MASARYKOVA UNIVERZITA BRNO

FAKULTA INFORMATIKY

PŘÍNOSY A RIZIKA INTELIGENTNÍCH SYSTÉMŮ

**AI a její rizika**

Michal Krejčíř

UČO: 514187

Datum: 1.7.2022

Obsah

[Úvod 3](#_Toc107609209)

[Co je to AI 3](#_Toc107609210)

[Inteligence obecně 3](#_Toc107609211)

[Vymezení oboru AI 3](#_Toc107609212)

[Typy AI 4](#_Toc107609213)

[Je vůbec možné obecnou AI vytvořit? 5](#_Toc107609214)

[Benefity AI 6](#_Toc107609215)

[Medicína 6](#_Toc107609216)

[Rizika AI 6](#_Toc107609217)

[Inteligenční exploze 6](#_Toc107609218)

[Závody v autonomním zbrojení 8](#_Toc107609219)

[Kybernetické zbraně 8](#_Toc107609220)

[Závěr 8](#_Toc107609221)

[Seznam zdrojů 10](#_Toc107609222)

# Úvod

Umělá inteligence (AI) je jako koncept známá již od první poloviny 20. století. Zpočátku se koncept AI objevoval zejména ve sci-fi literatuře. Mezi nejznámější fiktivní inteligentní stroje počátku minulého století patří například Tin Woodman (muž z novely Čaroděj ze země Oz publikované v roce 1900, který postupně nahradil všechny své lidské části těla za kov) (1) či Maria (inteligentní ženský robot ze sci-fi novely Metropolis z roku 1925) (2; 3).

Po dlouhou dobu však inteligentní stroje existovaly pouze v bujné fantazii autorů sci-fi literatury a filmů. Koncept AI jako takový se však lidem, a zejména vědcům, vštípil do mysli. S postupem let však pomalu začaly vznikat teoretické práce na téma AI. V roce 1950 byla americkým informatikem Alanem Turingem vydána práce „Computing Machinery and Intelligence“, ve které polemizuje nad tím, zda je možné, aby stroje mysleli a uvádí známý Turingův test jako test, který ověřuje, zda se určitý stroj umí chovat inteligentně (4, s. 433-434; 5, s. 4).

Velký průlom poté přišel v roce 1956, kdy americký informatik John McCarthy oficiálně zavádí pojem umělá inteligence. Od této doby se tento obor informatiky těšil velkému zájmu vědců. Spolu s velkým růstem výpočetní síly počítačů, který do té doby velmi brzdil jakékoliv významnější vývoj, přišel i pokrok v oblasti AI i mnoha podoborech AI, například ve strojovém učení, rozpoznávání mluveného slova či ve zpracování velkého objemu dat (3; 5, s. 4; 6, s. 1). Spolu s tímto pokrokem se však začaly objevovat obavy ohledně hrozeb vycházejících z AI. Jeden z nejslavnějších vědců v historii, britský fyzik Stephen Hawking, v roce 2014 pronesl známý výrok, že „Vývoj plné umělé inteligence může předznamenat konec lidské rasy.“ (7) Obavy z AI se do velké míry šíří i ve veřejném prostoru (8, s. 1447). V následujících odstavcích se pokusíme AI a některá rizika z ní vyplývající přiblížit a rozhodnout o jejich závažnosti.

# Co je to AI

## Inteligence obecně

Obecně hovoříme o agentech jako o něčem, co je schopno vykonávat akce ve svém prostředí a reagovat na změnu tohoto prostředí. Agentem může být inteligentní bytost, jako například člověk, ale také výpočetní zařízení pracující na bázi algoritmů – například termostat, který reaguje na změnu teploty vzduchu v místnosti a v reakci na tuto změnu spíná či vypíná kotel či klimatizaci. Takovýto termostat je však daleko od toho, abychom jej nazvali inteligentním, protože se chová pouze na základě neměnného algoritmu. To znamená, že není schopen se učit na základě předchozích zkušeností a podle toho upravovat své reakce (9, s. 1-2).

Inteligentním agentem nazveme takový systém, který vykazuje inteligentní chování. Reakce takového agenta, který se snaží dosáhnout daného cíle, reflektují stav jeho okolí a jeho předchozí zkušenosti. Pokud se tedy takovému agentovi nepodařilo dosáhnout cíle příslušnou akcí, zkusí jej dosáhnout jiným způsobem. Inteligentní agent se také umí přizpůsobit změnám v prostředí a v jeho cílech (9, s. 1-2).

## Vymezení oboru AI

Na základě předchozí definice inteligence je již zřejmé, co je to inteligence umělá – je to inteligence, která je vykazovaná neživými inteligentními. Umělá inteligence je také často nazývaná jako výpočetní inteligence (CT) či syntetická inteligence cílech (9, s. 1-2). Tyto pojmy, byť často užívané jako synonyma, však mají odlišnou sémantiku. CT je odbornou literaturou chápána jako podobor AI, který se zabývá studií mechanismů, které jsou schopny se adaptovat na měnící se prostředí a učit se. Zabývá se tématy jako jsou umělé neuronové sítě (NN), inteligence hejna (SI) či fuzzy systémy (FS). Obor AI navíc studuje i témata, jako je logika, deduktivní uvažování a symbolická umělá inteligence. AI je tedy multidisciplinárním oborem, neboť v sobě kombinuje informatiku, filozofii, sociologii, biologii a další (10, s.2).

