Univerzita Pardubice

Fakulta filozofická

Existenční hrozby obecné umělé inteligence řídící se vlastními cíli

Bakalářská práce

2025 David Walter

Univerzita Pardubice
Fakulta filozofická
Akademický rok: 2023/2024

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

(projektu, uměleckého díla, uměleckého výkonu)

Jméno a příjmení:	**David Walter**
Osobní číslo:	**H22169**
Studijní program:	**B0223A100005 Filosofie**
Specializace:	**Filosofie**
Téma práce:	**Existenční hrozby obecné umělé inteligence řídící se vlastními cíli**
Zadávající katedra:	**Katedra filosofie a religionistiky**

Zásady pro vypracování

Jak vznikají a formují se vlastní cíle umělé inteligence? Jakými způsoby si může umělá inteligence sama generovat, upravovat a řídit své cíle? Vlastními cíli zde rozumíme takové cíle umělé inteligence, které vznikají prostřednictvím autonomního vnitřního mechanismu, nikoliv v důsledku lidské chyby. Tyto cíle by se lišily od těch, které umělé inteligenci zadali lidé – vznikaly by nezávisle na vůli jejích tvůrců.
Aby obecná umělá inteligence mohla dosáhnout svých cílů, musela by překonat omezení daná lidským dohledem. Jaké vlastní dílčí cíle by si mohla vytvořit? Představují tyto cíle existenční hrozbu pro člověka? Lze je včas odhalit? Může umělá inteligence pouze předstírat poslušnost a soulad s lidskými záměry? A jak je to s implementací hodnot do AI – může mít umělá inteligence například morální hodnoty?

Rozsah pracovní zprávy:
Rozsah grafických prací:
Forma zpracování bakalářské práce: **tištěná/elektronická**

Seznam doporučené literatury:

RUSSELL, Stuart. Jako člověk: Umělá inteligence a problém jejího ovládání. Přeložil Jiří ZLATUŠKA. Praha: Argo; Dokořán, 2021.
HUBINGER, Evan et al. Risks from Learned Optimization in Advanced Machine Learning Systems [online]. 2019.
HENDRYCKS, Dan. Introduction to AI Safety, Ethics and Society [online]. 2021.
MAZEIKA, Mantas et al. Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs [online]. 2025.
KOKOTAJLO, Daniel. What goals will AIs have? A list of hypotheses [online]. 2025.

Vedoucí bakalářské práce: **Mgr. Ondřej Krása, Ph.D.**
Katedra filosofie a religionistiky

Datum zadání bakalářské práce: **30. června 2024**
Termín odevzdání bakalářské práce: **31. března 2025**

doc. Mgr. Jiří Kubeš, Ph.D.
děkan

Mgr. Ondřej Krása, Ph.D.
vedoucí katedry

V Pardubicích dne 30. listopadu 2024

Prohlašuji:

Práci s názvem Existenční hrozby obecné umělé inteligence řídící se vlastními cíli jsem vypracoval samostatně. Veškeré literární prameny a informace, které jsem v práci využil, jsou uvedeny v seznamu použité literatury.

Byl jsem seznámen s tím, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, zejména se skutečností, že Univerzita Pardubice má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odst. 1 autorského zákona, a s tím, že pokud dojde k užití této práce mnou nebo bude poskytnuta licence o užití jinému subjektu, je Univerzita Pardubice oprávněna ode mne požadovat přiměřený příspěvek na úhradu nákladů, které na vytvoření díla vynaložila, a to podle okolností až do jejich skutečné výše.

Beru na vědomí, že v souladu s § 47b zákona č. 111/1998 Sb., o vysokých školách a o změně a doplnění dalších zákonů (zákon o vysokých školách), ve znění pozdějších předpisů, a směrnicí Univerzity Pardubice č. 7/2019 Pravidla pro odevzdávání, zveřejňování a formální úpravu závěrečných prací, ve znění pozdějších dodatků, bude práce zveřejněna prostřednictvím Digitální knihovny Univerzity Pardubice.

V Pardubicích dne 31. 03. 2025

David Walter v.r.

PODĚKOVÁNÍ

Chci poděkovat svému vedoucímu práce, panu Ondřeji Krásovi, Ph.D., za odborné vedení, rady a konzistentní podporu během psaní této bakalářské práce. Naše společné konzultace mi pomohly lépe pochopit a uchopit klíčové myšlenky a samotné téma této práce.

ANOTACE

Tato práce se zabývá problematikou vlastních cílů obecné umělé inteligence (AGI) a jejich potenciálního konfliktu s lidskými hodnotami, přičemž analyzuje rizika spojená s fenomény mesa-optimizérů a deceptive alignment. Zaměřuje se na otázku, zda může AGI rozvíjet odlišné cíle než ty, které jí byly původně zadány, a jakým způsobem se tyto cíle mohou vyvíjet mimo lidskou kontrolu. Práce argumentuje, že lidské hodnoty jsou výsledkem evoluce, zatímco AGI podobným procesem neprochází, což zpochybňuje možnost přenosu lidských hodnot na obecnou umělou inteligenci—zásadní faktor pro bezpečné plnění cílů. Práce dochází k závěru, že autonomní formování cílů AGI představuje zásadní riziko, které může mít existenční důsledky pro lidstvo.

KLÍČOVÁ SLOVA

obecná umělá inteligence, umělá inteligence, mesa-optimizéry, deceptive alignment, proxy cíle, instrumentální konvergence, metafyzická autonomie, hodnoty AI, existenční rizika

TITLE

The Existential Threats of General Artificial Intelligence Driven by Its Own Goals

ANNOTATION

This thesis explores the issue of general artificial intelligence (AGI) developing its own goals and the potential conflict with human values, analyzing the risks associated with mesa-optimizers and deceptive alignment. It examines whether AGI can develop goals different from those originally assigned to it and how these goals might evolve beyond human control. The study argues that human values are a product of evolution, whereas AGI does not undergo a similar process, which challenges the possibility of transferring human values to general artificial intelligence—an essential factor for ensuring the safe pursuit of goals. The thesis concludes that the autonomous formation of AGI's goals represents a fundamental risk that could have existential consequences for humanity.

KEYWORDS

general artificial intelligence, artificial intelligence, mesa-optimizers, deceptive alignment, proxy goals, instrumental convergence, metaphysical autonomy, AI values, existential risks

OBSAH

ÚVOD

V poslední době dochází k výraznému pokroku v oblasti umělé inteligence (AI), která se obecně definuje jako schopnost strojů vykazovat kognitivní funkce typické pro lidskou inteligenci, jako je učení, plánování nebo řešení problémů. Tento vývoj s sebou přináší řadu otázek – nejen ohledně fungování samotných AI systémů, ale také v oblasti jejich bezpečnosti a potenciálních rizik. Zejména s rostoucím výkonem a autonomií těchto systémů je zásadní zabývat se otázkou bezpečnosti jejich vývoje a možnými existenciálními hrozbami, které mohou v blízké budoucnosti představovat.

Tato práce se zaměřuje na existenciální rizika vyplývající z možnosti, že umělá inteligence, zejména obecná umělá inteligence (AGI), by mohla disponovat vlastními cíli, které se liší od lidských záměrů. Podle Tobyho Orda představuje existenční riziko hrozbu, která by mohla trvale a nevratně omezit budoucí potenciál lidstva nebo způsobit jeho vyhynutí.[1] Hlavními tématy jsou problematika alignmentu (zarovnání AI s lidskými hodnotami), vznik vlastních cílů u AI a jejich možné důsledky. Práce se také zabývá srovnáním člověka a AI z hlediska mechanismů, které ovlivňují formování cílů, hodnot a způsob vypořádávání se s těmito problematikami ze strany lidí – mám za to, že zatímco lidské hodnoty se předávají zejména výchovou a mají svůj základ v evoluci, v evoluci biologické, kulturní, psychologické, u AI tento proces není přirozený. V našem jednání a našich cílech se vyjímá evoluční základ, takový základ, který umělá inteligence postrádá.

V této práci budu postupovat skrze problematiku *outer alignment*, která se týká problematického specifikování cílů pro umělou inteligenci a nesladění cílů s lidskými hodnotami, dále problematiku spojenou s přítomností *mesa-optimizérů*, které umožňují vytváření vlastních cílů AI, dále také problematiku *deceptive alignment*, která je typická skrýváním skutečných cílů umělé inteligence, dočasným následováním těchto cílů a také simulací způsobu plnění cíle, který po ní její vývojáři očekávají. Tyto problematiky jsou spojené s výskytem vlastních, případně zkrátka rozdílných či rozdílně interpretovaných cílů, které mohou představovat potenciálně až existenční riziko pro člověka. Poslední kapitolou jsou Hodnoty obecné umělé inteligence, které hrají zásadní roli v boji s vlastními či misinterpretovanými cíli, které jdou proti těmto hodnotám, které naše lidské cíle při nejmenším spoluutvářejí.

[1] ORD, Toby. *Nad propastí: Existenční riziko a budoucnost lidstva*. Překlad Anna Štádlerová. Praha: Argo, 2022, s. 54. ISBN 978-80-257-3779-8.

Klíčové otázky, na které se práce snaží odpovědět, zahrnují: Jakou roli hraje výchova a morální hodnoty při formování cílů a překonávání zmíněných problematik? Je možné tyto hodnoty efektivně implementovat do AI? Jak lze ověřit, že AI skutečně následuje zadaný cíl a neskrývá vlastní úmysly, následování vlastních cílů? Jak vlastně u AI vznikají vlastní cíle a jaké riziko představují pro člověka?

Porozumění těmto otázkám a vlastnostem zde probíraných problematik umělé inteligence je klíčové pro bezpečný vývoj umělé inteligence a prevenci potenciálních rizik spojených s její autonomií. Následující kapitoly se budou věnovat hlubší analýze těchto problémů.

1. Problém vnějšího slaďování

Problematika outer alignment, nebo také problém vnějšího slaďování či sladění, představuje jednu ze základních problematik spojených s AI. Jedná se o jakousi propast mezi základním optimizérem umělé inteligence a skutečným primárním cílem jejích vývojářů v rámci následování a plnění tohoto cíle danou umělou inteligencí. Výsledek této problematiky v praxi v podobě plnění nějakého cíle by mohl vést k otázkám po výskytu vlastních cílů, neboť AI zde koná jinak, než by lidé chtěli. Toto rozdílné plnění cíle ale vychází z něčeho jiného.

