UNIVERZITA KARLOVA

FAKULTA SOCIÁLNÍCH VĚD

Institut mezinárodních studií

Katedra severoamerických studií

Bakalářská práce

2024 **Michaela Mňuková**

UNIVERZITA KARLOVA

FAKULTA SOCIÁLNÍCH VĚD

Institut mezinárodních studií

Katedra severoamerických studií

Etická problematika implementace autonomních zbraňových systémů: perspektiva USA

Bakalářská práce

Autorka práce: Michaela Mňuková

Studijní program: Teritoriální studia

Vedoucí práce: PhDr. Jan Hornát, Ph.D.

Rok obhajoby: 2024

Prohlášení

1. Prohlašuji, že jsem předkládanou práci zpracovala samostatně a použila jen uvedené prameny a literaturu.
2. Prohlašuji, že práce nebyla využita k získání jiného titulu.
3. Souhlasím s tím, aby práce byla zpřístupněna pro studijní a výzkumné účely.
4. Při přípravě této práce autor použil ChatGpt za účelem kontroly pomoci s překlady termínů, jež nemají etablovaný či oficiální český překlad a tvorby bibliografických údajů. Po použití tohoto nástroje/služby autor obsah podle potřeby zkontroloval a upravil a přebírá plnou odpovědnost za obsah publikace.

V Praze dne 30. července 2024 Michaela Mňuková

Bibliografický záznam

MŇUKOVÁ, Michaela. *Etická problematika implementace autonomních zbraňových systémů: perspektiva USA*. Praha, 2024. 45 s. Bakalářská práce (Bc). Univerzita Karlova, Fakulta sociálních věd, Institut mezinárodních studií, Katedra severoamerických studií. Vedoucí bakalářské práce PhDr. Jan Hornát, Ph.D.

Rozsah práce: 72 934 znaků

Abstrakt

Bakalářská práce se zabývá etickou problematikou implementace smrtících autonomních zbraňových systémů ve Spojených státech amerických. Práce využívá odbornou literaturou i škálou pramenů, včetně dokumentů legislativního, strategického i analytického charakteru. Teoretická část představuje eticky problematické aspekty využití autonomních technologií v ozbrojených konfliktech. Důraz je kladen na tři zvolené problematiky: ospravedlnitelnost, předpojatost a zodpovědnost. Tyto aspekty jsou zkoumány z perspektivy potenciálních porušení mezinárodní humanitární právo. Praktická část práce analyzuje americkou strategii na implementaci smrtících autonomních zbraní. Zkoumá, do jaké míry tato strategie zohledňuje a řeší tři výše zmíněné etické problematiky. V rámci této části jsou analyzovány dokumenty zveřejněné americkým ministerstvem obrany a dalšími vládními subjekty. Analýza zkoumaných dokumentů prokázala, že americká strategie na implementaci smrtících autonomních zbraňových systémů zohledňuje všechny zmíněné etické problematiky. Zodpovědnost za činy smrtících autonomních zbraňových systémů připadá vždy na operátora této zbraně. Problematiku ospravedlnitelnosti americká strategie řeší skrze vzdělání operátorů a zajištění, že osoby operující autonomní zbraně rozumí jejím funkcím a jsou schopny vysvětlit jejich jednotlivá rozhodnutí a činy. Otázka předpojatosti není ve strategii adresována přímo, i přesto strategie problematiku zohledňuje. Všechny autonomní zbraně musí projít precizním testováním, aby byl zajištěn jejich bezchybný provoz. Využívané technologie jsou nadále monitorovány i během nasazení a útoky těchto zbraní podléhají schválení jejich operátora.

Abstract

The thesis explores the ethical implications of implementing lethal autonomous weapon systems in the USA. The thesis works with academic literature as well as a variety of sources, including legislative, policy and analytical documents. The theory part introduces the ethical concerns stemming from implementing autonomous technologies in armed conflicts. The emphasis is on three selected issues: justifiability, prejudice and responsibility. These issues are explored in terms of the possibility of them violating international humanitarian law. In the practical part of the thesis an analysis of the U. S. strategy to implement lethal autonomous weapons was conducted. It examines, to what extent and how, the three ethical

issues mentioned above are considered in this strategy. This part analyses documents published by the Department of Defense or other federal entities. The analysis demonstrated that the U. S strategy considers all of the mentioned ethical issues. The responsibility for the actions of lethal autonomous weapons is always held by its operator. The issue of justifiability is solved through education of the operators to ensure, that the personnel operating autonomous weapons understands its functions and is able to explain its individual decisions and actions. Although the issue of prejudice is not addressed directly, the strategy still considers the issue. Every autonomous weapon must undergo detailed testing before its fielding, to ensure its flawless function. The technologies keep on being monitored during its active use and its attacks must be verified by the operator.

Klíčová slova

Autonomní zbraně, vojenská etika, vojenská aplikace umělé inteligence, etické aspekty autonomních zbraní, smrtící autonomní zbraňové systémy (LAWS)

Keywords

Autonomous weapons, Military ethics, AI in military applications, Ethical implications of autonomous weapons, Lethal autonomous weapon systems (LAWS)

Title/název práce

Ethical Issues in the Implementation of Autonomous Weapon Systems: A U.S. Perspective

Poděkování

Ráda bych poděkovala svému vedoucímu Janu Hornátovi za pomoc při psaní této práce. Velice si vážím všech rad, které mi pomohly tuto práci dokončit.

Také bych ráda poděkovala své rodině za podporu. Jsem vděčná za vše, co pro mě dělají a vážím si možností, které díky nim v životě mám.

Obsah

Úvod

Inovativní technologie byly vždy využívány armádami po celém světě a jinak tomu není ani v kontextu implementace umělé inteligence a autonomních zbraňových systémů v dnešní době. V průběhu historie jsme, jako společnost, dosáhly závěru, že vedení ozbrojených konfliktů by se mělo řídit univerzálně uznávanými pravidly. Tato pravidla, známá jako mezinárodní humanitární právo, chrání civilisty a určují, jaké zbraně a jakým způsobem lze používat a jaké jsou naopak zakázané. Postupem času armády začaly tyto etické kodexy přizpůsobovat svým vlastním legislativám a doplňovat je o nová pravidla. Některé typy zbraní jsou z etického hlediska rizikovější než jiné.

Využití smrtících autonomních zbraní může být eticky velmi problematické, pokud nebudou podstoupeny kroky k zajištění jejich bezpečného a etického využití. Světové armády se řídí nejen mezinárodním humanitárním právem, ale i vlastními etickými normami. Tyto etická pravidla jsou individuální a každý stát si je definuje sám. Zhotovení jasně artikulovaných pravidel využití smrtících autonomních zbraní je klíčové pro zajištění ochrany základních lidských práv. Současně má tento krok potenciál vyvolat vyšší míru důvěry v tyto technologie mezi veřejností i mezi členy armády.

Spojené státy americké disponují jednou z největších armád na světě. Taktéž se jedná o první zemi na světě, která zveřejnila etický kodex pro využití umělé inteligence v ozbrojených konfliktech. Současně se řadí mezi vedoucí mocnosti v oblasti implementace umělé inteligence a adaptace autonomních systémů. V praktické části této práce budou zkoumány dokumenty mapující americkou strategii na etickou implementaci smrtících autonomních zbraňových systémů. Tyto dokumenty jsou právního, analytického a strategického charakteru. Všechny dokumenty byly publikovány buď americkým ministerstvem obrany či jinými vládními subjekty.

Cílem práce je zhodnotit, zdali a do jaké míry americká strategie na implementaci smrtících autonomních zbraňových systémů zohledňuje tři zvolené eticky problematické oblasti – zodpovědnost, předpojatost a ospravedlnitelnost. Tento cíl lze rozdělit do dílčích otázek. Práce se snaží zjistit, kdo ponese zodpovědnost za činy autonomních zbraňových systémů. Jaké kroky jsou podstoupeny, aby byla zajištěna ospravedlnitelnost těchto činů. A na závěr, jak plánuje americké ministerstvo obrany zajistit, že autonomní zbraňové systémy nebudou negativně ovlivněny předpojatostí. Analýza doufá v potvrzení, že se Spojené státy řadí mezi přední vojenské velmoci, jež respektují mezinárodní humanitární právo. Práce je

určena všem, kteří se o tématiku etické implementace autonomních zbraní zajímají.

1. Umělá inteligence a pojetí termínu v kontextu práce

Umělá inteligence se postupně, avšak nepochybně, stává součástí našich každodenních životů. Proniká do oblastí, kterým dříve dominovali pouze lidé, a představa, že by stroj mohl dosáhnout, nebo dokonce překonat lidské schopnosti, byla dlouho považována spíše za sci-fi fantazii technických nadšenců než za reálnou možnost. Díky technologickému pokroku se však tyto fantazie začínají přibližovat reálné situaci.[1] Tato rychle se rozvíjející, zdánlivě všestranná, technologie, jejíž kořeny sahají do poloviny 20. století, je i ze své samotné podstaty velice proměnlivá.[2] Tento fakt se odráží i v termínech a definicích s ní spjatých. Proto je tato úvodní kapitola věnována vymezení klíčových termínů a definic, aby se předešlo případným chybným interpretacím následujících částí práce.

V úvodu je klíčové rozdělit fakta od fikce. Umělá inteligence je atraktivním a často diskutovaným tématem, jež se pravidelně objevuje v médiích. Média mají často tendenci tématiku umělé inteligence mystifikovat a interpretovat umělou inteligenci jako potencionální hrozbu v podobě plně autonomní technologie, která by mohla ovládnout lidskou rasu, jak v jedné ze svých přednášek zmiňuje Stuart Russell.[3] Umělá inteligence se obvykle dělí na tzv. *narrow* či *weak*, tedy slabou umělou inteligenci a *general* či *strong*, tedy obecnou umělou inteligenci.

V současné situaci se v praxi využívá první ze zmíněných, tedy *narrow artificial intelligence (AI)*. Jak již samotný název naznačuje, jedná se o specializované technologie, které jsou schopny vykonávat určité úzce specifikované úkony autonomně, tedy bez zásahu člověka.[4] Těmito úkony mohou být například rozeznávání řeči či vizuálního materiálu, zároveň se však jedná o technologie využívané v autonomních vozidlech, bezpilotních letounech či autonomních zbraňových systémech. Naopak *general AI* lze chápat jako plně autonomní technologii, která se svými kognitivními schopnostmi vyrovná člověku, či dokonce předčí lidské schopnosti. Tato technologie není na člověku nijak závislá. Dodnes

[1] National Security Commission on Artificial Intelligence, *The Final Report* (Washington, D.C.: National Security Commission on Artificial Intelligence, March 1, 2021), https://assets.foleon.com/eu-central-1/de-uploads-7e3kk3/48187/nscai_full_report_digital.04d6b124173c.pdf. (staženo 2. dubna 2024) Str. 20-21

[2] Bottino, Andrea et al. 2021. „A Brief History of AI: How to Prevent Another Winter (A Critical Review).“ *PET Clinics* 16 (4): 449-469. https://doi.org/10.1016/j.cpet.2021.07.001. Zobrazeno ve verzi preprint zveřejněné 12. prosince 2022 (staženo 30. dubna 2024), str. 4-8

[3] Stuart Russell, „AI in Warfare,“ *The Reith Lectures*, BBC Radio 4, December 8, 2021, https://www.bbc.co.uk/programmes/m00127t9. (staženo 7. května 2023)

[4] Bartneck, Christoph, Christoph Lütge, Alan Wagner, and Sean Welsh. 2021. „What is AI?“ 2021. In *An Introduction to Ethics in Robotics and AI*, 5-16. Springer Briefs in Ethics. https://doi.org/10.1007/978-3-030-51110-4_2. (staženo 12. února 2024), str. 10

však této úrovně inteligence nedosáhl žádný systém. Vytvoření obecné umělé inteligence však zůstává i nadále cílem výzkumu řady vědců.[5]

Jak bylo již dříve zmíněno, umělá inteligence, i kvůli své komplexní a nejednotné podobě, stále není jednotně definovaný termín. Pro účely této práce byla zvolena definice vyplývající ze Zákonu o národní obraně na fiskální rok 2019 (*National Defense Authorization Act for Fiscal Year 2019*), jež umělou inteligenci vymezuje jako systém, který racionálně řeší úkoly v nestálých podmínkách bez výrazné lidské intervence.[6] Tato definice[7] byla zvolena, jelikož na ni odkazuje závěrečná zpráva (*The Final Report*) Národní bezpečnostní komise pro umělou inteligenci (*National Security Commission on Artificial Intelligence*), se kterou tato práce v následujících částech dále pracuje. Tento systém svůj výkon, díky různým inovativním postupům, jako je strojové učení, neustále zlepšuje a učí se na základě jemu dostupných dat. Zadané cíle řeší pomocí snímání, plánování, poznávání, učení, komunikace, rozhodování a jednání, a to buď jako inteligentní softwarový agent nebo jako robot. Je navržen tak, aby myslel či jednal jako člověk.[8]

Učení systémů, proces známý jako strojové učení, je umožněno díky samotné struktuře této technologie. Při vývoji umělé inteligence se vědci snažili napodobit neuronovou síť podobnou té v lidském mozku. Došlo tak k vytvoření digitální imitace neuronů, které, podobně jako ty v lidském mozku, přijímají, zpracovávají a dále předávají informace. Společně tvoří digitální neuronové sítě. Díky kombinaci implementace neuronových sítí a procesu známém též jako *deep learning*[9], tedy hluboké učení, vzniká umělá inteligence, jak ji známe dnes. Jedná se o technologii, která se učí díky zpracování dat, v kterých hledá vzorce repetitivních znaků. Získané znalosti systém dále implementuje ve svých rozhodnutích a při plnění zadaných úkolů. Zkušenosti a poznatky načerpané v průběhu plnění zadaných akcí systém dále využívá ke svému zlepšení.[10]

Dalším klíčovým termínem, jež je pro jasné porozumění práci třeba definovat jsou autonomní zbraně, specificky autonomní zbraňové systémy a polo-autonomní zbraňové

---

[5] Ibid, str. 10

[6] *National Defense Authorization Act for Fiscal Year 2019: Pub. L. No. H.R. 5515*. 2018. https://www.congress.gov/115/bills/hr5515/BILLS-115hr5515enr.pdf (staženo 18. března 2024) str. 62-63

[7] V současné době nejaktuálnější definicí umělé inteligence je definice Evropské unie formulována v EU Artificial Intelligence Act – viz Článek 3 (1) https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf

[8] National Defense Authorization Act for Fiscal Year 2019, str. 62-63

[9] Jedná se o zastřešující termín, pod nějž spadají specializované typy učení jako je *supervised, unsupervised* či *semi-supervised*. Zároveň se nejedná o jediný typ strojového učení, je však vnímán jako nejrozšířenější.

