19.
Roboti za mřížemi – je české trestní právo připraveno na rozvoj umělé inteligence?[350]

Robots behind bars – is Czech criminal law ready for the development of artificial intelligence?

Jiří Mulák / Jan Provazník

Tento příspěvek se zabývá otázkou trestní odpovědnosti za umělou inteligenci a současně také trestní odpovědností umělé inteligence. Autoři nejprve předestřou tři základní modely trestní odpovědnosti. Prvním je koncept umělé inteligence jako nástroje spáchání, druhým konceptem je selhání umělé inteligence jako způsobu naplnění objektivní stránky nedbalostních trestných činů a konečně třetím je pak model přímé a samostatné odpovědnosti umělé inteligence. Předestřené modely jsou pak konfrontovány se základy trestní odpovědnosti. Autoři upozorňují na skutečnost, že se vývoj umělé inteligence ubírá jiným směrem. V tomto ohledu bude potřeba redefinovat určitá paradigmata ohledně základů trestní odpovědnosti.

This paper deals with the issue of criminal liability for artificial intelligence and at the same time also the criminal liability of artificial intelligence. The authors first present three basic models of criminal liability. The first is the concept of artificial intelligence as a tool to commit; the second concept is the failure of artificial intelligence as a way of fulfilling the objective side of negligent crimes, and finally the third is the model of direct and independent responsibility of artificial intelligence. The presented models are then confronted with the basics of criminal liability. The authors point out that the development of artificial intelligence is moving in a different direction. In this respect, certain paradigms regarding the basis of criminal liability will need to be redefined.

Dovolíme si tvrdit, že většina příslušníků právnické obce nepatří současně do odborné obce zabývající se umělou inteligencí.[351]

350 Text byl v rozsahu podílu dr. Jiřího Muláka zpracován v rámci projektu Cooperatio na Univerzitě Karlově, vědní oblasti *Law*.

351 Z české perspektivy lze zmínit příkladmo KOLAŘÍKOVÁ, L. / HORÁK, F. *Umělá inteligence & právo*. Praha: Wolters Kluwer, 2020; ŠŤEDROŇ, B. *Právo a umělá inteligence*. Plzeň: Aleš Čeněk, 2020; KOLAŘÍKOVÁ, L. Odpovědnost (za) robota aneb právo umělé inteligence. *Bulletin advokacie*, 2018, č. 3, s. 11–19; MIKEŠ, S. Právo ve věku inteligentních strojů. *Bulletin advokacie*, 2018, č. 4, s. 17–22; POLČÁK, R. K otázce subjektivity autonomních systémů. In: RONOVSKÁ, K / HAVEL, B. / LAVICKÝ, P. a kol. *Pocta prof. Janu Hurdíkovi k 70. narozeninám. Základní otázky života, práva a vůbec…!* Brno: Masarykova univerzita, Právnická fakulta, 2021, s. 173–184; POLČÁK, R. Odpovědnost umělé inteligence a informační útvary bez právní osobnosti. In: ZOUFALÝ, V. (ed.) *XXVI. Karlovarské právnické dny*. Praha: Leges, 2018, s. 535 a násl.; SMEJKAL, V. / SOKOL, T. Trestněprávní aspekty robotiky – část I. In: *Právní prostor* [online]. Ostrava: ATLAS Consulting, 2018, 26. 9. 2018 [cit. 2022-03-29]. Dostupné z: https://www.pravniprostor.cz/clanky/trestni-pravo/trestnepravni-aspekty-robotiky; SMEJKAL, V. / SOKOL, T. Trestněprávní aspekty robotiky – část II. In: *Právní prostor* [online]. Ostrava: ATLAS Consulting, 2018, 2. 10. 2018 [cit. 2022-03-29]. Dostupné z: https://www.pravniprostor.cz/clanky/trestni-pravo/trestnepravni-aspekty-robotiky-cast-ii.

352 Evropský výbor pro řešení problémů kriminality: Studie možných nástrojů Rady Evropy pro řešení problémů umělé inteligence a trestního práva [CDPC(2020)3Rev]. Štrasburk, 4. 9. 2020 [cit. 2022-03-29].

Náš pohled na ni a souvislosti její právní regulace[352] může být a často i je poměrně zjednodušený či zploštělý. S trochou nadsázky lze říci, že máme tendenci vnímat umělou inteligenci jako něco, co se vyskytuje na ose jednoduchý program schopný obehrát nás v piškvorkách, který ale nic jiného neumí, přes chytrý mobilní telefon, který umí tak složité výpočetní operace, že si na něm můžeme přečíst e-maily, pustit dětem animovaný film, aby nás nechaly chvíli vydechnout, či rozpoznat náš obličej při odemykání či bankovních platbách, až po notebook, který je způsobilý tak komplikovaných výpočetních operací, že nám umožňuje hrát počítačové hry, které jsou graficky již téměř nerozpoznatelné od reality a svou komplexností již dosahují jednoduchých autonomních světů s vlastními pravidly, či osobní automobil, který nás ve výhledu ještě našich životů bude schopen převézt v běžném provozu z místa A do místa B. Osa končí robotem, který bude imitovat člověka ve všech směrech a bude od něj nerozeznatelný, ale toho se my už nedožijeme.

S výjimkou posledních dvou položek, které ovšem stále vnímáme jako hudbu budoucnosti, si současné projevy umělé inteligence zpravidla nespojujeme s úvahami o trestní odpovědnosti, nevnímáme-li je prostě jako nástroj, jímž lze spáchat trestnou činnost. V tomto textu však chceme nastínit, že trestní odpovědnost v souvislosti s umělou inteligencí již v současné době začíná být aktuální a že problémy, které v souvislosti s ní trestní právo může v budoucnu očekávat, mohou a pravděpodobně i budou mít jinou podobu, než jak se dnes jeví.