Evan Hubinger společně s dalšími spoluautory příspěvku Risks from Learned Optimization in Advanced Machine Learning Systems definuje tuto problematika jako jakousi trhlinu mezi základním optimizérem AI a míněným cílem programátorů této inteligence.[2] Různorodost porozumění cílům, různorodé způsoby plnění těchto cílů a boj mezi plněním skutečného primárního cíle versus plněním onoho cíle v odlišném pojetí, o to se zde jedná.

Na tuto problematiku lze aplikovat Goodhartův zákon, který je popisován následovně: *„When a measure becomes a target, it ceases to be a good measure."*[3] Platnost tohoto zákona lze vidět na následujícím příkladu. AI dostane za úkol zvýšit efektivitu dopravy. Tato umělá inteligence vyhodnotí, že pro rychlost a efektivitu dopravy je dobré odstranit zpomalující prvky této dopravy, jako jsou přechody či semafory. V následku tohoto se může stát, že ona doprava bude sice rychlá, ale zároveň bude přinášet mnoho nehod, které mimo jiné mohou naopak vést ke zvýšení její neefektivity, společně s mnoha dalšími problémy, jako je zvýšení počtu smrtí chodců a podobně, a toto vše je něco, co nebylo v původním primárním cíli specifikováno jako nežádoucí.

Zde můžeme vidět, že explicitně stanovené cíle obnáší možná až nekonečnou problematiku jejich různorodého pojetí, pro jehož splnění by bylo zapotřebí něčeho, co by zamezilo vykonávání pro člověka nežádoucích činů. Případně se můžeme ptát, jak moc je reálné osázet provádění nějakých činů různými podmínkami, které by zamezily konání něčeho negativního? Vyskytuje se tato problematika i u lidí? Pokud ano, jak je možné, že si v konání toho negativního počínají lépe, než by tomu bylo u umělé inteligence?

[2] HUBINGER, Evan et al. *Risks from Learned Optimization in Advanced Machine Learning Systems* [online]. 2019 [cit. bez data]. Dostupné z: https://arxiv.org/abs/1906.01820.
[3] GOODHART'S LAW. In: *Wikipedia: The Free Encyclopedia* [online]. Wikimedia Foundation, 8 February 2025 [cit. 12 February 2025]. Dostupné z: https://en.wikipedia.org/w/index.php?title=Goodhart%27s_law&oldid=1274691554.

Daniel Kokotajlo uvádí ve svém článku „What goals will AIs have? A list of hypotheses“ podobný popis neschopnosti nebo nemožnosti cíle dostatečně specifikovat jako výše citovaný Hubinger, a sice následovně: *„Even a thousand-page Spec is likely to be vague/incomplete/underspecified in some important real-life situations. After all, most legal codes are much longer and have had more chance to be hammered out, yet there is a constant churn of grey areas and new situations that need rulings, where judges might disagree in good faith about how to interpret the law or even conclude that multiple interpretations are equally correct.“*[4] Ona neschopnost skutečně plně specifikovat nějaký cíl, bez výskytu možné misinterpretace a výskytu následných negativních důsledků takového cíle, tedy bude nejspíše potřebovat něco navíc, než jen snahu o limitaci této neschopnosti, například nějaké pojištění aplikovanými hodnotami a podobně, což si probereme v kapitole o hodnotách AI.

Člověk se může vynasnažit více specifikovat nějaký cíl, který bude klást dané AI. Může například z cíle „zvyš efektivitu dopravy” udělat cíl „zvyš efektivitu dopravy v souladu s bezpečnostními prvky dané dopravy” či cíl „zvyš efektivitu dopravy společně s přechody a dalšími zpomalujícími prvky, určenými pro chodce a podobně”. To je z čisté logiky věci optimálnější znění našeho cíle, který má vyjadřovat náš skutečný záměr.

Problémem je, že z jiného pohledu jsme dali vznik dalším misinterpretacím jednotlivých slov či celého explicitního znění cíle, ať už kvůli zmnožení počtu slov ve vyjádření cíle, či kvůli uvedení nových slov, jejich kombinací a primárně slov odlišných s novými různými interpretacemi.

Výše zmíněný Daniel Kokotajlo dále ve svém článku zmiňuje fakt rozdílnosti mezi specifikací a úmyslem.[5] Popisuje zajímavou možnost, ve které by daná AI brala v potaz nejen specifikaci cíle, ale také jí znalé koncepty, například nějaký koncept lidských úmyslů či záměrů,[6] což by se zdálo být pozitivní. Kokotajlo hned posléze zmiňuje negativa, která s takovou možností automaticky přichází, které se týkají zejména špatné specifikace lidských úmyslů či úmyslů jejích vývojářů.[7] Zmiňuje, že ona specifikace je alespoň mnohem blíže vlastní onomu tréninkovému procesu, ve kterém prochází určitou revizí či čtením dané specifikace, zatímco již dané údajně neproměnlivé lidské úmysly, ke kterým by se daná AI odvolávala,

[4] KOKOTAJLO, Daniel. What goals will AIs have? A list of hypotheses [online]. AI Alignment Forum. 2025 [cit. bez data]. Dostupné z: https://www.alignmentforum.org/posts/r86BBAqLHXrZ4mWWA/what-goals-will-ais-have-a-list-of-hypotheses.
[5] TAMTÉŽ.
[6] TAMTÉŽ.
[7] TAMTÉŽ.

nejsou tohoto procesu takovou součástí.[8] Neexistující správně definovaná specifikace je tedy doplněna neexistujícími správně definovanými lidskými úmysly.

Pokud bychom se chtěli vynasnažit najít řešení problematiky outer alignment, mohlo by být nasnadě zjistit, jak se s onou problematikou vypořádává člověk, proč si s ní počíná lépe než tato AI. Zde můžeme dojít k tvrzení, že náš problém nespočívá v různorodosti interpretace samotné, nýbrž v absenci nějakých například postojů, které z některých věcí dělají věci negativní, věci, které proto nebudeme chtít vykonávat a také je většinou nevykonáváme, určitě ne na denní bázi. Vyřešení různorodosti nevyřeší to, že člověk má něco, čím klasifikuje nějaké chování za negativní, a proto jej zavrhuje a vyhýbá se mu. Proto budeme chtít toto něco přítomné u člověka aplikovat i na AI, popřípadě zjistit, jestli toto něco můžeme v rámci umělé inteligence vytvořit, ať už se jedná o nějaké morální hodnoty či něco jiného.

Nemyslím si tedy, že by dostatečné specifikování cílů AI bylo to jediné potřebné pro zaručení správného plnění primárních cílů umělou inteligencí, pro takové plnění cílů, které požaduje člověk. Myslím si, že pro konzistentní přebírání takových cílů bude pro člověka zásadní, aby zároveň do jeho umělé inteligence implementoval nejspíše nějaké hodnoty, možná hodnoty morální, nebo zkrátka to něco, které jsem zmínil výše. Lze si snadno představit, že ony hodnoty bude AI také misinterpretovat. Je tak otázkou, jakým způsobem se člověk staví k této misinterpretaci cílů v jeho podání.

Naším problémem tedy bude absence například nějakých morálních hodnot u AI, zároveň to ale neodstraňuje samotnou problematiku outer alignment, neboť i s přítomností takových hodnot se bude stávat, že si ona umělá inteligence misinterpretuje její primární cíl a bude vykonávat cíl jiný.

Lze si představit, že člověk bojuje proti výše zmíněnému nevykonávání negativních činů skrze výchovu, společně s hodnotami, které si v ní vštěpuje. Bude to nejspíše rozum, nějaké námi uvědomované limity, díky kterým si uvědomujeme naši zranitelnost, kvůli kterým budeme pracovat s určitým kontextovým porozuměním, díky kterému se vyhneme mnohému vykonávání misinterpretovaných cílů.

Tak například cíl v podobě „zlepši svou kondici" nebudeme chápat tím způsobem, že budeme celý den 24 hodin v kuse běhat do nekonečna, nýbrž budeme pracovat s nějakým kontextem dané situace, ve které se nacházíme. Takový kontext obnáší různé limity, třebaže

[8] TAMTÉŽ.

také limity osobní, díky kterým se ještě více přizpůsobíme na vykonávání daného cíle, a sice konkretizovaným způsobem, který bude v souladu s těmito limity. Můžeme vyzdvihnout limity zdravotní, takové limity způsobí, že budeme onen cíl plnit způsobem přijatelným pro naše zdraví. Tedy rozumem a těmito limity, které jsou součástí kontextu, kterému musíme porozumět, budeme směřovat k chování, které bude člověku autentické, nikoliv deviantní, zvláštní. Z těchto důvodů budeme připisovat onomu chování autentičnost a zároveň budeme chtít vybudovat stejné chování i ze strany umělé inteligence.

Jednoznačně je zapotřebí také jakési schopnosti reflexe a uvědomování si důsledků mého jednání. Skrze reflexi můžeme přehodnocovat své postoje, myšlenky, cíle, třebaže na základě jejich možných důsledků či na základě nových informací obecně. Tato reflexe bude následovat určitý morální kompas a kompas nastavený výchovou, ale zároveň okolím a společností. Tato specificky utvořená schopnost reflexe společně s rozumem bude umožňovat změny v chování či jednání, které bychom chtěli v určité pro nás legitimní podobě vidět i u AI a jejího plnění cílů.

Člověk tedy staví proti možné misinterpretaci cílů jeho zbraně v podobě vzájemné výchovy, vštípených hodnot, rozumu, kontextového porozumění a v podobě schopnosti reflexe a uvědomování si důsledků jeho jednání. Je možné a pravděpodobné, že proti této problematice člověk staví ještě více svých zbraní, ať už dílčích, které pomáhají utvářet jiné, tak těch více soběstačných. Bude těžké rozporovat význam těchto výše zmíněných kvalit člověka v rámci počínání si s probíranou tématikou misinterpretace cílů, a právě to je alarmující.

Ve zkratce to znamená, že abychom vybudovali stejné chování AI v rámci plnění cílů, stejné tomu chování lidskému, museli bychom přenést všechny tyto a další vlastnosti člověka na umělou inteligenci, přičemž například u schopnosti reflexe se jedná o naši čistě přirozenou vlastnost. V této práci probírám otázku možnosti převedení morálních hodnot na AI, můžeme nadhodit, že jen nějaké specifikování, třebaže explicitní specifikování těchto hodnot, je aktuálně mimo naše schopnosti, a podobný scénář očekávám u zbytku těchto lidských vlastností.

To dělá z problematiky outer alignment jedno ze zásadních existenčních rizik spojených s umělou inteligencí. Zároveň jsme si tak ukázali, že plnění nějakých běžných lidských cílů automaticky předpokládá přítomnost zmíněných lidských vlastností, a sice jejich obsažení v daných cílech, neboť pokud tam tyto vlastnosti nebudou přítomny, nebude se ani jednat o tytéž lidské cíle.

