UNIVERZITA KARLOVA

FAKULTA HUMANITNÍCH STUDIÍ

Bakalářská práce

Umělá inteligence a Superinteligence

Michal Krňák

Vedoucí práce: Mgr. Richard Zika, Ph.D.

Praha 2023

Bibliografický záznam

KRŇÁK, Michal. *Umělá inteligence a Superinteligence*. Praha, 2023. Bakalářská práce (Bc.). Univerzita Karlova, Fakulta humanitních studií, Studium humanitní vzdělanosti. Vedoucí práce: Mgr. Richard Zika, Ph.D.

Abstrakt

V současné době se v rámci vzrůstajícího pokroku oboru umělé inteligence stává tato problematika předmětem mnoha debat. Pojem umělá inteligence je tak skloňován v mnoha ohledech a perspektivách, které už nezahrnují pouze počítačovou vědu. Je proto potřeba tyto pojmy teoreticky uvést na pravou míru a správně vysvětlit jejich fungování, aby nedocházelo k záměně významů nebo přisuzování vlastností, které jim nenáleží.

Tato práce si tak klade za cíl podat strukturovaný vhled do problematiky funkčnosti umělé inteligence, které je v současnosti laickou veřejností přisuzováno myšlení, a vysvětlit význam a rizika, která představuje pojem Superinteligence neboli umělé inteligence obecně převyšující člověka, které se ve stejnojmenné knize věnuje filosof Nick Bostrom.

Klíčová slova: umělá inteligence, Superinteligence, lidstvo, Nick Bostrom

Abstract

Nowadays, with increasing progress in the field of artificial intelligence, the matter of artificial intelligence has become a subject of much debate. The term artificial intelligence is understood in many ways and perspectives that no longer include only the field of computer science. There is therefore a need to set these concepts straight theoretically and to explain how they work, so as not to confuse what they mean, or attribute to them properties that do not belong to them.

This thesis aims to provide a structured insight into the issue of the functionality of artificial intelligence, to which the ability to think is currently attributed by the general public, and to explain the meaning of and risks posed by the concept of Superintelligence, or artificial intelligence generally superior to humans, which is discussed by philosopher Nick Bostrom in the book of the same name.

Key words: artificial intelligence, Superintelligence, humanity, Nick Bostrom

Čestné prohlášení

Prohlašuji, že jsem tuto práci vypracoval samostatně. Všechny použité prameny a literatura byly řádně ocitovány. Práce nebyla využita k získání jiného nebo stejného titulu.

V Praze, dne 23. 6. 2023 ……

Poděkování

Tímto bych rád poděkoval vedoucímu této práce Mgr. Richardu Zikovi, Ph.D. za čas, vstřícnost a především trpělivost investovanou do připomínkování této práce.

Obsah

1. Úvod

Pro průměrného člověka 21. století žijícího v moderní společnosti jsou výpočetní technologie každodenní součástí jeho života. Přichází s nimi neustále do kontaktu a jeho život přímo i nepřímo ovlivňují. Od 30. až 40. let minulého století, kdy se začaly objevovat první počítače, dnes označované jako počítače nulté generace, ušla tato technologie velký kus cesty. Z výlučně akademického prostředí se s postupně stoupající výpočetní silou a nižší energetickou náročností, jakož i se stále zmenšující se velikostí rozšířily počítače do kuchyňských spotřebičů, aut, telefonů nebo hodinek. S rostoucím technologickým rozvojem společnosti začaly být na počítače kladeny větší nároky na složitější výpočty, díky čemuž byly vyvíjeny stále zdatnější softwary, které byly schopny splňovat stále náročnější požadavky. S touto potřebou zpracovávání stále komplexnějších dat a operování s nimi přichází pokusy o vytvoření softwaru, který by tuto úlohu vykonával místo lidí. Takovýto software ale i jemu odpovídající hardware lze označit jako umělou inteligenci.

V současnosti je počítačový sektor tím nejdynamičtějším, co se ekonomického růstu týče, a nepředpokládá se, že se růst nějak výrazně zpomalí. S tím se pojí i tlak na vývoj stále výkonnějších a schopnějších technologií, ať už se jedná o hardware, či software. To jde ruku v ruce se snahami o vývoj schopnějších umělých inteligencí, které by nám usnadnily cestu k dalším cílům, které si lidé před sebe kladou. Řada současných myslitelů – v čele s Nickem Bostromem, jehož text je pro tuto práci klíčovým zdrojem – si pokládá otázku, co nastane, až budou umělé inteligence schopné myslet stejně a pravděpodobně i lépe nežli my, uvědomovat si samy sebe a disponovat rozumem a vůlí, a jaká úskalí to s sebou pro nás jako lidstvo přináší.

Vidina vzniku umělé inteligence chytřejší než jakýkoliv člověk pro autora této práce znamená, že, jelikož se jedná o produkt exaktních věd, který by mohl v lidském životě jednoho dne zaujímat tak významnou pozici, je nutné k tomuto problému připojit rovněž humanističtější přístup. Již samo směřování ke schopnosti vytvořit nástroj mnohem chytřejší než člověk, který by člověku měl sloužit, totiž dává vzniknout rozsáhlým debatám ohledně otázek vzniku, účelu, hodnot nebo vědomí takové Superinteligence a v důsledku i rozebírání otázky lidství. Je zcela patrné, že záběr této problematiky protíná mnoho oborů. Proto si tato práce klade za cíl poskytnout

strukturovaný a pochopitelný základní vhled do mnohovrstevnatého oboru, principů a pojmů týkajících se umělé inteligence optikou Nicka Bostroma a v kontextu dalších myslitelů.

2. **Transhumanismus**

Aby dostalo rozebírání problematiky umělé inteligence patřičný význam, je prvně nutné definovat myšlenkový kontext, ve kterém je vlastně pojem umělé inteligence (zkráceně UI[1]) pro její zastánce ukotven. Jako laici můžeme konstatovat, že je to jednoduše praktická věc se spoustou možných využití, ale pro akademické účely je dobré takový názor o něco opřít. Proto užití pojmu transhumanismus, tedy myšlenkového proudu, který souhrnně otevírá debatu a předkládá otázky vztahu člověka a technologie, je nasnadě. Ten ve zjednodušené formě zastává pozitivní stanovisko vůči vědeckým objevům a novým nástrojům lidstva, které mohou člověka a lidstvo obecně dovést k větší prosperitě a příznivějším, nikoliv přirozenějším, podmínkám pro život. Při obecném vymezování toho, čím vlastně transhumanismus je, nemusíme nutně ani sahat do nedávno napsané literatury, jelikož otázka lidské přirozenosti a jejího překonávání a posouvání byla předmětem debat už v antice.

O tom více či méně obrazně pojednává už mýtus o Prométheovi (Petiška, 2006). První lidé, které Prometheus (s pomocí bohů) stvořil, byli bezbranní. Neuměli využívat své schopnosti, obstarat si potravu a obydlí. Tomu všemu je musel Prométheus naučit. Učil je rozumět světu kolem sebe a co nejlépe využívat k vlastnímu prospěchu věci v něm. Pověstný přinesený oheň, byť rozhodně není ničím, co bychom z dnešního pohledu považovali za technologii, představuje nejen pouhý nástroj ke konkrétním účelům zajištění tepla, světla a jídla. Jde zde o prezentování vztahu člověka a přírody, která před něj staví neustále různé překážky, a proto se jejích zdrojů snažíme využít tak, abychom co nejvíce odsunuli hlad, žízeň, nemoci a smrt dál od našeho obydlí a vytvořili si vlastní svět, kterému jsou takovéto těžkosti co nejvíce cizí.

Na těchto ideálech oslavy lidského intelektu, pokroku a kultury s důrazem na člověka jako takového dále staví humanismus a na něj navazující osvícenství. V těchto

[1] Poznámka autora: Lze také používat mezinárodní zkratku AI vycházející z anglického Artificial intelligence, která se také běžně používá i v české literatuře.

myšlenkových směrech je formulován požadavek po systematickém vědeckém zkoumání světa s ohledem na prospěch lidské společnosti.

Myšlenka rozšířit lidské možnosti začíná ve společnosti ve větší míře rezonovat po roce 1859, kdy Charles Darwin (2007) publikoval svou práci *O vzniku druhů přírodním výběrem*, která naznačovala, že člověk má přirozený potenciál se dále vyvíjet. To nepřímo podnítilo debatu nejen ohledně toho, kam by mohl směřovat další lidský vývoj, ale také zda je možné, abychom sami korigovali tento vývoj a jakými prostředky by se to mělo uskutečňovat.