1. Umělá inteligence – definiční přístup

V tomto příspěvku hojně operujeme s pojmem „umělá inteligence“. Na první pohled se může zdát, že obsah tohoto pojmu je vcelku intuitivní, resp. že jeho definice se bude opírat o poměrně jasná a pevně daná kritéria. Při bližším zkoumání ovšem zjistíme, že věc není tak jednoduchá, jak by se mohlo zdát, neboť definice pojmu „umělá inteligence“ je poměrně fluidní a lze se setkat s mnoha různými přístupy

k jejímu vytvoření. V nejširším slova smyslu lze za „umělou inteligenci“ označit jakoukoliv situaci, v níž dochází k řešení nějakého problému autonomním rozhodovacím procesem, který není vrozeným prvkem ani výsledkem vývoje vrozených dispozic určitého živočišného druhu. Klíčové prvky námi vytvořené pracovní definice jsou (1) autonomie (na rozdíl od pouhého spuštění dopředu naprogramovaného schématu či ovládání jinou inteligencí při provádění úkolu), (2) jiná entita, resp. skutečnost, že rozhodování provádí jiná entita, než pro kterou je proces autonomního řešení problémů přirozený v biologickém slova smyslu, (3) inteligence jako schopnost řešit problémy ve vnějším prostředí interakcí s ním.

Prvním (autonomií) a třetím (inteligencí) prvkem se umělá inteligence odlišuje od pouhé automatizace či robotizace určitého procesu, druhým od inteligence lidské či obecně živočišné (do jisté míry se autonomně rozhodují i zvířata).

Odborná komunita zpravidla však za umělou inteligenci nepovažuje jakýkoliv systém, stroj či entitu, která je schopna se v nějakém, byť nepatrném, rozsahu autonomně rozhodovat, ale až tehdy, jestliže snese určité srovnání s člověkem. Od samotného počátku byly úvahy o umělé inteligenci v moderní době zasazeny právě do tohoto rámce. Již od cca poloviny minulého století se jako referenční rámec pro určení, zda jde, či nejde o umělou inteligenci, uplatňoval tzv. Turingův test, tedy jednoduchá hra pro tři hráče, v níž jeden hráč (člověk) měl určovat, který ze zbývajících hráčů je také člověk a který je strojem, později rozvinutý do mnoha různých dalších variant, např. reverzního Turingova testu, v němž měl počítač určit, který subjekt, s nímž komunikoval, je člověk, a který počítač, případně sofistikovaná verze umělé inteligence v počítačové hře, jejíž herní chování mělo pro lidi být nerozlišitelné od lidského.[353] Umělá inteligence je tak stále často definována např. jako program, který bude v nahodilém světě fungovat nikoliv hůře než člověk,[354] případně jako něco, co je schopno plnit úkoly, které by vyžadovaly inteligenci, kdyby byly vykonávány člověkem.[355] Postupem času se však vývoj umělé inteligence vydal směrem, v němž srovnání s člověkem začíná být irelevantní. Sofistikovanost

353 Srov. např. HERNANDÉZ-ORALLO, J. Twenty Years Beyond the Turing Test: Moving Beyond the Human Judges Too. In: *Minds and Machines*, roč. 30, 2020, s. 535 a násl.

354 Např. DOBREV, D. A Definition of Artificial Intelligence. *Mathematica Balkanica, New Series*, roč. 19, 2005, Fasc. 1–2, s. 67–74.

355 Např. HOFFMANN, H. Is AI Intelligent? An assessment of artificial intelligence, 70 years after Turing. In: *Technology in Society*, roč. 68, 2022, článek č. 101893, s. 2.

a komplexnost lidského myšlení zatím žádná umělá inteligence uspokojivě napodobit nedokáže, avšak v parametrech dílčích, izolovaných, ale ucelených úkolů je dokáže hravě překonat (nejlepší dnešní šachoví velmistři již ani zdaleka nedokáží držet krok s nejlepšími šachovými programy;[356] počítač už dokázal porazit i nejlepšího lidského hráče ve hře Go,[357] v níž je možných mnohem více kombinací než v šachu; primát lidé nedrží ani v pokeru,[358] který často bývá dáván za příklad ještě složitější hry, neboť se odehrává v parametrech neúplných informací – protihráči musí zvolit vhodnou herní kombinaci při vzájemné neznalosti herních možností svých oponentů).

Je tedy zřejmé, že vývoj umělé inteligence nesměřuje zatím dominantně k tomu, aby se stroje vyrovnaly člověku ve všech aspektech jeho života (ostatně takový cíl by nutně vyvolával otázku, do jaké míry by vůbec byl pro lidstvo užitečný jinak než pro uspokojení ega, že něco takového dokážeme), ale aby prohlubovaly schopnost svého výkonu v dílčích izolovaných úkolech daleko za hranicemi lidských schopností. Nejde přitom ani zdaleka již jen o tak (relativně) výpočetně triviální úkoly, jako je výpočet kombinací v šachové partii, která se odehrává ve stabilním, předvídatelném a homogenním systému pravidel, ale i o úkoly natolik komplexní a kombinující nejisté proměnné v různých heterogenních rovinách navíc různého řádu, jako jsou třeba projekty autonomních vozidel. Je však třeba distingovat. Něco jiného je totiž výpočet optimální spotřeby paliva, něco jiného optimální výkon motoru k překonání určité složité nahodilé dopravní situace, něco jiného je výpočet optimální trasy průjezdu v daných podmínkách a ještě něco jiného je predikce chování jiného účastníka provozu (tzv. princip omezené důvěry),[359] který vykazuje nestandardní rysy,

356 Poprvé byl šachový velmistr pokořen v tradiční soutěži umělou inteligencí v roce 1996, kdy Garyho Kasparova porazil počítačový program DeepBlue – viz https://www.theguardian.com/sport/2021/feb/12/deep-blue-computer-beats-kasparov-chess-1996.

357 Stalo se tak v roce 2016, když počítačový program AlphaGo porazil světového mistra v této hře Lee Se-dola – viz https://www.bbc.com/news/technology-35761246.

358 Jako milník se udává rok 2017, když počítačový program Libratus porazil čtyři celosvětové lidské špičky této hry, údajně poslední, v níž ještě lidé počítačům dokázali vzdorovat – viz https://www.reuters.com/article/us-artificialintelligence-poker-idUSKBN15G5NP.