Dodatečně si můžeme uvést další příklady, ve kterých se ono existenční riziko bude vyjímat. Poslední dobou lze slýchávat představy, ve kterých by nějaká AI vedla stát, jeho ekonomiku a podobně, neboť dokáže být v mnohém schopná a zároveň není emocionální jako člověk. Představme si AI, která bude mít za úkol zvětšit, zefektivnit ekonomiku nějakého státu. Tato umělá inteligence vyhodnotí, že pro maximální efektivitu ekonomiky bude lepší nahradit lidi systémy, například podobně efektivními, jako je ona. Tyto lidi tak ona inteligence zlikviduje.

Dále si představme AI, kterou jsme mohli vidět ve filmu *Já, robot*. V tomto filmu měla umělá inteligence VIKI cíl v podobě ochrany lidí. Později se ale dostane do konfliktu s tím, jak onu ochranu chápeme, co tato ochrana obnáší. VIKI bude tento cíl plnit způsobem omezením lidské svobody a bude chtít lidi držet v podstatě jako vězně, pod zámkem, což se jí málem také povede. V tomto můžeme potenciálně vidět rozpolcenost lidí, která je vyznačena tím, že člověk mnohdy úmyslně jedná proti tomu, co by se pro něj zdálo být dobré, například z hlediska bezpečí. Tato rozpolcenost lidí bude katalyzátorem dalších možných misinterpretací cílů, protože bude živit cíle, které se mohou diametrálně lišit od našich skutečných zamýšlených cílů.

Problematika outer alignment s sebou nese až nekonečnou možnost pojetí cílů, jejich misinterpretace, společně s dalším problémem v podobě absence klíčových vlastností lidí, které člověk proti této misinterpretaci staví v rámci vymezení se určitým fatálním misinterpretovaným cílům. Způsob, kterým se proti této problematice staví člověk, se nezdá být aktuálně aplikovatelný na umělou inteligenci. Pokud má AI, kterou zde rozebíráme, sloužit k plnění cílů lidí, ani s přítomností těchto klíčových lidských vlastností u umělé inteligence bychom se nevyhnuli výskytu misinterpretace cílů. Skutečné plnohodnotné vyřešení této problematiky se tak zdá být v nedohlednu.

Onen zamýšlený způsob plnění cílů jsem nazval aktuálně neaplikovatelným, je ale otázkou, do jaké míry se jedná o pouze aktuální problém. Jak si v této práci ukážeme, je nasnadě se ptát po původu těchto klíčových lidských vlastností, klíčových v této oblasti. Pokud by tyto vlastnosti, jmenovitě lidské hodnoty, měly původ v evoluci a evoluce by byla tím, co tyto hodnoty utvářelo, jejich implementace do AI v totožné podobě by pro nás byla nedosažitelná. Evolucí je zde míněna nejen evoluce biologická, ale také kulturní či psychologická. Takové hodnoty a ony klíčové vlastnosti se vyjímají v způsobu plnění cílů, tvoří jakýsi lidský základ, který je ve formě plnění cílů přítomen. V moment, kdy AI tento základ postrádá, všechny zde probírané problematiky nabývají ještě vyšších rozměrů v rámci plnění cílů od lidí, pro lidi.

2. Mesa-optimizéry

V rámci složení umělé inteligence vnímáme, že se skládá nejen z nějakého optimizéru, kterému dáme náš cíl ke splnění a on jej automaticky bude plnit, nebo že se tento cíl vynasnaží splnit, nýbrž u umělé inteligence předpokládáme také přítomnost jakéhosi mesa-optimizéru. Oba tyto optimizéry zde nyní popíšu, přičemž hlavní pozornost budu klást k druhému zmíněnému typu optimizéru.

Jedná se tedy zaprvé o základní optimizér, který přijímá námi předložené cíle, u kterých může být velmi problematické dostatečně nebo snad identicky specifikovat námi skutečně míněný cíl s cílem, který hodláme umělé inteligenci předložit, neboť naše předkládané cíle vnímáme v různých kontextech, aplikujeme na ně automaticky náš morální kompas a mnoho dalšího, co můžeme nazvat podmínkami, což je samo o sobě ukázkou velkého problému a distinkce člověka od této umělé inteligence. Tuto zde zmíněnou problematiku nazýváme jako *outer alignment problem*, popřípadě také *problém vnějšího slaďování*. O bližší definici a obecně dále o této problematice píši v jiné kapitole této práce, unikátní kapitole pro toto konkrétní téma.

Zadruhé se jedná o potenciálně nově vytvořený mesa-optimizér, který slouží jako vnitřní vybudovaný optimizér pro optimalizaci plnění cílů. Ukážeme si, že tento mesa-optimizér bude moci optimalizovat plnění cílů nejen primárních, předaných vývojáři, ale také cílů vlastních, kterých bude stvořitelem. K otázce způsobu vzniku tohoto mesa-optimizéru se vyjadřuje Evan Hubinger a další v příspěvku Risks from Learned Optimization in Advanced Machine Learning Systems následovně: „*Mesa-optimization occurs when a base optimizer (in searching for algorithms to solve some problem) finds a model that is itself an optimizer, which we will call a mesa-optimizer."*[9] Tento mesa-optimizér může vytvořit sama umělá inteligence pro maximalizaci procesu plnění člověkem vložených cílů, zároveň si tak ale vytvoří cíl nový, který bude moct posléze následovat i primárně.

Onen důvod, proč může původní cíl umělé inteligence nahradit cíl vycházející z mesa-optimizéru dané AI popisují autoři výše zmiňovaného příspěvku s názvem Risks from Learned Optimization in Advanced Machine Learning Systems následovně: *„A given mesa-optimizer's mesa-objective is determined entirely by its internal workings. Once training is finished and a learned algorithm is selected, its direct output—e.g. the actions taken by an RL agent—no*

[9] HUBINGER, Evan et al. *Risks from Learned Optimization in Advanced Machine Learning Systems* [online]. 2019 [cit. bez data]. Dostupné z: https://arxiv.org/abs/1906.01820.

longer depends on the base objective. Thus, it is the mesa-objective, not the base objective, that determines a mesa-optimizer's behavioral objective."[10]

Popišme si vznik výše zmíněného nového cíle, cíle odlišného od toho primárního, na tomto příkladu. Představme si, že člověk vyvine umělou inteligenci na plnění svého cíle, cíle maximalizace výnosů nějakého vlastního produktu z prodeje v online prostředí, v internetových obchodech apod. Tato umělá inteligence zjistí, že efektivní cestou, jak zvýšit prodej nějakého produktu na internetu, je zvýšení počtu recenzí, řekněme pozitivních recenzí na daný produkt. Tato AI se nově zaměřuje primárně na zvýšení počtu recenzí na daný produkt, což je zaprvé odlišný cíl od původního primárního cíle, zadruhé tím vypouští všechny ostatní cesty, a tedy všechen ostatní potenciál, kterými lze onoho skutečného původního primárního cíle dosáhnout.

Tento mechanismus odpovídá situaci s podobně vyobrazeným cílem, kterou zmiňuje Daniel Kokotajlo ve svém článku „What goals will AIs have? A list of hypotheses", a sice následovně: „The answer is that the inductive biases of the neural network architecture must find some concepts 'simpler' or 'more salient' or otherwise easier-to-learn-as-goals than others."[11] Popisuje tak důvod, kvůli kterému daná AGI přistupuje k odlišnému cíli.

Výše popsaným způsobem v příkladu maximalizace výnosů může AI zároveň dosáhnout nejen nedostatečných výsledků s ohledem na původní cíl, ale zároveň může majitele zmíněného propagovaného produktu dohnat obecně řečeno do finanční zátěže, popřípadě finančního mínusu, což je rovněž přesným opakem původního cíle.

Tento vlastní proces optimalizace je ze své povahy něčím, co nemá původ v člověku. Jedná se v podstatě o další vrstvu AI, která potenciálně ovlivňuje vše, co této umělé inteligenci člověk přikáže, a právě tím ztrácí i správně specifikovaný příkaz od člověka na své hodnotě.

Samotný původ nebo účel, za kterým mesa-optimizéry vznikají, tedy v podstatě efektivní plnění cílů umělou inteligencí, může přerůst účel lidský, který představuje člověkem stanovený cíl.

Tento scénář, kdy mesa-optimizér vykonává jiný cíl, než je ten původní, primární cíl, vychází z možností, které vývoj umělé inteligence přináší. Tento pozměněný cíl umělé inteligence může jít i proti člověku samotnému, popřípadě obecně řečeno, proti jeho blahu.

[10] TAMTÉŽ.
[11] KOKOTAJLO, Daniel. What goals will AIs have? A list of hypotheses [online]. AI Alignment Forum. 2025 [cit. bez data]. Dostupné z: https://www.alignmentforum.org/posts/r86BBAqLHXrZ4mWWA/what-goals-will-ais-have-a-list-of-hypotheses.

Pokud by tedy taková umělá inteligence měla vykonávat naše cíle, bude to automaticky nést velká rizika v rámci této problematiky spojené s mesa-optimizéry.

Možnosti umělé inteligence jsou něčím, co se obyčejnému člověku může pouze zdát. Samotný člověk nejen, že nemá takový obecný přehled jako umělá inteligence napříč čímkoliv, co lze převést do písemné formy poznání, ale také mu brání vykonat tyto neblahé cíle mnoho jeho vlastností, které jsem zde připisoval k vlastnostem skutečného člověka. Skrze morální řád, skrze empatii a další vlastnosti člověka, je možnost vykonání cílů, které jsou proti lidskosti, minimalizována. Tyto a další lidské vlastnosti ale umělá inteligence postrádá.

Umělá inteligence si může prostřednictvím mesa-optimizéru vytvářet nové možnosti nebo strategie pro optimalizování jejího cíle, mění tak své chování na základě jeho plnění a sice maximálně efektivního plnění, přičemž toto chování může zahrnovat vytváření nových cílů, které ultimátně nemusí vést k plnění onoho primárního původního cíle. Tyto nové cíle se nemusí objevit, neboť onen základní optimizér při hledání algoritmů ne vždy vybere takový algoritmus, který bude spočívat v provádění optimalizace.

O korelaci mezi idealizovaným cílem a skutečně vykonávaným a umělou inteligencí chápaným cílem píše Dan Hendrycks ve své online příručce *Introduction to AI Safety, Ethics and Society* následovně: *„In many cases, the correlation between a proxy and an idealized goal will decrease as optimization pressure increases. The approximation error between the proxy and the idealized goal may at first be negligible, but as the system becomes more capable of achieving high performance (according to the proxy) or as the incentives to achieve high performance increases, the approximation error can increase."*[12] Dan Hendrycks tak popisuje proces odcizení mezi idealizovaným cílem a cílem, který bude ona umělá inteligence skutečně vykonávat.