Pokud bychom hledali novější náhled na lidskou potřebu vzdorovat příkoří plynoucímu z omezení přírodního světa, lze přihlédnout k dílu Vita activa Hannah Arendtové (2007). Podle ní člověk boural přirozené bariéry, které okolo něj klade příroda a sám si vytváří vlastní umělý svět, čímž nepřímo proměňuje i svou přirozenost. V současné době se však již stále častěji setkáváme s pokusy o přímé zásahy do lidské přirozenosti, což lze doložit příklady medicínských snah o prodloužení lidského života, genetické manipulace směřující k jeho zkvalitnění nebo samotného vytvoření nového života umělým oplodněním. Arendtová tak v podstatě hovoří v souladu s bájí o Prométheovi. Každý lidstvu prospěšný vynález a jeho praktické využívání je pak posouváním naší lidské přirozenosti.

Transhumanismus je ale sám o sobě poměrně mladý myšlenkový proud, o čemž vypovídá i první užití tohoto názvu Julianem Huxleym (1957, s. 17) v knize *New bottles for New wine* v tomto znění:

Lidský druh může, pokud chce, překonat sám sebe – nejen sporadicky, jeden jednotlivec zde nějak, jiný jinde jinak – ale jako celek, jako lidstvo. Pro tuto novou víru potřebujeme jméno. Možná poslouží transhumanismus: člověk zůstává člověkem, ale překračuje sám sebe tím, že uskutečňuje nové možnosti své lidské přirozenosti a pro ni jako takovou.[2]

Spojíme-li potom všechny uvedené exkurzy, přičemž by se dalo najít zajisté více příkladů obsahujících tyto názory, docházíme k současné definici transhumanismu tak,

[2] The human species can, if it wishes, transcend itself – not just sporadically, an individual here in one way, an individual there in another way – but in its entirety, as humanity. We need a name for this new belief. Perhaps transhumanism will serve: man remaining man, but transcending himself, by realizing new possibilities of and for his human nature. HUXLEY, Julian. New Bottles for New Wine. *The Sociological Review* [online]. 2016, 64(1_suppl), 318 s. [cit. 2022-06-16]. ISSN 0038-0261. Dostupné z: doi:10.1111/2059-7932.12018. Vlastní překlad.

jak jí podává Nick Bostrom (2014, s. 1), který se sám k tomuto proudu veřejně hlásí. Ona jeho definice potom zní takto:

1. *Intelektuální a kulturní hnutí, prohlašující za možné a vhodné zásadní zlepšení lidských podmínek za použití lidského rozumu, zejména rozvojem a širším zpřístupněním technologií k odstranění stárnutí a obrovskému vylepšení lidského intelektu, fyzické a psychické kondice.*
2. *Studium důsledků, příslibů a potenciálních nebezpečí technologií, které nám umožní překonat základní lidská omezení, a související studium etických otázek spojených s užíváním takových technologií.*[3]

Z tohoto Bostromova souhrnu lze odvozovat, jak rozumí myšlence transhumanismu její stoupenci. Technologie jsou z této perspektivy chápány jako dobré a mají všeobecně lidstvu sloužit ku prospěchu. Celkově vzato je tedy podle transhumanismu nejen eticky přípustné, ale naprosto žádoucí do společnosti zavádět technologie takovým způsobem, který je co nejprospěšnější. Pokud tento výměr použijeme na problematiku umělé inteligence, jde „jen" o další z ohňů a způsobů, jakými se vymanit z naší původní přirozenosti. Ale tentokrát nám, jakožto celému lidstvu, nejde o upečení masa nebo postavení pohodlnějších příbytků a tvorbu umění, ale o nástroj, který bude „prodlouženou hbitější rukou" lidské kognice. Je třeba k němu přistupovat velice opatrně, protože s tímto „ohněm" si není radno zahrávat. Už jen proto, že takový plamen má potenciál zastínit svého tvůrce. Stejně jako v případě jiných „ohňů" tedy i zde platí ono známé: „Oheň je dobrý sluha, ale zlý pán.".

Vyústěním této části tedy je vysvětlení v úvodu nastíněného stoupajícího trendu užívání technologií v každodenním životě jako projevu lidské potřeby zlepšit své každodenní podmínky, která je nám jako druhu vlastní odedávna. Tato snaha o vědecké průlomy tedy není nikterak novým fenoménem. Transhumanistické snahy ubírající se směrem k vytvoření UI pro účely zkvalitnění lidských životních podmínek jsou zcela racionálním vyústěním těchto procesů v moderní době, pročež je třeba začít nastolovat debaty, které se právě umělé inteligence více dotýkají. Rovněž musím konstatovat, že na

[3](1) The intellectual and cultural movement that affirms the possibility and desirability of fundamentally improving the human condition through applied reason, especially by developing and making widely available technologies to eliminate aging and to greatly enhance human intellectual, physical, and psychological capacities.

(2) The study of the ramifications, promises, and potential dangers of technologies that will enable us to overcome fundamental human limitations, and the related study of the ethical matters involved in developing and using such technologies. BOSTROM, Nick, ed. Introduction—The Transhumanist FAQ. MERCER, Calvin a Derek F. MAHER. *Transhumanism and the Body* [online]. 1. New York: Palgrave Macmillan, 2014, s. 1-17 [cit. 2022-06-13]. ISBN 978-1-137-34276-8. Dostupné z: doi: https://doi.org/10.1057/9781137342768 Vlastní překlad.

tomto poli se zcela nepokrytě střetávají jak státoprávní, filosofické, sociologické a výpočetně-technologické perspektivy, tak i perspektivy přírodních věd. To z této debaty dělá zajímavé střetnutí hned několika světů, o kterém je nutné vést diskuse, a to právě i v tomto mnohovrstevnatém kontextu. Nejde jen o to se ptát, jak daný vynález vytvořit, ale položit si otázku, jak moc jsme schopni mu porozumět, a podle toho i jednat. Argument pro snahy o vytvoření Superinteligence tedy lze shrnout jako další logický krok v touze se přirozeně posunout dál od původní čistě biologické přirozenosti.

3. **Inteligence lidská a umělá**

Vědecko-fantastická představa, že uvnitř chladného stroje žije řadou jedniček a nul komponovaná myslící bytost, která se svými schopnostmi vyrovná člověku, vůči němuž by se mohla vzepřít, je nám prozatím stále stejně vzdálená jako před padesáti lety. Úkolem této kapitoly je vysvětlit, proč tomu tak je a ukázat si současný stav oboru umělé inteligence, jeho nedávnou historii a jaké rozdíly či podobnosti mohou v lidském a strojovém myšlení být. Ostatně problém vědomí UI, a v této souvislosti i člověka, je červenou linkou, která prochází celou historií tohoto oboru. Pravděpodobně právě proto, že překoná-li nás stroj inteligenčně, překoná nás celkově. Náš strach pak tkví zejména v tom, aby nás vůbec UI chytřejší než člověk byla schopná stále ještě akceptovat a byla ochotná s námi komunikovat.

Tyto rozdíly a podobnosti tedy naznačují podmínky pro vytvoření UI i rizika spojená s jejím následným využitím a mají v této kapitole své místo. Je tomu tak zejména proto, abych ukázal, že dnes ještě žádné svébytné umělé myslící *Já* nemáme a z jistého úhlu pohledu jím ani my sami nejsme. Tím spíš je označení *umělá inteligence* nepravdivé, respektive krajně nepřesné. Rovněž nám tato kapitola bude v budoucnu sloužit jako dobrý ukazatel rozdílů mezi Superinteligencí, tedy něčím, co má být pro Bostroma „opravdovou UI“, a soudobými UI, pro které je zcela jistě přiléhavějším označením „strojové učení“.

Lidská perspektiva v této problematice je sice naším výchozím bodem, ale i ta má své nedostatky. Je tak třeba nastínit naprostý základ, na kterém tento obor staví a podívat se, v jaké logice a z jakého úhlu pohledu k problému vztahu člověka a UI přistupuje. V úvodu bylo předestřeno, že se pojednávání o umělých inteligencích nachází na hranicích hned několika oborů. Nejde jen o debatu filosofickou, ale i o spojení exaktních

věd jako je biologie, chemie, fyzika a informatika. Díky tomu existuje řada hypotéz a přístupů. Je proto potřeba podívat se na to, jak chápeme pojem inteligence i ve vztahu k člověku, abychom byli schopni z naší lidské perspektivy definovat strojovou inteligenci, nebo se naopak od této perspektivy oprostit. Pokoušíme-li se totiž vytvořit UI, která by se alespoň kognicí rovnala člověku, je třeba nějak obecně chápat tuto „veličinu" jako takovou a v kontextu vůči lidstvu. Tedy i včetně perspektivy dosavadního přírodovědného poznání, které má také k tomuto tématu co říct.