359 Nejvyšší soud již dříve vymezil definici tohoto principu např. v usnesení Nejvyššího soudu sp. zn. 6 Tdo 143/2011 ze dne 29. listopadu 2011, které uvádí: „Zásada tzv. omezené důvěry v dopravě (vyjádřená např. v rozhodnutí č. 43/1982 Sb. rozh. tr.), na kterou obviněný v dovolání upozorňuje, znamená, že řidič motorového vozidla může spoléhat na dodržení dopravních předpisů

přičemž všechny tyto aspekty musí být autonomní vozidlo schopno sloučit v jednu optimální jízdní strategii. Ačkoliv i u takto komplikovaného a komplexního úkolu umělá inteligence nepochybně dokáže předčit člověka, nebude už schopna současně alespoň srovnatelně s lidským průměrným výkonem vyřešit i další obdobně komplexní úkoly, jako je odstranění partnerského konfliktu, výchova potomstva či rozhodování o strategii tvorby a zhodnocení rodinných úspor.

Obecně lze rozlišit čtyři druhy (současně vlastně i generace) umělé inteligence: (1) reaktivní (*reactive*), (2) s omezenou pamětí (*limited memory*), (3) s teorií mysli (*theory of mind*), a (4) s vědomím sebe sama (*self-aware*).[360] Reaktivní umělá inteligence je schopna pouze podle naprogramovaných algoritmů řešit v reálném čase předem definované problémy. Nemá tak vědomí času, neučí se podle předchozích výsledků své činnosti, pouze uplatňuje svoji jednou pro vždy danou kapacitu. Příkladem může být již zmíněný DeepBlue, počítač, který dokázal porazit šachového velmistra. Nebyl si tak vědom, že hraje nějakou hru a každý svůj tah vyhodnocoval zvlášť bez jakéhokoliv přihlédnutí k tomu, jak táhl v předchozích kolech či jak v předchozích tazích táhl jeho protihráč.

Umělá inteligence s omezenou pamětí je schopna sama sebe aktualizovat na základě předchozích řešení. Je tedy schopna se v omezeném rozsahu učit na podkladě svých minulých výkonů a optimalizovat tak své výkony do budoucna. Většina moderních modelů umělé inteligence v současné době patří právě do této kategorie (kupř. autonomní vozidla či všechny systémy, které fungují na principu tzv. *deep learning*).

Teorie mysli je v současné době toliko teoretický, byť v praxi již rozostatními účastníky provozu na pozemních komunikacích, nevyplývá-li z konkrétní situace opak. Tato zásada se ale neuplatňuje v případech, kdy ze situace v provozu na pozemních komunikacích vyplývá povinnost dbát zvýšené opatrnosti nebo s předstihem reagovat na situaci, aby bylo zabráněno kolizi (na komunikacích nebo v jejich blízkosti se pohybují děti, osoby těžce zdravotně postižené, přestárlé, zjevně volně pobíhající zvířata nebo to vyplývá z existence instalovaných dopravních značek). Řidič motorového vozidla je povinen zachovávat potřebnou míru opatrnosti vůči chodcům, kteří vstoupili do vozovky nebo se pohybují v její těsné blízkosti. Důvodně spoléhat na to, že tito účastníci silničního provozu dodrží pravidla silničního provozu, může jen v případě, pokud z konkrétních okolností neplyne obava, že tomu tak nebude. Mohou však nastat situace, kdy i chodec vytvoří řidiči svým náhlým neočekávaným a nepředvídatelným vstoupením do vozovky překážku, jež může být pro řidiče i objektivně nezvládnutelná. V posuzované trestní věci se ale o takovou situaci nejedná."

360 Srov. např. JOSHI, N. 7 Types of Artificial Intelligence. In: Forbes [online]. Dostupné z: https://www.forbes.com/sites/cognitiveworld/2019/06/19/7--types-of-artificial-intelligence/.

víjený koncept. Umělá inteligence s teorií mysli by musela být schopna rozpoznávat člověka jako jednotlivce a chápat všechny složky jeho vnitřního života jako jsou emoce, postoje, představy atd.

Nejvyšším a v praxi zatím ani teoreticky nedosažitelným modelem je umělá inteligence, která si uvědomuje sama sebe. Oproti teorii mysli by tak umělá inteligence vnímala jako individualitu nejen jednotlivce ve vnějším světě, ale i sebe sama.

Druhým vcelku dobře uchopitelným přístupem je rozdělení umělé inteligence na úzkou umělou inteligenci (*narrow AI*), obecnou umělou inteligenci (*general AI*) a umělou superinteligenci (*artificial superinteligence*).[361] Úzká umělá inteligence je podle svého názvu schopna řešit jen úzký obor aktivit, při nichž člověk obecně užívá inteligenci, tedy neimituje všechny myšlenkové či dokonce duševní pochody člověka. Jde tedy svým způsobem o „fachidiota", který je schopen již v současnosti v úzkém oboru své činnosti člověka předčit (např. spočítat dráhu vesmírného letu), avšak není schopen učinit nic jiného (mít ze své práce radost, uvědomovat si její význam pro lidstvo atd.).

Obecná umělá inteligence je dosud toliko teoretickým konceptem, jehož měřítkem je člověk – měla by tedy být schopna všeho, čeho je schopen on.

Umělá superinteligence je pak předpokládaným dalším krokem ve vývoji, tedy bude schopna vykonávat vše, co vykonává člověk, lépe než on.

Pro úplnost na tomto místě můžeme dodat, že byť koncepty obecné umělé inteligence i umělé inteligence jsou nepochybně přitažlivé (a možná i dost děsivé) a jistě nelze vyloučit, že se v budoucnu dočkáme i jejich realizace, přinejmenším v současné době se nezdá, že půjde o prioritní směr vývoje umělé inteligence. Schopnost vytvořit dokonalou umělou kopii člověka je totiž sice nepochybně lákavá a nelze pochybovat o tom, že se o to někdo pokusí byť i jen z prestižních důvodů, ovšem smysluplná ekonomická kauza takového počinu v zásadě chybí. S jen mírnou hyperbolou lze říci, že planeta Země se určitě nepotýká s nedostatkem lidí, takže astronomické náklady na vyvinutí a výrobu funkčního umělého exempláře jsou stále dosti zbytečnou investicí, jejíhož přínosu lze s nepoměrně menšími náklady dosáhnout prostě tím, že na daný úkol bude najat jako pracovní síla člověk.