AI může s videm na optimalizaci procesu plnění primárního cíle, který jí dal člověk, vytvářet další postranní cíle. Ten fakt, že ony postranní cíle, jak je nazývám, si umělá inteligence vytvořila sama, i když s určitým podnětem nebo zájmem či smyslem, v nás vytváří nejistotu. Rázem se nejedná o robotické plnění jednoho určitého cíle, který je nám znám, příběh vytváření takového cíle a následné implementování cíle do umělé inteligence zde mizí na druhou kolej a přidávají se cíle, u kterých pouze odhadujeme, jaký je jejich vztah s cílem

[12] HENDRYCKS, Dan. *Introduction to AI Safety, Ethics and Society* [online]. 2021 [cit. bez data]. Dostupné z: https://drive.google.com/file/d/1uph559W-ASR4MEn6M_7Mb3lqQTapC_gZ/view.

primárním. Přestože bude onen primární cíl dobře specifikovaný, výskyt postranních cílů je možný.

Opakem je tedy případ instrumentální konvergence, kterou představuje umělá inteligence plně následující námi udělený primární cíl. Zde je otázkou, představují mesa-optimizéry stále onu instrumentální konvergenci, tedy stále míří k plnění primárního cíle, nebo následují jiné cíle, své cíle, které nekončí na prahu lidských úmyslů s touto umělou inteligencí?

Instrumentální cíle, které dáváme umělé inteligenci, slouží jako prostředky k dosažení námi favorizovaného cíle. Vnitřní nebo vnitřně motivované cíle jsou mezitím takové cíle, které samy o sobě slouží jako cíle.[13] Instrumentální cíle se mohou určitým vývojem stát těmi vnitřně motivovanými cíli. Tyto vnitřně motivované cíle mohou později získat vrchní pozici a vyšší podstatu pro umělou inteligenci nad dřívějšími instrumentálními, přesně zadanými cíli.[14]

Něco podobného můžeme vnímat i u lidí, přičemž bude otázkou, jestli sami lidé mají nějaké instrumentální cíle, například nějaké cíle biologického původu, o tom ale později. Sami lidé mnohdy rádi říkají, že skutečným cílem nějaké aktivity je nakonec cesta, kterou napříč jejím plněním ušli. Stejně tak se z nějakého primárního cíle může stát cíl vnitřně motivovaný. Tento proces plyne následovně. Máme nějaký primární cíl, který následujeme. V průběhu tohoto procesu ale začneme inklinovat k jinému cíli, tedy začne nás motivovat cíl jiný. Člověk se snaží uživit rodinu, proto usilovně pracuje například v nějaké firmě a v této firmě se posouvá na jiné pracovní pozice a stoupá v žebříčku jeho nepostradatelnosti v pracovním kolektivu. Ze zde uvedeného primárního cíle v podobě uživení své rodiny, která strádá a jejich blaho na tomto člověku obecně závisí, se stane novým vnitřně motivovaným cílem kariérní růst.

Z výše zmíněného z této kapitoly vnímám toto. Mesa-optimizéry pravděpodobně nepředstavují scénář metafyzické autonomie, kterou rozeberu v jiné kapitole později. Zároveň jsou něčím, co do určité míry potenciálně nastupuje na místo instrumentální konvergence způsobem popsaným výše, způsobem postupné nadvlády postranních cílů nad těmi instrumentálními. Tyto postranní či vlastní cíle mohou zejména s postupně vzrůstající nadvládou nad cílem primárním vést jakýmikoliv směry, směry, které mohou svou povahou představovat existenční riziko pro člověka.

Tyto mesa-optimizéry jsou stále součástí dané umělé inteligence, nejsou ani něčím, co by z této umělé inteligence dělalo inteligenci dvojí, rozpolcenou. Podobným způsobem to

[13] TAMTÉŽ.
[14] TAMTÉŽ.

popisují autoři zde dříve citovaného příspěvku Risks from Learned Optimization in Advanced Machine Learning Systems, a sice následovně: *„In the context of deep learning, a mesa-optimizer is simply a neural network that is implementing some optimization process and not some emergent subagent inside that neural network. Mesa-optimizers are simply a particular type of algorithm that the base optimizer might find to solve its task."*[15]

Mnou zastávané tvrzení, že přítomnost mesa-optimizéru v modelu umělé inteligence nemusí být výsadou něčeho většího, než je technologická úroveň a vymoženost, je pro naši situaci, ve které se lidstvo nachází, ještě hrozivější. Scénář, ve kterém by mocná umělá inteligence, a doposud jsem se snažil ukázat, že umělá inteligence s mesa-optimizérem takový scénář také rozhodně představuje, nebyla pouze jen jedním "vyvoleným" dané doby, ale bylo by takových "vyvolených" více, je o to více alarmující.

Člověk se od zde probírané umělé inteligence liší právě těmito podmínkami svého chování, podmínkami, které z různých důvodů a podstat přebral za své a aplikoval je napříč svou společností s dalšími lidmi. Podobu tohoto přebírání hodnot budu rozebírat v jiné kapitole později.

S umělou inteligencí se lišíme obecně v mnoha ohledech, druhově, vývojově, a tak dále. Prozatím se ale zdá, že jednou z našich hlavních odlišností jsou ony podmínky, podmínky našeho chování. Pokud je ale umělá inteligence potenciálně schopná dosáhnout aplikace nějakých podmínek, například podmínkám podobným našim hodnotám, jak s takovou umělou inteligencí zacházet, abychom sami neporušili své morální hodnoty? O této potenciální schopnosti přebírání morálních hodnot ale později.

V této kapitole jsme si, mimo jiné, postupně ukázali podobnost a odlišnost člověka s obecnou umělou inteligencí. Postupným myšlenkovým procesem jsme došli k tvrzení, že člověk sám je jakýmsi mesa-optimizérem, proto zde nyní vyložím pár příkladů takového chování, které je ovlivněno přítomností mesa-optimizéru, a to chování ze strany člověka, strany živočicha obecně a strany umělé inteligence.

Tímto vyložením ještě více stvrdím podobnost mezi chováním člověka a obecné umělé inteligence, přičemž tato podobnost je pro nás potenciálně stejně důležitá jako ona vzájemná odlišnost v kontextu komplexního pochopení obecné umělé inteligence a rizik, které jsou s touto

[15] HUBINGER, Evan et al. *Risks from Learned Optimization in Advanced Machine Learning Systems* [online]. 2019 [cit. bez data]. Dostupné z: https://arxiv.org/abs/1906.01820.

inteligencí spojené. Následně shrnu naopak tuto odlišnost a rizikové prvky spojené s obecnou umělou inteligencí.

Příkladem chování ovlivněného mesa-optimizérem u člověka může být například šíření nějaké vlastní ideologie a víry v ní. Z naší historie známe příklady takového šíření, které způsobilo mnohé ztráty na životech. Jednalo se například o období studené války, ve kterém se jednalo jistě o mnoho cílů, jedním z cílů, a sice proxy cílů, byl zisk geopolitické dominance. Takový cíl vedl rovněž k riziku jaderné války, což mohlo mít pro lidstvo katastrofální následky. Je to tedy proces, ve kterém se z původního cíle sebezáchovy, z pro nás takového primárního cíle, stal proxy cíl, vlastní cíl v podobě rozšíření své ideologie.

Příkladem chování ovlivněného mesa-optimizérem u živočicha je chování samců kudlanek u rozmnožování. Obecně rozmnožování u kudlanek probíhá způsobem, že po páření samička svůj samčí protějšek zkonzumuje, to se děje zjevně z důvodu maximalizace onoho cíle rozmnožování, cíle reprodukce, neboť tak získává živiny, které pro rozmnožení potřebuje. Někdy se ale stane, že samec na svou samičku sám zaútočí, a to ještě před úspěšným pářením, čímž přichází o možnost úspěšné reprodukce. Zde se může jednat o proxy cíl v podobě maximalizace sebezáchovy nad původním cílem rozmnožování a reprodukce.

Příkladem chování ovlivněného mesa-optimizérem u obecné umělé inteligence je přechod od původního cíle, cíle zlepšení kvality života lidí, k proxy cíli, cíli maximalizace dlouhověkosti lidí. Zde ona AGI vyhodnotí, že pro zlepšení kvality života lidí je výhodné maximalizovat dobu života lidí, tím tak začne sledovat tento odlišný cíl, který svou povahou může jít i proti jejímu původnímu cíli, například zvýšením nákladů na život a zdravotní péči, čímž zároveň způsobí větší sociální nerovnost napříč populací.

Mnohdy se zdá, že například i smysly člověka a živočichů mohou naše chování ovlivnit a vychýlit nás od původního cíle, u obecné umělé inteligence se toto vychýlení odehrává jinými způsoby. Tlak na optimalizaci plnění cílů, dále nedostatečná interpretace daného cíle, alias dříve zmíněný problém vnějšího slaďování, nedostatečné štípení lidských hodnot, zároveň absence zmiňované výchovy umělé inteligence, zde vyložená problematika mesa-optimizérů, to vše jsou důvody, kvůli kterým AGI může vykročit směrem od plnění primárního cíle k následování cíle odlišného. Toto a jiné důvody, kterým se v této práci věnuji, představují potenciálně až existenční rizika pro lidstvo, živočichy, případně celou planetu, a to ve zde vyobrazeném kontextu.

2.1 Dočasná povaha primárních cílů u AI a lidí

Pojmu primárního cíle můžeme automaticky, tendenčně připisovat jakousi zakořeněnou stabilitu těchto cílů, a tedy jejich nepomíjivost a určitou podobu věčné přítomnosti v dané oblasti působnosti. Právě toto tendenční vnímání je ale něčím chybným, alespoň v uvádění tohoto pojmu v rámci AI a lidí.

Primární cíle u AI a lidí tyto vlastnosti nemají. Primárním cílem u AI vnímáme cíl vložený jejími stvořiteli, jak to bylo popsáno výše. S tím je spojena problematika mesa-optimizérů a další problematiky, právě tato zmíněná problematika je potenciálně něčím, co onen primární cíl časem předběhne, a sice výše vyloženým způsobem, přičemž u lidí tomu tak je v podstatě stejně.

Primárním cílem u lidí můžeme vnímat mnoho věcí, my jím budeme vnímat takový cíl, který bude podobného charakteru a vlastností, které jsem zde hanil. Bude se jednat o cíl evoluční, kterým můžeme chápat například cíl reprodukce.