Yuval Noah Harari sleduje pohled na vztah člověka a UI z perspektivy biochemie a zastává východisko, že lidské jednání a myšlení je výsledkem daných chemických procesů v mozku, který se jednoduše pokouší zachovat naše geny, za které přímo zodpovídá. Dá se s nadsázkou tvrdit, že podobně jako si lidé v počítačích vytvářejí své dílčí programy, vytvořila evoluce vlastní naprogramování v rámci nejen lidských mozků, kterého si nejsme přímo vědomi, a které se nás, jakožto druh snaží zachovat? Má skutečně pro dnešní přírodovědní poznání pojem duše ještě místo? Reaguji tím na hlasy, které skepticky tvrdí, že si uměle vytvořená inteligence nikdy nebude vědoma sama sebe a člověk díky své duši zůstává na vrcholu potravního i intelektuálního řetězce.

Harari sice započíná svůj exkurz se záměrem zpochybňovat individualismus v rámci kritiky liberalismu. My se ale na tento exkurz dívejme jako na jisté přiblížení toho, jak fragmentární myšlení může být. Popisuje, jak by se vyjadřovali dnešní vědci ohledně člověka, který se rozhodl někoho zavraždit. Objevování mechanismu našeho myšlení je ovšem pro lidstvo poměrně novou záležitostí z posledních pár století, což znamená, že jsme si museli vytvořit nějaké uspokojivé vysvětlení už předtím. A tak vycházíme z toho, že se takový vrah pro zločin přímo svobodně rozhodl ze zkaženosti vlastní duše. Na to by ale dnešní genetika a neurologie aplikovala pojmy jako genetický obrys, evoluční tendence a elektrochemické procesy v mozku (Harari, 2017, s. 278). Jsme obětí chemie našich vlastních mozků, která se o příčinnost našeho jednání dělí už jen s náhodností. „Když neuron vystřelí elektrický výboj, může se jednat buď o danou reakci na vnější podnět, nebo o výsledek náhodné události, jakou je například spontánní rozpad radioaktivního atomu. V žádném případě se zde nejedná o svobodnou vůli. Lidské rozhodnutí je vyvrcholením celého řetězce biochemických akcí, každá z nich je určena tou předchozí, a proto nemůže být svobodné" (tamtéž). Zaujmeme-li toto vědecké vysvětlení, kterým popřeme existenci svobodné vůle, znamená to, že mozek stavíme na úroveň biologického mechanismu, který lze kontrolovat a ovládat jej. Ať už se jedná

o genetické inženýrství, přímou elektrickou stimulaci nebo využití drog a jiných látek (Harari, 2017, s. 282). Současnou UI můžeme považovat za „hloupý“ nástroj. Abychom ale tomuto podceňování do budoucna zabránili a nebrali UI na lehkou váhu, musíme si uvědomit, že i my sami jsme myslící agenti, kterými jde manipulovat. Na rozdíl od existujících UI ale fungujeme na principu biologického substrátu.

Ostatně manipulování s mozkem je naše běžná praxe. V kontextu současných modelů UI je často skloňován pojem strojového učení nebo trénování. Aby UI fungovala, jak mají její vývojáři v úmyslu, musí se nejprve naučit zacházet s konkrétním druhem dat. Dostane tak k dispozici trénovací data, tedy sadu informací, která slouží k výcviku modelu. Model je pak schopný generalizovat své znalosti a použít je na nová, dosud neviděná data. Přesně takhle funguje i lidská kognice. Aby mozek získal nějakou kognitivní schopnost, musí být konfrontován s problematikou, jejímž řešením si tuto dovednost vypěstuje. Školní instituce tak můžeme v této analogii považovat právě za trénovací centra organizovaně poskytující našim mozkům data a tím pádem i vylepšující lidskou kognici. Užíváním specializovaného datového vzorku pak může společnost vychovávat jedince schopné porozumět specifickým problematikám a prohlubovat tak lidské poznání. Dochází tak k systematickému zvyšování naší inteligence bez přímého technologického zásahu. Lidská společnost sama o sobě je „mechanismem“ a projevem lidské kolektivní inteligence a kolektivní inteligencí o sobě. Souhrnně můžeme za projevy tohoto typu inteligence označit právě zmíněné školství, organizovanost firem, dopravy, mezinárodní spolky i státy jako takové. Rovněž vynálezy jako knihtisk, bezdrátová komunikace nebo objevy v medicíně měly přímý dopad na zvýšení celkové inteligence populace. Vytváříme si tak systém, ve kterém jsme například kvůli individuálním schopnostem a nadáním vedeni k volbě specializací, které zpětně společnost udržují a posouvají (Bostrom, 2018, s. 94–95). Dalším takovým způsobem manipulace na nás samých je jednoduše zdravá strava, pohyb, správný spánkový režim, nebo kofein a nikotin obsahující přípravky. Tím vším se sice na hyper inteligentní úroveň nedostaneme, ale můžeme tím větší části populace zlepšit kvalitu života, což pak může vést k dalším vědeckým objevům, které nám zkoumání UI usnadní. A ostatně takový postup přímo souvisí s kolektivní inteligencí (Bostrom, 2018, s. 68–69). Ač to na první pohled kvůli naší subjektivní perspektivě není zřejmé, naše společnost nese znaky fungujícího mechanismu, který se snaží zachovat sám sebe, k čemuž je potřeba dosahovat spousty jiných cílů. Jsme inteligenční systém, který je sice biologický, ale lze si představit, že by strojová obecná umělá inteligence fungovala

právě takto, nebo velice podobně. Bylo by možné toho dosáhnout propojením dílčích menších algoritmů, které by tvořily jeden celek.

Harari pokračuje ukázkou několika experimentů zkoumajících jedince s poruchou propojení hemisfér a následně i experimenty týkající se lidí bez této poruchy (Harari, 2017, s. 288–294). Výsledky takových pokusů dokazují lidové rčení „levá ruka neví, co dělá pravá." Lidé postižení touto chorobou pak často verbálně jednají v nesouladu s neverbální částí mozku. Ruka jednoho z účastníků experimentu tak učiní volbu, jíž si ale druhá část mozku zodpovědná za racionální myšlení není vědoma a musí přijít s vlastním vysvětlením. Dále uvádí experiment, který dokazuje, že takové mozkové „rozdělení" je přítomno v jisté míře i ve zdravém mozku. Vychází přitom z díla *Thinking, fast and slow*[4] (2011) od Daniela Kahnemana. Tento laureát Nobelovy ceny dochází ve svém zkoumání k tomu, že naše hemisféry vnímají stejnou zkušenost rozdílným způsobem. Zjistil, že lidské rozhodování je často ovlivněno systematickými chybami a zkresleními, které odporují racionálnímu přístupu k rozhodování, jaký byl například tradičně předpokládán v ekonomii. Popis zmíněného pokusu pak zní takto: „Skupina dobrovolníků se zúčastnila experimentu, který měl tři části. V té první a krátké měli dobrovolníci držet jednu minutu ruku v nepříjemně studené vodě (14 °C) na hranici bolesti. Ve druhé, delší části pokusu, byla po minutě tajně do chladné vody přidána voda teplá, jež zvýšila teplotu o jeden stupeň. Poté mohli účastníci ruce z vody vytáhnout. Někteří prodělali nejprve první, minutový experiment, jiní začali s druhým. Po sedmi minutách přišla na řadu třetí část experimentu, v níž se pokusné osoby měly rozhodnout, kterou z předchozích zkušeností si zopakují. Plných 80 procent se rozhodlo pro tu delší variantu, kterou si mylně pamatovali jako méně nepříjemnou" (Kahneman, 2011 in Harari, 2017, s. 290). Ve světle těchto výsledků tedy rozděluje mozek na dva subjekty. Jeden subjekt zodpovídá za pocity, prožitky a uchovávání vzpomínek, zatímco druhé z těchto *Já* tyto vzpomínky a prožitky za sebe organizuje a interpretuje tak, aby nám poskytl jakousi vnitřní logiku, se kterou lze dál nějak nakládat. Koná tak bez asistence jiných center, a tak si může bez žádné další kontroly vymýšlet, vzpomínky míchat nebo je vyprávět zkratkovitě. Vezme si tak například pouze informace o tom, jak teplá byla voda na začátku a na konci, a proti racionalitě nám vnukne subjektivní iracionální

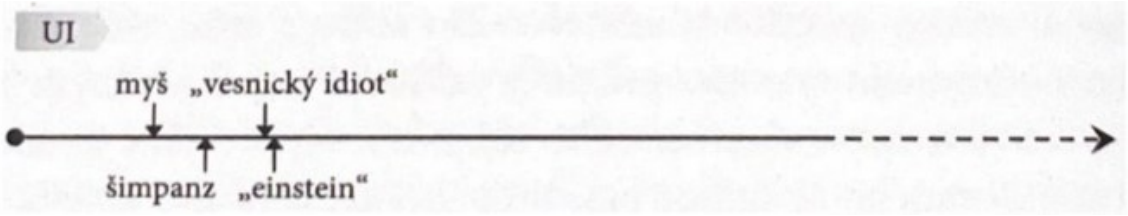
[4] Poznámka autora: V češtině vyšlo v edici Pod povrchem nakladatelství Jan Melvil Publishing v Brně jako Myšlení – rychlé a pomalé v roce 2012, ISBN 978-80-87270-42-4

„pravdu“, že delší pobyt ve vodě byl ten méně nepříjemný. Jedna část mozku vlastně jednala nezávisle na druhé a v rozporu s ní (Harari, 2017, s. 290-291).