Jako mnohem pravděpodobnější alespoň v krátkém a střednědobém horizontu lze očekávat prudký rozvoj úzké umělé inteligence až do

361 Srov. např. JOSHI, N. 7 Types of Artificial Intelligence. In: Forbes [online]. Dostupné z: https://www.forbes.com/sites/cognitiveworld/2019/06/19/7--types-of-artificial-intelligence/.

podoby, již označujeme jako úzkou umělou „superinteligenci“, tedy do podoby takové umělé inteligence, která sice zůstane ve svých jasně definovaných omezených mantinelech, avšak její schopnosti v těchto mezích dosáhnou takové úrovně pokročilosti, že už ani ti nejinteligentnější lidé s ní neudrží krok a bude zcela mimo jejich chápání, co taková úzká superinteligence vlastně dělá. Jelikož půjde stále o úzkou umělou inteligenci, nebude možno korekci možných negativních důsledků její činnosti svěřit ani nějakým jejím vlastním morálním korektivům, neboť ty by se nacházely zcela mimo její úzce vymezené pole činnosti. K problémům, které tento možný model bude přinášet pro sféru práva obecně a trestního práva zvláště, se ještě vrátíme, neboť je považujeme v současné době za ty nejpalčivější, které paleta palčivých problémů umělé inteligence přináší.

2. Modely trestní odpovědnosti – nástin problematiky

S ohledem na výše nastíněné by se tak daly současné představy o umělé inteligenci s jistou mírou zjednodušení promítnout v zásadě do tří modelů trestní odpovědnosti: (1) umělá inteligence jako nástroj spáchání, (2) selhání umělé inteligence jako způsob naplnění objektivní stránky nedbalostních trestných činů, (3) umělá inteligence jako subjekt trestního práva hmotného.

3. Umělá inteligence jako nástroj spáchání (první model)

V prvém modelu je umělá inteligence toliko nástrojem jakožto složkou objektivní stránky trestného činu, její využití je tedy toliko způsob, jímž pachatel jedná.[362] Relevance umělé inteligence z hlediska trestní odpovědnosti je tak stejná, jako v případě užití jakéhokoliv jiného nástroje. Vliv na právní kvalifikaci tak může mít jen z hlediska naplnění některých fakultativních znaků skutkové podstaty (např. spáchání činu se zbraní), z hlediska právních následků může představovat toliko obecně přitěžující okolnost (např. kvůli zvláštní sofistikovanosti provedení činu či obsáhlé přípravě), případně může umožňovat uložení některých sankcí proti majetku (trest propadnutí věci podle § 70 tr. zákoníku, ev. ochranné opatření zabrání věci podle § 101 tr. zákoníku). Půjde zejména o ty případy, v nichž člověk naplňuje objektivní stránku trestného činu sám s využitím umělé inteligence tak, že ji ovládá, byť by jejím prostřednictvím prováděl i dílčí operace, jichž by sám člověk nebyl schopen (například zde může být např. počítačový hacker ovládající škodlivý software, jehož prostřednictvím se snaží prolomit zabezpečení a získat přístup k datům; lékař, který při operaci využívá sofistikovaného robota; statik, který s pomocí počítače provádí složité výpočty ohledně budovy; programátor, který prostřednictvím škodlivého software vytváří za účelem spáchání pomluvy tzv. *deep fake*).[363] Tu lze umělou inteligenci považovat toliko za nástroj. Tento model by pokrýval jak případy trestných činů úmyslného zneužití umělé inteligence ke spáchání trestného činu, tak trestných činů kulpózních, v nichž by se zavinění ve formě nedbalosti pachatele vztahovalo k neodbornému použití umělé inteligence. Oproti druhému modelu (viz níže) by tak umělá inteligence fungovala bezchybně a při správném použití by ke škodlivému následku nedošlo, avšak k pochybení došlo při jejím ovládání pachatelem (např. lékař využívající operačního robota by si řádně neprostudoval manuál a bez potřebného proškolení by zadal robotovi špatný pokyn, v jehož důsledku by došlo k ublížení na zdraví).

Obdobně by pod tento model bylo možno zařadit případy, v nichž by umělá inteligence představovala samostatně fungující nástroj ke spáchání dle úmyslu pachatele (níže zmíněný případ s automatickým rozesíláním škodlivého obsahu na sociálních sítích, naprogramování dronu tak, že bude schopen po aktivaci identifikovat cíl, naletět na něj a odpálit nálož, která je k němu připojena). I v takovém případě je umělá inteligence toliko nástrojem, podobně jako výše uvedené cvičené zvíře či třeba důmyslně zkonstruované odpalovací zařízení.

Tento model (podobně jako model následující) vychází z předpokladu, že umělá inteligence nemá lidské atributy. Hledáme tedy trestně odpovědného pachatele. Umělá inteligence je zde chápána jako nevinný zprostředkovatel (*innocent agent*), u kterého absentují obecné znaky (věk, příčetnost, případně rozumová a mravní vyspělost). Podle stávající právní úpravy bude umělá inteligence pojímána jako věc. S ohledem na to, že umělá inteligence bude patrně nadána i určitou „vůlí“, bude zde možné vést určitou paralelu se zvířetem (§ 134 tr. zákoníku) a bude tak *mutatis mutandis* použitelná rozhodovací praxe

362 Lze se setkat i s označením modelu odpovědnosti „spáchání prostřednictvím jiného“ (the Perpetration-via-Another Liability Model).