[10] Lee, Kai-Fu. 2018. *AI Superpowers: China, Silicon Valley, and the New World Order*. Houghton Mifflin Harcourt. Str. 16-17

systémy, tedy technologie, které jsou schopny, díky umělé inteligenci, určité cíle plnit autonomně, tedy, jak již bylo zmíněno, bez zapojení člověka. Práce využívá definice stanovené směrnicí amerického ministerstva obrany 3000.09 Autonomie zbraňových systémů (*Department of Defense Directive 3000.09 Autonomy in Weapon Systems*). Tato směrnice definuje autonomní zbraňový systém jako zbraňový systém, který po aktivaci dokáže vybrat a zasáhnout cíle bez další intervence operátora. Tyto autonomní systémy jsou navrženy tak, aby umožnily zásah operátora, který má možnost akce přerušit. Zbraňové systémy jsou však schopny fungovat i plně autonomně a na operátorech nejsou závislé.[11] Polo-autonomní zbraňové systémy jsou směrnicí definované jako zbraňové systémy, které jsou po aktivaci určeny pouze k zasahování jednotlivých cílů nebo konkrétních skupin cílů, které byly vybrány operátorem. Tyto zbraňové systémy se dále dělí na ty, které jsou autonomní ve funkcích spjatých s interakcí s cílem, kupříkladu při identifikaci a sledování potenciálních cílů či při pomoci operátorovi s přesným načasováním výstřelů. Druhým typem je specializovaná munice, která funguje na principu *fire and forget*, čili vystřel a zapomeň. Tento typ střel je schopen zasáhnout, jim předem stanovené, cíle i bez další pomoci operátora.[12]

[11] *Department of Defense Directive 3000.09 Autonomy in Weapon Systems. 2023*. https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf. (staženo 29. března 2024) str. 21

[12] DoD Directive 3000.09 Autonomy in Weapon Systems, str. 23

2. Etická otázka využití umělé inteligence v ozbrojených konfliktech

Umělá inteligence se pomalu, ale jistě, stává součástí našich každodenních životů. Tato nová technologie se stává předmětem veřejných debat, korporace ji začínají implementovat do svých procesů s cílem zvýšení zisku a vlády po celém světě začínají zpracovávat své vlastní národní strategie na implementaci umělé inteligence.[13] Jinak tomu není ani ve vojenském sektoru. S příchodem umělé inteligence se mění styl vedení boje a vzniká nový koncept řízení války. V této nové informační éře budou nejdůležitějšími válečnými prostředky nikoliv lidé a stroje, ale propracované algoritmy poháněné umělou inteligencí. Rozhodujícím faktorem se stává množství, kvalita a rychlost zpracování dat a efektivnost komplexně provázaných sítí tvořené zbraněmi i systémy, které jsou vzájemně propojené díky umělé inteligenci.[14]

Umělá inteligence se užívá napříč celým armádním sektorem – od plánování, přes záchranné a zásobovací mise, až po aktivní zapojení do vojenských akcí, včetně přímého útoku[15]. Využití této technologie a její implementace se však řídí určitými pravidly. Tato pravidla, spadají, mimo jiné, pod mezinárodní humanitární právo. Jsou implementována nejen aby poskytovala ochranu obětem konfliktů, ale zároveň aby stanovila zásady vedení vojenských operací. Mezinárodní humanitární právo se nevztahuje pouze na osoby, které se, přímo či nepřímo, účastní těchto konfliktů, ale i na prostředky a zbraně které jsou v konfliktu využívány.[16]

I přesto, že se využití umělé inteligence v ozbrojených konfliktech do určité míry řídí mezinárodním humanitárním právem, je její využití stále spjato s řadou etických otázek. Především problematické jsou smrtící autonomní zbraně. V současné situaci existuje řada zemí, které vyzvali k preventivnímu zákazu těchto technologií právě kvůli jejich etické rizikovosti.[17] Tato práce se konkrétně zaměří na tři vybrané eticky problematické oblasti –

[13] Lee. AI Superpowers: China, Silicon Valley, and the New World Order, str. 6

[14] National Security Commission on Artificial Intelligence, The Final Report, str. 77

[15] Bartneck, Christoph, Christoph Lütge, Alan Wagner, and Sean Welsh. 2021. „Military Uses of AI.“ In An Introduction to Ethics in Robotics and AI, 93-99. Springer Briefs in Ethics. https://doi.org/10.1007/978-3-030-51110-4_11. (staženo 12.2.2024), str. 97

[16] Melzer, Nils, and Etienne Kuster. 2019. International Humanitarian Law A Comprehensive Introduction. International Committee of the Red Cross. https://doi.org/10.1017/S1816383117000091. (staženo 3. dubna 2024) str. 17

[17] Stop Killer Robots. „UN Head Calls for a Ban.“ Last modified December 11, 2018. https://www.stopkillerrobots.org/news/unban/. (citováno 12. července 2024)

ospravedlnitelnost, předpojatost a zodpovědnost. Tyto oblasti jsou již ze své podstaty úzce provázané a navzájem se ovlivňují. Je proto nezbytné o problému uvažovat komplexně a nenahlížet na tyto oblasti izolovaně.

Při uvažování o rizicích spjatých s využitím umělé inteligence, je důležité si uvědomit, že umělá inteligence je pouhým prostředkem. Jak Mark Scott řekl v podcastu EU Confidential od Politico: „Je [umělá inteligence] pouhým nástrojem. Není ani dobrým nástrojem, ani špatným nástrojem. Je to pouze nástroj. Záleží na tom, jak ho využíváme.“[18] Na tento Scottův výrok zároveň navazuje myšlenka Kai-Fu Leeho, že budoucnost umělé inteligence je tvořena námi, lidmi, a bude odrážet naše rozhodnutí a činy.[19] Na základě těchto tvrzení můžeme usoudit, že implementace umělé inteligence do zbraňových systémů je možná, protože záleží na našich rozhodnutích a pravidlech, která pro tento účel stanovíme. Jedná se o podobný proces jako u jiných nových technologií, které byly v průběhu dějin začleněny do vojenské sféry, i přesto že tato vyžaduje více pozornosti.

Je důležité si uvědomit, že narozdíl od jiných nám známých prostředků, jež lze jasně definovat a jejichž podoba je do určité míry neměnná, v případě umělé inteligence se jedná spíše o komplexní síť zahrnující škálu různých technologií.[20] Tak, jako již jiné inovativní technologie v minulosti, přináší umělá inteligence a autonomní zbraně řadu výhod. Implementace umělé inteligence zajistí rychlejší a přesnější rozhodnutí a za předpokladu, že algoritmus projde dostatečným testováním, může být tento způsob efektivnější nežli rozhodnutí, které provede člověk. Systém je totiž, především v krizových situacích, často schopen provést rozhodnutí rychleji, informovaněji a objektivněji nežli člověk. Zároveň v případě rozhodnutí, jež provede umělá inteligence, nehrozí, že bude ovlivněno emocemi či jinými vnějšími vlivy.[21] Otázkou však zůstává, jak zajistit, že rozhodnutí umělé inteligence budou etická.

Častým názorem je, že by součástí rozhodovacího procesu měl být člověk. Zapojení člověka do rozhodovacího procesu může být prospěšné z celé řady důvodů, mimo jiné lepší zhodnocení rizika a příčin výskytu civilistů v zóně aktivního konfliktu. Ačkoliv zapojení

[18] Sarah Wheaton, „How Russian Disinformation Could Skew EU Election — and Whether Europe Can Fight It,“ *EU Confidential*, podcast episode, POLITICO, May 17, 2024, https://www.politico.eu/podcast/eu-confidential/how-russian-disinformation-could-skew-eu-election-and-whether-europe-can-fight-it/. (staženo 30. května 2024)

[19] Lee, AI Superpowers: China, Silicon Valley, and the New World Order, str. 8

[20] National Security Commission on Artificial Intelligence, The Final Report, str. 7

[21] Rowe, Neil C. 2022. „The comparative ethics of artificial-intelligence methods for military applications.“ *Frontiers in Big Data* 5 (991759). https://doi.org/https://doi.org/10.3389/fdata.2022.991759. (staženo 9. července 2023), str. 3

lidského elementu do rozhodovacího procesu řeší určité problémy, s kterými se v rámci etické otázky využití umělé inteligence v ozbrojených konfliktech setkáváme, především pak problematiku zodpovědnosti, nezaručuje, že daná rozhodnutí budou etičtější. Důvodem může být nedostatečná informovanost operátorů, kteří se na rozhodnutí podílejí. Současně existuje riziku, že tito operátoři mohou být ovlivněni vnějšími vlivy, jako je propaganda či emoce. V neposlední řadě je také důležité pracovat s variantou, že některá lidská rozhodnutí jsou již ze své podstaty cíleně neetická.[22]

I přesto, že využití umělé inteligence a autonomních systémů může přispět k bezpečnější a etičtější společnosti, je klíčové, aby existovala pravidla, jež budou upravovat využití těchto technologií. Bez kritického zhodnocení etických i bezpečnostních rizik, jež jsou spjatá s využitím autonomních zbraní, může dojít k narušení základních lidských práv. Světové armády, přinejmenším Spojené státy a jejich partneři, by proto měly přijmout mezinárodně uznávané etické normy, jež by definovaly, jak navrhovat a využívat autonomní technologie.[23]

2.1. Eticky problematické oblasti využití umělé inteligence v ozbrojených konfliktech

Eticky problematických oblastí spjatých s využitím autonomních zbraní je celá řada. Vedle, v této kapitole diskutované, předpojatosti, ospravedlnitelnosti a zodpovědnosti, se mezi ně dále řadí i lidská důstojnost, či etické aspekty spjaté s vývojem, akvizicí a testováním těchto technologií.[24]

Předpojatost

Předpojatost je otázka, jež se týká rovným dílem autonomních zbraní i člověka, a to i přesto, že se na první pohled může zdát, že se jedná primárně o problematiku spjatou s rozhodnutím člověka. Pokud rozhodnutí provádí člověk, je zde jisté riziko, že tato rozhodnutí budou

[22] Rowe, The comparative ethics of artificial-intelligence methods for military applications, str. 3

[23] Galliott, Jai. 2021. „Toward a Positive Statement of Ethical Principles for Military AI.“ In *Lethal Autonomous Weapons: Re-Examining the Law and Ethics of Robotic Warfare*, 121-135. Oxford University Press. https://books.google.cz/books?hl=cs&lr=&id=3PYTEAAAQBAJ&oi=fnd&pg=PA121&dq=Galliott,+J.+(2021).+%E2%80%9CToward+a+positive+statement+of+ethical+principles+for+military+AI,%E2%80%9D+in+Lethal+Autonomous+Weapons,+eds+Galliott,+J.,+MacIntosh,+D.,+and+Ohlin,+J.+(Oxford:+Oxford+University+Press)&ots=OTGxl9TLlR&sig=y8qGZhV1gUloHigyV2H1e1QvNj4&redir_esc=y#v=onepage&q&f=false. (staženo 15. července 2024), str. 121-126

[24] Amoroso, Daniele, and Guglielmo Tamburrini. 2020. “Autonomous Weapons Systems and Meaningful Human Control: Ethical and Legal Issues.” *Current Robotics Reports* 2020 (1): 187-194. https://link.springer.com/content/pdf/10.1007/s43154-020-00024-3.pdf. (staženo 3.dubna 2024), str. 188-191

ovlivněna vnějšími vlivy. Mezi tyto vlivy se řadí propaganda, emoce a různé osobní zkušenosti a názory, jež mohou negativně ovlivnit rozhodnutí daného člověka. Příkladem mohou být osobní postoje vůči určitým národnostem, etnikům nebo politickým či náboženským skupinám, které jsou způsobeny předešlou negativní zkušeností či vlivem výchovy nebo prostředí v kterém se daná osoba dlouhodobě nachází. Tyto negativní postoje vůči specifickým skupinám mohou u člověka negativně ovlivnit rozhodovací proces, vést k neadekvátní reakci či špatnému odhadnutí rizik.[25]