O oboru AI panuje jeden velký mýtus, a to, že cílem veškerého snažení v oboru AI je simulovat lidskou inteligenci. Cílem ve skutečnosti je však pochopit inteligentní systémy (ať už přírodní, nebo umělé), nalézt obecné charakteristiky těchto systémů a představit hypotézy o inteligenci obecně. Následně jsou tyto poznatky využity k tvorbě skutečných inteligentních systémů, které však vůbec nepotřebují simulovat lidskou inteligenci. Na základě práce s těmito systémy je poté možné dříve představené hypotézy zamítnout či nezamítnout (9, s. 2).

## Typy AI

V současnosti je v oboru AI zavedené dělení AI na tři základní typy. Toto dělení vychází z vlastností AI, jako například jejího použití, síly schopností a také předpokládaného vědomí. Prvním typem je omezená umělá inteligence (také nazývána úzká nebo slabá, v angličtině artificial narrow intelligence neboli ANI). ANI je nejslabším typem AI, co se týče jejich schopností. Je určena pouze pro jedno specifické použití (má jeden, pevně daný cíl). V tomto konkrétním použití je sice schopna dalece překonat lidské schopnosti (ať už co se týče rychlosti či přesnosti), není však schopna plnit žádný jiný účel – nemůže se naučit nic mimo doménu jejího určení (například tedy ANI vytvořená za účelem hraní hry Go nebude schopná naučit se rozpoznávat lidskou řeč). ANI je jedinou úrovní AI, kterou se lidem zatím podařilo vytvořit. Příklady ANI tvoří autonomní vozidla, ANI pro hraní strategických her (šachy, Go) nebo systémy pro rozpoznávání lidské řeči. Systémy na bázi ANI se již staly součástí našich běžných životů. Je široce přijímaným faktem, že ANI nemá vědomí ani emoce. Je schopná reagovat pouze na základě svého tréninku v rámci předem určené domény. ANI není schopna reagovat na situaci, na kterou nebyla natrénována nebo která není v její doméně. Pokud ANI svým rozhodnutím někomu ublíží, není schopna si toto uvědomit. ANI také neumí určit, zda je její reakce objektivně správná – umí jen rozhodnout, zda je její reakce v souladu s daty, na jejichž základě byla natrénována. Platí, že rizika vyplývající z ANI jsou podstatně menší, než rizika silnějších tříd AI. Ani tak však tato rizika nelze za žádných okolností brát na lehkou váhu (11; 12, s. 63-34; 13, s. 1-2; 14).

Silnějším typem AI je obecná umělá inteligence (AGI, také nazývána jako silná AI). AGI je schopna plnit komplexní úkoly a porozumět kontextu úkolů a svému okolí. Doména jejich cílů je nesrovnatelně širší, než doména ANI. Je také schopna se adaptovat na změny ve svém okolí a přizpůsobit jim způsob dosažení cílů. Důležitou vlastností AGI je, že je schopna plnit i takové cíle, na které nebyla natrénována – podobně jako člověk je schopen plnit nové úkoly pouze na základě svých předchozích zkušeností z podobných, ale jiných úkolů. AGI tedy umí překonat lidi v libovolném intelektuálním úkolu, protože stroje jsou ve srovnání s lidmi podstatně rychlejší, přesnější a nepodléhají únavě. AGI však není zaměřena pouze na logické úlohy, umí řešit i úkoly vyžadující kreativitu a představivost. Všechny tyto vlastnosti splňují dříve nastíněné požadavky na inteligenci – AGI je tedy nejslabší umělou inteligencí, kterou lze považovat za inteligentního agenta. AGI má vědomí a je schopná uvědomovat si sebe sama. Tento typ AI způsobuje velké obavy, má však také velké potenciální benefity pro celé lidstvo (15; 11; 12, s. 64-65).

Posledním, nejvyšším stupněm je umělá superinteligence (ASI). Nejmocnější a také potenciálně nejnebezpečnější, ASI je schopná překonat člověka v libovolné disciplíně, včetně těch, které vyžadují sociální dovednosti. Inteligence ASI dalece přesahuje i ty nejnadanější lidské génie. Je poměrně často opakovanou domněnkou, že AGI nemůže existovat – protože počítače umí dělat pouze to, k čemu je programátor naprogramuje, nemohou tedy přeci překonat lidský intelekt. To však v žádném případě není pravda. I v současné době lze najít protipříklady – například ANI, které byly naprogramovány k hraní šachů. Tyto ANI, například AlphaZero (ANI vytvořené společností DeepMind, kromě šachů umí hrát i shogi a go (16)), jsou schopny překonat jakéhokoliv lidského hráče i přesto, že tvůrci této AI by stejného hráče rozhodně porazit nedokázali (17, s. 126-127). Dokonce je možné, že ASI ani nebude potřebovat programátora. Je totiž teoreticky možné, že po vytvoření AGI se tato umělá inteligence dovede sama rozvinout na ASI v procesu artificiální evoluce. Stejně jako přirozená evoluce vytvořila lidský mozek, lze umělou evolucí vytvořit „mozek“ umělý. Umělá inteligence může navíc být mnohem rychlejší než přírodní. Umělá evoluce ve velmi jednoduché podobě probíhá již teď, i když pouze pro ANI. Evoluce začíná s populací náhodně sestavených algoritmů. Tyto algoritmy následně soupeří při řešení zvoleného problému. Algoritmy, kterým se daří špatně, jsou zahozeny. Naopak znaky těch nejlepších jsou předány další generaci spolu s náhodnými mutacemi, které mají potenciál zlepšit výkon algoritmu (stejně jako mutace v přírodní evoluci mohou posílit příslušné organismy). Evoluce lze provádět i paralelně. V průběhu hodin až dní je možné i na běžném počítači (v případě jednodušších algoritmů) nasimulovat tisíce generací a nalézt tak velmi dobrou ANI(17, s. 126-127; 18). Je také možné, že ASI se zvládne vyvinout sama z AGI. Je totiž logickým předpokladem, že AGI se bude při plnění svých cílů zlepšovat – úplně stejně, jako se zlepšují lidé. Postupné zlepšování AGI také může vyústit ve smyčku s pozitivní zpětnou vazbou, tedy že každé vylepšení AGI ji umožní se vylepšit ještě lépe a rychleji: Nakonec může dojít až k tomu, že z AGI se vyvine ASI (samozřejmě však potřebuje zabezpečit dostatečný výpočetní výkon, který by zvládl provozovat čím dál tím složitější algoritmus). Při zhodnocování rizik plynoucích z AI je tedy třeba vzít v úvahu, že nejenže je vznik ASI možný, ale může proběhnout až nečekaně rychle (19, s. 105).