363 Deep fake je označení pro realistickou úpravu videa – především tváří zobrazených osob – která umožňuje např. velmi věrohodně změnit mimiku, tedy i samotnou řeč jednotlivých aktérů videa. V praxi tak můžeme osobám na videu vkládat do úst věty, které nikdy nepronesly, provádět záměny postav, obličejů atd. Deep fake využívá pokročilého počítačového zpracování dat s využitím umělé inteligence (využívá neurální sítě se schopností učit se) a kvalita výsledného produktu – upraveného videa – se neustále zvyšuje.

ohledně poštvaných zvířat, byť zcela jistě sofistikovanějším (důmyslnějším) způsobem.

Ilustrativním případem může být situace, kdy určitá fyzická osoba (či právnická osoba prostřednictvím právnické osoby) vytvoří nebo si pořídí umělou inteligenci a této později uloží (nařídí), aby vykonala skutek, který bude vykazovat znaky trestného činu. Takto si lze např. představit, že umělé inteligenci bude nařízena kontinuální kontrola příspěvků na sociálních sítích s tím, že pokud se objeví nějaký příspěvek (komentář) od osoby odlišného vyznání či rasy, aby umělá inteligence bezprostředně zahájila „kulometnou palbu" komentářových příspěvků, které budou nenávistné, zastrašující či jinak dehonestující,[364] případně aby např. pod příspěvky ohledně trestného činu agrese[365] publikovala popírající, zpochybňující, schvalující nebo ospravedlňující komentáře.[366] O jakých osobách však bude možné uvažovat jako o pachatelích? Patrně budeme uvažovat o trestní odpovědnosti programátora či o trestní odpovědnosti uživatele umělé inteligence, případně v různých modifikacích. Trestní odpovědnost internetového providera je pak dovozována z ustanovení § 112 tr. zákoníku, stanovícího prameny zvláštní povinnosti konat, zakládající odpovědnost za následek.[367]

4. Selhání umělé inteligence jako způsob naplnění objektivní stránky nedbalostních trestných činů (druhý model)

Ve druhém modelu by určitá fyzická či právnická osoba byla trestně odpovědná za selhání umělé inteligence, které nedbalostně zavinila. Šlo by tedy o model postihující nežádoucí následky použití „vadně fungující" umělé inteligence, kterým bylo možno při zachování potřebné míry opatrnosti předejít, avšak nestalo se tak. Tento model by tak v zásadě byl totožný jako současná trestněprávní reakce na společensky škodlivé následky, k nimž došlo složitým kauzálním průběhem při porušení potřebné míry opatrnosti, zejména v důsledku nedodržení potřebných bezpečnostních norem, profesionálních standardů, zvláštní povinnosti péče atd.[368]

364 V úvahu bude připadat právní kvalifikace podle § 355 odst. 1, 2 písm. b) tr. zákoníku.

365 Toto může být relevantní v souvislosti s invazí Ruska na Ukrajinu v rámci hybridní a informační války.

366 V úvahu bude připadat právní kvalifikace podle § 405 tr. zákoníku.

367 K tomu blíže ŠÁMAL, P. a kol. *Trestní zákoník. Komentář. I. díl.* Praha: C. H. Beck, 2012, s. 1260 – 1263.

368 Například zřícení mostu v důsledku chyby v projektové dokumentaci; dopravní nehoda v důsledku chyby při provádění servisu vozidla; průmyslová havárie v důsledku zanedbání povinné periodické bezpečnostní kontroly, která by odhalila předčasné a snadno řešitelné opotřebení materiálu, kvůli němuž k havárii došlo, atd.

Odlišností tohoto modelu oproti modelu předchozímu je to, že v úvahu připadající trestně odpovědné osoby nemají úmysl spáchat prostřednictvím umělé inteligence, ale jsou současně angažováni například na kontrole, dohledu, dozoru, servisu umělé inteligence. V případě autonomních vozidel může jít příkladmo o řidiče, homologátora, výrobce auta, dodavatele hardware a dodavatele software.

Bavíme se tedy o nedbalostním zavinění. Měřítkem (kritériem) nedbalosti je zachovávání určité míry opatrnosti. Nedbalostně jedná ten, kdo: (a) nedodržuje míru opatrnosti, ke které je v rámci okolností povinen, a (b) podle svých subjektivních možností je také schopen ji dodržovat. Míru požadované opatrnosti však žádný trestní zákon pochopitelně nestanoví. Pouze jen naznačuje, jak ji zjistit. Trestněprávní nauka i aplikační praxe pak vysvětlují, že tato míra opatrnosti je dána spojením objektivního a subjektivního vymezení, a jaký je jejich vzájemný vztah.[369] Nedbalostně tedy jedná ten, kdo nedodržuje míru opatrnosti, ke které je v rámci okolností povinen (objektivní hledisko) a podle svých možností schopen (subjektivní hledisko). Odpovědnost za nedbalost je dána spojením hlediska objektivního a subjektivního.[370] [371] Uvedené patrně předpokládá (ve vztahu k objektivnímu hledisku) rejstřík práv a povinností, jak s umělou inteligencí nakládat.

Tento model by vzhledem ke své podstatě byl využitelný pro poměrně široké spektrum v úvahu připadajících situací až po škodlivé

369 SOLNAŘ, V. *Základy trestní odpovědnosti*. Praha: Academia, 1972, s. 237; Srov. též č. 5/2013 Sb. rozh. tr.

370 JELÍNEK, J. a kol. *Trestní právo hmotné. Obecná část. Zvláštní část*. Praha: Leges, 2022, s. 244 a násl.

371 Tomu konvenuje i navazující myšlenka, že „Hranice okolností, jež pachatel může či nemůže předvídat, nelze vymezovat jen v hypotetické rovině (neboť pak by musel každý předvídat v podstatě cokoliv), ale je zapotřebí vždy vycházet z existujících objektivních okolností vyplývajících z určité životní situace, která může být charakterizována celou řadou faktorů, jež pachatel vnímá svými smysly a může je pak hodnotit podle svých znalostí i dalších subjektivních dispozic. Z hlediska nedbalostního zavinění (§ 5 trestního zákona) to znamená, že kromě míry povinné opatrnosti vyplývající z obecných pravidel bezpečného chování zde existuje i subjektivní vymezení, které spočívá v míře opatrnosti, kterou je pachatel schopen vynaložit v konkrétním případě. Přitom o zavinění z nedbalosti může jít jen tehdy, pokud povinnost a možnost předvídat porušení nebo ohrožení zájmu chráněného trestním zákonem jsou dány současně." z usnesení Ústavního soudu sp. zn. II.ÚS 728/02 ze dne 20. května 2004 (U 33/33 SbNU 529).