Využití umělé inteligence se vzdáleně může zdát jako řešení této problematiky, ovšem i zde se setkáváme s problémy. Jak již bylo zmíněno v předešlé kapitole, umělá inteligence je velmi inovativní technologií, díky své schopnosti se učit, a to skrze zpracovávání enormního množství dat v rámci procesu známém jako strojové učení. Autonomní systémy si tak mohou, ač nechtěně, naučit podobné diskriminativní názory jako lidé. Právě data, jež jsou využívána k strojovému učení mohou být problematická, a to ze dvou důvodů. Zaprvé, je důležité zajistit, že užívaná data nabízí dostatečnou škálu variabilních příkladů. Data, využívaná pro strojové učení systému, by měla představit dostatečně diverzifikovanou skupinu zástupců z různých skupin, aby s nimi byl systém seznámen. Nejenže tato data musí nabídnout systému dostatečně širokou škálu možností, aby nehrozilo, že se systém rozhodne špatně kvůli nedostatečné informovanosti, ale zároveň musí být zajištěno, že jednotlivé skupiny budou zastoupeny rovným dílem. Pokud se systém bude učit na datech, jež obsahují proporcionálně větší zastoupení některé ze skupin, znamená to že, pro tuto skupinu si vytvoří více propracované vzorce, načerpá o ní více informací a snadněji ji rozezná. Naopak nedostatečné množství vzorků mohou znamenat, že systém bude mít problém osobu zařadit do správné skupiny, jelikož mu nebylo poskytnuto dostatek informací, s kterými může pracovat. Nedostatečné znalosti systému mohou potenciálně vést ke špatným rozhodnutím.[26] Toto se potvrdilo i ve výzkumu *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, na nějž ve svém článku *Artificial Intelligence, Robotics, Ethics, and the Military: A Canadian Perspective* odkazují Thorpeová a Wasilowá, který se zaměřoval na rozpoznávání obličejů umělou inteligencí. Výzkum prokázal, že umělá inteligence chybně rozpoznávala určité skupiny obyvatel, protože během procesů strojového učení nebyla vystavena dostatečnému množství

[25] Rowe, The comparative ethics of artificial-intelligence methods for military applications, str. 4

[26] Thorpe, Joelle B., and Sherry Wasilow. 2019. „Artificial Intelligence, Robotics, Ethics, and the Military: A Canadian Perspective.“ *AI Magazine* 40 (1): 37-48. https://doi.org/10.1609/aimag.v40i1.2848. (staženo 5. dubna 2023) str. 40

diverzifikovaných informací. Úspěšnost systému se snižovala ve vztahu k pohlaví a barvě pleti – čím tmavší pleť, tím horší úspěšnost. Horší úspěšnost se ukázala i u žen. Černošské ženy byly špatně rozpoznány v až 34 % případů, na rozdíl od bílých mužů, kde byla chybovost systému jen 0,3 %.[27] Podobná míra chybovosti v případě rozeznávání civilistů od vojáků může mít tragické dopady.

Podobné problémy spjaté s typem použitých dat pro strojové učení mohou být rozdíly specifické pro určité oblasti. Často se jedná o detaily, které si na první pohled neuvědomíme. Mohou jimi být kupříkladu průměrná výška vojáků, kteří byli použiti jako reference pro systém. V momentě, kdy systém učíme, kdo je nepřítel a kdo spojenec, a členové spřátelených armád jsou vyššího vzrůstu, může se stát, že systém začne osoby nižšího vzrůstu automaticky vyhodnocovat jako nepřítele. Tyto detaily, které systém zachytí, aniž bychom si to my, lidé, uvědomili, způsobují vznik zkreslených informací, jež nemusí být objektivní a působit negativně na fungování autonomních systémů. Vznik těchto předpojatostí je neúmyslný, v množství dat, které systémy v průběhu strojového učení zpracovávají, se jim však předchází velice obtížně. Je proto důležité, aby rozhodnutí autonomních systémů byla transparentní a aby systémy byly vyvíjeny takovým způsobem, aby byly své kroky schopny vysvětlit, či aby byli srozumitelné jejich operátorům.[28] Tento koncept dále rozvíjí podkapitola týkající se ospravedlnitelnosti.

Dalším problémem spjatým s daty využívanými pro strojové učení. Cílem je, aby data byla eticky získaná. Zde jsou značně znevýhodněny státy, jež chrání data svých obyvatel a kde jsou firmy vázány k ochraně dat jejich uživatelů.[29] Naopak státy, jako je Čína, které mají přístup k obrovskému množství dat svých obyvatel, jsou v tomto ohledu značně zvýhodněny.[30]

Ospravedlnitelnost

Ospravedlnitelnost je vedle předpojatosti další klíčovou etickou otázkou, a to obzvláště v diskuzi týkající se smrtících autonomních zbraní či zbraňových systémů. Z obecného hlediska má každý autonomní systém jiný účel, plní tedy jiné cíle, a tudíž i samotný

[27] Buolamwini, Joy, and Timnit Gebru. 2018. „Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.“ *Conference on Fairness, Accountability, and Transparency*: 1-15. https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf. (staženo 15. dubna 2023) str. 11-12

[28] Rowe, The comparative ethics of artificial-intelligence methods for military applications, str. 4

[29] Thorpe and Wasilow, Artificial Intelligence, Robotics, Ethics, and the Military: A Canadian Perspective, str. 43

[30] Lee, AI Superpowers: China, Silicon Valley, and the New World Order, str. 27

rozhodovací proces daného systému bude odlišný a do určité míry specifický. Abychom byli schopni rozhodnout, zdali jednotlivé kroky a daná rozhodnutí byly či nebyly etické, je potřeba vědět proč a jakým způsobem k danému rozhodnutí došlo. Skrze pochopení, jak systém operuje a jak k rozhodnutím dochází, nabývá celé využití umělé inteligence jistou míru transparentnosti.[31] Pokud tedy jsme schopni vysvětlit, proč se systém rozhodl tak, jak se rozhodl, jsme schopni toto rozhodnutí také právně ospravedlnit. Tím je rozuměno, že každý má právo na vysvětlení rozhodnutí umělé inteligence, k němuž došlo na základě algoritmu. Tento proces je také znám jako *Right to(an) explenation*.[32]

V případě autonomních zbraní a zbraňových systémů je důležité se zaměřit na to, jakým způsobem volí své cíle. Jak již bylo zmíněno, různé autonomní systémy se rozhodují odlišnými způsoby. Jedním z nejsnadněji ospravedlnitelných je *logical reasoning*, tedy logické uvažování. Tento způsob rozhodování je nejsnadněji představitelný, jelikož se podobá tomu, jak se rozhodujeme my, lidé. Systém v tomto případě zhodnotí vstupní data a na jejich základě rozhodne, zdali se jedná o vhodný cíl, či nikoliv.[33] V praxi se tak může jednat např. o rozlišení, zdali se jedná o civilisty či naopak členy nepřátelské armády. V tomto případě bude systém hledat jasná specifika, jež lze připsat civilistům, či naopak vojákům. Jasným ukazatelem, že se jedná o nepřátelské armádní složky může být detekování skupiny osob, které nesou zbraně, pohybují se v oblasti aktivního konfliktu a mají na sobě specifické uniformy, jež implikují, že se jedná o nepřátelské jednotky. Reakce systému bude nadále odpovídat účelu, pro který byl systém zvolen, ať už je to pouhé upozornění, že se ve sledované oblasti objevily nepřátelské jednotky, či přímo jejich eliminace.[34]

Důležité v tomto případě je, jak systém rozhodl o tom, že se jedná o nepřátelskou jednotku, tedy jak byl schopen zvolit svůj cíl. Pokud systém rozhodnutí dosáhl na základě *logical reasoning* je tato odpověď relativně jednoduchá, jelikož specifikem tohoto typu rozhodování je, že se systému můžeme specifickými dotazy zeptat, jak postupoval při

---

[31] Rowe, The comparative ethics of artificial-intelligence methods for military applications, str. 3

[32] Atkinson, Katie, Trevor Bench-Capon, and Bollegala Danushka. 2020. „Explanation in AI and law: Past, present and future." *Artificial Intelligence* 289 (103387). https://doi.org/https://doi.org/10.1016/j.artint.2020.103387. (staženo 8. května 2024), str. 2-4

[33] Rowe, The comparative ethics of artificial-intelligence methods for military applications, str. 3

[34] Margulies, Peter. 2017. „Making autonomous weapons accountable: command responsibility for computer-guided lethal force in armed conflicts." In *Research Handbook on Remote Warfare*, 405-442. Cheltenham UK: Edwar Elgar Publishing. https://www.elgaronline.com/edcollchap/edcoll/9781784716981/9781784716981.00024.xml. (staženo 8. května 2024), str. 409

rozpoznávání nepřítele a lze tak sledovat celý postup volby cíle.[35] Porozumění celému rozhodovacímu procesu je tak v tomto případě poměrně jednoduché.

Dalším typem rozhodování je *numerical calculations* neboli rozhodování na základě matematických výpočtů. Tento způsob rozhodování je mnohem obtížnější vysvětlit, jelikož výpočty, jimiž systém k rozhodnutí došel jsou obvykle natolik komplexní, že je lidský mozek nedokáže adekvátně zpracovat, a to i proto, že systémy provádí několik matematických operací současně.[36] Tato rozhodnutí pak jsou eticky riziková, jelikož jsou v přímém rozporu s *Right to explanation*.[37]

V neposlední řadě je potřeba mít na mysli, že *Right to explanation* může být v přímém rozporu s ochranou dat. Především ve vojenské sféře hrají v rozhodovacím procesu důležitou roli data a informace které se řadí mezi (přísně) utajované. Ne všechna data, jež hrála v rozhodování roli je tak možné odtajnit za účelem vysvětlení určitých rozhodnutí, jelikož by mohla způsobit bezpečnostní riziko či stát znevýhodnit v budoucnosti. Neil C. Rowe ve svém článku nabízí jako jedno z možných řešení tohoto problému využití již předešlých precedent, jež se projednávanému případu podobají.[38] Otázkou však je, zdali takové vysvětlení bude dostatečné.

Zodpovědnost

Otázka zodpovědnosti se zabývá problematikou, jak legislativně ošetřit, kdo za činy autonomních systémů ponese zodpovědnost, a to jak právní, tak morální. Zejména v kontextu otázky zodpovědnosti je integrace lidského prvku do rozhodovacího procesu zásadní.

Zapojení lidského operátora do rozhodovacího procesu vyžaduje například i Montrealská deklarace, dokument zveřejněný v roce 2018, zabývající se etickým, udržitelným a zodpovědným vývojem umělé inteligence. Tato deklarace tvrdí, že veškerou zodpovědnost za činy autonomních systémů nese člověk. Toto tvrzení je založeno na předpokladu, že o finálních krocích umělé inteligence, především pak o krocích, které přímo ovlivní či ukončí lidský život, musí vždy rozhodovat pouze člověk, a to na základě

[35] Atkinson et al., Explanation in AI and law: Past, present and future, str. 6-7
[36] Rowe, The comparative ethics of artificial-intelligence methods for military applications, str. 3
[37] Atkinson et al., Explanation in AI and law: Past, present and future, str. 2
[38] Rowe, The comparative ethics of artificial-intelligence methods for military applications, str. 3-4

informovaného a svobodného rozhodnutí.[39]

Pro uživatele a operátory je klíčová důvěra v systém. Jedním z faktorů, jež udává, jak moc lidé technologii důvěřují je spolehlivost. Pokud systém neodvádí správné výsledky, není možné, aby se na něj operátor spolehl, a tudíž v takovýto systém nemůže chovat důvěru. V případě, že člověk operuje zbraň, jež funguje na principu umělé inteligence, vznikají dva aspekty této důvěry – funkční a etický. Nejenže musí operátor důvěřovat, že systém bude fungovat správně a nedojde k funkční chybě, v kontextu této práce se může jednat kupříkladu o špatně zaměřený cíl či vypálení střely ve špatný čas, ale je důležitá důvěra i z etického aspektu.[40] Operátor si musí být jistý, že systém neprovede rozhodnutí, jež by bylo v rozporu s obecně platným etickým kodexem, v našem případě se může jednat o cílené zaměřování civilistů či nemocnic, což je v rozporu s mezinárodním humanitárním právem.[41]

Situaci komplikuje tvrzení, jež autoři zmiňují v *An Introduction to Ethics in Robotics and AI*, kdy tvrdí, že autonomie je přímo spjatá s morální zodpovědností. Toto tvrzení podporují faktem, že člověk, v momentě, kdy pozbývá autonomii a svobodnou vůli, pozbývá současně i morální zodpovědnost. Jedná se o etablovaný koncept v západní filosofii a sami autoři v textu odkazují na myšlenky Immanuela Kanta.[42] Problém tak nastává v momentě, kdy se tyto principy pokusíme aplikovat na autonomní zbraně, protože stroje a autonomní systémy nemohou v současné situaci nést zodpovědnost za své činy.[43] Proto je pro otázku zodpovědnosti klíčové zapojení člověka, jelikož v momentu, kdy se do rozhodovacího procesu zapojí člověk, dochází automaticky k přenosu zodpovědnosti na tohoto operátora.