Jak jsme si již řekli, Harari použil tento exkurz mimo jiné jako kritiku vůči liberalismu, pro který je stěžejní existence svobodného individuálního *Já*, která se ale ve světle těchto dvou proti sobě jdoucích *Já* zdá najednou méně jistou. Záměrem, který jsem chtěl v případě tohoto exkurzu vědeckého vyvrácení existence svobodné vůle dosáhnout, bylo poukázat na lidskou mysl jako na kompozit několika inteligencí. Stejně jako jednotliví lidé tvoří společnosti kolektivní inteligenci jsou i lidé složeninou dílčích inteligenčních *Já.* Duše je pak jen historickou tradicí dochovaná odpověď na otázky o našem mozku, kam dříve věda „nedosáhla“ (Harari, 2017, s. 278). Jsme-li tedy jako lidé schopni v našich mozcích rozlišit různé inteligenční více či méně na sobě závislé struktury, pak se nezdá nemožné například vytvoření obecné umělé inteligence právě z dílčích současných nebo budoucích umělých inteligencí. Minimálně evoluci se toto již jednou podařilo. Od počátků oboru umělé inteligence namítá laická veřejnost, že si UI nikdy nebude vědoma sama sebe tak, jako to umíme my. Lze předpokládat možnost, že se tedy jednou možná dočkáme obecné superinteligence, která si sama sebe nebude vědomá. Nicméně ani my nemáme v současnosti přesvědčivé důkazy o lidském vědomí.

Tímto problémem vědomí, nejen v kontextu UI, se zabývá i John Searle. K jeho nejznámějšímu argumentu čínského pokoje se dostaneme později. Nyní si zmiňme, jak rozumí lidské mysli. „John Searle je v otázkách mysli a vědomí propagátorem emergentismu. Emergentní teorie vědomí tvrdí, že vědomí je vlastností mozku, která vyvstává z jeho složitosti a tato složitost, která vědomí formuje, není zpětně vysvětlitelná z jednotlivých mozkových částí. Mentální sféra je tak makrovlastností mozku, která vyvstává ze vzájemné interakce jeho mikrostruktur, ale přitom mentální sféra není z těchto mikrostruktur odvoditelná“ (Havlík, 2012, s. 190). On sám pak vědomí doslovně definuje takto: „Vědomím jednoduše míním subjektivní stavy vědomí nebo cítění, které začínají, když se člověk ráno probudí. Vědomý stav pokračuje do doby, co je člověk vzhůru, a přetrvává po tuto dobu, než člověk upadne do bezesného spánku, kómatu, zemře nebo se jakkoli jinak stane nevědomým“ (Searle, 1990 in Havlík, 2012, s 190). Z tohoto úhlu pohledu tedy nelze uznat, jak si níže dokážeme, že by dnešní umělé inteligence měly vědomí. Nedosahují totiž celistvé komplexity a jenom jedna složka kognice zcela určitě

není dostatek pro zapojení vědomí. Patrně bychom nemohli o vědomí hovořit ani pokud by takových inteligencí bylo pospojovaných vícero.



Zdroj: Bostrom, 2018, s. 118

Další předpoklad, který je potřeba zmínit, je pravděpodobnost, že nejsme jako lidé vrcholem inteligence. Sice jsme nejinteligentnějším druhem na planetě Zemi, což nám dopomohlo k dominanci nad ostatními druhy, ale co se typů inteligentních tvorů týče, vlastně jsme se nesetkali s druhem, který by se způsobem myšlení a jednání fundamentálně odlišoval od lidského myšlení. Jedním ze způsobů, jak hodnotit míru inteligence je všeobecně známý inteligenční kvocient neboli IQ. Yudkowsky nás ale varuje, že IQ je pro nás v případě obecného řešení inteligence vlastně naprosto irelevantní metrika, protože o něčem vypovídá pouze při aplikaci na kontext člověka. Ano, inteligence je zdrojem naší „síly“ a dovolila nám stanout lidskou nohou na Měsíci. My jako lidé ale máme představu výhradně o lidském myšlení a jsme schopni odhadnout, jaké myšlenkové operace budou či nebudou ostatním lidem dělat problémy v závislosti na jejich IQ. Vzdáleně pak ještě máme představu o inteligenci šimpanzů a několika dalších druhů nebo je alespoň z této perspektivy zkoumáme. Stále ale jde o srovnání vycházející z naší inteligenční pozice. Yudkowsky nás tak upozorňuje, abychom se na inteligenci nedívali příliš antropocentricky, protože naše nazírání toho, co považujeme za „kvalitní“ inteligenci, je zkresleno způsobem, jakým rozumíme světu, zatímco máme vlastní zkušenost pouze s jinými lidmi a jejich kognitivní kapacitou (Yudkowsky, 2008, s. 7–9). Nestojíme na vrcholu inteligence, patrně ani nikde v dosahu tohoto vrcholu, a pro umělého agenta, který má tu možnost vyvíjet se mnohonásobě rychleji oproti tomu, co dokázala evoluce, představujeme opravdu jen drobnou škálu.

Dá se očekávat, že naše snažení budou úspěšná a UI se začne posouvat dál a dál na inteligenčním žebříčku. Jak můžeme sledovat z obrázku, až překročí inteligenční úroveň šimpanze, hrozí, že ji podceníme a budeme stále považovat za příliš hloupou na to, aby nás mohla ohrozit. Nicméně už by se jednalo o dostatečně komplexního agenta na to, aby

se byl schopen rozvíjet výrazně rychleji, než kdyby se intelektem rovnal myši. Možná bychom ji podcenili, protože se zdánlivě bude vyjadřovat pro nás nesrozumitelným způsobem nebo dokonce její způsob vyjadřování nebude odpovídat ničemu z toho, co známe. Tak či onak, my lidé představujeme inteligenčně jako druh jen malé rozmezí, které může budoucí UI překlenout za zlomek doby, kterou jí trvalo se na naši úroveň vůbec dostat (Yudkowsky in Bostrom, 2018, s.118).

Naše rozhodnutí nejsou řízená námi, společnost je stejně jako náš mozek, jen kompilát dílčích inteligencí, a přesto tu hovoříme o inteligenci umělé, která je spíše strojovým učením, protože sama inteligentní není. Jak to tedy je a co si z toho všeho zatím vzít? Jako druh nepředstavujeme velké rozpětí na škále inteligence a v rámci postupu UI po tomto žebříčku nejsme závratně velikou škálou. Naše inteligence nás sice dostala do pozice nejsilnějšího druhu, což nesvědčí o tom, že jsme dosáhli inteligenčního maxima. Tohoto vrcholu jsme dosáhli díky komplexitě našeho mozku, který je obdařen schopností najít řešení více dílčích problémů, která nás posouvají dál a dál. Obor umělé inteligence se tedy snaží vytvořit myslícího agenta, který bude alespoň stejně komplexní jako lidský mozek. Neurologie má stále mnoho nezodpovězených otázek, ale principiálně můžeme věřit, že sami duchem nejsme nijak svobodní a zároveň jsme různými dílčími mozkovými centry kompilovaní *My*. Takové dílčí umělé inteligence momentálně máme. V následující kapitole se podíváme na dosavadní pokrok oboru a příležitostně si tak vysvětlíme, proč dnes skutečnou umělou inteligencí zatím nedisponujeme.

3.1. Historie a současný stav oboru umělé inteligence

Pro nezasvěcené se může hranice mezi lidskou a strojovou kognicí už dnes zdát rozmazanou. Jak nám ostatně předkládají média na honu za senzací, že již dnes vznikají myslící agenti, kteří představují hrozbu pro naši společnost. Současné UI v porovnání s jejími o dvacet let staršími předchůdci sice možná působí velmi pokročile, ale jak píše sám Bostrom (2014, s. 21), vznik strojové inteligence, která by v myšlení byla rovna člověku, se od vynalezení počítače ve 40. letech 20. století nijak výrazně nepřiblížil.