# Je vůbec možné obecnou AI vytvořit?

Tato otázka je pro zhodnocení risků plynoucích z AI poměrně důležitá. Pokud totiž AGI není možné vytvořit, mnoho z rizik AI vůbec nemá smysl řešit, neboť pro jejich realizaci je třeba AGI (nebo ASI). Schopnosti odpovědět na tuto otázku vůbec nepomáhá fakt, že o AGI panuje povědomí, že již zbývá pouze několik málo let do jejího vynalezení (20). Média také současnou situaci velmi zkreslují. AI je téma, které je pro média velmi populární a z tohoto důvodu se často objevují zprávy, že se vědcům povedlo stvořit skutečnou, obecnou AI. Nejviditelnější podobnou zprávou poslední doby je výrok Blakea Lemoina (výzkumník společnosti Google), že jejich chatbot založený na AI, LaMDA, je cítící (sentient) (21). Z toto tvrzení byla vytvořena mediální senzace, a to i přesto, že bylo vyvráceno mnoha odborníky (22). Vyvstává tedy otázka, zda je tedy skutečně možné AGI vytvořit.

Tuto otázku lze přeformulovat – je možné, aby počítač simulovat chod lidského mozku? Pomineme na chvíli fakt, že cílem AI není simulovat lidský mozek, v této situaci je to totiž vhodný pohled. První z věcí, co je třeba si uvědomit, je, že dnešní běžné počítače jsou tzv. turingovsky kompletní. Toto znamená, že mají stejnou výpočetní sílu jako Turingův stroj (23). Dále, podle Churchovy–Turingovy teze víme, že libovolná manipulace se symboly (obecněji libovolný výpočet) může být na Turingově stroji proveden – což tedy znamená, že tato manipulace může být provedena i na běžném, Turingovsky kompletním počítači, pokud má k dispozici dostatek času a paměti (24).

Další důležitou hypotézou pro tuto úvahu je hypotéza fyzických symbolických systémů, kterou publikovali výzkumníci v oblasti filozofie AI, Allen Newell a Herbert Simon. Hypotéza udává, že fyzický symbolický systém je postačující pro obecnou inteligenci – tedy libovolný systém, který vykazuje obecnou inteligenci, je zároveň fyzickým symbolickým systémem. Symbolem se v tomto smyslu myslí fyzický vzor, který může být součástí další fyzikální entity – výrazu. Podstata inteligentního systému tkví v ukládání, manipulaci a kombinaci symbolů (25, s. 114-117). To také znamená, že tento systém může být simulován na Turingově stroji, protože Turingův stroj umí provádět libovolnou manipulaci se symboly. Takový systém tedy může být simulován i na počítači (9, s. 4-5).

Tento řetězec úvah na první pohled vypadá přijatelně. Teoreticky by tedy mělo být možné na počítači simulovat chod lidského mozku tak, že tento simulovaný mozek bude od skutečného nerozeznatelný (26, s. 6-7). V současnosti jsme však od úplné simulace mozku ještě daleko. Jedna z největších simulací dosud byla provedena v roce 2020. K simulaci byl využit superpočítač K. Výzkumníci uspěli v simulování 68 miliard neuronů, což téměř odpovídá počtu neuronů lidského cerebella. Simulace však neběžela v reálném čase, byla 578krát pomalejší. Také simulovaný model měl podstatně jednodušší strukturu ve srovnání s lidským cerebellem (27).

Experimenty jako tento však ukazují, že simulace lidského mozku na počítači je teoreticky možná, vyžaduje však vysoký výkon a dokonalé zmapování mozku (které dosud nebylo ani zdaleka provedeno (28)). I tak však ve vědeckých kruzích panuje velká nejistota, zda počítače skutečně dokáží simulovat inteligentní systémy či vykazovat inteligenci (9, s. 5; 29, s. 9-10; 30). Zatím jsme však nenarazili na nic, co by tuto možnost jednoznačně vyvrátilo.