Vedle rozhodnutí, kdo ponese zodpovědnost za činy umělé inteligence ve standardních situacích, je potřeba také stanovit, kdo ponese zodpovědnost za činy, jež budou ovlivněny závadou systému – ať už způsobenou nedostatečným testováním či například vlivem kybernetického útoku.[44] Montrealská deklarace v tomto případě tvrdí, že za tato rozhodnutí a chyby nesmí být viněn nikdo z vývojářů systému ani operátor autonomního

---

[39] Abrassart, Christophe et al. „Montréal Declaration for a Responsible Development of Artificial Intelligence 2018," 2018. https://declarationmontreal-iaresponsable.com/wp-content/uploads/2023/04/UdeM_Decl-IA-Resp_LA-Declaration-ENG_WEB_09-07-19.pdf. (staženo 6. července 2023), str. 16

[40] Bartneck, Christoph et al. 2021. „Trust and Fairness in AI Systems." In *An Introduction to Ethics in Robotics and AI*, 27-39. Springer Briefs in Ethics. doi.org/10.1007/978-3-030-51110-4_4. (staženo 12. února 2024), str. 27-28

[41] Melzer and Kuster, International Humanitarian Law A Comprehensive Introduction, str 12-20

[42] Bartneck et al., „Trust and Fairness in AI Systems.", str. 31

[43] Bartneck et al., „Military Uses of AI.", str. 96

[44] Thorpe and Wasilow, Artificial Intelligence, Robotics, Ethics, and the Military: A Canadian Perspective, str. 43

systému, nenabízí již však řešení, kdo za tyto činy ponesee zodpovědnost.[45] V tomto případě se nabízí prostor pro jednotlivé státy, aby tuto problematiku legislativně ošetřily a nabídly řešení.

2.2. Předcházení etickým problémům

Testování autonomních systémů

Bavíme-li se o využití umělé inteligence a autonomních systémů v ozbrojených konfliktech, je taktéž potřeba adresovat, jak stěžejní je důkladné testování těchto systémů. Testování je klíčové nejen aby se ověřilo, že produkty fungují tak, jak mají, ale záro veň i proto, aby byly odbourány určitá rizika. Tato rizika vyplývají primárně ze skutečnosti, že developeři těchto autonomních systémů v naprosté většině případů nemají s reálným vedením konfliktů žádné zkušenosti. Nedostatek znalostí v této oblasti může potencionálně negativně ovlivnit finální produkt, kterým je v našem případě (smrtící) autonomní systém. Jednou z nejčastějších chyb, jež můžeme u systémů sledovat, je využití smrtící síly v momentě, kdy nebyla potřeba.[46]

Komplexita těchto systémů a obrovské množství nejen vstupních dat, ale i potencionálních reakcí na různé situace, však celé testování poněkud komplikuje. Mezi jeden z nejčastějších způsobů testování se řadí tzv. *fuzzing* či *fuzz testing*.[47] Jedná se o autonomní nebo polo-autonomní proces testování využívající software, který systém testuje v náhodných situacích a sleduje, jak se v nich systém zachová. Specifikem tohoto testování je využívání chybných informací a stavění systému do situacích, do kterých by se správně neměl ani dostat. Cílem tohoto testování je odhalit možné chyby systému. Díky specifickému využívání matoucích a nesprávných informací je možné odhalit velké množství chyb, které by nemusely být při jiných typech testování[48] odhaleny.[49] Testování může dále taktéž nabídnout hlubší porozumění systémů, což může mít vliv na vysvětlení, a tedy i ospravedlnitelnost, jeho činů.

---

[45] Abrassart, Christophe et al. „Montréal Declaration for a Responsible Development of Artificial Intelligence 2018," str. 16

[46] Rowe, The comparative ethics of artificial-intelligence methods for military applications, str. 3

[47] Ibid

[48] Mezi další typy testování se řadí např. Modified Condition Decision Coverage, Penetration, Scenario-based Testing či Black box a White box testing. Více informací viz např. Testing the Input Timing Robustness of Real-time Control Software for Autonomous Systems, David Powell et al., 2012

[49] Azar, Kimia Zamiri et. al. 2022. „Fuzz, Penetration, and AI Testing for SoC Security Verification: Challenges and Solutions." *Future microelectronics security research series*, no. 394. https://eprint.iacr.org/2022/394.pdf (staženo 7. května 2024)

Etické autonomní systémy

Řešení, jež by mohlo zajistit etičtější rozhodování autonomních systémů je implementace etického kodexu, jež by vycházel z mezinárodního humanitárního práva, přímo do návrhu těchto zbraňových systémů. Etické aspekty těchto zbraňových systémů by měli být zohledněny developery již během jejich vývoje.[50]

Jeden z konceptů, které mají potenciál zajistit etičtější autonomní zbraně, je naučit tyto autonomní zbraňové systémy rozeznávat specifické symboly. Díky pokroku v oblasti strojového učení můžeme tyto zbraňové systémy navrhnout tak, aby byly schopny své okolí vizuálně analyzovat a naučit je význam některých důležitých symbolů. Jedním z těchto symbolů může být červený kříž, mezinárodně uznávaný symbol značící zdravotnická zařízení. Systém můžeme naučit, že pokud rozezná tento symbol v lokalitě, kde má proběhnout aktivní útok, plnění své mise přeruší, aby neporušil mezinárodní humanitární právo. Takto lze systém naučit škálu chráněných symbolů či charakteristik. Mohou mezi ně patřit děti, či bílá vlajka. Autonomní zbraně je zároveň možné navrhnout takovým způsobem, že mohou odmítnout splnit jim zadaný a autorizovaný úkol (útok na zvolený cíl), pokud v lokalitě rozeznají jakýkoliv z předem definovaných chráněných objektů.[51]

Systémy tedy je možné navrhnout způsobem, aby jejich činy byly v souladu s mezinárodně platným humanitárním právem či jinými etickými normami.

[50] Galliott, Toward a Positive Statement of Ethical Principles for Military AI, str. 128
[51] Galliott, Toward a Positive Statement of Ethical Principles for Military AI, str. 130-131

3. Analýza americké strategie implementace etických autonomních zbraní

Praktická část této práce se zabývá analýzou americké strategie implementace umělé inteligence do vojenských misí v podobě autonomních a polo-autonomních smrtících zbraní. Konkrétně se soustředí na tři, dříve zmíněné, problematické oblasti – ospravedlnitelnost, předpojatost a zodpovědnost.

Spojené státy americké byly zvoleny jako předmět této analýzy z několika důvodů. Armáda Spojených států amerických se řadí mezi největší armády světa. Spojené státy byly prvním, kdo zveřejnil etické principy pro využití umělé inteligence v armádní sféře.[52] Současně se očekává, že Spojené státy budou v budoucnu dominovat v množství bezpilotních letounů určených pro armádní využití.[53] Americké ministerstvo obrany zůstává ve vedoucí pozici v oblasti etické problematiky využití umělé inteligence v obraně, práce proto zkoumá, do jaké míry americká strategie na implementaci umělé inteligence zohledňuje možná etická rizika. [54] I přesto, že tyto dosavadní úspěchy i lichotivé statistiky jsou motivující, mají potenciál zkreslovat stav reálné situace. V současnosti Spojené státy nevyužívají smrtící autonomní zbraně ani přímo neplánují jejich vývoj.[55] Jedná se tedy o teoretické úvahy, jejichž cílem není implikovat, že takto budou Spojené státy k reálné implementaci tohoto typu zbraní přistupovat.

V rámci této analýzy je zkoumána řada dokumentů různého typu. Vybraná škála zkoumaných materiálů byla zvolena s cílem zachytit postupně se měnící narativ vládních orgánů. Dalším aspektem, proč byla zvolena tato škála vzorků, je snaha pokrýt různé perspektivy a fáze implementace. Zkoumané materiály jsou právního, strategického, informativního a analytického charakteru. Z právních předpisů byla zvolena směrnice Ministerstva obrany, přesněji směrnice 3000.09 Autonomie zbraňových systémů (*Directive*

---

[52] Stanley-Lockman, Zoe. 2021. *Responsible and Ethical Military AI: Allies and Allied Perspectives*. Center for Security and Emerging Technology. https://cset.georgetown.edu/wp-content/uploads/CSET-Responsible-and-Ethical-Military-AI.pdf. (staženo 10. července 2023) str. 2

[53] Araya, Daniel, and Meg King. 2022. „The Impact of Artificial Intelligence on Military Defense and Security.“ *CIGI Papers* March 2022 (263). https://www.econstor.eu/bitstream/10419/299735/1/cigi-paper263.pdf. Odkazuje na McCarthy, Niall. „The Countries Set To Dominate Drone Warfare.“ Statista. November 19, 2019. https://www.statista.com/chart/20005/total-forecast-purchases-of-weaponized-military-drones/ (staženo 27. června 2023), str. 10

[54] Stanley-Lockman, Responsible and Ethical Military AI: Allies and Allied Perspectives. str. 2-14

[55] Congressional Research Service. *Defense Primer: U.S. Policy on Lethal Autonomous Weapon Systems*. February 1, 2024. https://crsreports.congress.gov/product/pdf/IF/IF11150. (staženo 15. února 2024), str. 1

3000.09 Autonomy in weapon systems). Dokument strategického charakteru, jenž práce analyzuje je Strategický a implementační plán pro odpovědné využívání umělé inteligence Ministerstva obrany Spojených států (*U.S. Department of Defense Responsible Artificial Intelligence Strategy and Implementation Pathway*). Informativním dokumentem je Přehled obranné politiky: Americká politika týkající se smrtících autonomních zbraní (*Defense Primer: U.S. Policy on Lethal Autonomous Weapon Systems)* zveřejněné Výzkumnou službou Kongresu (*Congressional Research Service*). Posledním zkoumaným dokumentem, tedy materiálem analytického charakteru, je závěrečná zpráva Národní bezpečnostní komise pro umělou inteligenci. Vzhledem k rozsahu této publikace byly zvoleny pouze dvě kapitoly, jež jsou pro tuto práci relevantní, a to kapitola třetí, Umělá inteligence ve válečných konfliktech (*AI in Warfare*), a kapitola čtvrtá, Autonomní zbraňové systémy a rizika spjatá s využitím umělé inteligence ve válečných konfliktech (*Autonomous Weapon Systems and Risks Associated with AI-Enabled Warfare*). Vzhledem k již dříve zmíněné proměnlivé charakteristice zkoumaného tématu byl kladen důraz na to, aby zkoumané dokumenty byly, pokud možno, co nejaktuálnější. Nejstarším zkoumaným materiálem je závěrečná zpráva Národní bezpečnostní komise pro umělou inteligenci, která byla zveřejněna v březnu 2021, naopak nejaktuálnějším materiálem je Přehled obranné politiky Výzkumné služby Kongresu, jež byl zveřejněn v únoru 2024.

Zvolené dokumenty, především pak ty právního, strategického a informativního charakteru, na sebe často vzájemně odkazují a tvoří tak komplexní síť nařízení a pravidel. Nejedná se však o jedinou síť stanovující meze a pravidla pro využití a implementaci umělé inteligence. Zkoumané dokumenty často odkazují na další směrnice, jež však nejsou pro tuto práci relevantní, a proto nebudou zmíněny. Zároveň často odkazují na tzv. *Law of War*, tedy na mezinárodní humanitární právo, mezinárodně uznávaná pravidla vedení ozbrojeného konfliktu a další nadřazené dokumenty, mezi nimi i Ústavu Spojených států amerických.[56]

Jednotlivé analýzy zkoumaných materiálů jsou řazeny v chronologickém pořadí jejich zveřejnění. Na úvod je představen dodatečný dokument, Etické principy využití umělé inteligence ministerstva obrany (*Department of Defense AI Ethical Principles*). Tento dokument nebude kriticky analyzován, jako zbylé dokumenty, v něm představené principy

[56] U.S. Department of Defense. *Responsible Artificial Intelligence Strategy and Implementation Pathway*. Washington, D.C.: U.S. Department of Defense, June 2022, https://media.defense.gov/2022/Jun/22/2003022604/-1/-1/0/Department-of-Defense-Responsible-Artificial-Intelligence-Strategy-and-Implementation-Pathway.PDF (staženo 7. dubna 2024), str. 4

však budou použity v závěrečné diskuzi, kde podpoří závěry vyvozené ze zkoumaných dokumentů.