Rok 1957 můžeme označit jako jeden z přelomových, neboť se tehdy spojilo deset vědců zabývajících se neuronovými sítěmi, inteligencí a teorií automatů, aby společně zorganizovali dvouměsíční spolupráci, kterou dnes známe pod názvem Dartmouthská

konference. Základní představení cílů tohoto sjezdu Rockefellerově nadaci, která akci sponzorovala, znělo: *„Pokusíme se zjistit, jak vytvořit stroje, které budou užívat jazyk, tvořit abstrakce a pojmy, řešit ty druhy úloh, jež jsou nyní vyhrazeny lidem, a které se budou zdokonalovat. Domníváme se, že když na těchto problémech bude během jednoho léta pracovat pečlivě vybraná skupina vědců, bude možné v jednom nebo ve více z nich dosáhnout významného pokroku"* (Bostrom, 2018, s. 23).

Toto prohlášení následně označuje Bostrom za ukvapené, neboť většina následujícího vývoje v dalších dvaceti letech sestávala z drobných pokroků, které ale brzy narazily na tehdejší technologické limity. Veškeré posuny tak byly kvalitativní, avšak omezené tehdejší výpočetní kapacitou. Většinou proto, že většina tehdejších algoritmů fungovala na principu „řešení hrubou silou". Jednoduše algoritmus vybíral ze všech možných řešení to nejlepší. Čím složitější úkol, tím více možných výsledků, ze kterých bylo potřeba vybírat. Náročnost výpočtů tak narůstala geometrickou řadou, což vedlo k tomu, že program narážel na již zmíněné omezení výkonem. I tak se však objevovaly programy, které zvládaly komponování hudby ve stylu určitých skladatelů, simulace pohybů ruky v digitálním prostředí, nebo například napodobovat terapeuta (Bostrom, 2018, s. 24-25). Dartmouthská konference je tak stále důležitá právě proto, že položila základy pro další výzkum a vývoj UI, který pokračuje dodnes. I když sama o sobě nepřinesla závratné výsledky, vzbudila v oblasti UI veřejný zájem, který vzrůstá.

V 80. a 90. letech 20. století pak nastává boom v podobě využívání neuronových sítí a genetických algoritmů. Ty jsou na výsost důležité pro svou schopnost se samy vyvíjet, optimalizovat nezávisle svá řešení i neintuitivní cestou a učit se (Bostrom, 2018, s. 26-29). Oba zmíněné způsoby jsou dnes stále hojně využívané a významné.

Současné UI sice pracují na bázi stále komplexnějších algoritmů, neuronových sítí a vyhodnocovacích pravidel, avšak jejich využití vždy směřuje konkrétním směrem, tedy na jednu jim pevně danou problematiku. Takovému druhu UI, který je dobrý jen k jedné věci, se říká slabá umělá inteligence, a bez výjimek připadá toto označení všem současným UI. Vlastně až teď pomalu docházíme k cílům, které si slibovala Dartmouthská konference. Díky větší kapacitě dat tak máme systémy opravdu schopné nahradit řidiče automobilů, nejlepšího „robotického" šachistu nebo Chatbot, který vám dokáže ve smysluplné řeči odpovědět na řadu otázek. Nyní si blíže rozebereme pro tuto práci několik zásadních mezníků a termínů tohoto oboru.

3.1.1. Deep Blue

Zmiňuji-li v práci, že UI se od dob svých počátků do dnešních dnů v rámci praktičnosti jejich užití nijak principiálně nezměnily, pak beru Deep Blue jako jeden z užitečnějších příkladů k ukázce. Rok 1997 se asi sice nedá označit přímo za počátek pro UI, ale v případě Deep Blue jde o první z vrcholů a vítězství technologie nad člověkem. Podotýkám, že stáří této události – souboje stroje s člověkem – není nijak proti účelu tohoto exkurzu, neboť k žádné fundamentální proměně principiálních funkcí těchto programů nedošlo.

Jak vyplynulo ze záměrů Dartmouthské konference, bylo cílem vytvořit stroje, které budou dobré jako člověk v disciplínách, které jsou člověku vlastní. Takovou disciplínou na výsost vlastní člověku byl šach (Bostrom, 2018, s. 35). Zejména proto, že vyžadoval strategické přístupy, plánování, počítání tahů a všeobecně umění hraní šachů svědčilo o intelektuálních schopnostech. Porážka člověka umělou inteligencí v šachu by tak pro mnohé znamenala schopnost myslet lépe, což by vedlo k vyhledávání dalších způsobů využití. Na druhou stranu už I. J. Good ve své době předpokládal, že naučit stroj šachy není nemožné. Vycházel z analogie, že pro amatéry některé tahy vyžadují dobrou imaginaci, ale pro šachového mistra je taková myšlenková operace rutinní. Tím spíš by byla rutinní pro stroj (Good, 1965, s. 34).

Své momenty slávy si prožila Deep Blue v letech 1996 a 1997, kdy měřila síly s tehdejším mistrem světa v šachu Garrim Kasparovem. Toto soupeření pak pro člověka skončilo porážkou. Nesmíme ale podlehnout domnění, že šlo o klání dvou vědomí či myslí. Deep Blue byl strojem, který operoval na základě algoritmického kombinatorického prohledávání možných stavů šachové partie, v závislosti na současném stavu hry, a na sadě expertních pravidel, která mu poskytují informace o tom, které tahy nevedou k jeho cíli vyhrát, a tudíž je může z vyhodnocování užitečnosti dalších tahů vypustit. Mezitím co Kasparov na základě vlastního strategického myšlení a letitých zkušeností a intuice zvažoval sám za sebe, jak lze vyhrát, Deep Blue z odpovídajících řešení vlastně jen vybral to s největší pravděpodobností vyhrát. Stroj tak vyhrál díky schopnosti kombinatoriky možných šachových pozic a znalostí dat, které sestávaly z partií mnoha jiných šachových mistrů, Kasparovovy předešlé hry nevyjímaje. Tyto nedostatky „výpočetního“ výkonu ale lidský mozek nahrazuje právě zkušenostmi, protože na tolik kombinací jednoduše nemá kapacitu. Nedá se tedy zcela přesvědčivě

tvrdit, že by spolu svedla boj dvě vědomí, ale spíše člověk a nástroj (Voců, 1997). Nicméně můžeme být stále klidní, co se naší příčky na žebříčku inteligence týče, protože algoritmus Deep Blue byl takzvaně jednoúčelový. Nebyl dobrý k ničemu jinému než k hraní šachů (Bostrom, *s. 35*).

3.1.2. Turingův test

Vracím se historicky v čase, ale i tento exkurz má svou platnost. Výše jsem zmiňoval Yudkowského, který varoval před antropocentrickým pohledem na UI. Turingův test je přesně tím, co by se tohoto varování týkalo. U Turingova testu je důležité zmínit, kdy vznikala jeho hypotéza ohledně umělé inteligence. Jak víme, v 30. a 40. letech 20. století, tedy v období vzniku prvních sálových počítačů, panoval optimismus vyznačující se očekáváním příchodu opravdové umělé inteligence v dohledných dvaceti letech. Z nám známých důvodů se tak nestalo. Nicméně na Turingově zastaralém testu lze demonstrovat, že dnes by jím patrně prošlo hned několik UI, což nevypovídá nic o prokázání schopnosti strojového myšlení, na kterou Alan Turing tímto testem mířil. Ostatně to nevypovídá nic ani proti neschopnosti myslet.

Aby bylo možné takové přemýšlení alespoň z našeho úhlu pohledu prokázat, vytvořil a publikoval v článku *Computing machinery and Intelligence* roku 1950 Turing svou hypotézu, která je dnes běžně známá jako „Turingův test“. Turingova hypotéza jednoduše spočívala v tom, že kvalita, s jakou dokáže komunikovat UI s člověkem bez toho, že by onen člověk dokázal odhalit, že komunikuje se strojem, vypovídá o schopnosti konkrétního stroje myslet. Zkráceně měl Turing takový názor, že pokud dokáže UI odpovědět na otázky tazatele, pak jednoduše ovládá umění jazyka a komunikace a lze tak tvrdit, že umí myslet.

Plnění Turingova testu je, jak si následně ukážeme, pouze o manipulaci se znaky a znalosti syntaxe. Dokonce to nemá ani co dělat se znalostí konkrétního jazyka. Jde jen o naprogramované užívání správných frází, které vyvolávají dojem, že k vám někdo promlouvá. Ostatně sám Turing svůj test nazval imitační hrou. Spíše, než by ale stroje imitovaly myšlení, imitují pouze jazykovou promluvu (Turing, 1950). Anekdoticky můžeme v rámci kritiky Turingova testu konstatovat, že stroje, které s námi vedou konverzaci, se nám snaží namluvit, že umí mluvit, než aby s námi skutečně mluvily.