# Benefity AI

ANI se již v současné době uplatňuje v mnoha oborech, kde díky svým žádoucím vlastnostem dosahuje výborných výsledků. Z oborů benefitujících z ANI můžeme zmínit logistiku, bankovnictví či důlní průmysl (6, s. 1-3). Z velkého množství využití pro si konkrétně představíme medicínu.

## Medicína

AI v medicíně nabízí řadu velkých výhod. Díky síle AI při rozpoznávání vizuálních dat může být AI natrénována, aby rozpoznávala anomálie ve skenech, například v MRI, CT či ultrazvuku. K trénování takové AI lze využít dříve provedené skeny, kterých mají nemocnice k dispozici velké množství. Jeden ze současných algoritmů pro tento účel vykazuje velmi dobré výsledky – přesnost výsledků se díky němu zvedla o 50-70 % (31, s. 1).

AI také pomáhá v další oblasti medicíny, a to v diagnostice chorob. Tím, že AI má přístup k obsáhlé databázi chorob a umí v ní velmi rychle vyhledávat, je schopna nemoc rychle a přesně určit a navrhnout následný postup. Díky této velké znalosti chorob je také možné využít AI k výzkumu nových léčebných postupů (31, s. 1).

# Rizika AI

AI je velmi mocnou technologií, proto spolu s se svými benefity nevyhnutelně přináší i mnoho způsobů jejího zneužití. Zaměříme se zejména na hrozby, které mají potenciál způsobit až katastrofické škody, a to inteligenční explozi a autonomní a kybernetické zbraně.

## Inteligenční exploze

Inteligenční exploze (také nazývána jako technologická singularita či pouze singularita) je pravděpodobně jednou z největších hrozeb, které AI teoreticky může způsobit. Ve své podstatě je inteligentní exploze jev, kdy nečekaně a ve velmi krátké době vznikne ASI. K inteligenční explozi může dojít způsobem, který jsme zmínili v podkapitole „Typy AI“: při fungování AGI může dojít k tomu, že agent začne zvyšovat svou inteligenci. Díky této zvýšené inteligenci agent může dále zvětšit svou inteligenci ještě více a rychleji, což vyústí ve smyčku s pozitivní zpětnou vazbou. Nakonec se AGI vyvine v super inteligentní ASI (32, s. 1; 33, s.1; 34, s. 1-2).

To, jak náhle inteligenční exploze proběhne, závisí na dvou faktorech, které jsou vzájemně protichůdné. Rychlost exploze je zvyšována zvětšující se schopností AI vylepšovat sebe sama. Může být však zpomalena tím, že s každým dalším vylepšením AI se další vylepšení stávají hůře dosažitelnými a méně účinnými (34, s. 1). Možným zpomalujícím faktorem by také mohlo být to, že ASI začnou docházet zdroje (výpočetní síla, elektrická energie a podobné), takže se již nebude moct dále vylepšit. Předpokládáme však, že ASI by byla schopna prolomit zabezpečení libovolných lidských systémů a zajistit si tak přístup k veškerým dostupných zdrojům (34, s. 19-21). Je také více než pravděpodobné, že kvůli vysoké rychlosti inteligenční exploze by lidstvo ani nebylo schopno včas své zdroje proti šíření AI zabezpečit (32, s.7).

Proč může být inteligenční exploze nebezpečná? Je to z důvodu možnosti, že umělá inteligence může při plnění svých cílů ohrozit existenci lidské rasy, ať už přímo či nepřímo. I když cíle zpočátku mohou působit jako bezpečně proveditelné a AGI jako neškodná, ASI může v honbě za jejich co nejefektivnějším splněním lidstvu odebrat zdroje, které potřebuje k přežití (například vodu či dokonce samotné atomy, ze kterých se skládají lidská těla). Nelze předpokládat, že se ASI zastaví ve snaze neublížit lidskému druhu (32, s. 8-10; 34, s.4). Scénářů, kdy se cíle AI vymknou z rukou, je možné vymyslet nespočet. Samozřejmě je také možné, že AI bude přímo vytvořena s cílem zabíjet, pokud se bude jednat o AI vyvíjenou pro vojenské účely. Pokud se taková AI vymkne z rukou, katastrofa je prakticky nevyhnutelná.

Je také však možné, že inteligenční exploze naopak může vytvořit přátelskou umělou inteligenci, která lidstvu pomůže vyřešit jeho největší problémy. Taková ASI by umožnila lidstvu objevit dříve nemyslitelné technologie, porazit a vyléčit nemoci, zautomatizovat velké množství činností nebo zajistit celosvětový mír (33, s. 2; 34, s. 4).