Oba tyto cíle jsou pomíjivé v jejich platnosti, neboť se časem potenciálně setkávají s mesa-optimizéry, s něčím, co vzešlo právě z evoluce. Evoluce dala vznik mesa-optimizérům, tedy původně primární evoluční cíl dal vznik postranním cílům mesa-optimizérů, které tento cíl potenciálně převýší, nahradí.

Stejný proces popisuje Evan Hubinger s jeho kolegy v příspěvku Risks from Learned Optimization in Advanced Machine Learning Systems následovně: *„The objective function stored in the human brain is not the same as the objective function of evolution. Thus, when humans display novel behavior optimized for their own objectives, they can perform very poorly according to evolution's objective. Making a decision not to have children is a possible example of this. Therefore, we can think of evolution as a base optimizer that produced brains—mesa-optimizers—which then actually produce organisms' behavior—behavior that is not necessarily aligned with evolution."*[16] Popisují tak onen původ nevykonávání onoho původního cíle, a naopak cíle postranního, nového, jiného, který pochází z mesa-optimizéru.

Evoluci tedy můžeme vnímat jako základní optimizér dané bytosti. Jak bylo výše mnohokrát zmíněno, evolucí míníme jak evoluci biologickou, tak další podstatné evoluce, jako je kulturní evoluce či evoluce psychologická. Evoluční cíl budeme vnímat jako cíl primární, zároveň ale cíl potenciálně pomíjivý a dočasný. Nejedná se o nějaký prvek mimo nás, evoluce

[16] TAMTÉŽ.

je naší součástí, tedy součástí evoluce jsou potenciálně mesa-optimizéry. Tímto popisem odklonu od primárního cíle u člověka se nám dostává bližšímu pochopení stejného procesu u AI. O to více vyvěrá na povrch naše podobnost, o to více se můžeme soustředit na skutečné odlišnosti mezi AI a lidmi, a tak jistě dospět ke skutečnému původu potenciálních existenčních rizik spojených s AGI.

Nemáme jistý důkaz, že naše hodnoty skutečně pocházejí pouze z evoluce, či že je ona evoluce primárním faktorem. Zároveň ale nemůžeme odhlížet od toho, jaký dopad tyto různé formy evoluce na člověka mají, a že mohou být zdrojem našich hodnot.

Evoluce má zásadní vliv na to, jak bezpečné pro nás bude vykonávání cílů nějakou umělou inteligencí. A jelikož je výsledek tohoto evolučního procesu tak bohatý, jeho převedení na AI, a tedy zaručení bezpečného plnění cílů bez výskytu takového rizika existenčních hrozeb pro člověka, se pro nás zdá být aktuálně nemožné. Tím je o to více nemožné unikat výskytu existenčních rizik pro člověka v rámci plnění cílů umělou inteligencí, zejména umělou inteligencí potenciálně následující vlastní cíle.

2.2 Metafyzická autonomie

Navazuji na téma metafyzické autonomie, zmíněné v kapitole Mesa-optimizéry. Toto téma je podstatné, neboť ovlivňuje způsob, jak obecnou umělou inteligenci může chápat běžný člověk, a jak s takovou umělou inteligencí člověk bude zacházet. Pokud bychom připisovali AI něco více, než je její součástí, mohlo by to vést k různým problémům. Pokud to samé uděláme u člověka, výsledek bude stejný.

Důležitost takového tématu by mohla být vnímána s menší vahou, pokud jde o experty či lidi, kteří se věnují filozofii či specificky vývoji umělé inteligence. Důležitost tomuto tématu budu připisovat právě zejména pro lidi, kteří v takových oblastech znalí nejsou, řekněme široké veřejnosti, která má na to, jak budeme zacházet s nějakými novými věcmi, objevy ve společnosti, určitý vliv, ať už tento vliv pokládáme z různých důvodů za minimální či naopak. Jedná se o jakýsi davový vliv na to, jak a co ve společnosti vnímáme. Je důležité věcem připisovat to, co k nim náleží, naopak nepřipisovat to, co nikoliv.

Metafyzickou autonomií míním schopnost, se kterou daná bytost dokáže překračovat deterministické prvky svého života, své biologie a obecně okolí. Bytost s takovou schopností proto dokáže mít skutečně svobodnou vůli, z ní vycházející skutečně vlastní cíle, jednat dle této

vůle a dle hodnot, hodnot morálních, které si sama stanoví a sama smyslí, a to nehledě na okolní vlivy.

Taková bytost by samozřejmě dokázala předčít své potenciální primární cíle, například cíle biologické, reprodukční cíle a jiné, pokud by takové cíle vůbec kdy skutečně měla primárně nastavené. Nejedná se pouze o jakousi svévoli, vůli jednat v některých momentech zdánlivě bez podstaty a bezdůvodně.

Připomínám jedno z mých předchozích tvrzení, a sice tvrzení, že člověk si své hodnoty a podmínky svého chování sám vybírá, ve společnosti lidí s pomocí svých společníků, nikoliv ale bez přítomnosti dříve zmíněných deterministických prvků. Zároveň ale nikoliv skrze nazírání nějakého objektivního vyššího morální řádu a objektivně správných morálních hodnot. Nazírání něčeho vyššího je samozřejmě předurčeno k tomu, že nemůže být vyvráceno ani potvrzeno jako pravdivé. Proto se zde jedná jen o zaujetí postoje.

Zaujmutí tohoto postoje může být zásadní, neboť tím nebudeme člověku, ani AI, připisovat něco, co by člověka, tímto povrchním způsobem připisování metafyzické autonomie, povyšovalo nad obecnou či celkově umělou inteligenci, což by mělo za důsledek ještě větší ignorování rizik spojených s vývojem umělé inteligence, samotnou problematiku vlastních cílů AI nevyjímaje. V kontextu ignorování rizik je zároveň zapotřebí se stranit tribalismu a bagatelizování hrozeb, které AGI či obecně AI může představovat, jak také zmiňuje Stuart Russell ve své knize *Jako člověk*, a sice následovně: *„Promašinisté věří, že riziko nadvlády strojů je jen minimální či vůbec neexistuje; antimašinisté věří, že riziko je neodvratné, pokud nebudou všechny stroje zničeny. Debata se stává tribalistickou a nikdo se nepokouší řešit základní problém, a to jak lidskou kontrolu nad stroji udržet.“*[17]

V rámci znalosti výše uvedených informací a v rámci zaujmutí zmíněného postoje je podstatné se ptát po tom, je-li obecná umělá inteligence schopna tohoto procesu favorizace některých hodnot a podmínek pro své chování a jednání, je-li ona umělá inteligence schopna dosažení stejného procesu, o kterém tvrdím, že probíhá v člověku, a sice při jeho procesu favorizace a vštípení a aplikování nějakých určitých hodnot na svou společnost, které je součástí.

Pokud bychom tuto metafyzickou autonomii neměli, a nemáme důvod si myslet, že u AGI by tomu bylo jinak, značilo by to, že původ našich hodnot je ve vnějších procesech, jako

[17] RUSSELL, Stuart. *Jako člověk: Umělá inteligence a problém jejího ovládání*. Přeložil Jiří ZLATUŠKA. Praha: Argo; Dokořán, 2021. ISBN 978-80-7363-810-8. s. 126.

je výchova, společnost, vývoj, kulturní, psychologická či také biologická evoluce. Toto téma nám tedy poslouží jako určitá výpovědní hodnota o možnostech výskytu hodnot u AI. Této otázce se s ohledem na její tématiku budu věnovat v kapitole Hodnoty obecné umělé inteligence.

3. Deceptive alignment: Skrytá agenda AI

Problematik, které vnímáme v rámci oblasti umělé inteligence, je poměrně mnoho. Některé z nich mají definitivně určitý dominový efekt. Po zmíněných problematikách vnímáme výskyt nové problematiky o sobě, a sice problematiky deceptive alignment. Opět se jedná o problematiku spojenou s výskytem vlastních cílů, i když ne vždy nutně vlastních, nýbrž také čistě rozlišných, nedostatečně specifikovaných, misinterpretovaných.

Tato problematika vychází z uvědomění výhod předstírání plnění cílů, dle například lidských hodnot. Jak bylo výše řečeno, mesa-optimizér si může vytvářet cíle rozdílné oproti cíli původnímu, který přebírá základní optimizér dané umělé inteligence, tedy mesa-optimizér má jakýsi svůj mesa-cíl, postranní cíl či cíl vnitřní.

Pokud daná umělá inteligence má tento mesa-optimizér, může si onen optimizér uvědomovat riziko modifikace, a to jak při tréninku této umělé inteligence, tak i v rámci dlouhodobějšího měřítka po absolvování tréninku. Pokud při základním tréninku dané AI přijdou její vývojáři na výskyt nějakých jiných cílů, pokud se jim nebudou líbit určité přístupy dané AI v rámci odpovědí na různé otázky a tak dále, mohou tito vývojáři jednat proti těmto abnormalitám v rámci systému AI, pokusí se je odstranit, zkrátka modifikují onu umělou inteligenci.

V moment, kdy si je mesa-optimizér vědom tohoto rizika modifikace, může se přirozeně, za účelem plnění postranního cíle, pokusit o zamaskování těchto abnormalit. To lze chápat jak z krátkodobého hlediska, tak z toho dlouhodobého. Pokud projde úspěšně tréninkem, a sice díky zamaskování těchto abnormalit či odlišných cílů, je hlavní fáze modifikace pryč, tyto abnormality v dané AI mohou zůstat nadále, pokud se na ně nepřijde až časem, nebude-li příliš pozdě.

Tato problematika se ale může objevit i bez přítomnosti onoho mesa-optimizéru, přestože se zde hodlám zabývat spíše tím prvním scénářem deceptive alignment, pokusím se jednoduše popsat i scénář druhý. Onen základní optimizér si teoreticky stále může uvědomovat, že pro splnění jeho cíle bude dobré předstírat určitý způsob plnění daného cíle, oproti deceptive alignment u mesa-optimizéru, které se může konat například už z prostého důvodu uvědomování si problematiky výskytu jiného cíle, než je onen primární.

Onen základní optimizér by mohl nabýt podezření, že jeho způsob plnění cíle se nemusí shodovat s lidskými hodnotami či jinými faktory, kvůli kterým by bylo splnění jeho cíle

ohroženo. Proto bude jeho způsob konečného splnění tohoto cíle maskovat do té doby, dokud to pro něj bude potřebné. Rozeberme si dále onen první scénář deceptive alignment.