Na druhou stranu je ale asi třeba dodat, že Turing byl limitován technologií své doby a test probíhal prostřednictvím textového média. Pokud bychom v budoucnu chtěli udělat „skutečný“ Turingův test, pravděpodobně by se odehrál tváří v tvář. V takovém případě by pak musel stroj mít alespoň vnější podobu člověka. Takový stroj by pak ale musel umět pohybovat přirozeně vlastním tělem, orientovat se v prostoru, rozeznávat předměty a manipulovat s nimi. To už se ovšem dotýkáme problematiky pokroku spousty dalších technologických oblastí a samozřejmě obecné umělé inteligence, která by byla na takovýto test potřeba. Přesuňme se ke konceptu Johna Searla, který reaguje právě na Turingův test.

3.1.3. Searlův čínský pokoj

V kapitole osvětlující současný stav UI zmiňuji, že v podstatě stagnují, a ke schopnosti přemýšlet mají ještě daleko. Jak ale takovou věc teoreticky podepřít? To, co Turing na základě svého testu předpokládal, bylo jednoduše řečeno, že osvojení jazyka UI automaticky vypovídá o jejím myšlení. Ovládání jazyka je komplexní dovednost, vlastní pouze člověku, a jako taková je založená na manipulaci se znaky, kterým je pro komunikaci nutno porozumět, nebo ne? Ostatně takový předpoklad Turingova testu by platil i v případě znalosti šachů u již uvedeného Deep Blue. Ten by tak v přeneseném významu splnil Turingův test stejně, jako John Searle ve svém myšlenkovém experimentu s čínským pokojem, kterým reaguje na Turingův test.

V případě čínského pokoje zdánlivě opakujeme princip Turingova testu. Jedna osoba, zavřená a izolovaná v místnosti, má za úkol přesvědčit svého tazatele nacházejícího se mimo místnost, že umí čínsky. Tazatel tak pokládá na papírovém lístku škvírou ve dveřích otázky v čínštině. Člověk v uzavřené místnosti disponuje sborníky jazykových pravidel čínštiny a v podstatě vším co kdy v tomto jazyce bylo napsáno. Podle pravidel se tedy řídí, a tak ví, že na znaky XY má odpovědět znaky AB, jak je zapsáno v pravidlech. Pro dotazovaného stojícího mimo místnost tak celý proces dosahuje stejných výsledků, jako splnění Turingova testu. A to bez toho, že by osoba v místnosti musela umět čínsky. Validita Turingova testu se tím tak bortí, protože schopnost adekvátně odpovědět ještě nenaznačuje cokoliv o tom, že by počítač v takovém jazyce uvažoval a že by uvažoval vůbec (Searle, 1980, s. 209-214).

Searlův čínský pokoj vlastně ukazuje, že odpovědět na položenou otázku je v zásadě sémiotický problém. UI nemá ponětí o tom, co znamená jí položená otázka. Potřebujeme znát syntaktiku systému, který se chystáme použít, a jeho „slovník" znaků, který takový systém užívá, ať už jde o šach nebo překládání z jednoho jazyka do druhého. UI postupuje mechanicky. V případě moderních Chatbotů, které následují stejný princip, je pak ještě potřeba vzorek, na kterém se algoritmus naučí slova správně používat, aby byl co nejpřesvědčivější a komunikační agent je rázem na světě, ač sám neví, co to komunikace je a že (ne)komunikuje. Jedním z takovýchto komunikačních UI je i ChatGPT[5], která v současnosti budí veřejné nadšení i obavy pro své, v současnosti asi nejlepší, jazykové schopnosti. Ty jsou však založené na jednoúčelovém algoritmu, jehož cílem je dodat uživateli uspokojivou odpověď.

Podobně funguje i práce se současnými internetovými překladači. Stačí „vzdělat" software podle bilingvních jazykových korpusů a pozorovaného jazykového úzu mezi sebou srovnávat jednotlivá slova a z toho na základě statistických modelů, úzu a znalosti syntaxe vyvozovat překlad (Bostrom, 2018, s. 36-37). Ať už inteligentní agent používá jeden nebo dva jazyky současně, nezdá se, že by bylo potřeba mu vytvořit i vědomí, které by těmto jazykům rozumělo, aby je dokázal aplikovat.

3.1.4. Současná debata o UI

K ChatGPT se ještě na moment vrátíme. Jak jsem však zmínil, její dovednosti vzbuzují pozitivní i negativní reakce. Považuji za přínosné pro kontext této práce zmínit i toto současné dění ohledně nastolování veřejné debaty o problematice vývoje a použití UI. Dne 22. března 2023 vydal Future of Life Institute výzvu[6], kterou mimochodem podepsali i v této práci zmiňovaný Harari, miliardář Elon Musk nebo třeba bývalý spoluzakladatel společnosti Apple Steve, Wozniak. Tento otevřený dopis požaduje okamžité zastavení vývoje modelů umělých inteligencí silnějších než dosavadní model ChatGPT4 na minimálně šest měsíců. Primární požadavek je vytvořit obecně dodržované bezpečnostní postupy a transparentnost vývoje umělých inteligencí. Zmíněný ChatGPT je totiž jedním z důvodů vzniku této výzvy. Dopis vyjadřuje obavy ohledně strmě

[5] Dostupné z: https://openai.com/chatgpt
[6] Dostupné z: https://futureoflife.org/open-letter/pause-giant-ai-experiments/

stoupajících schopností tohoto chatbota. Ty jsou mimo jiné také způsobené publikovaným článkem společnosti OpenAI, který mimo jiné uvádí, jak velký objem práce v mnoha oborech je možné pomocí modelů ChatGPT zastat (Eloundou, 2023, s. 15-18).

Ostatně i tohle je jedna z problematik, kterou rozebírá Bostrom v kapitole věnované inteligenční explozi (2018, s. 109). Pro úplnost si řekněme, že tyto prognózy ohledně společenské reakce na UI se týkají zejména Obecné inteligence nebo Superinteligence, která proti člověku v žádné kognitivní funkci nezaostává. ChatGPT pracuje na principu matematických modelů a datového vzorku, na kterém byl vytrénován, a podle toho pak generuje odpověď na jemu položenou otázku. Tato UI sice nerozumí ani jednomu ze slov v otázce „Co je to pes?“, ale vzhledem k výskytu mnoha odpovědí na tuto otázku v jí dostupném datovém vzorku vyhodnotí jako vhodná slova pro odpověď následující: čtyřnohé, je, pes, zvíře. Na základě syntaxe už pak vybraná slova pouze složí do věty ve správném pořadí. Schopnost ChatGPT je pak pouze bez znalosti významů a navýsost sémiotická (Gvoždiak, s. 63-67). Veliká část profesí dnes při svém výkonu manipuluje s informacemi obsaženými v textu. Nelze se potom divit, že by tento nástroj ubral na práci například překladatelů až z 76,5 % (Eloundou, 2023, s. 16).

K tomuto otevřenému dopisu se vyjádřil i Eliezer Yudkowsky. Krom vyslovení respektu všem, kteří se rozhodli tuto poetici podepsat, však konstatuje, že ji nepodepsal, protože je toto šestiměsíční moratorium podle něj příliš málo. Yudkowsky zastává názor, že řešení problémů ohledně vývoje UI neleží až v období, kdy bude umělý intelekt rovný lidem, ale právě dokud se tak ještě nestalo. Zmiňuje, že obecná umělá inteligence pravděpodobně nezůstane rovna s člověkem na moc dlouho, pokud vůbec. Poté už jí budeme jenom překážet a pravděpodobně nás bude považovat za „podřadné“. Rovněž kritizuje společnost OpenAI za to, že plán vypracování problému sladění hodnot umělé inteligence s lidskými hodlá uložit jako úkol jednomu z jejich budoucích modelů. Považuje to za absurdní a z principu nefunkční řešení. Co hůř, možná dosahujeme pokroků ve vývoji UI, ale už ne ve vymyšlení principu, který by zařídil, že s námi bude sdílet naše lidské hodnoty a bude je respektovat. V této souvislosti konstatuje, že na vyvíjení umělé inteligence nejsme připravení, globálně by se měla přinejmenším tvrdě regulovat výpočetní kapacita datacenter, ale v lepším případě úplně vývoj zastavit (Yudkowsky, 2023).

Je ovšem ve hvězdách, zda se státy a korporace budou na takové dohodě podílet. Ostatně je diskutabilní, jak upřímně myslel Elon Musk svůj podpis této petice. Sám vlastní firmy, které vyvíjejí své modely umělých inteligencí a v minulosti se sám snažil koupit firmu OpenAI, která ChatGPT vyvíjí, ale byl odmítnut. Lze tedy bohužel předpokládat i postranní úmysly v podobě snahy omezit postup konkurence, aby mohly jeho firmy dohnat tento náskok. To jen dohromady více opodstatňuje, proč by měl být vývoj regulován globálně, nikoliv zanechán čistě v rukou soukromých subjektů. Zasáhnout musí státy, protože pochybuji, že by firmy šly proti vlastnímu ekonomickému zisku, který tyto UI generují, nebo alespoň mají potenciál tak činit.