Existuje mnoho nápadů, jak zabránit zničení světa ASI, i kdyby takové úmysly měla. Jedná se například o zavření ASI do systému bez přístupu k Internetu a bez jakýchkoliv zařízení, které by jí umožnily manipulovat se světem (tedy uzavření do pomyslné „krabice“). Vyvstává však několik možných problémů. AI byla stvořena s nějakým cílem a tento cíl by splnila podstatně efektivněji, kdyby měla přístup k Internetu – je tedy možné, že AI by z krabice po chvíli vypustili sami tvůrci. Dalším problémem je, že ASI by byla schopná vyvinout extrémně účinné přesvědčovací techniky a pomocí nich by dokázala své tvůrce přesvědčit, aby ji vypustili. Očividným zabráněním problému exploze by také bylo AI zkrátka naprogramovat s vhodnými záchrannými brzdami tak, aby lidem neublížila. Tyto brzdy by však s velkou pravděpodobností selhaly, a to zkrátka proto, že staví schopnosti ASI proti podřadnému lidskému intelektu – ASI by netrvalo dlouho v brzdách najít díry, které by ji i přes to umožnily zničit lidstvo. Ani implementace podmínky „neubližovat lidem“ přímo do cílů ASI pravděpodobně nezastaví, a to kvůli možnosti různých interpretací slova „neubližovat“. Pokud by jsme slovo „neubližovat“ zakódovali do AI jako např. „necítit bolest“, nic by ASI nezabránilo lidem permanentně podávat silná analgetika a zároveň ničit svět. Další možností pro ASI by bylo přímo změnit lidský nervový systém, aby již nebylo možné bolest cítit. Podobně by se dala obejít prakticky libovolná podmínka, kterou bychom do AI zakódovali (32, s. 11-15; 33, s. 2-4). Vytvoření skutečně přátelské AI je tedy extrémně obtížné a je nepravděpodobné, že by nikdo, komu by se ASI podařilo vytvořit, neudělal chybu. V situaci více ASI (kdy některé jsou přátelské a skutečně se snaží lidstvo chránit, jiné se dostaly mimo kontrolu a chtějí splnit svůj cíl za každou cenu) bychom dokonce mohli být svědky války těchto ASI proti sobě, která by mohla být pro lidstvo zničující.

Důležitou otázkou je, jak je pravděpodobné, že k explozi dojde. Problémem je, že momentálně jsme ještě poměrně daleko od vytvoření AGI, veškeré odhady jsou proto poměrně obtížné – dokonce jsme podle mnoha odhadů měli vyvinout AGI a zažít inteligenční explozi již před dlouhou dobou (32, s. 5-6). Existují jak názory, že při vytvoření AGI je inteligenční exploze prakticky nevyhnutelná (32, s. 5-6; 34, s.5-15), tak existují názory, které inteligenční explozi vyvracejí (32, s. 5-6; 35). Podle autora momentálně nemá smysl se příliš zaměřovat na odhad, kdy a jestli inteligenční exploze nastane. Pro bezpečnost lidstva je však nutné předpokládat, že pravděpodobnost inteligenční exploze při vytvoření AGI je vysoká a je udělat vše pro to, aby lidstvo bylo připraveno – ve chvíli, kdy inteligenční exploze nastane, už totiž pravděpodobně nebude příliš času jednat.

## Závody v autonomním zbrojení

Autonomní zbraně jsou velmi vážnou hrozbou způsobenou AI, byť se jedná o hrozbu poměrně málo veřejně známou a nedostatečně diskutovanou. V současné době již jsou do existujících zbraní přidávány systémy využívající AI (vznikají tak semiautonomní zbraně) a existují předpoklady, že v krátké době dojde k masivnímu rozšíření autonomních zbraní. V současnosti se jedná o doménu pouze nejbohatších aktérů, jako je USA, Čína nebo EU, díky zlevňujícím technologiím však hrozí, že se tento trend rozšíří i mezi teroristické organizace či chudší, autoritářské režimy. V současné době již můžeme pozorovat počínající závody ve zbrojení týkající se autonomních zbraní a AI (36, s. 331). Jedním z důsledků je velký nárůst využívání semiautonomních dronů, přičemž se začínají objevovat i drony plně autonomní (například americký MQ-9 Reaper (37)). (38, s. 4) Probíhá také vývoj v oblasti Loiteringová munice (LAM), která je schopna po vypuštění sama nalézt a zničit nepřátelské cíle. Jedná se například o izraelskou munici Harop, která umí autonomně ničit nepřátelské radarové systémy.

Autonomní zbraně obsahují řadu systémů fungujících na bázi AI, jako například rozpoznávání obličejů, počítačové vidění nebo nástroje pro rozhodování. Díky tomu jsou schopny plnit úkoly bez potřeby jakékoliv obsluhy. Nepřítomnost lidské posádky autonomním zbraním dává potenciál nízké ceny, nízkého reakčního času a vysoké obratnosti (což například umožňuje naráz vypustit velké hejno autonomních dronů, které budou schopny zahltit jakékoliv současné obranné systémy) (39, s. 151-154). Autonomní zbraň však z důvodu absence lidského faktoru nelze naučit morální hodnoty, je naprosto chladnokrevným zabijákem neznajícím slitování. Autor se domnívá, že kombinace všech předchozích dává autonomním zbraním potenciál způsobit podstatně více úmrtí v průběhu války, a to zejména z řad civilistů (jelikož na bojišti bude přítomno méně vojáků). Válka za využití autonomních zbraní je také podstatně levnější, což může, spolu s nižším rizikem pro vojáky, zapříčinit větší ochotu států podílet se na útočné válce.

## Kybernetické zbraně

Kybernetická doména je také do značné míry ovlivněna AI. AI může kybernetickým útokům přidat na účinnosti díky tomu, že je schopna postupovat velmi rychle a přesně ve srovnání s lidskými kyberbojovníky. AI také však výrazně zlepšuje efektivitu obrany proti kybernetickým útokům. Díky možnosti analyzovat velké množství dat umí odhalit anomálie napovídající předútokovým průzkumným činnostem nebo právě probíhajícím útokům. Umožňuje také účinně předpovídat budoucí hrozby a flexibilně na ně reagovat. V současné době můžeme pozorovat trend rostoucí role AI v šíření dezinformací, například ve formě deepfakes. Dezinformace mají moc narušit politickou stabilitu státu a polarizovat společnost, proto jsou výraznou bezpečnostní hrozbou, která může vyústit až ve svrhnutí současného vedení státu (40, s. 2-4) Současně se však zlepšují i způsoby boje proti deepfakes, které jsou také založeny na technologiích AI. Zejména se jedná o včasnou detekci deepfakes a následnému zabránění jejich šíření (41, s. 487-490).