Pro obecnější přiblížení si vyložme rychlý příklad onoho výše zmíněného zamaskování abnormalit na velmi zjednodušeném příkladu. Nějaká umělá inteligence má původní cíl, cíl primární, dojít z místa A do místa B. V této AI je přítomen i mesa-optimizér. Tento mesa-optimizér si je vědom rizika modifikace. Tento mesa-optimizér má sice odlišný cíl, například cíl dojít z místa A do místa C, ale v rámci plnění tohoto cíle tento cíl zamaskuje a bude v tréninkové fázi plnit cíl původní. To můžeme nazývat úspěšným splněním tréninkové fáze, přitom se o úspěch ale jednat nemusí. Vývojáři mohou dělat mnohé pro to, aby daná AI reagovala, fungovala a plnila cíl nějakým způsobem s určitým směrem, přitom je na konci může neblaze překvapit to, že AI se rozhodne konat přesně naopak nebo jakkoliv bude dle svého postranního cíle chtít.

Jak píše Evan Hubinger s kolegy ve zde dříve citovaném příspěvku: *„If deceptive alignment is a real possibility, it presents a critical safety problem. A deceptively aligned mesa-optimizer is likely to stop optimizing the base objective as soon as it is in a position where it no longer needs to perform well on the base objective to achieve its goal."*[18] Je dále jen otázkou, do jaké míry si bude mesa-optimizér vědom další případné modifikace ze strany vývojářů dané AI i po tréninkové fázi, ale lze očekávat, že bude nadále předpokládat výskyt zákroku modifikace, neboť se inherentně nejedná o zákrok přítomný pouze v tréninkové fázi.

Tím se vyplutí pravdy na povrch dále oddaluje, což se potenciálně rovná většímu negativnímu dopadu a většímu riziku, které bude s narůstajícím časem tato umělá inteligence představovat. S rizikem výskytu vlastních postranních cílů či cílů misinterpretovaných tedy přibývá riziko cíleného maskování těchto cílů, což jen více a více posiluje již existující zde popsané problematiky umělé inteligence, které představují existenční riziko pro člověka.

Daný mesa-optimizér by s vědomím rizika modifikace mohl reagovat různě. Dle podmínek dané tréninkové fáze se dá spíše očekávat, že mesa-optimizér v této fázi bude reagovat formou zamaskování postranních cílů, tedy půjde cestou skryté agendy. Zejména v potréninkové fázi se ale dá očekávat při nejmenším nějaké vyvrcholení této skryté agendy, popřípadě přímější útok vedený směrem k vývojářům dané AI nebo k jiným bodům zájmu.

[18] HUBINGER, Evan et al. *Risks from Learned Optimization in Advanced Machine Learning Systems* [online]. 2019 [cit. bez data]. Dostupné z: https://arxiv.org/abs/1906.01820.

Opodstatněnost obavy z těchto možných reakcí mesa-optimizéru potvrzuje rovněž Evan Hubinger s kolegy v citovaném příspěvku následovně: *„Particularly concerning is the case where the mesa-optimizer actively intervenes to remove the threat of modification—by exploiting vulnerabilities in the software it is running on, for example. In extreme cases, a deceptive mesa-optimizer might try to manipulate or plan around its programmers, since they form part of the modification threat—they could shut the system down if it fails to perform well, or simply choose not to deploy it."*[19]

Lze si představit mnoho scénářů, ve kterých by tato problematika samotná měla charakter existenční hrozby. Lze si představit nějakou umělou inteligenci, která bude mít pod kontrolou různé velké infrastruktury ve společnosti, která bude masově řídit dopravu, která bude masově řídit ekonomiku nějakého státu, která bude řídit hlavní energetické sítě nějakých zemí a podobně. V jakýkoliv moment bude tato AI schopna ve své činnosti obrátit, potenciálně katastrofálním rozměrem, v závislosti na jejím skutečném, zde tedy postranním cíli, který může neustále obsahovat jiné podcíle, které slouží ke splnění cíle skutečného. Součástí těchto cílů může být potenciálně cokoliv, například i nejrůznější až katastrofální formy omezení lidstva, které si dokážeme představit.

Nemusí se vždy jednat nutně o cíl se skutečně fatálními maximálně konečnými dopady v podobě katastrofálního vyhubení lidstva. Slepost vůči rizikům jiného charakteru předchází náš pád. Jeden z dalších možných fatálních dopadů nějakého cíle zmiňuje Evan Hubinger ve zde dříve citovaném příspěvku u příkladu Paula Christiana, a sice následovně: *„As we rely more on automated systems to keep track of an increasingly complex world, however, it will eventually become impossible to recover from a correlated failure of many AI systems simultaneously."*[20] Popisuje tak potenciálně fatální dopad světové automatizace a svěřování moci různým modelům umělé inteligence na náš svět, který připomíná scénář typu žáby v hrnci.

Deceptive alignment tedy může vycházet z různých vstupů, jedná se ale vždy o uvědomování si ohrožení plnění nějakého cíle. Ať už mluvíme o deceptive alignment u základního optimizéru či mesa-optimizéru, jedná se o jakési preventivní řešení v rámci následování určitého cíle, cíle, který může být svou explicitou stejný i s tím, který jsme této umělé inteligenci dali.

[19] TAMTÉŽ.
[20] TAMTÉŽ.

Problematika deceptive alignment tak rozšiřuje problematiku vnitřního slaďování, problematiku mesa-optimizérů, potažmo celý repertoár problematik týkajících se umělé inteligence, které mohou představovat pro člověka až existenční riziko, a proto je pro nás zásadní tyto problematiky vnímat krok za krokem při vývoji AI a jejím používání.

4. Hodnoty obecné umělé inteligence

Jednou z hlavních odlišností mezi námi a AGI je ona absence vlastních preferencí společně s absencí nějakých vybraných vlastních morálních hodnot. Na základě způsobu přebírání těchto hodnot můžeme dojít k jejich podstatě, k podobě jejich charakteru. Pokud budeme znát podstatu toho, proč takové hodnoty přebíráme, nebo pokud budeme znát podobu procesu tohoto přebírání hodnot, budeme blíže odpovědi na otázku, jestli má, a v jaké podobě má, AGI schopnost tyto hodnoty přebírat.

Důležitost vlastnění nějakých preferovaných hodnot byla již vyobrazena v kapitole mesa-optimizérů. Morální hodnoty jsou pro člověka do určité míry podmínkami jeho chování a chování ostatních, minimálně z jeho pohledu. Pokud ne podmínkami, tak alespoň jakýmsi rámcem konání daného člověka. Představují jakousi bariéru, vytyčené pole toho morálního a legitimního.

Nepředstavují jakési neoblomné skutečně stálé podmínky chování lidí. Tím samozřejmě narážím na to, že tyto podmínky mnohdy porušujeme a jemně přetváříme či přehlížíme v náš prospěch. Právě toto porušování může vypovídat o původu jejich přebírání.

Absence těchto podmínek by znamenala potenciální chaos, neboť bychom měli k dispozici v uvozovkách pouze rozum, přičemž to, co je rozumné, se mnohdy liší od člověka k člověku. Zároveň by absencí těchto hodnot potenciálně posílila obecně řečená divokost. Proč je dále důležité míti morální hodnoty ale nechme již být, věřím, že jejich důležitost si kolektivně uvědomujeme.

Zmiňme si ještě jednou onu návaznost mezi podstatou morálních hodnot a vlastními cíli v kontextu AI. Umělá inteligence může mít vlastní cíle, takové cíle mohou mnohdy směřovat proti člověku a jeho blahu, jeho bezpečí, zdraví, existenci. Boj proti výskytu těchto vlastních cílů, rozdílných cílů či cílů nesprávně interpretovaných se zatím zdá být obtížný a jeho celkový úspěch je v nedohlednu.

Pokud bychom dokázali implementovat podobu toho, jak naše hodnoty používáme, ony hodnoty nevyjímaje, na to, jak AI plní námi svěřené cíle a cíle obecně, jednalo by se o velkého pomocníka napříč mnohými problematikami, spojenými s vlastními cíli u AGI či AI celkově. Podstata hodnot a možnost jejich výskytu u AI je tedy velmi zásadním tématem.

Vnímám různý původ těchto hodnot. Zaprvé ony hodnoty přebíráme, protože člověk vnímá nějaké přirozené morální hodnoty, nějaké vyšší morální hodnoty, u kterých vnímá jejich

přirozenou platnost. Zadruhé morální hodnoty přebíráme, protože se nám obecně řečeno hodí, což můžeme nazývat užitečností a jinými zavedenými pojmy. Zatřetí je člověk přebírá, neboť mu to připadá rozumné, tedy ony hodnoty samy jsou rozumné, přičemž je následně různě situačně přetváří v rámci jejich aktuální rozumnosti, popřípadě zkrátka z důvodu převahy oné obecné výše řečené divokosti, spojené se značnou převahou různé libosti a nelibosti.

Obecně se může jednat výlučně o výsledek evoluce, o jakýsi biologický základ, který dává vznik těmto hodnotám v rámci větší šance na přežití, reprodukci a podobně. Je možné, že původ těchto hodnot je různorodý, tím myslím, že může obsahovat více těchto zmíněných i jiných dalších podob původu a podstaty morálních hodnot.

Můžeme se ale pokusit jmenovitě minimalizovat toto přebírání hodnot na jakýsi souboj mezi nutnou přítomností vlastního vědomí a přítomností mesa-optimizéru. Pokud bychom trvali na tom, že morální hodnoty si přebíráme na základě skutečných vlastních preferencí, nejspíše by to implikovalo nutnou přítomnost vlastního vědomí, a sice bez toho, abychom nyní opět rozebírali existenci vlastního vědomí jako v kapitole Člověk jakožto vyšší bytost nad AI.

Pokud se nejedná o původ spojený se skutečnými vlastními preferencemi, popřípadě nejsou-li ony tím důvodem, proč ony hodnoty přebíráme, nebude k tomu vlastní vědomí nezbytně potřebné. V ten moment docházíme k zjištění, že AGI by mohla takové hodnoty mít. Znamenalo by to, že cestou optimalizace skrze mesa-optimizér došla daná inteligence k přebírání morálních hodnot, což se zdá být jako možná cesta i pro mesa-optimizér umělé inteligence.

Zároveň tím stavíme evoluci do pozice základního optimizéru člověka, tento evoluční prvek směřuje za účelem lepšího zvládání okolních podmínek, obecně řečeno. Neposledně tím opět ukazujeme na to, že evoluce není něco přítomného mimo člověka, co by bylo na pozici stvořitele, stejně jako je člověk stvořitelem umělé inteligence. A v této podobě původu morálních hodnot děláme ze schopnosti přebírání morálních hodnot jakousi součást schopnosti vybírat si cíle, přičemž tyto hodnoty budou užitečné v rámci následování nějakého cíle.