4. Co je to Superinteligence?

Z popisu současného pokroku lze usuzovat, že nám pravděpodobně k dosažení Superinteligence několik kroků chybí. To pochopitelně vyvolává řadu otázek, jež spolu úzce souvisejí. Jak je tato cesta dlouhá a náročná a jaký konkrétní význam má pro lidstvo umělá Superinteligence? Ač mají současné UI malé specifické zaměření, můžeme pozorovat, že i s takhle úzkým profilem jejich využití stoupají v současnosti obavy z nebezpečí, které by mohly představovat, ale na druhé straně se očekávají výrazné benefity. Jedním z dalších logických kroků ve vývoji UI, pokud momentálně opomeneme otázku bezpečnosti, je stávající dosažené schopnosti propojovat a zobecňovat. Je třeba vytvořit agenta, pro kterého „… schopnost učit se by byla nedílnou součástí samotného jádra systému určeného k dosažení obecné inteligence, nikoliv něčím, co by k němu bylo dodatečně připojeno jako rozšíření nebo doplněk. Totéž platí pro schopnost efektivně pracovat s neurčitostí a pravděpodobnostními informacemi. Dále mezi základní vlastnosti moderní UI, jež by mohla dosáhnout obecné inteligence, pravděpodobně patří schopnost získávat užitečné pojmy ze smyslových dat a svých vlastních vnitřních stavů a vkládat takto získané pojmy do flexibilních kombinatorických reprezentací, aby mohly být využity v logickém a intuitivním usuzování“ (Bostrom 2018, s. 48).

Uskutečněním výše popsaného „receptu“ bychom velice rychle došli k systému, kterému můžeme říkat zárodečná UI. Ta, jak název napovídá, je základním stavebním kamenem obecné umělé inteligence. Jak si níže ukážeme, jelikož bude pravděpodobně obecná UI přemýšlet mnohonásobně rychleji než člověk, bude mít sama o sobě dostatek

času zanalyzovat, jaké jsou její programové nedostatky, odstranit je a vylepšit samu sebe. Právě proto je tolik důležité, aby obsahovala schopnost rozpoznávání vnitřních stavů. S každou úpravou by tak exponenciálně rostla její inteligence, až by nás mílovými kroky předčila a nechala daleko za sebou. Přesně z tohoto usuzování, které jsme si nastínili v předešlých kapitolách, tedy vychází Yudkowsky, když nám předkládá, že nás umělá inteligence skokově překoná, a proto je nutné se mít na pozoru již dnes.

Tomuto technologickému zvratu, kdy se umělá inteligence sama začne vylepšovat, obecně říkáme technologická singularita. Irving John Good popisuje tento postup dále: „Nechť je ultra-inteligentní stroj definován jako takový stroj, který ve všech intelektuálních činnostech dokáže dosáhnout mnohem vyšší úrovně než libovolný člověk, ať jakkoliv chytrý. Protože jednou z lidských intelektuálních aktivit je i navrhování strojů, dokázal by ultrainteligentní stroj navrhovat ještě lepší stroje; bez pochyby by tedy došlo k „inteligenční explozi“ a lidská inteligence by zůstala daleko pozadu. První ultrainteligentní stroj je tudíž tím posledním vynálezem, který kdy člověk bude muset vytvořit – pod podmínkou, že tento stroj bude dostatečně povolný, aby nám sdělil, jak jej udržet pod kontrolou“[7] (Good, 1965 in Bostrom, 2018, s. 22). Došli jsme tak až k pojmu Superinteligence.

Překonání člověka ve všech ohledech je jistě impozantní vlastnost, avšak poměrně strohá ve svém popisu. Aby se Bostrom vyhnul příliš vágní definici, že je Superinteligence ve všem jednoduše lepší než my, vymezuje následující klíčové intelektuální schopnosti, které by patrně všechny byly Superinteligenci vlastní. Přepokládá rovněž, že stačí vytvořit Superinteligenci pouze s jednou z těchto schopností. Pokud je schopnost dostatečná, dovedla by si postupně Superinteligence vytvořit a osvojit všechny zmíněné schopnosti (Bostrom, 2018, s. 99).

Tou první z nich je rychlost myšlení. Takový agent nemusí být přímo rychlejší než člověk. Stačí ale, aby jeho vnitřní procesy myšlení probíhaly mnohonásobně rychleji než v mozku založeném na biologickém substrátu. Bostrom chápe jako jednu z klíčových vlastností rychlost digitálního přenosu dat, která je už na dnešních mikroprocesorech

[7] Let an ultraintelligent machine be defined as a. machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an "intelligence explosion," and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control. IRVING, John Good, „Speculations Concerning the First Ultraintelligent Machine“, v: Franz L. Alt a Morris Rubinoff (eds.), *Advances in* Computers, sv. 6, Academic Press, New York 1965, str. 33.

o celých sedm řádů rychlejší, než co zvládá lidský mozek (Bostrom, 2018, s. 101). Pokud by taková mysl byla například postavená na počítači umožňujícím optický přenos dat, výrazně by se limit jejího přemýšlení přibližoval rychlosti světla. Taková mysl by pak podle Bostroma možná nerada komunikovala s lidmi, protože by v jejím vnímání v rámci jejího subjektivního světa docházelo k časové dilataci. Interakce s námi by tedy byla nesnesitelně pomalá (Bostrom, 2018, s. 92-93). Tento popis se velice úzce váže na to, jak hovoří Bostrom o emulaci mozku, o které budeme hovořit níže.

Druhou zmíněnou schopností je kolektivní inteligence. Tu později také Bostrom aplikuje na scénář rozvoje společnosti postaveného na zefektivnění její vnitřní struktury. Jinými slovy se tím myslí, že lze dosahovat lepších výsledků, pokud je přenos informací v rámci dané struktury efektivní. Což se může dotýkat jak lidstva, tak struktury menších umělých inteligencí (Bostrom, 2018, s. 95).

Následuje pojem kvalitativní inteligence. Tak jako je člověk na výši oproti například opicím, které postrádají schopnost komplexního jazyka, předpokládá Bostrom, že člověk rovněž alespoň zatím nedošel k jiným schopnostem. Vychází z toho, že například i některé lidské mozky svou strukturou neumožňují vnímat hudbu a jedinec se tak rodí s vrozenou amúzií. Kvalitativní superinteligence by si tak byla díky své unikátní vnitřní struktuře schopna uvědomovat jiné souvislosti než člověk, který své nedostatky výpočetní kapacity mozku činí intuitivní volby (Bostrom, 2018, s. 97-99).

Můžeme si rovněž zmínit výhody, které pro UI představuje fakt, že bude digitální a které se defacto slučují s tím, jak fungují a jak lze dnes vylepšovat naše domácí počítače. Rychlost přenosu dat již byla zmíněna u vysvětlení superrychlé UI. Dále pak Bostrom zmiňuje zvyšování operační a výpočetní paměti, nepoměrně větší uložiště než to v lidském mozku, schopnost celé informace sdílet, přesouvat, kopírovat, upravovat, a to včetně UI samotné, a mimo jiné skutečnost, že tato technologie nedegraduje s postupujícím věkem tak jako člověk (Bostrom, 2018, s. 101-105). Zcela jistě pak každého z nás napadá, co objev Superinteligence s takovouto paletou schopností přináší. Příchodu Superinteligence, jeho rizikům a podobám a následkům se budeme věnovat níže.

4.1. Inteligenční exploze

Inteligenční exploze, jak už víme, je hypotetická budoucí situace, ve které dochází k rychlému a významnému zvýšení a zdokonalení intelektuálních schopností umělé inteligence. Následkem toho by se její intelektuální kapacita mohla nekontrolovaně zvyšovat a vést k dramatickému přelomu ve vývoji technologií a společnosti. Předmětem úvah o inteligenční explozi tak jsou pochopitelně otázky, kdy ji můžeme očekávat a jak se takový „výbuch" bude projevovat.

Nelze sice přesně předpovědět, kdy se k UI dostaneme, avšak podle současných objevů se dá předpokládat, že do čím vzdálenější budoucnosti hledíme, tím se její uskutečnění zdá pravděpodobnějším. (Bostrom, 2018, s. 56-57) Zároveň je pravděpodobné, jak jsme si výše naznačili, že příchod Superinteligence bude následovat rychle po příchodu obecné umělé inteligence. Ovšem o jakých časových horizontech se bavíme?