3.1. DoD AI Ethical Principles

Etické principy využití umělé inteligence ministerstva obrany byly zveřejněny v únoru 2020. Tyto principy se vztahují na využití umělé inteligence na bojišti i mimo něj. Ministerstvo obrany vytyčilo pět principů, jež by měla umělá inteligence (a systémy fungujících díky jejímu využití) adoptovat. Umělá inteligence by měla být zodpovědná, spravedlivá, sledovatelná, spolehlivá a spravovatelná. Tento dokument je stěžejní především protože stanovuje jakousi základní linii, od níž se odvíjí zbylé dokumenty.[57]

Princip zodpovědnosti amerického ministerstva obrany tvrdí, že zaměstnanci musí při využívat umělé inteligence či autonomních systémů být vůči těmto technologiím dostatečně kritičtí a chování těchto technologií hodnotit na základě jejich (lidského) úsudku. Zaměstnanci ministerstva taktéž nesou plnou zodpovědnost za vývoj, využití a nasazení těchto technologií. Princip spravedlnosti zaručuje, že ministerstvo podnikne všechny dostupné kroky proto, aby zajistilo, že rozhodnutí umělé inteligence ani autonomních systémů nebudou nijak zkreslené ani ovlivněné předpojatostí. Princip sledovatelnosti říká, že tyto technologie musí být vyvíjeny tak, aby relevantní zaměstnanci rozuměli vývoji, procesům a operacím spjatých s využitím těchto technologií. To současně znamená, že tito zaměstnanci budou schopni vysvětlit, proč se systémy chovají tak, jak se chovají a co vedlo k jejich jednotlivým rozhodnutím. Princip spolehlivosti zavazuje ministerstvo k vytvoření jasně definovaných procesů pro využití těchto nových technologií. Tyto technologie musí současně projít testování, aby se zajistila bezpečnost jejich využití a že jsou skutečně schopny vykonávat jim předem stanovené úkoly. Poslední z principů, tedy princip spravovatelnosti či ovladatenlosti, říká, že ministerstvo bude navrhovat technologie využívající umělou inteligenci tak, aby operátor či jiná relevantní osoba, byli schopni odhalit chybné jednání systémů a systém zastavit, než dojde k nežádoucím výsledkům.[58]

[57] U. S. Department of Defense. „*Implementing Responsible Artificial Intelligence in the Department of Defense*.“ Memorandum for Senior Pentagon Leadership, Commanders of the Combatant Commands, Defense Agency, and DoD Field Activity Directors. Washington, DC: Deputy Secretary of Defense, 2021. https://media.defense.gov/2021/May/27/2002730593/-1/-1/0/IMPLEMENTING-RESPONSIBLE-ARTIFICIAL-INTELLIGENCE-IN-THE-DEPARTMENT-OF-DEFENSE.PDF. (staženo 7. dubna 2024), str. 1

[58] Ibid

3.2. The Final Report

Závěrečná zpráva Národní bezpečnostní komise pro umělou inteligenci je analytický dokument zveřejněný v roce 2021. Úkolem Národní bezpečnostní komise pro umělou inteligenci je dávat doporučení prezidentovi Spojených států amerických a americkému Kongresu v otázkách týkajících se umělé inteligence, strojového učení a technologií ve vztahu k národní bezpečnosti a obraně. Cílem tohoto rozsáhlého dokumentu čítajícím 16 kapitol je představit jasnou strategii implementace umělé inteligence. Závěrečná zpráva zároveň obsahuje doporučení a jasný plán kroků, jež by vláda Spojených států amerických měla adoptovat.[59] Pro účely této práce byly zvoleny dvě kapitoly této publikace, které jsou pro téma nejvíce relevantní, přesněji se jedná o kapitoly Umělá inteligence ve válečných konfliktech (*AI and* Warfare) a Autonomní zbraňové systémy a rizika spjatá s využitím umělé inteligence ve válečných konfliktech (*Autonomous Weapon Systems and Risks Associated with AI-Enabled Warfare).* Kapitoly jsou koncipovány stejným způsobem – nejprve představují diskutovaný koncept relevantní pro danou kapitolu a následně nabízí doporučení určená ministerstvu obrany.

AI and Warfare

Kapitola se, jak již název napovídá, věnuje využitím umělé inteligence v ozbrojených konfliktech. Zabývá se spíše obecnými plány na implementaci a na kroky, jež je potřeba podstoupit nežli konkrétně na určitý typ technologie. Koncept, který kapitol představuje je plán nesoucí název *AI Ready by 2025*. Cílem tohoto plánu je zajistit, že vojenský personál a celý systém bude připravený na intenzivnější implementaci umělé inteligence napříč sektorem, tedy nejen přímo na bojišti ale například i ve finanční frakci. V rámci *AI Ready by 2025* by měl vojenský personál dosáhnout dostatečné úrovně digitální gramotnosti, jež je klíčová pro úspěšnou integraci umělé inteligence do výcviku a operací. Kapitola klade velký důraz na shora řízenou inovaci, díky níž se zajistí že implementace nových technologií bude úspěšná a vojenské složky budou moci naplno využít jejich potenciál.[60]

Kapitola nepřináší informace, jež by výrazně ovlivnily etické aspekty jakýchkoliv legislativ či budoucích strategii implementace autonomních zbraní, co je však důležité je, že zmiňuje rizika. Národní bezpečnostní komise pro umělou inteligenci si uvědomuje, že komplexní integrace umělé inteligence a autonomních systémů je spjatá s určitými riziky a

[59] National Security Commission on Artificial Intelligence. 2021. „The Final Report." https://reports.nscai.gov/final-report/. (citováno 30. května 2024)
[60] National Security Commission on Artificial Intelligence, The Final Report, str. 76-81

vždy je potřeba situaci vyhodnotit a zvážit přínosy i potenciální negativa.[61] Přínosy využití těchto moderních technologií jsou nezpochybnitelné – od informovanějších a objektivnějších rozhodnutí, až po přesnější výběr cílů a načasování útoků. Využití tohoto typu technologií je však spjato i s negativy – rychlejší a agresivnější útoky, které nenechávají prostor pro mediaci, kybernetické útoky zaměřené na tyto typy zbraní či způsobení ztráty strategické převahy v konfliktu v důsledku selhání systémů.[62] Národní bezpečnostní komise pro umělou inteligenci však dochází k závěru, že jsou situace, kdy je potřeba volit riskantnější možnosti. Nadále podporuje ministerstvo obrany v pokračování adaptace těchto technologií, pod podmínkou, že budou zohledňovat s nimi spjatá rizika.

Kapitola se také věnuje roli nových technologií a člověka. Dle názoru Národní bezpečnostní komise pro umělou inteligenci by ve vojenských akcích měla umělá inteligence a autonomní systémy pouze doplňovat roli člověka, nikoliv ji plně nahradit. To však neznamená, že velitelé nebudou moci delegovat určité úkoly na autonomní systémy. Oblastí, kde může umělá inteligence a různé typy autonomních systému asistovat člověku, je zpracování dat. Moderní technologie mohou sbírat, zpracovávat a analyzovat data a následně distribuovat získané informace. Tato data mohou být dále využívána pro trénování a zdokonalování autonomních zbraňových systémů.[63] Díky využití autonomních systémů a umělé inteligence ke sběru a zpracování dat lze omezit předpojatost. Speciální senzory využívající umělou inteligenci mají možnost sesbírat větší množství kvalitních dat a v kombinaci s dalšími autonomními technologiemi získávat data i z míst, kam se běžně člověk nedostane. Systémy se díky tomu mohou učit na mnohem variabilnější škále dat.

Mimo sběru dat se předpokládá, že budou umělá inteligence a autonomní systémy využívány i přímo během bojových misí. Kapitola však nenaznačuje, že očekává využití plně autonomních zbraňových signálů a jsou zmíněny pouze pozice, kde by technologie asistovaly člověku. Tyto funkce zahrnují pomoc při koordinaci přesunů či zaměřování cílů. Využití autonomních a polo-autonomních systémů zajistí přesnější, lépe načasované, a tudíž efektivnější útoky, které mají potenciál zajistit strategickou převahu.[64]

Kapitola taktéž nabízí řadu konstruktivních doporučení pro ministerstvo obrany, jež usnadní proces implementace nových technologií. V rámci podpory shora řízeného vývoje

[61] Ibid, 79
[62] Ibid str. 91
[63] Ibid, str. 81
[64] Ibid, str. 81

navrhla Národní bezpečnostní komise pro umělou inteligenci vytvoření pozice, jež je v dokumentu označena jako Zástupce pro operativní nasazení umělé inteligence (*AI Operational Advocate*). Tento člověk by měl být odborníkem na autonomní systémy využívající umělou inteligenci a měl by radit velitelům a dalším zaměstnancům v otázkách týkajících se schopností a limitů daných autonomních systémů a současně dohlížet, že jsou systémy využívány v souladu s platnými pravidly a mezinárodním humanitárním právem.[65] Nebyly nalezeny žádné informace o vzniku této pozice, do jisté míry však tuto pozici zastává Ředitelka pro digitální technologie a umělou inteligenci (*Chief Digital and Artificial Intelligence Officer*).

V rámci doporučení Národní bezpečnostní komise pro umělou inteligenci také navrhla několik sfér, na něž by se ministerstvo obrany mělo zaměřit včetně krátkodobých a dlouhodobých cílů výzkumu. Mezi tyto cíle se řadí kupříkladu vývoj odolných adaptivních autonomních systémů, které se učí na základě dat, které samy sbírají. Současně by tyto systémy měly být schopny samostatně plnit komplexní dlouhodobé úkoly, na kterých mohou spolupracovat s dalšími autonomními systémy. Tyto aspirace jsou samozřejmě podpořeny plánem na zhotovení procesu na testování plně autonomních systémů využívajících umělou inteligenci.[66]

Autonomous Weapon Systems and Risks Associated with AI-Enabled Warfare

Hlavním tématem této kapitoly závěrečné zprávy Národní bezpečnostní komise pro umělou inteligenci je, zdali jsou autonomní zbraňové systémy zákonné, bezpečné a etické. Využití umělé inteligence a autonomních zbraních fungujících na principu umělé inteligence umožní velitelům činit rychlejší a informovanější rozhodnutí. Tyto technologie mají zároveň potenciál dosáhnout rychlosti, výkonu a rozeznávání cílů, jež výrazně předčí schopnosti člověka. Využití těchto technologií by však mělo být podmíněno detailním testováním, aby bylo zajištěno, že jsou tyto technologie navrženy a používány v souladu s pravidly mezinárodního humanitárního práva. Kapitola také zmiňuje velice důležitou informací, a sice že s využitím tohoto typu technologií se pojí riziko eskalace konfliktů. Využití umělé inteligence konflikty zrychluje, a tudíž zkracuje čas, kdy je možné konflikt de-eskalovat a dojít k řešení nenásilným způsobem. Současně také dochází k jistému distancování od

[65] Ibid, str. 84
[66] Ibid, str. 85

aktivní zóny konfliktu, což vede k vytvoření iluze bezpečí, což může agresory dále motivovat k agresivnějším a riskantnějším útokům.[67] Těmto eskalacím se však Spojené státy snaží vyvarovat.

V souvislosti s diskutovanou tématikou Národní bezpečnostní komise pro umělou inteligenci konzultovala členy občanské společnosti, akademické organizace i vládní agentury, aby zjistila, jaký je jejich názor na využití autonomních zbraňových systémů, včetně, s nimi spjatými, etickými riziky. Výsledkem těchto diskuzí vznikly čtyři závěry, k nimž byla formulována doporučení pro zákonodárce. Oblasti, kterých se tyto závěry týkají jsou zapojení člověka do rozhodovacího procesu, vývoj a využití zbraní v souladu s mezinárodním humanitárním právem, přístup konkurence k vývoji autonomních zbraňových systémů a vyjádření k snaze globálně zakázat vývoj a využití těchto technologií.