následky způsobené umělou inteligencí v podobě téměř autonomních robotů. V úvahu by připadal jak pro již dnes řešené situace technicky složitých kauzálních dějů (havárie v komplikovaném výrobním provozu), v nichž může trestně odpovědných být vícero subjektů každý za „svou" nedbalost (ředitel provozu za nedostatečnou kontrolu, mistr za nenahlášení rizikového faktoru, dodavatel zabezpečovacích prvků za jejich nižší než předepsanou účinnost atd.), tak pro situace zanedbání dozoru nad fakticky relativně samostatnými, právně však nesamostatnými aktéry (ublížení na zdraví útokem zaběhlého psa, jehož majitel nedbalostně dopustil jeho útěk; způsobení dopravní nehody dezorientovanou nesvéprávnou osobou, nad níž v té době měl někdo vykonávat dozor, ale nedbalostně dopustil její útěk atd.).

Přenositelný by tak byl docela snadno aplikovatelný i na obdobné situace, v nichž by složitý kauzální průběh či relativně autonomní aktér byl představován právě autonomním systémem.

5. Umělá inteligence jako subjekt trestního práva hmotného

Třetí model si lze hypoteticky představit, jak shora nastíněno, po uznání právní osobnosti umělé inteligence v budoucnu. Patrně poměrně jednoduše by to bylo možné tam, kde by šlo o umělou inteligenci tzv. typu android (co nejdokonalejší imitace člověka),[372] která by byla vybavena ve všech ohledech nejen možnostmi pohybu, ale zejména i psychickými a emocionálními vnitřními procesy relevantními pro trestní právo (složka rozumová i volní, vnitřní ustálený hodnotový systém, morální referenční rámec atd.). V takovém případě by bylo možno uplatňovat trestní odpovědnost na takovou umělou inteligenci stejně jako na fyzickou osobu.

Tento model tedy nepředpokládá odpovědnost jiné osoby (jako je tomu u předchozích dvou modelů), která může být v současné době trestně odpovědná, ale zaměřuje se přímo na vlastní trestní odpovědnost umělé inteligence. Výhodou tohoto modelu je určitá eliminace situací, kdy by za určité společensky závažné jednání umělé inteligence, na které nemá fyzická osoba žádný vliv (umělá inteligence jednala na základě vlastních výpočtů a zhodnocení situace), byla odpovědná jiná osoba nebo také vůbec nikdo. Příkladem budiž srážka autonomní vozidla s „běžným" vozidlem, které do křižovatky vjede na oranžovou barvu semaforu, a způsobení smrtelného zranění řidiče „běžného" vozidla.

372 Označení „*Android*" pochází z řeckých slov „*andros*" (muž, člověk) *a „-eides* (stejného druhu, podobný).

Nevýhodou tohoto modelu by byla patrně jednak podstatná změna paradigmatu (ohledně subjektivity umělé inteligence), jednak potřeba důsledné novelizace soukromoprávních i trestních předpisů.

Zjednodušeně řečeno to předpokládá subjektivitu umělé inteligence, tedy stanovení určitých práv a povinností. Jde tedy o vytvoření třetí formy osoby – elektronické osoby, která by stála na pomezí mezi fyzickou a právnickou osobou. Odlišnost od fyzické osoby by spočívala v tom, že umělá inteligence je uměle vytvořená a postrádá kvality lidské bytosti. Současně by nebyla tak „prázdná" jako právnická osoba, která je sice nositelem práv a povinností dle českého právního řádu, ale fakticky vlastním jednáním se nemůže k právům a povinnostem zavazovat – jedná za ni fyzická osoba. Právě tím, že umělá inteligence je schopna na základě podnětů a vlivu jevů z okolního světa utvářet vlastní jednání, by se zase výrazně odlišovala od právnické osoby. Mikeš v této souvislosti navrhuje, aby v právním řádu byla určitá definice umělé inteligence, kdy by mohlo jít např. o „útvar odlišný od člověka nadaný schopností samostatně rozhodovat a jednat, potažmo jako umělý útvar nadaný právní osobností od svého vzniku do svého zániku, obdobně jako je tomu u právnické osoby",[373] kdy z hlediska časového, tj. otázky vzniku a zániku by pak mohla být řešena např. zápisem do rejstříku umělé inteligence, obdobně jako u právnických osob.

Přiznání subjektivity umělé inteligence a připuštění její přímé trestní odpovědnosti nás přivede na několik dalších úvah a myšlenek. Bude případně správné distingovat trestní odpovědnost v závislosti na vývojové generaci umělé inteligence? Je možná nějaká forma trestné součinnosti, tj. zejména účastenství (včetně spolupachatelství) u dvou trestně odpovědných umělých inteligencí nebo trestně odpovědné umělé inteligence a trestně odpovědné fyzické či právnické osoby? Bylo by žádoucí zavést další typizovanou okolnost vylučující protiprávnost v kontextu posuzování situací, pokud by umělá inteligence měla určitou virovou nálož? Bylo by to důvodem pro zavedení okolnosti vylučující protiprávnost či nějaké obdoby trestného činu „opilství" dle § 360 odst. 1 tr. zákoníku? V návaznosti na to, jestliže by umělá inteligence měla virus (např. od jiné agresivní umělé inteligence) a současně by to chápáno (konstruováno) jako okolnost vylučující protiprávnost, bylo by možné onu agresivní umělou inteligenci označit jako nepřímého pachatele (§ 22 odst. 2 tr. zákoníku)? Mohla by umělá inteligence projevit svoji účinnou lítost? Nebo z jiného po-

373 MIKEŠ, S. Právo ve věku inteligentních strojů. *Bulletin advokacie*, 2018, č. 4, s. 18.

hledu a subsystému trestního práva, mohla by prohlásit, že spáchala určitý skutek, doznat se ke spáchání trestného činu či prohlásit svoji vinu? Jaké druhy sankcí (trestů) by bylo možné umělé inteligenci uložit, aby byl zachován tradičně pojímaný účel sankce (trestu)?

