacks on Deep Learning Visual Classification. [online]. Dostupné na internete: https://arxiv.org/pdf/1707.08945.pdf. DOI: https://doi.org/10.1109/CVPR.2018.00175

FAZEKAS, F. 2021. AI and Military Operations' Planning. In VISVIZI, A., BODZIANY, M. (eds). Artificial Intelligence and Its Contexts. Cham: Springer, 2021, s. 79–91. DOI: https://doi.org/10.1007/978-3-030-88972-2_6

Federal Office for Information Security, "AI Security Concerns in a Nutshell" [online]. Dostupné na internete: https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/Practical_AI-Security_Guide_2023.pdf?__blob=publicationFile&v=5

FISCHER, S. C. 2022. Military AI Applications: A Cross-Country Comparison of Emerging Capabilities. In: Armament, Arms Control and Artificial Intelligence. Cham: Springer, 2022, s. 39-55. ISBN 978-3-031-11042-9. DOI: https://doi.org/10.1007/978-3-031-11043-6_4

FLOURNOY, M., HAINES, A., CHEFITZ, G. 2020. Building Trust through Testing: Adapting DOD's Test & Evaluation, Validation & Verification (TEVV) Enterprise for Machine Learning Systems, Including Deep Learning Systems. [online]. Dostupné na internete: https://cset.georgetown.edu/wp-content/uploads/Building-Trust-Through-

GOODFELLOW, I., BENGIO, Y., COURVILLE, A. 2016. Deep Learning. Cambridge: MIT Press, 2016. 800 s. ISBN 978-0262035613.

GOODFELLOW, I., Shlens, J., Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. Mountain View: arXiv preprint arXiv:1412.6572, 2015. 11 s. doi: https://doi.org/10.48550/arXiv.1412.6572

HOFFMAN, W. 2021. AI and the Future of Cyber Competition. [online]. Dostupné na internete: https://cset.georgetown.edu/publication/ai-and-the-future-of-cyber-competition

ISO 31000:2018 Risk Management Guidelines. [online]. Dostupné na internete: https://www.iso.org/obp/ui/en/#iso:std:iso:31000:ed-2:v1:en

JANJEVA, A. et al. 2023. Strengthening Resilience to AI Risk: A guide for UK policymakers. [online]. Dostupné na internete: https://cetas.turing.ac.uk/publications/strengthening-resilience-ai-risk

LEE, J. D., SEE, K. A. 2004. Trust in Automation: Designing for Appropriate Reliance. Human Factors: The Journal of the Human Factors and Ergonomics Society. [online]. Dostupné na internete: https://journals.sagepub.com/doi/epdf/10.1518/hfes.46.1.50_30392. DOI: https://doi.org/10.1518/hfes.46.1.50.30392

LOHN, A. 2020. Hacking AI: A Primer for Policymakers on Machine Learning Cybersecurity. [online]. Dostupné na internete: https://cset.georgetown.edu/publication/hacking-ai/

LOHN, A. J. 2020. Estimating the Brittleness of AI: Safety Integrity Levels and the Need for Testing Out-of-Distribution Performance. [online]. Dostupné na internete: https://arxiv.org/pdf/2009.00802.pdf.

MINISTERSTVO OBCHODU USA, Národný inštitút pre štandardy a technológie, "Artificial Intelligence Risk Management Framework". [online]. Dostupné na internete: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

MORGAN, F. E. 2020. Military Applications of Artificial Intelligence: Ethical Concerns in an Uncertain World. [online]. Dostupné na internete: https://www.rand.org/pubs/research_reports/RR3139-1.html. DOI: https://doi.org/10.7249/RR3139-1

OECD, "Advancing Accountability in AI. Governing and managing risks throughout the lifecycle for trustworthy AI". [online]. DOI: https://doi.org/10.1787/2448f04b-en

Our Common Agenda. Policy Brief 9. A New Agenda for Peace. 19. júl. 2023 [online]. Dostupné na internete: https://www.un.org/sites/un2.un.org/files/our-common-agenda-policy-brief-new-agenda-for-peace-en.pdf.

PAPERNOT, N. et al. 2016. Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks. [online]. Dostupné na internete: https://arxiv.org/pdf/1511.04508. DOI: https://doi.org/10.1109/SP.2016.41

PARASURAMAN, R., SHERIDAN, T. B., WICKENS, C. D. 2000. A Model for Types and Levels of Human Interaction with Automation. New York : Institute of Electrical and Electronics Engineers, 2000. 286-297 s. ISSN 1083-4427. DOI: https://doi.org/10.1109/3468.844354

pdf/2112.03570.pdf.

RUDNER, T. G. J. a H. Toner. 2021. Key Concepts in AI Safety: Robustness and Adversarial Examples [online]. Dostupné na internete: https://cset.georgetown.edu/publication/key-concepts-in-ai-safety-robustness-and-adversarial-examples/

RUSSELL, S., NORVIG, P. 2022. Artificial Intelligence: A Modern Approach. Harlow : Pearson, 2022. kapitola 11 a kapitola 26, ISBN 978-0-13-604259-4.

SHARKEY, N. E. 2008. Killing Made Easy: From Operational to Moral Implications of Automating War.. [online]. Dostupné na internete: https://www.dhi.ac.uk/san/waysofbeing/data/governance-crone-sharkey-2012b.pdf

SCHARRE, P. 2018. Army of None: Autonomous Weapons and the Future of War. New York: W. W. Norton & Company, 2018. 448 s. ISBN 978-0393608984.

SIVA KUMAR, R. 2019. Failure Modes in Machine Learning Systems. [online]. Dostupné na internete: https://arxiv.org/pdf/1911.11034

STYBER, M. 2023. Warfare in the Age of AI: A Critical Evaluation of Arkin's Case for Ethical Autonomy in Unmanned Systems. In Artificial Intelligence Research. SACAIR2023. Cham: Springer Nature Switzerland, 2023, s. 62-78. DOI: https://doi.org/10.1007/978-3-031-49002-6_5

SZEGEDY, C., ZAREMBA, W., SUTSKEVER, I., et al. 2014. Intriguing properties of neural networks. [online]. Dostupné na internete: https://ar5iv.labs.arxiv.org/html/1312.6199

Testing.pdf .

von BRAUN, J., ARCHER, M., REICHBERG, G.M., SÁNCHEZ SORONDO, M. 2021. Robotics, AI, and Humanity: Science, Ethics, and Policy. Cham: Springer, 2021. 269 s. ISBN 978-3-030-54172-9.

WANG, B. 2019. Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks. [online]. Dostupné na internete: https://people.cs.uchicago.edu/~huiyingli/publication/backdoor-sp19.pdf. DOI: https://doi.org/10.1109/SP.2019.00031

WICKENS, C. D., HOLLANDS, J. G. 2000. Engineering Psychology and Human Performance. Upper Saddle River : Prentice Hall, 2000. 573 s. ISBN 9781315665177

WOJTON, H. M., PORTER, D. J., and DENNIS, J. W. Test & Evaluation of AI-enabled and Autonomous Systems: A Literature Review. Institute for Defense Analyses, 9 March 2021. [online]. Dostupné na internete: https://testscience.org/wp-content/uploads/formidable/20/Autonomy-Lit-Review.pdf.

YAMPOLSKIY, R. V. 2018. Artificial Intelligence Safety and Security. Boca Raton: CRC Press, 444 s. ISBN 9781138547086. DOI: https://doi.org/10.1201/9781351251389

kpt. JUDr. Milan KUSÁK, PhD.
Borovianska cesta 1, 960 01 Zvolen
kusakmilan@gmail.com