Nejdůležitějším závěrem je první[68], který říká, že pokud jsou činy autonomních systémů schváleny člověkem, mohou být využívány v souladu s normami mezinárodního humanitárního práva. Důležitým aspektem mezinárodního humanitárního práva je zde zásada přiměřenosti, která zakazuje útoky, jež by způsobily nepřiměřené množství ztrát na životech civilistů.[69] Autonomní systémy využívající umělou inteligenci, pokud by byla dodržena podmínka dostatečného testování, mohou být přesnější a výrazně snížit nevyžádané zásahy civilistů. Jsou však situace, kdy se obětem na straně civilistů nelze vyhnout. Autonomní systémy by v sobě měly mít zabudovaný etický kodex, jež se shoduje s mezinárodním humanitárním právem, i přesto však morální zodpovědnost za rozhodnutí, kdy se musí zvážit váha získaného strategického přínosu ve vztahu k potenciální újmě civilistů, vždy nese člověk.[70]

Tato kapitola taktéž výslovně říká, že zodpovědnost za vývoj, použití a chování autonomních systému by měl nést člověk. Tak jako Montrealská deklarace, i závěrečná zpráva Národní bezpečnostní komise pro umělou inteligenci tvrdí, že v rozhodnutí o použití smrtící síly musí být vždy zapojen člověk. Zároveň však Národní bezpečnostní komise pro

[67] Ibid, str. 91-97

[68] Rozbor zbylých závěrů byl vyhodnocen pro práci nepřínosný, pro vice informací k nim viz The Final Report str. 93-96 https://assets.foleon.com/eu-central-1/de-uploads-7e3kk3/48187/nscai_full_report_digital.04d6b124173c.pdf

[69] National Security Commission on Artificial Intelligence, The Final Report, str. 92, odkazuje na International Committee of the Red Cross, „Proportionality“, https://casebook.icrc.org/a_to_z/glossary/proportionality. (původní použitá citace z 15. ledna 2021)

[70] National Security Commission on Artificial Intelligence, The Final Report, str. 92

umělou inteligenci zohledňuje, že situace, kdy dojde k využití autonomních zbraňových systémů se mohou značně lišit a míra zapojení člověka, nutná pro dodržení principu zodpovědnosti, je vždy individuální. To znamená, že ve stabilním prostředí, kde je nízké riziko výskytu civilistů bude vyžadovaný dohled operátora nižší nežli v nestálých situacích, kde se mohou vyskytovat civilisté. V případě bojů v odlehlých oblastech může stačit, že operátor jen zadá a schválí úkol a autonomní systém ho vykoná sám bez další asistence či zásahu operátora. Naopak při konfliktu v městských oblastech s vysokým množství výskytu civilistů, je potřeba, aby operátor periodicky kontroloval a schvaloval kroky autonomního systému, aby došlo k vyvarování se nechtěným útokům na civilisty. Tyto principy by měly být přímo zakomponovaný do návrhů jednotlivých autonomních systémů. Obecné pravidlo, které by podmiňovalo každé využití smrtelné síly autorizovat operátorem, však hodnotí jako kontraproduktivní, i kvůli zmíněnému individuálnímu charakteru situací, v nichž budou tyto systémy nasazovány.[71]

3.3. U.S. Department of Defense Responsible Artificial Intelligence Strategy and Implementation Pathway

Strategický a implementační plán pro odpovědné využívání umělé inteligence ministerstva obrany Spojených států byl zveřejněn v červnu 2022. Je součástí dlouhodobého plánu na implementaci etické umělé inteligence ministerstva obrany a navazuje na memorandum zveřejněné v květnu 2021. Toto memorandum nastínilo obsah samotného Strategického a implementačního plánu pro odpovědné využívání umělé inteligence a připomnělo Etické principy využití umělé inteligence ministerstva obrany, jež byly představeny v úvodu praktické části.

Ministerstvo obrany klade velký důraz na důvěru. Tato důvěra má více aspektů – veřejnost, spojenci, ale i samotné ministerstvo. Ministerstvo obrany doufá, že díky představení tohoto strategického plánu a ujištění, že implementace umělé inteligence a autonomních technologií bude v souladu s platnými etickými normami, přesvědčí veřejnost, že využití těchto technologií zajistí Spojeným státům do budoucna strategickou převahu vůči konkurencí a že jsou tyto technologie bezpečné. Vedle veřejnosti tímto krokem vysílá ministerstvo jasnou zprávu i svým spojencům – Spojené státy jsou připraveny na inovaci, která bude etická a v souladu s mezinárodním humanitárním právem a od svých partnerů očekává stejný přístup. Současně však doufá v získání důvěry ze strany vlastních

[71] Ibid, str. 92-93

zaměstnanců. Tato důvěra je z mnoha důvodů důležitější nežli důvěra veřejnosti a spojenců. Pokud ministerstvo doufá v úspěšnou integraci nových technologií, je potřeba, aby v ni lidé, kteří tuto technologii budou využívat, chovali důvěru. Jak již bylo vysvětleno v teoretické část práce, pokud lidé nebudou těmto technologiím důvěřovat, je nižší pravděpodobnost, že je reálně budou využívat. Ministerstvo obrany doufá, že jasná a transparentní komunikace plánů na etickou a odpovědnou umělou inteligenci jim zajistí důvěru všech dříve zmíněných stran.[72]

Ministerstvo v tomto strategickém plánu taktéž stanovuje řadu cílů, k nimž postupně aspiruje a jejichž dokončení je kritické pro úspěšnou implementaci autonomních technologií. Mezi ty, pro tuto práci relevantní, se řadí obezřetný postup při vývoji a adaptaci nových technologií, nové, nejen zbraňové, systémy využívající umělou inteligenci by měly být kriticky hodnoceny od začátku, aby se zamezilo využití či vývoji chybně navržených technologií. Autonomní technologie by měly být využívány, pokud to daná mise umožňuje a vždy by měla být zohledněna rizika spjatá s využitím těchto technologií. Posledním z relevantních cílů je zajištění, že všichni zaměstnanci, kteří přichází do kontaktu s těmito technologiemi jim rozumí, ví, jak fungují, a jsou srozuměny s jejich limity a riziky, jež využití těchto technologií přináší.

Dosažení zodpovědné umělé inteligence vyžaduje návrh, vývoj a využití umělé inteligence a autonomních systémů způsobem, který zajišťuje bezpečnost a dodržování platných etických norem. Aby byly dodrženy tyto podmínky, je úkolem ministerstva zajistit, že tyto technologie budou dostatečně testovány, aby byla zajištěna bezpečná implementace. Zároveň je kladen důraz na přímé zapojení člověka do rozhodovacích procesů. Systémy je důležité pravidelně kontrolovat, aby byl zachován princip zodpovědnosti a bylo zajištěno, že systém nevykazuje známky závady. Strategie zdůrazňuje, že činy autonomních systémů musí být vysvětlitelné.[73] Jak již bylo zmíněno v teoretické části práce, pokud jsme schopni činy autonomních systémů vysvětlit, jsme je zároveň schopni ospravedlnit.

I tato strategie je rozdělena do dílčích oblastí zájmu, jež byly zveřejněny již v memorandu. Těmito oblastmi jsou odpovědná správa umělé inteligence, důvěra vojáků, akviziční a životní cyklus umělé inteligence a autonomních systémů, schvalování požadavků, odpovědná globální síť využití umělé inteligence a pracovní tým pro využití umělé inteligence. Každá z těchto oblastí má své specifické cíle, v jejichž dosažení

[72] DoD. Responsible Artificial Intelligence Strategy and Implementation Pathway. str. 1-7
[73] Ibid str. 6

ministerstvo doufá.[74] Ne všechny tyto oblasti jsou pro práci relevantní, proto se rozbor bude týkat jen některých z nich.

V rámci odpovědné správy umělé inteligence ministerstvo doufá v modernizaci struktury a procesů dohledu na využívání technologií. Díky této modernizaci by bylo možné neustále sledovat autonomní systémy a kontrolovat, zdali dodržují Etické principy využití umělé inteligence ministerstva obrany. To by pomohlo v odhalení rizikových situací spjatých s využitím autonomních systémů, kterým by se v budoucnu dalo vyvarovat, pokud budou tato nově odhalená rizika zohledněna ve vývoji nových autonomních systémů. Je proto v zájmu ministerstva, aby jakékoliv obavy či problémy spjaté s implementací autonomních systémů, ať už ze strany vývojářů či uživatelů, byly nahlášeny a vyřešeny, než dojde k plnému nasazení těchto systémů.[75]

Důvěra vojáků se zaměřuje na vzdělávání osob, které budou s autonomními systémy aktivně pracovat. Cílem je, aby tito lidé dosáhly dostatečné úrovně vzdělání, díky které budou moci objektivně hodnotit rozhodnutí autonomních systémů a budou plně srozuměni s limity a schopnostmi těchto systémů. Pokud budou operátoři srozuměni s tím, jak systémy fungují, je pravděpodobnější, že jim budou důvěřovat a budou si jistější při jejich nasazení. Vzdělání operátorů autonomních systémů by mělo probíhat adaptivně a být kontinuální, aby se zajistilo, že jsou operátoři srozuměni s nejaktuálnějšími informacemi a vývojem. Tato důvěra je podpořena detailním testováním autonomních systémů a jejich neustálým monitorováním, aby se podchytilo jakékoliv nechtěné chování či selhání v průběhu aktivního nasazení autonomních zbraňových systémů ve válečných misí.[76]

V rámci akvizičního a životního cyklus umělé inteligence a autonomních systémů se strategický plán zaměřuje na kontrolu návrhů i již aktivně využívaných technologií. Snahou je zamezit aktivnímu využití jakkoliv závadných technologií. Je proto důležité důsledně kontrolovat systém nejen ve fázi vývoje a testování, ale i po aktivním nasazení, jelikož chyby se mohou objevit, i přesto, že byl systém testován. Je totiž možné, že byly autonomní zbraňové systémy poškozeny v průběhu nasazení, či byly obětí kybernetického útoku. Neustálá kontrola je proto kritická, aby nedošlo k porušení platných etických kodexů či mezinárodního humanitárního práva.[77]

[74] Ibid, str. 9-11
[75] Ibid, str. 9-21
[76] Ibid, str. 9-24
[77] Ibid, str. 10-27

Schvalováním požadavků autonomních systémů před tím, než systém tyto akce provede povede k omezení množství chybných výsledků těchto technologií. Zároveň, pokud jsou jednotlivé akce systémů schvalovány, dochází k nezpochybnitelnému přenesení zodpovědnosti na operátora tohoto systému. Současně, díky rozdělení plnění cíle na individuálně schvalované kroky dosahujeme vyšší transparentnosti fungování systému a jeho jednotlivé kroky lze snadněji vysvětlit.[78] Tento postup by značně snížil možné komplikace spjaté s ospravedlnitelností a zodpovědností.

Implementace zodpovědné a etické umělé inteligence a autonomních systémů je komplikovaný proces, jež vyžaduje dostatek času. Ministerstvo proto implementaci rozděluje do čtyř fázi, které jsou dále děleny na dílčí kroky. Těmito fázemi jsou návrh, vývoj, nasazení a použití. Ministerstvo zohledňuje, že adaptace různých typů autonomních systémů vyžaduje různé postupy, které budou trvat různě dlouho. Plošná pravidla, jak postupovat při adaptaci, jsou hodnocena jako kontraproduktivní a riziková. Každý systém proto prochází tímto procesem individuálně, aby byly zohledněny veškerá rizika, technické aspekty a splněny podmínky stěžejní pro úspěšnou adaptaci.[79]

3.4. Department of Defense Directive 3000.09 Autonomy in Weapon Systems

Směrnice 3000.09 amerického ministerstva obrany o autonomii zbraňových systémů (*DoD Directive 3000.09 Autonomy in Weapon Systems*) byla schválena 25. ledna 2023 a nahradila tak předešlou verzi z roku 2012. Vztahuje se na návrh, vývoj, akvizici, testování a využití autonomních a polo-autonomních zbraňových systémů. Zároveň se vztahuje i na využití smrtící síly.[80] Směrnice má 3 cíle – prvním je stanovit zásady a přiřadit zodpovědnost za vývoj a využití autonomních a polo-autonomních zbraňových systémů. Druhým cílem je vytvořit pokyny určené k minimalizaci pravděpodobnosti selhání a minimalizaci možných následků selhání autonomních a polo-autonomních zbraňových systémů. Posledním z cílů je vytvoření pracovní skupiny zabývající se problematikou autonomních zbraňových systémů.[81] Z těchto cílů jsou pro tuto práci nejvíce relevantní první dva.

Úvodní část této směrnice se zabývá zásadami využití autonomních zbraňových

[78] Ibid, str. 10-28
[79] Ibid, str. 14-17
[80] DoD Directive 3000.09 Autonomy in Weapon Systems. str. 3
[81] Ibid, str.1

systémů. První pravidlo 1.2 a.[82] spjaté s využitím těchto technologií se týká jejich návrhu. Autonomní a polo-autonomní zbraňové systémy musí být navrženy tak, aby umožnily jejich operátorům zhodnotit využitou sílu. To v praxi znamená, že úkolem operátora je vyhodnotit, jestli intenzita a typ plánované použité síly bude adekvátní vůči situaci, v níž se obě strany nachází, jestli je v souladu s platnými normami a dodržuje mezinárodní humanitární právo. Tito kvalifikovaní operátoři tak mají učinit na základě jejich (lidského) úsudku.[83] Otázkou je, zdali toto pravidlo implikuje, že zodpovědnost za činy těchto autonomních zbraní ponese daný operátor, protože je to právě on, kdo vyhodnocuje míru využití síly a má možnost zasáhnout.

Tuto myšlenku podporuje i druhý bod 1.2 b., který říká, že osoby, které autorizují či řídí použití nebo operují autonomní, či polo-autonomní, zbraňový systém, tak musí činit s náležitou opatrností a v souladu s mezinárodním humanitárním právem, platnými dohodami, bezpečnostními zásadami a pravidly pro použití síly. Dalším bodem, který toto tvrzení podporuje je bod 1.2 a. (3), který specifikuje, že operátor činí informovaná a adekvátní rozhodnutí v případě zasahování cílů.[84] Pozdější ze zmíněných bodů je v přímém souladu se zněním Montrealské deklarace. Lze tedy usoudit, že podle doporučení Montrealské deklarace ponese za činy autonomních systému zodpovědnost právě ta osoba, jež jeho použití autorizuje, řídí či ho operuje.