Dosud jsme řešili otázku subjektu (trestně odpovědného pachatele – umělé inteligence). Otázku objektu trestného činu (právem chráněných zájmů; tj. otázku věcné působnosti trestních zákonů), resp. výčet trestných činů, kterých se umělá inteligence může (pozitivní taxativní výčet) či nemůže dopustit (negativní taxativní výčet), případně konstituování nových trestných činů v tomto ohledu ponecháváme stranou.

Přejděme nyní k otázce objektivní stránky (zejména projevení samostatné vůle jednat) a subjektivní stránky (zavinění) trestného činu. Zde předesíláme, že jestliže připustíme možnost projevení i samostatné „vůle" umělé inteligence, bude koncepce trestní odpovědnosti spíše tendovat k trestní odpovědnosti fyzických osob než k trestní odpovědnosti právnických osob.

Ohledně objektivní stránky je klíčovou otázkou to, zda je umělá inteligence schopna samostatně jednat, aniž by ji někdo (fyzická osoba) dala určitý výchozí pokyn. Jde tedy o schopnost umělé inteligence mít svoji vlastní vůli a tuto posléze projevit ve vnějším světě. Zde především musíme opustit bezvýjimečnou ideu, že vůli může projevovat jen fyzická osoba, neboť např. trestní odpovědnost právnických osob je řešena pomocí konceptu přičitatelnosti jednání určité fyzické osoby za právnickou osobu (§ 8 odst. 1, 2 ztopo). Koncept přičitatelnosti u tohoto modelu trestní odpovědnosti umělé inteligence je však značně problematický. Předně totiž umělé inteligence (tím spíše vyšší generace)[374] budou schopné jednat i na základě vlastních podnětů, které re-

374 V průmyslu se rozlišují různé úrovně automatizace řízení (norma SAE J3016_201401). Při úrovni „0" vozidlo ovládá pouze řidič (tj. bez systému moderních asistencí), při úrovni „1" (hands on) řidič a automatický systém sdílejí kontrolu nad vozidlem (adaptivní tempomat či parkovací asistent, kdy je otáčení kol řízeno automaticky, zatímco rychlost ovládá řidič; při úrovni „2" (hands off) automatizovaný systém plně ovládá vozidlo (zrychlování, brzdění a řízení). Řidič však musí sledovat řízení a být připraven k okamžitému zásahu, pokud systém nereaguje správně; při úrovni „3" (eyes off) řidič může bezpečně odvrátit pozornost od jízdních úkolů, např. psát textové zprávy nebo sledovat film. Vozidlo zvládne situace vyžadující okamžitou reakci, jako je nouzové brzdění. Řidič však musí být i nadále připraven k zásahu během určitého časového limitu, který specifikuje výrobce; při úrovni „4" *(mind off)* je to obdobné jako u úrovně 3, avšak pozornost řidiče již není nutná vůbec, takže řidič může bezpečně jíst, spát nebo opustit sedadlo řidiče. Platí, že autonomní jízda je podporována pouze ve vymezených oblastech nebo za zvláštních okolností, jako jsou dopravní zácpy. Mimo tyto oblasti nebo okolnosti musí být vozidlo schopné bezpečně přerušit jízdu, tj. zaparkovat, pokud se řidič neujme kontroly nad vozidlem. Při úrovni „5" se autonomní systém plně věnuje řízení bez jakéhokoliv zásahu ze strany lidského faktoru.

gistrovaly při směřování za určitým cílem. Dále je sporné i definování odpovědných osob, ale i originární složka objektivní stránky („v rámci činnosti umělé inteligence" či „v zájmu umělé inteligence"). Řešení by bylo možné spatřovat v tom, pokud by interní procesy v rámci programu umělé inteligence byly koncipovány jako „uměle vytvořená vůle".

Pokud jde o projevení vůle umělé inteligence, jako druhé složky jednání, bude zde potřeba, aby umělá inteligence pohnula některou svojí částí fyzické (pevné) složky umělé inteligence a daný pohyb bude v přímém a příčinném vztahu s následkem, tedy s porušením (poruchové delikty) či ohrožením (ohrožovací delikty) právem chráněného zájmu. Tímto však odhlížíme od kybernetických trestných činů, kdy se následek (a zpravidla i jednání) projevují ve virtuálním prostoru.[375] Z hlediska jednání pak rozeznáváme komisivní trestné činy a omisivní trestné činy. Výše popsané bude patrně dopadat na komisivní trestné činy. U omisivních trestných činů pak postačí, aby umělá inteligence sama (nikoliv na základě určitého příkazu fyzické osoby) vyhodnotila určitou situaci a na ni reagovala klidovým režimem (opomenutí přikázaného chování), což by mělo za následek porušení nebo ohrožení objektu trestného činu. Rozlišování pravých a nepravých omisivních trestných činů nás vede k úvaze, zda by umělá inteligence měla obecnou a zvláštní povinnost konat. To by pravděpodobně organicky souviselo s otázkou rejstříku práv a povinností umělé inteligence.