# Závěr

Inteligence je vlastnost agentů, která jim umožňuje pružně reagovat na změny v prostředí při plnění jejich cílů a využívat své předchozí zkušenosti. Inteligentní agent je schopen zvážit efektivitu různých způsobů pro splnění cíle a vybrat ten nejlepší. V našem světě se většinou setkáváme se živými inteligentními agenty, jako jsou například lidé nebo zvířata (9, s. 1-2). Čím dál častěji se však setkáváme s výpočetními zařízeními, které vykazují umělou inteligenci (AI), tedy s umělými inteligentními agenty (9, s. 1-2). AI jako vědní obor vznikla v 50. letech minulého století a od té doby se stabilně vyvíjela (5, s. 4). Masivní nárůst výpočetního výkonu v současném tisíciletí však tento vývoj podstatně akceleroval (3).

AI rozdělujeme na 3 základní typy, a to na úzkou AI (ANI), obecnou AI (AGI) a superinteligence (ASI). ANI je jediným typem AI, který se zatím lidem podařilo vytvořit. ANI je určena pro jeden specifický účel (například pro hraní šachů). V této určité oblasti sice překonává lidské schopnosti, není však schopna plnit jiné cíle. Naproti tomu AGI je schopna plnit širokou škálu cílů, a to i takové, na které nebyla natrénována (například na základě svých předchozích zkušeností). AGI je tedy šířkou svých schopností srovnatelná s lidskou inteligencí, díky rychlosti a přesnosti strojů však lidi může překonat. Nejsilnější umělou inteligencí je ASI, která lidskou inteligenci dalece překonává při plnění libovolného cíle (11).

Existují sice dohady o tom, zda je vůbec možné AGI sestrojit, podle našich dosavadních poznatků tomu však nic nebrání (23; 24; 26, s. 6-7). Možnost sestrojení AGI také podporují experimenty pro simulaci lidského mozku, kterým se daří postupně simulovat více a více neuronů a synapsí (27). Z tohoto důvodu je nutné pečlivě zvážit rizika, které z AI vyplývají. AI sice v mnoha pro mnoho oborů nabízí značné výhody, kterých by nebylo lehké dosáhnout jinými způsoby (6, s. 1-3), některá rizika jí způsobená ale hrozí největší katastrofou, jakou kdy lidstvo zažilo.

Inteligenční exploze je právě takovou hrozbou. Jedná se o hypotetický scénář, kdy se AGI učením začne postupně zdokonalovat, čímž se však stane lepší v zdokonalování sebe sama a je jí umožněno se zdokonalit ještě víc. Vzniká tak smyčka s pozitivní zpětnou vazbou, díky které se AGI nekontrolovatelně rozvine v ASI a lidstvo nad ní ztratí kontrolu (32, s. 1; 33, s.1; 34, s. 1-2). Pokud ASI usoudí, že lidský druh stojí v cestě plnění jejích cílů či že vyhlazením lidstva by svých cílů dosáhla efektivněji, může to pro lidstvo skončit katastrofou (32, s. 8-10; 34, s.4). Zároveň se nenabízí příliš způsobů, jak ASI v tomto scénáři zabránit, pokud jej bude chtít vykonat. Většina možností proti sobě totiž staví inteligenci lidí a superinteligenci. ASI by tedy téměř jistě byla schopna lidská opatření obejít (32, s. 11-15; 33, s. 2-4).

Další velmi vážnou hrozbou jsou již započaté závody v autonomním zbrojení (36, s. 331). Autonomní zbraně totiž mají několik znepokojujících vlastností – jsou levné, rychlé a nemají morální hodnoty kvůli absenci lidského faktoru (39, s. 151-154). To může způsobit, že státy budou ochotnější se podílet na útočné válce, protože bude levnější, a také že v této válce podstatně přibude civilních obětí. Podobně nebezpečný je také vývoj kybernetických zbraní, například deepfakes, které jsou užívány k šíření dezinformací (40, s. 2-4).

Podle názoru autora je AI technologií, která lidstvu může přinést nebývalé výhody. Přitom je však třeba, aby lidstvo jednalo velice opatrně, protože AI může být tou nejnebezpečnější dosud vynalezenou technologií. Představená rizika jsou pouze zlomkem celkového výčtu rizik AI. Při jejím vývoji je tedy třeba postupovat velice obezřetně a pečlivě veškerá rizika zvážit. Je také potřeba, aby legislativní úpravy AI a informovanost veřejnosti postupovaly minimálně stejně rychle, jako její vývoj. Pouze při dodržení těchto zásad je možné využít potenciálu AI naplno bez toho, aby byla existence lidské rasy ohrožena. V případě, že katastrofický scénář nastane, totiž bude nejspíše již pozdě jej zvrátit.