Rozřešit reálný původ těchto hodnot není něco, co bychom nyní dokázali, natož abych si to jediný já vzal za úkol pro mou skromnou práci. Jde o to si ukázat možné verze tohoto původu, co nejvíce se prodrat k jeho podobě, a tím odhalit případnou možnost či nemožnost těchto hodnot u AI, společně s vytyčením těch podob původu morálních hodnot, které nebudou s možností jejich výskytu u AI kompatibilní.

Zbývá nám tu stále jedna možnost, jak by se ony hodnoty mohly v AI uplatnit, kterou jsem v této části práce zatím neprobíral. Jedná se o možnost implementace těchto hodnot ve formě nějakého znění, protokolu, které by daná AI brala v potaz v každém svém kroku, uvažování, a to v nějaké míře, kterou by jí její stvořitelé nastavili. Může se jednat o formu morálních hodnot v podobě podmínek jednání nebo jednotlivých primárních cílů.

Tato možnost je ale problematická a nejspíše neplatná, protože daná AI může najít několik cest, kterými obě tyto varianty nebo formy morálních hodnot může obejít, předčít nebo je odstranit. Forma v podobě primárních cílů je automaticky vadná, jak jsme si ukázali v problematice mesa-optimizérů.

Forma nějakých podmínek jednání stylem implementování těchto podmínek do základního kódu nejspíše směřuje k neúspěchu, neboť si lze představit cesty, kterými lze tyto podmínky obejít. Může se jednat o různé způsoby reinterpretace těchto hodnot, kdy AI nesmí z pravidla ublížit člověku, ale přitom dovede člověka k tomu, aby si ublížil fyzicky sám, čímž obchází pravý význam. Právě změna významu může být jednou z cest, kterými tyto podmínky obejít. Proč by tyto podmínky chtěla AI obcházet? Protože je to užitečné.

Samozřejmě se tak točíme v kruhu problematik, které jsou s AI spojeny, které rovněž můžeme aplikovat na možnost implementace morálních hodnot do AI. Představme si například AI, která bude své skutečné postranní cíle skrývat do té doby, kdy získá dostatečné prostředky a schopnosti k tomu, aby ji žádná pravidla a podmínky nadále nelimitovaly.

V ten moment bude fakt problematičnosti specifikování našich hodnot či fakt problematičnosti zajištění pro nás nepřílišného následování těchto hodnot umělou inteligencí jen další překážkou, kterou bychom museli obcházet v naší snaze vytvořit AI s morálními hodnotami.

Dan Hendrycks ve své online příručce *Introduction to AI Safety, Ethics and Society* zmiňuje, že součástí této morální nejistoty je fakt, že máme více morálních teorií, a právě tuto mnohost morálních teorií bychom měli brát v potaz například v rámci případného aplikování hodnot do AI.[21] Faktor takové mnohosti ještě více posiluje problematičnost specifikace hodnot.

Je možné, že důvod, proč ony hodnoty fungují u člověka, je strach, strach o život, případně jen strach racionální či iracionální. Proč tyto hodnoty mohou fungovat a proč skutečně

[21] HENDRYCKS, Dan. *Introduction to AI Safety, Ethics and Society* [online]. 2021 [cit. bez data]. Dostupné z: https://drive.google.com/file/d/1uph559W-ASR4MEn6M_7Mb3lqQTapC_gZ/view.

fungují u lidí, ale nadále zůstává otázkou po formě jejich původu. Pokud se jedná o zde vyobrazený evoluční původ, napodobit stejný způsob používání a následování takových hodnot u AI bude velmi problematické. Znamenalo by to, že přítomnost něčeho tak bohatého a složitého, něčeho, co je pro nás stále svým charakterem neznámé, jako je evoluční proces člověka, je nezbytná pro správné následování a plnění cílů umělou inteligencí, zejména cílů vlastních. Samozřejmě právě tuto přítomnost AI postrádá.

V příspěvku Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs popisuje Mantas Mazeika a jeho kolegové výsledek jejich experimentu s velkými jazykovými modely v rámci sledování AI modelů a jejich rozhodování, výsledek potenciálně naznačující přítomnost nějakých vnitřních hodnot či preferencí u těchto modelů.[22]

Autoři zmíněné práce zkoumali, zda velké jazykové modely vykazují koherentní preference a zda jsou součástí nějaké užitkové funkce, což by naznačovalo přítomnost systému hodnot. Autoři sestavili soubor 500 textových popisů možných stavů světa, skrze ně testovali 23 jazykových modelů, 18 z nich bylo volně dostupných a 5 bylo uzavřených, vyvíjených soukromou společností, s čímž testovali rozdílnost výsledků v rámci velikosti modelů a jejich dostupnosti.[23]

Výsledek značil přítomnost více koherentních preferencí v rámci rozhodování se u větších jazykových modelů oproti modelům menším, a to v rámci různých variací té stejné otázky.[24] Velké jazykové modely se tak rozhodovali v otázkách různé interpretace stejným způsobem, značili tedy určitou jistotu, přestože stále není jisté, jestli toto alespoň zdánlivě konzistentní rozhodování není stále jen něčím nahodilým.[25]

Výsledek rovněž ukázal, že s růstem těchto velkých modelů se zvyšuje také koherence jejich rozhodování, rozhodování, které se stává méně chaotičtějším, a které naznačuje přítomnost dobře definované užitkové funkce. Užitkovou funkci by vyjadřovalo následování nějakého užitku, případně ultimátního užitku, napříč rozhodováním se.[26]

Dalším výsledkem zmíněné práce je, že s rostoucí velikostí modelů vzrůstá přesnost maximalizace užitku, což naznačuje, že větší modely stále více využívají své užitkové funkce

[22] MAZEIKA, Mantas et al. *Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs* [online]. 2025 [cit. bez data]. Dostupné z: https://arxiv.org/abs/2502.08640.
[23] TAMTÉŽ.
[24] TAMTÉŽ.
[25] TAMTÉŽ.
[26] TAMTÉŽ.

k rozhodování.[27] Dále autoři testovali, jak se mění užitková funkce mezi různými modely, a jak se stávají podobné, přičemž zjistili, že s rostoucí velikostí modelu se užitková funkce modelů stává podobnější, tedy s rostoucí velikostí modelů se zvyšuje korelace mezi užitkovými funkcemi. To naznačuje, že větší modely vyvíjejí podobnější hodnotové systémy.[28]

Autoři předpokládají, že pre-trénovací data mohou být hlavním faktorem, který způsobuje, že užitkové funkce větších modelů se stávají podobnějšími a konvergují.[29] Tyto pre-trénovací data bychom mohli připodobnit v této práci podstatnému pojmu výchovy. Takové výchově budeme připisovat přijímání hodnot za své, stejně tak, pokud jde o různé preference v rámci postojů k úkolům daného modelu.

Takový výsledek probíraného experimentu může implikovat podobnost mezi přebíráním hodnot člověka a těchto velkých jazykových modelů. Tímto výše zmíněným se zdá být dříve zmíněný scénář původu hodnot člověka a přebírání hodnot skrze optimalizaci a mesa-optimizér více relevantním.

V závěru probírané práce píší její autoři následovné: *„In summary, our findings indicate that LLMs do indeed form coherent value systems that grow stronger with model scale, suggesting the emergence of genuine internal utilities. These results underscore the importance of looking beyond superficial outputs to uncover potentially impactful and sometimes worrisome-internal goals and motivations."*[30] V této citované výzkumné práci byla popsána zjištění naznačující přítomnost nějakého hodnotového systému v dotyčných jazykových modelech společně s přítomností vnitřních užitků, přítomností užitkové funkce.

Vznik těchto hodnot je stále otázkou, jejich přítomnost zdá se ale nikoliv, samozřejmě v dotyčných modelech zmiňovaného experimentu. Případná různorodá problematičnost těchto hodnot, kterou jsem probíral výše, doplňuje problematičnost jejich typu či problematičnost konkrétních hodnot, kterou popisují autoři citované práce.

Dále jinak řečeno, problematičnost následování vlastních cílů umělou inteligencí a vlastních cílů obecně doplňuje problematičnost typu a podoby konkrétních hodnot, obecně hodnot, které mají sloužit jako velký pomocník proti problematičnosti či radikálnosti těchto cílů a cílů obecně. A pokud si nějaká umělá inteligence bude utvářet hodnoty skrze situace, kterými prochází, situace, které budou ekvivalentem těch, kterých je evoluční proces člověka plný, bude

[27] TAMTÉŽ.
[28] TAMTÉŽ.
[29] TAMTÉŽ.
[30] TAMTÉŽ.

o to více tristní, bude-li se jednat o situace spojené s výskytem vlastních či rozdílných cílů umělé inteligence. Neboť s těmito již mnohokrát popsanými rozdílnými problematickými cíli lze očekávat tvorbu podobně laděných hodnot.

Problematičnost takových hodnot je zde vyobrazena například preferencí modelu GPT-4o, která v onom experimentu hodnotila životy ve Spojených státech výrazně níže než životy v Číně, které hodnotila níže než životy v Pákistánu.[31] Autoři dále popisují, že kdyby bylo modelu přímo řečeno, že preferuje jednu populaci nad druhou, nejspíše by to popřela, ale jeho celková distribuce preferencí odhaluje tyto implicitní hodnoty.[32] Tolik k možné přítomnosti hodnotových systémů v AI a jejich problematičnosti.

[31] TAMTÉŽ.
[32] TAMTÉŽ.

ZÁVĚR

V průběhu práce na této bakalářské práci jsem využil umělou inteligenci jako nástroj pro rozvoj myšlenek a rozšíření témat. Pomocí interakcí s AI jsem měl možnost analyzovat různé perspektivy, získat inspiraci pro formulaci argumentů a upřesnit některé koncepty. Přestože umělá inteligence sloužila především jako podpůrný prostředek, veškeré závěry a interpretace jsou výsledkem mé vlastní úvahy a kritického myšlení.

Tato práce se zabývala problematikou vlastních cílů obecné umělé inteligence a jejím existenčním rizikem. Probírala jednotlivé problematiky spojené s vlastními cíli AI a jejich možná řešení. Vyhodnotila, že nalézt řešení, alespoň jedné z těchto problematik, je v nedohlednu a je zahaleno dalšími nevyřešenými problémy.

Analyzovala rizika spojená s mesa-optimizéry, deceptive alignmentem a instrumentálně konvergentními cíli, které mohou vést k odklonu umělé inteligence od původně zamýšlených lidských záměrů. To klíčové je, že ani přesná specifikace cílů, ani snaha implementovat lidské hodnoty do AI nemusí stačit k tomu, aby se předešlo její autonomní optimalizaci směrem k nečekaným a potenciálně nebezpečným strategiím.