Subjektivní stránka trestného činu je dalším typovým znakem trestného činu. Zavinění, jako obligatorní znak subjektivní stránky, je tradičně vykládáno jako „vnitřní (psychický) vztah člověka k určitým skutečnostem, jež zakládají trestný čin, ať již vytvořeným pachatelem nebo objektivně existujícím bez jeho přičinění již v době činu".[376] České trestní právo rozeznává dvě základní formy zavinění, úmysl a nedbalost, z nichž každá má dvě odvozené formy (úmysl přímý a úmysl nepřímý; nedbalost vědomá a nedbalost nevědomá). V rámci těchto forem můžeme tzv. formy subodvozené (rozmysl; předchozí uvážení; hrubá nedbalost).[377] Jako nejproblematičtější se z našeho pohledu *pri-*

375 K tomu nejnověji blíže SMEJKAL, V. *Kybernetická kriminalita*, 3. vydání. Plzeň: Aleš Čeněk, 2022.

376 JELÍNEK, J. a kol. *Trestní právo hmotné. Obecná část. Zvláštní část*. Praha: Leges, 2022, s. 230.

377 KRATOCHVÍL, V. *Trestní právo hmotné. Obecná část.* Praha: C. H. Beck, 2012, s. 298 a násl.

ma vista jeví zejména to, zda je umělá inteligence schopna si vytvořit určitý psychický vztah ke skutečnostem, které zakládají trestný čin. Jde tedy o otázku emocí. Může mít umělá inteligence emoce? Je možné ve vztahu k umělé inteligenci chápat onu absenci emocí (chladný kalkul / formální „uvažování“) jako „psychiku“ *sui generis species*? Další těžkostí je otázka procesu dokazování subjektivní stránky trestného činu (§ 89 odst. 1 tr. řádu), které bude rovněž odvislé od generace umělé inteligence.[378]

S problematikou zavinění souvisí otázka omylu v trestním právu. Omyl v trestním právu je tradičně vykládán jako neshoda pachatelova „vědění“ (jeho vnímání, jeho představy) se skutečností, ať již proto, že si pachatel vůbec neuvědomil nějakou skutečnost, nebo proto, že měl o ní představu nesprávnou. Tato neshoda se může týkat jak okolností skutkových, tak ustanovení právních. Z tohoto hlediska distingujeme omyl skutkový[379] a omyl právní. Z jiného hlediska (v jakém směru se člověk mýlí, zda si o relevantní okolnosti myslí, že není či je) rozlišujeme omyl negativní (osoba nezná okolnost podmiňující trestní odpovědnost) a omyl pozitivní (osoba se mylně domnívá, že taková okolnost tu je). Zde se nabízí to, zda umělá inteligence může jednat v omylu, resp. zda se v jejím procesoru může odehrát určitá neshoda.

6. Kritika (nedostatečnost) modelů trestní odpovědnosti

Shora naznačené rozdělení modelů trestní odpovědnosti by však bylo nedostačující, protože je příliš simplicistní. Výše uvedená aktuální i výhledová časová osa totiž neodpovídá zcela recentnímu přístupu k vývoji v umělé inteligenci. Dobře však demonstruje, že tento pojem je poměrně široký a v představách neodborné veřejnosti nestejnoměrně ohraničený.

Z hlediska posuzování trestní či jakékoliv jiné právní odpovědnosti za škodlivý následek způsobený umělou inteligencí je tak třeba zvažovat míru nejen její sofistikovanosti, ale rovněž jejího přiblížení člověku, neboť v obojím lze nalézt značnou míru variability. Nepochybně

378 HALLEVY, G. The Criminal Liability of Artificial Intelligence Entities – from Science Fiction to Legal Social Control. *Akron Intellectual Property Journal* [online]. 2010, (Vol. 4: Iss. 2, Article 1.), s. 171–201 [cit. 29. 3. 2022]. Dostupné také z: https://ideaexchange.uakron.edu/cgi/viewcontent.cgi?article=1037&context=akroniNtellectualproperty, s. 187.

379 Za zvláštní případy skutkového omylu jsou omyl v předmětu útoku (*error in objecto – in personam*), aberace (odchýlení rány) a *dolus generalis* (generální úmysl).

i v budoucnu bude dále existovat a rozvíjet se i umělá inteligence toliko jako součást nástrojů sloužících k přímému a samostatnému použití člověkem k vykonání více či méně náročných úkolů (a to až do úrovně úzké umělé „superinteligence“ – viz výše). V těchto případech tak stále bude možno shora uvedené první dva modely trestní odpovědnosti.

6.1 Model samostatné trestní odpovědnosti umělé inteligence a jeho kritika

Problematický je však již sám o sobě třetí model, který by fungoval logicky jen na dokonalé lidské kopie, přičemž na určité vysoce sofistikované, ale člověku zcela nepodobné formy umělé inteligence nebude postačovat ani jeden z nich, a dokonce si lze představit, že umělá inteligence může hypoteticky značně fatálně nabourat i stávající „tradiční“ model trestní odpovědnosti aplikovatelný na fyzické osoby.

Nejprve k prvému problému. Model samostatné trestní odpovědnosti identický s tím, který postihuje lidi (fyzické osoby), by smysl dával jenom tehdy, kdyby byly splněny dva zamlčené předpoklady.

Prvním z nich je, že androidé by se lidem vyrovnali, ale v trestněprávně relevantních charakteristikách je nepředčili. Pokud by takový android jednou měl několikanásobně vyšší inteligenci než nejchytřejší člověk na světě, případně byly-li by vylepšeny i jeho morálně-volní vlastnosti (či by nebyly vylepšeny, ale prostě by fungovaly jinak), samo zavinění jako esenciální předpoklad trestní odpovědnosti (*nullum crimen sine culpa*) by byl zpochybněn, resp. musel by být „přebudován“ tak, aby vyhovoval androidům. Takové změny paradigmatu by pak byli schopni patrně zase jen androidé. I když půjdeme v našich úvahách dále (pustíme uzdu fantazie zcela na volno), již kvůli vstupní premise, že by intelektové a morálně volní schopnosti androidů dalece převyšovaly ty lidské, se nemůžeme odvážit ani odhadnout, jak by takový systém vypadal.

Druhým zamlčeným předpokladem je, že by si takto pokročilé formy umělé inteligence zachovávaly vlastní identitu v tom smyslu, jak ji vnímáme u lidí. Ačkoliv se v průběhu života každý z nás mění a podléháme i momentálním vlivům vnitřním i vnějším, myšlenka, že každý člověk je kontinuálně progresivní identita,[380] je neoddělitelná od

380 V tom smyslu, že byť se naše já v čase mění, takže na začátku i na konci života může být naprosto rozdílné ve všech aspektech, s nimiž si zpravidla spojujeme lidskou jedinečnost, rozdíl mezi naším dnešním a zítřejším já je tak nepatrný, že sami sebe vnímáme, a i svým okolím jsme vnímáni jako jedna a tatáž osoba.