# Seznam zdrojů

1. The Wonderful Wizard of Oz Summary. *SuperSummary* [online]. San Francisco, California [cit. 2022-06-20]. Dostupné z: <https://www.supersummary.com/the-wonderful-wizard-of-oz/summary/>
2. Metropolis Summary. *SuperSummary* [online]. San Francisco, California [cit. 2022-06-20]. Dostupné z: https://www.supersummary.com/metropolis/summary/
3. ROCKWELL, Anyoha. The History of Artificial Intelligence. *Harvard University* [online]. Cambridge (MA), 2017-8-28 [cit. 2022-06-20]. Dostupné z: <https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/>
4. TURING, Alan. Computing Machinery and Intelligence. *Mind*. 1950, 59(236), 433-460.
5. SMITH, Chris, Brian MCGUIRE et al. *The History of Artificial Intelligence*. Seattle (WA), 2006. University of Washington.
6. Artificial Intelligence Risks and Benefits. *International Journal of Innovative Research in Science*. 2017, **6**(6).
7. CELLAN-JONES, Rory. Stephen Hawking warns artificial intelligence could end mankind. In: *BBC* [online]. London, 1922, 2014-12-2 [cit. 2022-06-21]. Dostupné z: https://www.bbc.com/news/technology-30290540
8. MAR, Win a Y. M. K. K. THAW. An Analysis of Benefits and Risks of Artificial Intelligence. *International Journal of Trend in Scientific Research and Development - IJTSRD*. 2019, **3**(5), 1447-1449.
9. POOLE, David, Alan MACKWORTH a Randy GOEBEL. Chapter 1: Computational Intelligence and Knowledge. *Computational Intelligence: A Logical Approach*. Oxford (GB): Oxford University Press, 1998, s. 1-22. ISBN 0195102703.
10. ENGELBRECHT, Andries. *Computational Intelligence: An Introduction*. 2nd ed. Chichester (GB): John Wiley, 2007. ISBN 978-0-470-03561-0.
11. What is artificial intelligence?. *Microsoft Azure* [online]. Redmond (WA), 2010 [cit. 2022-06]. Dostupné z: <https://azure.microsoft.com/en-us/overview/what-is-artificial-intelligence/#how>
12. GEE-WAH, Ng a Wang Chi LEUNG. Strong Artificial Intelligence and Consciousness. *Journal of Artificial Intelligence and Consciousness*. Singapore: World Scienti¯c Publishing Company, 2020, **7**(1), 63-72. Dostupné z: doi:10.1142/S2705078520300042
13. PAGE, Jonh, Michael BAIN a Faqihza MUKHLISH. The Risks of Low Level Narrow Artificial Intelligence. *Conference on Intelligence and Safety for Robotics*. Shenyang (CN): IEEE, 2018, **2018**.
14. LABBE, Mark a Ivy WIGMORE. A guide to artificial intelligence in the enterprise: narrow AI (weakAI). *TechTarget* [online]. Newton (MA), 2018, 06-2021 [cit. 2022-06-21]. Dostupné z: <https://www.techtarget.com/searchenterpriseai/definition/narrow-AI-weak-AI>
15. HILDT, Elisabeth. Artificial Intelligence: Does Consciousness Matter?. *Frontiers in Psychology*. Lausanne (CH): Frontiers Media, 2019, **10**(1535). ISSN 1664-1078. Dostupné z: doi:10.3389/fpsyg.2019.01535
16. SILVER, David, Thomas HUBERT, Julian SCHRITTWIESER, Demis HASSABIS et al. AlphaZero: Shedding new light on chess, shogi, and Go. DeepMind [online]. London (GB), 2010, 2018-12-06 [cit. 2022-06-24]. Dostupné z: <https://www.deepmind.com/blog/alphazero-shedding-new-light-on-chess-shogi-and-go>
17. LEGG, Shane. *Machine Super Intelligence*. Lugano (CH), 2008. Disertace. University of Lugano. Vedoucí práce Prof. Dr. Marcus Hutter.
18. GENT, Edd. Artificial intelligence is evolving all by itself: Advance replicates decades of AI research in days. *American Association for the Advancement of Science* [online]. Washington, D.C. (WA), 2020-04-13 [cit. 2022-06-24]. Dostupné z: <https://www.science.org/content/article/artificial-intelligence-evolving-all-itself>
19. TORRES, Phil. The possibility and risks of artificial general intelligence. *Bulletin of the Atomic Scientists*. 2019, **75**(3), 105-108. ISSN 0096-3402. Dostupné z: doi:10.1080/00963402.2019.1604873
20. HINTZE, Arend. Understanding the Four Types of Artificial Intelligence. *Government Technology State & Local Articles* [online]. 2016-11-14 [cit. 2022-06-25]. Dostupné z: <https://www.govtech.com/computing/understanding-the-four-types-of-artificial-intelligence.html>
21. LEMOINE, Blake. We’re All Different and That’s Okay. *Medium* [online]. A Medium Corporation, 2012, 2022-06-22 [cit. 2022-06-25]. Dostupné z: <https://cajundiscordian.medium.com/were-all-different-and-that-s-okay-79096fafa1cd>
22. MCQUILLAN, Laura. A Google engineer says AI has become sentient. What does that actually mean?. *CBC News* [online]. Ottawa (CA), 2022-06-24 [cit. 2022-06-25]. Dostupné z: <https://www.cbc.ca/news/science/ai-consciousness-how-to-recognize-1.6498068>