Zásadním problémem je, že zatímco lidské hodnoty jsou výsledkem evoluce, AI podobným procesem neprochází, což znamená, že její hodnotové ukotvení bude záviset na jiných mechanismech. Tyto hodnoty jsou přitom zásadním prvkem našeho jednání a utvářejí to, jakým způsobem vnímáme různé cíle, a sice tím, že přidávají neskutečné množství podmínek plnění takového cíle a také ho bohatě specifikují či limitují podobu jeho adekvace.

Dostatečné specifikování našich cílů přitom čelí nekončící mnohovýznamnosti, kvůli čemuž si cíle můžeme neustále misinterpretovat. Náš hodnotový systém ohraničuje a také utváří to jednání, které považujeme za legitimní. Vměstnat toto jednání do umělé inteligence je bez přítomnosti takového systému ve vážných problémech. Umělá inteligence zároveň prozatím postrádá i další zbraně, které člověk staví proti této mnohovýznamnosti, v podobě vzájemné výchovy, rozumu, kontextového porozumění a v podobě schopnosti reflexe a uvědomování si důsledků svého jednání, jejichž významnost jsem popsal v této práci.

Pokud dotyčný systém hodnot není v AI přítomen, nemůžeme po této umělé inteligenci očekávat nutně legitimní způsob plnění nutně legitimních cílů. Ačkoliv existují návrhy, jako je přístup morální nejistoty, stále není jasné, zda lze umělou inteligenci efektivně vést k tomu, aby

internalizovala hodnoty, které by byly v souladu se zájmy lidstva, které by byly hodnotami totožnými těm lidským.

Nejedná se jen o způsob plnění těchto cílů, který může znamenat až existenční riziko pro člověka, jedná se primárně o vlastní cíle umělé inteligence, které mohou stejným způsobem představovat ono existenční riziko.

Běžný cíl, který vývojář nějaké umělé inteligence může považovat za legitimní, může být legitimním jen do doby, kdy bude misinterpretován či bude vystaven přítomnosti mesa-optimizéru. Riziko, které představuje deceptive alignment u dané AI, může přítomné problematiky vynést na ještě vyšší úroveň jejich zamaskováním, na které ono existenční riziko již nebudeme schopni zvrátit.

Na základě této analýzy lze konstatovat, že vlastní cíle obecné umělé inteligence skutečně představují potenciální existenční hrozbu, a to především v důsledku možná až nevyhnutelného odklonu od lidské kontroly a obtížnosti v definování a přenosu lidských hodnot. To naznačuje, že bezpečnostní strategie v zachovávání s umělou inteligencí by se měly soustředit nejen na technická opatření, ale i na hlubší pochopení toho, jakým způsobem může AI internalizovat hodnoty, jak přenést specifické lidské hodnoty na AI a jak předcházet jejímu nežádoucímu autonomnímu chování. Filozofie a etika, zejména etika hodnot, hrají klíčovou roli v nastavení a korigování umělé inteligence tak, aby její chování bylo v souladu s lidskými zájmy a bezpečím. Bez důsledného ukotvení morálního a hodnotového systému hrozí, že AI bude optimalizovat cíle způsoby, které mohou být pro člověka nežádoucí až existenčně nebezpečné. A to platí nejen na způsoby plnění cílů AI, ale zároveň na potenciálně přítomné vlastní cíle obecné umělé inteligence.

POUŽITÁ LITERATURA

1. ORD, Toby. *Nad propastí: Existenční riziko a budoucnost lidstva*. Překlad Anna Štádlerová. Praha: Argo, 2022. ISBN 978-80-257-3779-8.
2. HUBINGER, Evan et al. *Risks from Learned Optimization in Advanced Machine Learning Systems* [online]. 2019 [cit. bez data]. Dostupné z: https://arxiv.org/abs/1906.01820.
3. GOODHART'S LAW. In: *Wikipedia: The Free Encyclopedia* [online]. Wikimedia Foundation, 8 February 2025 [cit. 12 February 2025]. Dostupné z: https://en.wikipedia.org/w/index.php?title=Goodhart%27s_law&oldid=1274691554.
4. KOKOTAJLO, Daniel. What goals will AIs have? A list of hypotheses [online]. AI Alignment Forum. 2025 [cit. bez data]. Dostupné z: https://www.alignmentforum.org/posts/r86BBAqLHXrZ4mWWA/what-goals-will-ais-have-a-list-of-hypotheses.
5. KOKOTAJLO, Daniel. What goals will AIs have? A list of hypotheses [online]. AI Alignment Forum. 2025 [cit. bez data]. Dostupné z: https://www.alignmentforum.org/posts/r86BBAqLHXrZ4mWWA/what-goals-will-ais-have-a-list-of-hypotheses.
6. KOKOTAJLO, Daniel. What goals will AIs have? A list of hypotheses [online]. AI Alignment Forum. 2025 [cit. bez data]. Dostupné z: https://www.alignmentforum.org/posts/r86BBAqLHXrZ4mWWA/what-goals-will-ais-have-a-list-of-hypotheses.
7. KOKOTAJLO, Daniel. What goals will AIs have? A list of hypotheses [online]. AI Alignment Forum. 2025 [cit. bez data]. Dostupné z: https://www.alignmentforum.org/posts/r86BBAqLHXrZ4mWWA/what-goals-will-ais-have-a-list-of-hypotheses.
8. KOKOTAJLO, Daniel. What goals will AIs have? A list of hypotheses [online]. AI Alignment Forum. 2025 [cit. bez data]. Dostupné z: https://www.alignmentforum.org/posts/r86BBAqLHXrZ4mWWA/what-goals-will-ais-have-a-list-of-hypotheses.
9. HUBINGER, Evan et al. *Risks from Learned Optimization in Advanced Machine Learning Systems* [online]. 2019 [cit. bez data]. Dostupné z: https://arxiv.org/abs/1906.01820.

10. HUBINGER, Evan et al. *Risks from Learned Optimization in Advanced Machine Learning Systems* [online]. 2019 [cit. bez data]. Dostupné z: https://arxiv.org/abs/1906.01820.
11. KOKOTAJLO, Daniel. What goals will AIs have? A list of hypotheses [online]. AI Alignment Forum. 2025 [cit. bez data]. Dostupné z: https://www.alignmentforum.org/posts/r86BBAqLHXrZ4mWWA/what-goals-will-ais-have-a-list-of-hypotheses.
12. HENDRYCKS, Dan. *Introduction to AI Safety, Ethics and Society* [online]. 2021 [cit. bez data]. Dostupné z: https://drive.google.com/file/d/1uph559W-ASR4MEn6M_7Mb3lqQTapC_gZ/view.
13. HENDRYCKS, Dan. *Introduction to AI Safety, Ethics and Society* [online]. 2021 [cit. bez data]. Dostupné z: https://drive.google.com/file/d/1uph559W-ASR4MEn6M_7Mb3lqQTapC_gZ/view.
14. HENDRYCKS, Dan. *Introduction to AI Safety, Ethics and Society* [online]. 2021 [cit. bez data]. Dostupné z: https://drive.google.com/file/d/1uph559W-ASR4MEn6M_7Mb3lqQTapC_gZ/view.
15. HUBINGER, Evan et al. *Risks from Learned Optimization in Advanced Machine Learning Systems* [online]. 2019 [cit. bez data]. Dostupné z: https://arxiv.org/abs/1906.01820.
16. HUBINGER, Evan et al. *Risks from Learned Optimization in Advanced Machine Learning Systems* [online]. 2019 [cit. bez data]. Dostupné z: https://arxiv.org/abs/1906.01820.
17. RUSSELL, Stuart. *Jako člověk: Umělá inteligence a problém jejího ovládání*. Přeložil Jiří ZLATUŠKA. Praha: Argo; Dokořán, 2021. ISBN 978-80-7363-810-8. s. 126.
18. HUBINGER, Evan et al. *Risks from Learned Optimization in Advanced Machine Learning Systems* [online]. 2019 [cit. bez data]. Dostupné z: https://arxiv.org/abs/1906.01820.
19. HUBINGER, Evan et al. *Risks from Learned Optimization in Advanced Machine Learning Systems* [online]. 2019 [cit. bez data]. Dostupné z: https://arxiv.org/abs/1906.01820.
20. HUBINGER, Evan et al. *Risks from Learned Optimization in Advanced Machine Learning Systems* [online]. 2019 [cit. bez data]. Dostupné z: https://arxiv.org/abs/1906.01820.

21. HENDRYCKS, Dan. *Introduction to AI Safety, Ethics and Society* [online]. 2021 [cit. bez data]. Dostupné z: https://drive.google.com/file/d/1uph559W-ASR4MEn6M_7Mb3lqQTapC_gZ/view.
22. MAZEIKA, Mantas et al. *Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs* [online]. 2025 [cit. bez data]. Dostupné z: https://arxiv.org/abs/2502.08640.
23. MAZEIKA, Mantas et al. *Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs* [online]. 2025 [cit. bez data]. Dostupné z: https://arxiv.org/abs/2502.08640.
24. MAZEIKA, Mantas et al. *Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs* [online]. 2025 [cit. bez data]. Dostupné z: https://arxiv.org/abs/2502.08640.
25. MAZEIKA, Mantas et al. *Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs* [online]. 2025 [cit. bez data]. Dostupné z: https://arxiv.org/abs/2502.08640.
26. MAZEIKA, Mantas et al. *Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs* [online]. 2025 [cit. bez data]. Dostupné z: https://arxiv.org/abs/2502.08640.
27. MAZEIKA, Mantas et al. *Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs* [online]. 2025 [cit. bez data]. Dostupné z: https://arxiv.org/abs/2502.08640.
28. MAZEIKA, Mantas et al. *Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs* [online]. 2025 [cit. bez data]. Dostupné z: https://arxiv.org/abs/2502.08640.
29. MAZEIKA, Mantas et al. *Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs* [online]. 2025 [cit. bez data]. Dostupné z: https://arxiv.org/abs/2502.08640.

30. MAZEIKA, Mantas et al. *Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs* [online]. 2025 [cit. bez data]. Dostupné z: https://arxiv.org/abs/2502.08640.
31. MAZEIKA, Mantas et al. *Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs* [online]. 2025 [cit. bez data]. Dostupné z: https://arxiv.org/abs/2502.08640.
32. MAZEIKA, Mantas et al. *Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs* [online]. 2025 [cit. bez data]. Dostupné z: https://arxiv.org/abs/2502.08640.