Tak jako Montrealská deklarace, ani tato směrnice však nestanovuje, kdo nese zodpovědnost za činy autonomních a polo-autonomních zbraňových systému v případě jeho selhání či jinak chybného jednání. Je však nutno podotknout, že směrnice klade značný důraz na testování autonomních systémů ve snaze předejít těmto momentům. Dle bodu 1.2 a. (1), systémy projdou detailním testováním, a to jak hardwaru, tak softwaru. V průběhu tohoto testování se ověřuje, že systém bude jednat tak, jak se očekává. Testování probíhá v realistickém operačním prostředí, kde bude systémy nasazeny proti realistickým protivníkům, kteří budou reagovat a přizpůsobovat se zvolenému typu útoku, aby se byly autonomní systémy nuceny reaktivně rozhodovat a neustále vyhodnocovat danou situaci. Zároveň je potřeba ověřit, že je systém schopen dokončit předem zadané úkoly ve

[82] Jedná se o typ značení jednotlivých bodů užívaný v diskutované směrnici. Číslo značí část kapitoly, následuje písmeno pro jednotlivé body, v případě nutnosti dalšího dělení jednotlivých bodů následuje opět číslice, tentokrát v závorce, o úroveň níže písmeno v závorce

[83] DoD Directive 3000.09 Autonomy in Weapon Systems, str. 3

[84] Ibid, str. 4

vymezeném čase a zadané geografické pozici. Důležitým dodatkem tohoto bodu 1.2 a. (1) (b) je, že pokud systém nebude schopen zadaný úkol splnit v souladu s těmito podmínkami, ukončí svou aktivitu, pokud nezíská dodatečný souhlas operátora k pokračování plnění úkolu. Jedná se o funkci, jež doporučovala zakomponovat Národní bezpečnostní komise pro umělou inteligenci ve své závěrečné zprávě.[85] Toto rozhodnutí dále podporuje obecně přítomný preventivní přístup, kdy se ministerstvo obrany snaží snížit riziko selhání autonomních zbraňových systému na naprosté možné minimum.

Tuto domněnku dále utvrzuje bod 1.2 a. (1) (c), podle nějž se v průběhu testování také ověřuje, že jsou zbraňové systémy dostatečně odolné vůči vnějším vlivům, aby bylo zajištěno minimální riziko selhání a jeho případných dopadů. Dalším podpůrným bodem, jež se snaží vyvarovat potenciálnímu selhání systémů, je 1.2 a. (3) (c), díky němuž má kvalifikovaný operátor možnost systém kdykoliv deaktivovat, pokud shledá, že systém nepracuje dle předpokladů.[86]

Tento trend preventivního přístupu potvrzuje i samotný proces vývoje a nasazení autonomních a polo-autonomních zbraňových systémů. Každá zbraň využívající autonomní technologie musí projít minimálně dvěma schvalovacími procesy[87]. První z těchto schvalovacích procesů probíhá ještě před tím, než je zahájen vývoj této zbraně. Druhý schvalovací proces probíhá před nasazením zbraně. Zároveň mohou být vyžadovány dodatečné schvalovací procesy v případě, že se v průběhu testování či vývoje ukáže, že systém vyžaduje změny.[88] Tento důraz na testování systémů a ověření, že jednají tak, jak bylo zamýšleno pomáhá operátorům těmto zbraňovým systémům více důvěřovat. Důvěra je v tomto případě důležitým aspektem, jak již bylo zmíněno v předešlé části práce, především kvůli zodpovědnosti, jež operátoři za činy těchto systémů, s největší pravděpodobností, nesou.

Směrnice také upřesňuje určité detaily, jež jsou nápomocné v otázce ospravedlnitelnosti. Bod 1.2 a. (3) (a), říká, že činy autonomních a polo-autonomních zbraňových systémů budou kvalifikovaným operátorům snadno srozumitelné, stejně jako akce, jež systém plánuje provést a akce, které systém očekává, že provede operátor. Bod 1.2 a. (3) (b) od systému

[85] Ibid, str. 4
[86] Ibid, str. 4
[87] Více informací o procesu schvalování a seznam osob, jež musí toto schválení udělit viz Section 2: Responsibilities DoD Directive 3000.09 či bod1.2 c. stejného dokumentu.
[88] DoD Directive 3000.09 Autonomy in Weapon Systems. str. 4-5

vyžaduje, aby poskytoval transparentní zpětnou vazbu svého statusu. To znamená, že operátor je neustále informován o stavu systému a zdali je plně funkční a nepoškozený. Současně ho systém informuje o potenciálních cílech, které je systém schopen zasáhnout a informace které se tohoto útoku týkají, jako stav munice či podmínky, které by útok mohly negativně ovlivnit, jako je například počasí, pozice cíle či možné překážky.[89] Tyto dva body zajišťují, že kvalifikovaný operátor, který daný systém operuje, mu rozumí a je kdykoliv schopen vysvětlit, proč se systém rozhodl právě tak, jak se rozhodl. Současně, díky dříve zmíněným možnostem systém deaktivovat má operátor šanci zamezit neetickému jednání systému, i díky tomu, že je neustále informován o tom, jaké kroky systém plánuje provést.

S těmito body souvisí vzdělání těchto operátorů. Je klíčové, aby každý operátor systému skutečně dobře rozuměl a znal jeho schopnosti a možnosti. Směrnice zde navazuje na jeden z cílů stanovených v Strategickém a implementačním plánu pro odpovědné využívání umělé inteligence Ministerstva obrany, jež se zaměřoval na zajištění, že zaměstnanci ministerstva budou autonomním zbraňovým systémům a jiným technologiím využívajících umělou inteligenci rozumět a obdrží odpovídající vzdělání. Proto body 2.9 b. (6), 2.9 b. (7) a 2.9 b. (8) stanovují, že systémy musí být vyvíjeny tak, aby jim kvalifikovaní operátoři rozuměli. Současně musí operátoři projít specifickým kvalifikačním vzdělávacím programem, aby bylo zajištěno, že jsou srozuměni s funkcemi systému a s rolí, jež budou ve vztahu k tomuto systému plnit. S tím souvisí nejen znalost technologie, ale také schopnost objektivně posuzovat činy systému na základě jejich uvážení. Tato školení jsou periodicky opakovaná, aby se zajistilo, že operátoři vždy pracují s aktuálními informacemi.[90] Směrnice tak reflektuje neustále se měnící charakter nových technologií, které se stále vyvíjí, a přizpůsobuje tomu vzdělávací plány. Důraz na informovanost a kvalifikovanost operátorů těchto technologií se v intenzitě téměř vyrovnává snaze předejít jakémukoliv selhání těchto systémů. Tento důraz na vzdělání operátorů je však naprosto logický, vzhledem k roli, kterou operátoři zastávají. Díky operátorům lze potenciálně velkému množství chybných rozhodnutí a selhání systému předejít.

3.5. Defense Primer: U.S. Policy on Lethal Autonomous Weapon Systems

Tento dokument byl zveřejněný v únoru 2024 Výzkumnou službou Kongresu

[89] Ibid, str. 4
[90] Ibid, str. 11

(*Congressional Research Service*) a jedná se tak o nejaktuálnější zkoumaný materiál. Současně se jedná o dokument informativního charakteru určený primárně jako podklad pro americké zákonodárce. Již ze své podstaty se nejedná o dokument, který by přinášel zásadní inovativní myšlenky, ale spíše utvrzuje určité dříve zmíněné domněnky. Jak samotný název napovídá, tento Přehled obranné politiky se zabývá otázkou smrtících autonomních zbraňových systémů. Dokument hned v úvodu definuje smrtící autonomní zbraňové systémy jako speciální typ zbraňových systémů, které využívají senzory a algoritmy za účelem identifikace, zaměření a útoku na cíl, a to bez zásahu člověka. Dokument také specifikoval, že definici pro autonomní zbraňové systémy užitou ve směrnici 3000.09 Autonomy in Weapon Systems, lze aplikovat i na smrtící autonomní zbraňové systémy.[91]

Dokument potvrdil, že Spojené státy americké v současné době nevlastní smrtící autonomní zbraňové systémy, zároveň však nerozporoval potenciální budoucí vývoj těchto zbraní. Článek odkazuje na nejmenované vedoucí pracovníky armády a ministerstva obrany, jež sdělili, že Spojené státy budou v budoucnosti nuceny vyvíjet zbraně tohoto typu, pokud se touto cestou vydají jejich protivníci.[92] Tito protivníci nebyli jmenovaní, s největší pravděpodobností se však jedná o Čínu a Rusko, na než odkazuje nejen Závěrečná zpráva Národní bezpečnostní komise pro umělou inteligenci, ale i další nevládní publikace.[93]

Dokument také komentoval problematiku diskutovanou ve směrnici 3000.09, jež se týká role operátora těchto zbraní. Opět zopakoval, že zapojení tohoto operátora se značně liší, v souvislosti s danou situací. Situace, kdy musí operátor zasáhnout se liší případ od případu a není tak možné formulovat jasný postup, jež by bylo možné aplikovat v každé situaci. Opět je kladen obrovský důraz na vzdělání a znalosti tohoto operátora. Člověk, který operuje daný autonomní, či polo-autonomní, zbraňový systém mu musí rozumět. Je stěžejní, aby znal jeho schopnosti a hranice a byl srozuměn s riziky, jež jsou s využitím zbraně tohoto typu svázané. Tato znalost autonomního systému zajistí, že operátor rozumí, proč systém zvolil ty kroky, které zvolil.[94] Právě toto porozumění je extrémně důležité z pohledu ospravedlnitelnosti. Pokud lidský operátor rozumí systému a je schopen jeho kroky interpretovat, jsme schopni rozhodnutí tohoto systému objektivně zhodnotit a v případě

[91] Congressional Research Service, Defense Primer: U.S. Policy on Lethal Autonomous Weapon Systems. str. 1
[92] Ibid
[93] National Security Commission on Artificial Intelligence, The Final Report, str..77
[94] Congressional Research Service, Defense Primer: U.S. Policy on Lethal Autonomous Weapon Systems, str. 1

vyžádání dojde k splnění podmínek procesu *Right to(an) explenation*. I přesto, že rozhodnutí systému není transparentní a čitelné všem, operátor v těchto situacích zastupuje roli překladatele. Aby byla zajištěna ospravedlnitelnost, není důležité, aby rozhodnutím rozuměl každý, ale aby systému dedikovaný operátor byl schopen činy systému popsat a vysvětlit v jakýkoliv momentu v průběhu operace.

Závěr

Analýzy zkoumaných dokumentů prokázaly, že americká strategie na implementaci smrtících autonomních zbraňový systémů zohledňuje všechny tři zkoumané, eticky problematické, oblasti. Jednotlivé dokumenty na sebe navazují. Koncepty, které představila závěrečná zpráva, rozpracoval strategický plán a na závěr došlo k jejich začlenění do směrnice.

Otázku zodpovědnosti adresují již samotné Etické principy využití umělé inteligence ministerstva, které stanovují, že zodpovědnost za činy autonomních systémů nese člověk. Přesunutí zodpovědnosti za činy autonomních systémů na člověka dále podporuje i závěrečná zpráva Národní bezpečnostní komise pro umělou inteligenci, Strategický a implementační plán pro odpovědné využívání umělé inteligence ministerstva obrany Spojených států i Směrnice 3000.09 Autonomie zbraňových systémů. Přenos zodpovědnosti je také podpořen řadou povinností, které operátor musí plnit. Operátor autonomních zbraňových systémů musí každé rozhodnutí tohoto systému kriticky zhodnotit a rozhodnout, zdali není v rozporu s mezinárodním humanitárním právem. Současně operátor dohlíží na činnost daného autonomního systému a dle potřeby autorizuje jeho činy. Míra a intenzita dohledu je individuální, ne každé využití smrtící síly je nutné autorizovat. Zároveň má operátor možnost zakročit a systém deaktivovat, pokud jeho rozhodnutí vyhodnotí jako chybné.

Otázka ospravedlnitelnosti byla také adresována v Etických principech využití umělé inteligence. Dle nich totiž musí být autonomní systémy navrženy tak, aby jejich činy byly srozumitelné jejich operátorům. Aby byla tato podmínka splněna, je potřeba zajistit, že operátoři projdou kvalitním vzdělávacím programem. Jejich vzdělání je kontinuální, čímž se zajišťuje, že neustále pracují s nejaktuálnějšími informacemi. Operátoři musí být experty v oblasti autonomních zbraní a je jejich povinností být schopen vysvětlit jednotlivé činy těchto systémů. Tuto problematiku je potřeba ošetřit i po technické stránce a autonomní zbraňové systémy jsou proto navrhovány tak, aby operátora neustále informovaly o svém technickém stavu a o informacích relevantních pro plnění misí, jako je registrace nepřítele v okolí. Autorizace jednotlivých kroků nepomáhá řešit jen problematiku zodpovědnosti, ale usnadňuje i roli operátora ve vztahu k problematice ospravedlnitelnosti. Frekventované

schvalování činů autonomních zbraňových systémů usnadňuje vysvětlení jejich akcí. Právě kvůli schopnosti vysvětlit činy autonomních systémů kladou všechny zkoumané dokumenty velký důraz na proces vzdělávání operátorů. Strategie se zaměřuje na schopnost operátorů vysvětlit činy systémů, protože, jak bylo zmíněno v teoretické části, pokud jsme schopni činy autonomních zbraňových systémů vysvětlit, jsme je také schopni ospravedlnit.