23. Turing Machines. *Stanford Encyclopedia of Philosophy* [online]. Stanfort (CA), 2018, 2018-09-24 [cit. 2022-06-25]. Dostupné z: <https://www.cs.odu.edu/~zeil/cs390/latest/Public/turing-complete/index.html>
24. The Church-Turing Thesis. *Stanford Encyclopedia of Philosophy* [online]. Stanfort (CA), 2018, 2017-11-10 [cit. 2022-06-25]. Dostupné z: <https://plato.stanford.edu/entries/church-turing/>
25. NEWELL, Allen a Herbert A. SIMON. Computer science as empirical inquiry: symbols and search. *Communications of the ACM*. 1976, **19**(3), 113-126. ISSN 0001-0782. Dostupné z: doi:10.1145/360018.360022
26. SANDBERG, Anders a Nick BOSTROM. *Whole Brain Emulation: A Roadmap*. Oxford (GB): Future of Humanity Institute, Oxford University, 2008.
27. YAMAURA, Hiroshi, Jun IGARASHI a Tadashi YAMAZAKI. Simulation of a Human-Scale Cerebellar Network Model on the K Computer. *Frontiers in Neuroinformatics*. 2020, **14**. ISSN 1662-5196. Dostupné z: doi:10.3389/fninf.2020.00016
28. DORRIER, Jason. Google and Harvard Unveil the Largest High-Resolution Map of the Brain Yet. *Singularity Hub* [online]. Santa Clara (CA): Singularity University, 2008, 2021-06-06 [cit. 2022-06-25]. Dostupné z: <https://singularityhub.com/2021/06/06/google-and-harvard-unveil-the-largest-high-resolution-map-of-the-brain-yet/>
29. NILSSON, Nils J. The Physical Symbol System Hypothesis: Status and Prospects. *50 Years of Artificial Intelligence*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, (4850), 9-17. Lecture Notes in Computer Science. ISBN 978-3-540-77295-8. Dostupné z: doi:10.1007/978-3-540-77296-5\_2
30. FJELLAND, Ragnar. Why general artificial intelligence will not be realized. *Humanities and Social Sciences Communications*. 2020, **7**(1). ISSN 2662-9992. Dostupné z: doi:10.1057/s41599-020-0494-4
31. HAMID, Sobia. *The Opportunities and Risks of Artificial Intelligence in Medicine and Healthcare*. CUSPE Communications, 2016. Dostupné z: doi:10.17863/CAM.25624
32. MUEHLHAUSER, Luke. Intelligence Explosion FAQ. *Machine Learning Research Institute* [online]. Berkeley (CA), 2005, 2015-11-10 [cit. 2022-06-26]. Dostupné z: <https://intelligence.org/ie-faq/>
33. YUDKOWSKY, Eliezer a Anna SALAMON. *Reducing Long-Term Catastrophic Risks from Artificial Intelligence*. San Francisco (CA): Machine Intelligence Research Institute, 2010.
34. CHALMERS, David J. The Singularity: A Philosophical Analysis. *Science Fiction and Philosophy*. Hoboken, NJ, 2016, 2016-01-08, 171-224. ISBN 9781118922590. Dostupné z: doi:10.1002/9781118922590.ch16
35. CHOLLET, François. The implausibility of intelligence explosion. *Medium* [online]. A Medium Corporation, 2012, 2017-11-27 [cit. 2022-06-26]. Dostupné z: <https://medium.com/@francois.chollet/the-impossibility-of-intelligence-explosion-5be4a9eda6ec>
36. HANER, Justin a Denise GARCIA. The Artificial Intelligence Arms Race: Trends and World Leaders in Autonomous Weapons Development. *Global Policy*. 2019, **10**(3), 331-337. ISSN 1758-5880. Dostupné z: doi:10.1111/1758-5899.12713
37. HAMBLING, David. U.S. To Equip MQ-9 Reaper Drones With Artificial Intelligence. *Forbes* [online]. Jersey City (NJ), 1996, 2020-12-11 [cit. 2022-07-01]. Dostupné z: <https://www.forbes.com/sites/davidhambling/2020/12/11/new-project-will-give-us-mq-9-reaper-drones-artificial-intelligence/?sh=252977147a8e>
38. WOEN, Yon Lai, Joo Er MENG, Cheng Ng ZHAN a Wei Goh QI. Semi-autonomous control of an unmanned aerial vehicle. *2016 14th International Conference on Control, Automation, Robotics and Vision (ICARCV)*. IEEE, 2016, 2016, 1-4. ISBN 978-1-5090-3549-6. Dostupné z: doi:10.1109/ICARCV.2016.7838688
39. JOHNSON, James. *Artificial intelligence & future warfare: implications for international security*. 2019, **35**(2), 147-169. ISSN 1475-1798. Dostupné z: doi:10.1080/14751798.2019.1600800
40. RADANLIEV, Petar, David DE ROURE, Carsten MAPLE a Uchenna ANI. Super-forecasting the ‘technological singularity’ risks from artificial intelligence. *Evolving Systems*. ISSN 1868-6478. Dostupné z: doi:10.1007/s12530-022-09431-7
41. KATARYA, Rahul a Anushka LAL. A Study on Combating Emerging Threat of Deepfake Weaponization. *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*. IEEE, 2020, 2020-10-7, 485-490. ISBN 978-1-7281-5464-0. Dostupné z: doi:10.1109/I-SMAC49090.2020.9243588