Poslední zkoumanou problematikou byla předpojatost. I přesto, že předpojatost není adresována tak jednoznačným způsobem, jako dříve diskutované problematiky, lze i v tomto případě potvrdit, že je ve strategii zohledněna. Všechny zkoumané dokumenty kladou velký důraz na testování a kontrolu autonomních systémů. Všechny texty vykazují známky preventivního přístupu, panuje tedy snaha vyvarovat se jakémukoliv selhání systémů. I proto lze předpokládat, že v případě, kdy by jakékoliv předpojatosti u systému byly rozpoznány, byly by podstoupeny adekvátní kroky k odstranění tohoto defektu. Současně, i v případě že by se závadný zbraňový systém dostal do zóny aktivního konfliktu a došlo by k jeho nasazení během mise, jsou tyto zbraňové systémy doprovázeny operátory, kteří hodnotí, zdali jsou činy tohoto systému v souladu s platnými etickými kodexy. V případě, že by tento operátor shledal jakoukoliv předpojatost, jež by ovlivňovala zaměřování cílů či plnění mise, má možnost zakročit a systém deaktivovat.

Po zhodnocení všech těchto aspektů lze říci, že Spojené státy si uvědomují rizika spjatá s implementací smrtících autonomních systémů a podnikají adekvátní kroky, aby zajistily že bude jejich budoucí adaptace probíhat v souladu s mezinárodním humanitárním právem a tyto zbraně budou bezpečné a nebudou etickým rizikem. To však neznamená, že do budoucna nebude tato strategie vyžadovat změny. Technologický vývoj může odkrýt nová rizika, jež si není neuvědomujeme.

Práce přináší komplexní analýzu vzájemně propojených dokumentů. Dokumenty v této kombinaci dosud nebyly analyzovány. Současně se jedná o unikátní pojetí analýzy strategie vybrané země ve vztahu k zvoleným zkoumaným problematikám – ospravedlnitelnosti, předpojatosti a zodpovědnosti. Práce proto otevírá diskuzi a podněcuje k dalšímu výzkumu této problematiky.

Summary

The analysis of the examined documents demonstrates that the U. S. strategy to implement

lethal autonomous weapon systems considers all three ethically concerning aspects. Individual documents are interconnected and follow up on each other's concepts. The concepts introduced in The Final Report then got elaborated on in the Strategy pathway and later on were included in the directive.

The question of responsibility is already addressed in the Department of Defense AI Ethical Principles, where the responsibility for actions of autonomous weapon systems is assigned to the human operator. This shift of responsibility towards the human is promoted in The Final Report by the National Security Commission on Artificial Intelligence, U.S. Department of Defense Responsible Artificial Intelligence Strategy and Implementation Pathway as well as in the Department of Defense Directive 3000.09 Autonomy in Weapon Systems. This transfer of responsibility is also supported by a number of duties the operator must fulfil. The operator of the weapon system must judge each of the systems' decisions and decide whether the decisions are a violation of the international humanitarian law. In addition, the operator supervises the activity of the autonomous system and authorizes its action when needed. The intensity and amount of supervision needed is evaluated on individual basis, not every deployment of lethal force needs to be authorized. In addition to this, the operator has the option to step in and disable the system once its actions are deemed faulty.

The issue of justifiability is addressed in the Department of Defense AI Ethical Principles as well. According to these principles, autonomous systems must be designed in a way that makes their actions understandable to the operators. To ensure this condition is met, it is important to ensure that the operators complete a high-quality educational program. Their education is continual which ensures the operators always work with the most recent available information. The operators must be experts in autonomous weapons and its their responsibility to be able to explain the individual actions of these weapons. This issue needs to be addressed in terms of the design of the technology itself. Autonomous systems must be designed to provide feedback on the system status including information relevant for the ongoing mission as for example the identification of the adversary in the area. The authorization of individual steps also helps when addressing the issue of justifiability. Frequent authorizations of the decisions and keeping track of individual steps makes it easier to explain the actions of weapon systems. Due to the ability to explain the actions of autonomous systems, all the documents emphasize the education of the operators. The

strategy is so focused on the explainability aspects because, as was explained in the theory part, once we are able to explain the actions of autonomous weapon systems, we are also able to justify them.

The last of the explored concerns is prejudice. Although the strategy doesn't address this issue directly, as the above discussed concerns, we can confirm that it was taken into consideration. All of the examined documents put a huge emphasis on testing and control of the autonomous weapon systems. For this reason, all the texts show signs of a precautionary approach, which means they are trying to avoid the failure of these systems in any way. Therefore, we can assume that if any signs of prejudice were discovered, there would be adequate steps taken to ensure this defect was eliminated. At the same time, even if the flawed weapon system got fielded and was deployed to be used during an active armed conflict, the operator accompanying the weapon system has the option to deactivate the system if they deem its steps a violation of the ethical code of conduct.

After the evaluation of all the relevant aspects, I feel confident in confirming that the United States of America understand the risks that accompany implementation of lethal autonomous weapon systems. They are taking the appropriate steps to ensure the future adaptation of these technologies will be carried out in compliance with international humanitarian law and these weapons will be safe and won't pose an ethical risk. That doesn't mean that the strategy won't need to be updated in the future. Technical development can uncover new challenges which are currently hidden from us.

The thesis presents a comprehensive analysis of interconnected documents. These documents have not yet been analyzed in this combination before. At the same time, it is a unique approach to an analysis of the strategy of a selected country, in relation to the chosen issues under study - justifiability, prejudice and responsibility. The thesis therefore opens discussion and encourages further research on this issue.

Použitá literatura

Primární zdroje

Congressional Research Service. *Defense Primer: U.S. Policy on Lethal Autonomous Weapon Systems.* February 1, 2024. https://crsreports.congress.gov/product/pdf/IF/IF11150.

Department of Defense Directive 3000.09 Autonomy in Weapon Systems. 2023. https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf.

Lee, Kai-Fu. 2018. *AI Superpowers: China, Silicon Valley, and the New World Order*. Houghton Mifflin Harcourt.

National Defense Authorization Act for Fiscal Year 2019: Pub. L. No. H.R. 5515. 2018. https://www.congress.gov/115/bills/hr5515/BILLS-115hr5515enr.pdf

National Security Commission on Artificial Intelligence. 2021. „The Final Report." https://reports.nscai.gov/final-report/.

National Security Commission on Artificial Intelligence, „AI and Warfare" in *The Final Report,* 75-88, (Washington, D.C.: National Security Commission on Artificial Intelligence, March 1, 2021), https://assets.foleon.com/eu-central-1/de-uploads-7e3kk3/48187/nscai_full_report_digital.04d6b124173c.pdf.

National Security Commission on Artificial Intelligence, „Autonomous Weapon Systems and Risks Associated with AI-Enabled Warfare" in *The Final Report*, 89-106, (Washington, D.C.: National Security Commission on Artificial Intelligence, March 1, 2021), https://assets.foleon.com/eu-central-1/de-uploads-7e3kk3/48187/nscai_full_report_digital.04d6b124173c.pdf.

Sarah Wheaton „How Russian Disinformation Could Skew EU Election — and Whether Europe Can Fight It," *EU Confidential*, podcast episode, POLITICO, May 17, 2024, https://www.politico.eu/podcast/eu-confidential/how-russian-disinformation-could-skew-eu-election-and-whether-europe-can-fight-it/.

Stop Killer Robots. „UN Head Calls for a Ban." Last modified December 11, 2018. https://www.stopkillerrobots.org/news/unban/. (citováno 12. července 2024)

Stuart Russell, „AI in Warfare," The Reith Lectures, *BBC Radio 4*, December 8, 2021, https://www.bbc.co.uk/programmes/m00127t9. (staženo 7. května 2023)

U. S. Department of Defense. „*Implementing Responsible Artificial Intelligence in the Department of Defense*." Memorandum for Senior Pentagon Leadership, Commanders of the Combatant Commands, Defense Agency, and DoD Field Activity Directors. Washington, DC: Deputy Secretary of Defense, 2021. https://media.defense.gov/2021/May/27/2002730593/-1/-1/0/IMPLEMENTING-RESPONSIBLE-ARTIFICIAL-INTELLIGENCE-IN-THE-DEPARTMENT-OF-DEFENSE.PDF.

U.S. Department of Defense. Responsible Artificial Intelligence Strategy and Implementation Pathway. Washington, D.C.: U.S. Department of Defense, June 2022. https://media.defense.gov/2022/Jun/22/2003022604/-1/-1/0/Department-of-Defense-Responsible-Artificial-Intelligence-Strategy-and-Implementation-Pathway.PDF

Sekundární literatura

Abrassart, Christophe et al. „Montréal Declaration for a Responsible Development of Artificial Intelligence 2018," 2018. https://declarationmontreal-iaresponsable.com/wp-content/uploads/2023/04/UdeM_Decl-IA-Resp_LA-Declaration-ENG_WEB_09-07-19.pdf.

Atkinson, Katie, Trevor Bench-Capon, and Bollegala Danushka. 2020. „Explanation in AI and law: Past, present and future." *Artificial Intelligence*. 289 (103387). https://doi.org/https://doi.org/10.1016/j.artint.2020.103387.

Amoroso, Daniele, and Guglielmo Tamburrini. 2020. „Autonomous Weapons Systems and Meaningful Human Control: Ethical and Legal Issues." *Current Robotics Reports* 2020 (1): 187-194. https://link.springer.com/content/pdf/10.1007/s43154-020-00024-3.pdf.

Araya, Daniel, and Meg King. 2022. „The Impact of Artificial Intelligence on Military Defense and Security." *CIGI Papers* March 2022 (263). https://www.econstor.eu/bitstream/10419/299735/1/cigi-paper263.pdf.

Azar, Kimia Zamiri et. al. 2022. „Fuzz, Penetration, and AI Testing for SoC Security Verification: Challenges and Solutions." *Future microelectronics security research series*, no. 394. https://eprint.iacr.org/2022/394.pdf

Bartneck, Christoph, Christoph Lütge, Alan Wagner, and Sean Welsh. 2021. „Military Uses of AI." In *An Introduction to Ethics in Robotics and AI*, 93-99. Springer Briefs in Ethics. https://doi.org/10.1007/978-3-030-51110-4_11.

Bartneck, Christoph, Christoph Lütge, Alan Wagner, and Sean Welsh. 2021. „Trust and Fairness in AI Systems." In *An Introduction to Ethics in Robotics and AI*, 27-39. Springer Briefs in Ethics. doi.org/10.1007/978-3-030-51110-4_4.

Bartneck, Christoph, Christoph Lütge, Alan Wagner, and Sean Welsh. 2021. „What is AI?" 2021. In *An Introduction to Ethics in Robotics and AI*, 5-16. Springer Briefs in Ethics. https://doi.org/10.1007/978-3-030-51110-4_2.

Buolamwini, Joy, and Timnit Gebru. 2018. „Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." *Conference on Fairness, Accountability, and Transparency*: 1-15. https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf.

Bottino, Andrea et al. 2021. „A Brief History of AI: How to Prevent Another Winter (A Critical Review)." *PET Clinics* 16 (4): 449-469. https://doi.org/10.1016/j.cpet.2021.07.001.

Galliott, Jai. 2021. „Toward a Positive Statement of Ethical Principles for Military AI.“ In *Lethal Autonomous Weapons: Re-Examining the Law and Ethics of Robotic Warfare*, 121-135. Oxford University Press. https://books.google.cz/books?hl=cs&lr=&id=3PYTEAAAQBAJ&oi=fnd&pg=PA121&dq=Galliott,+J.+(2021).+%E2%80%9CToward+a+positive+statement+of+ethical+principles+for+military+AI,%E2%80%9D+in+Lethal+Autonomous+Weapons,+eds+Galliott,+J.,+MacIntosh,+D.,+and+Ohlin,+J.+(Oxford:+Oxford+University+Press)&ots=OTGxl9TLlR&sig=y8qGZhV1gUloHigyV2H1e1QvNj4&redir_esc=y#v=onepage&q&f=false.

Margulies, Peter. 2017. „Making autonomous weapons accountable: command responsibility for computer-guided lethal force in armed conflicts.“ In *Research Handbook on Remote Warfare*, 405-442. Cheltenham UK: Edwar Elgar Publishing. https://www.elgaronline.com/edcollchap/edcoll/9781784716981/9781784716981.00024.xml.

Melzer, Nils, and Etienne Kuster. 2019. International Humanitarian Law A Comprehensive Introduction. International Committee of the Red Cross. https://doi.org/10.1017/S1816383117000091.

National Security Commission on Artificial Intelligence, *The Final Report* (Washington, D.C.: National Security Commission on Artificial Intelligence, March 1, 2021), https://assets.foleon.com/eu-central-1/de-uploads-7e3kk3/48187/nscai_full_report_digital.04d6b124173c.pdf.

Rowe, Neil C. 2022. „The comparative ethics of artificial-intelligence methods for military applications.“ *Frontiers in Big Data* 5 (991759). https://doi.org/https://doi.org/10.3389/fdata.2022.991759.

Stanley-Lockman, Zoe. 2021. *Responsible and Ethical Military AI: Allies and Allied Perspectives*. Center for Security and Emerging Technology. https://cset.georgetown.edu/wp-content/uploads/CSET-Responsible-and-Ethical-Military-AI.pdf.

Thorpe, Joelle B., and Sherry Wasilow. 2019. „Artificial Intelligence, Robotics, Ethics, and the Military: A Canadian Perspective.“ *AI Magazine* 40 (1): 37-48. https://doi.org/10.1609/aimag.v40i1.2848.