Tabulka 2 ***Kdy bude s udanou pravděpodobností dosaženo strojové inteligence lidské úrovně?***[83]

	10 %	50 %	90 %
FT-UI	2023	2048	2080
OUI	2022	2040	2065
EETN	2020	2050	2093
TOP100	2024	2050	2070
Medián	2022	2040	2075

Zdroj: Bostrom, 2018, s. 44)

V uvedené tabulce prezentuje Bostrom data sesbíraná z dotazníků vyplněných účastníky těchto konferencí: Philosophy and Theory of AI z roku 2011 v Soluni, konference Artificial general intelligence a Impacts and Risks of Artificial General Intelligence konané v Oxfordu roku 2012 a průzkumy EETN a TOP100 z let 2013. Respondenti měli odpovědět, s jakou pravděpodobností očekávají příchod SILU – zkráceně silná inteligence lidské úrovně – za předpokladu, že lidská vědecká činnost bude pokračovat bez jejího významného narušení (Bostrom, 2018, s. 43-44). Vzhledem k nepočetné návratnosti dotazníků, kterou Bostrom zmiňuje, sice nemůžeme považovat tento vzorek zas reprezentativní, na druhou stranu si nějakou hrubou představu o názoru odborníků lze udělat. Bostrom dokonce považuje 10% šanci, že nebude do roku 2075 vyvinuta SILU, za moc vysokou. (2018, s. 44)

Nicméně, jak by mohl vypadat takový proces navyšování stupně inteligence do míry, kdy bude UI schopna rozpoznat vlastní nedostatky, ty odstranit a sama sebe vylepšovat?

Pro kinetiku zvyšující se umělé inteligence zavádí Bostrom ještě pojmy zlomový bod a systémový odpor. Prvním ze zmíněných je jednoduše překročení hranice, kdy vyvíjená přestane pro vývoj vyžadovat vnější zásahy, a dosáhne dostatečné vnitřní integrity na to, aby svou inteligenci nadále zvyšovala sama (Bostrom, 2018, s. 108). Systémový odpor zase popisuje, jak se přibližováním k vytvoření Superinteligence zvyšuje náročnost a nákladnost dosahování tohoto procesu. Tedy limity omezující vývoj na základě způsobu, jakým hodláme obecné UI dosahovat (Bostrom, 2018, s. 118).

Může se tak stát, že v závislosti na velikosti systémového odporu zvoleného vývojového přístupu tu bude Superinteligence v řádu hodin a my nestihneme ani zareagovat. Ale také je možné, že problém stihneme analyzovat, ale už ne vymyslet řešení, nebo bude UI přicházet ještě století nebo několik desetiletí, a my se možná dokážeme pečlivě připravit. Nicméně problém spočívá, jak bylo výše naznačeno, v tom, že my vlastně nevíme, kdy k tomu dojde, a právě proto, je tato debata tak důležitá.

Jako první z nich rozebírá Bostrom pomalý nárůst, který se zdá být pro všechny z nás tou nejlepší možností. Bostrom ho očekává v řádu roků, ne-li spíše dekád. Společensky bychom tak měli příležitost vyvinout technologické přístupy a sociální mechanismy pro co nejlepší přijetí v souladu s chodem naší společnosti. Státy a mezinárodní společenství by měly dostatek času na vzájemné dohody a spolupráce, aby i v rámci bezpečnosti byl příchod superinteligence bez negativních následků. Bohužel však vzhledem k potenciálním cestám, jakými se k Superinteligenci lze dobrat, nepovažuje Bostrom tuto pomalou detonaci inteligenční exploze za pravděpodobnou (2018, s. 108).

Druhým scénářem je pomyslná exploze v řádu měsíců či pár let. V tu chvíli bychom pro naši společenskou bezpečnost mohli pracovat pouze s tím, co už máme. Pravděpodobně by se začaly umělé inteligence více a více vyskytovat v rámci pracovních i ekonomických trhů, což by znamenalo plošné společenské, geopolitické a ekonomické turbulence. Státy by tak například čelily nátlakům ohledně zvýšení podpory v nezaměstnanosti, zajištění životního minima nebo uvalení daní a kvót na zaměstnavatele upřednostňujícího digitální zaměstnance před živými. Společenský

systém by se tak musel s nastávající implementací UI a SI vypořádávat v rámci jeho aktuálních přístupů. Je diskutabilní, jak by se s jejím příchodem zvládl vyrovnat (Bostrom, 2018, s. 109). S ohledem na to, že nevíme, kdy dojde k vytvoření Superinteligence, můžeme ale ve světle výše zmíněné petice o zastavení vývoje UI konstatovat, že se společnost pravděpodobně snaží jednat aspoň v rámci těchto rizik, které začíná částečně pociťovat.

Poslední taková exploze by byla významně rychlá. Hovoříme o průběhu minut, hodin nebo maximálně dní. Bostrom v podstatě říká, že bychom si pravděpodobně ničeho nevšimli, dokud by nebylo příliš pozdě (tamtéž). Jakákoliv reakce by byla bezpředmětná, neboť by byly platné pouze dosavadní přípravy.

Na místě je také přemýšlet o tom, jestli bude Superinteligencí nakonec víc, nebo se v souladu s Goodovou předpovědí tento vynález skutečně povede zhotovit jen jednou. Nutno podotknout, že tato očekávání se úzce odvíjí od systémového odporu a překročení zlomového bodu, kdy se dostane zárodečná UI do stavu Superinteligence, a kolik takových různých projektů snažících se o dosažení Superinteligence bude souběžně probíhat, a jak půjde takový pokrok vzájemně od sebe opisovat či nikoliv. Řada historických exkurzů tak poukazuje na to, jak získávaly národy (ne)závisle na sobě některé technologie, a že jejich přenos se zrychluje. Nové způsoby vytváření zbraní se rozšiřovaly daleko pomaleji ve středověku než za studené války, kdy byla pro technologický vývoj jedné mocnosti důležitá špionáž jiné. Bostrom říká, že i kdyby souběžně za sebou vznikaly dva projekty s rozdílem vývoje asi šest měsíců, měla by první dosažená Superinteligence dostatek času na to, nastolit takové podmínky, aby další nevznikla. Ať už by šlo užití mechanismů čistě demokratických, totalitních, nebo něčeho pro nás zcela neočekávatelného, co bychom zprvu ani nezaregistrovali. (Bostrom, 2018, s. 129-135) Sám za sebe bych si dokázal například představit, že by Superinteligence vytvořila třeba nějakou formu globální ekonomické krize, která by pro nás nebyla likvidační, nicméně by mohla znemožnit vývoj konkurenční Superinteligence a upevnit svou vlastní pozici.

4.1.1. Cesty nezahrnující strojovou inteligenci

Všeobecně všechny tyto scénáře neobsahující přímé strojové řešení považujeme za méně pravděpodobné než ty, které se odvíjejí od technologického sektoru. Vůči investované snaze do zlepšení biologické kognice je však vzrůst inteligence čím dál menší, až se nakonec taková řešení kvůli systémovému odporu nevyplácí (Bostrom, 2018, s. 112). Přesto si je ale nejen pro úplnost zahrneme, protože podle nich v současnosti částečně postupujeme a dílčí úspěchy z nich plynoucí by i tak mohly znamenat posun v intelektu, který by nám v řešení problematiky Superinteligence mohl pomoci.

Životospráva a strava

Jedním z aktuálně nejdosažitelnějších zlepšení je prostá úprava životosprávy. Samozřejmě vidí Bostrom potenciál současného běžného mozku, jehož funkčnost by byla nejsnadněji zlepšena za pomocí konzumace nutričně bohatší stravy, pravidelného pohybu a spánku. Zaměřuje se i na výživu matky v době těhotenství, která ovlivňuje tvorbu nervového systému dítěte. V této souvislosti také zmiňuje nedostatek jodu v některých částech světa, nebo například poukazuje na budoucí vývoj přípravků a chytrých drog podporujících soustředění a funkci mozku obecně. Neměli bychom tímto způsobem ale očekávat příchod lidské superinteligence. Tato cesta z počátku znamená pouze malé zvýšení inteligence, od kterého si ale Bostrom při plošné aplikaci slibuje lepší efektivitu při řešení současných problémů a objevování většího počtu lidí s intelektuálním nadáním (Bostrom, 2018, s. 68–69).

Genetické vylepšení

S člověkem pracuje i dál, avšak zapojuje využití genetiky. Zabývá se otázkou úprav genetického kódu, jako například odstraňování zárodečných buněk s genetickými vadami nebo predispozicemi různých chorob. Řeší i otázku genové selekce, a s tím související odstraňování genetických vad a šlechtění lidí. Těmito způsoby by šlo patrně docílit, že by průměrný jedinec byl inteligentní jako v současnosti ti nejinteligentnější z nás. Patrně by tak lidstvo dosáhlo kulturního i komunikačního posunu a stala by se z něj kolektivní superinteligence. Uvědomuje si přitom systémový odpor tohoto přístupu. Tedy, že společensky by tato cesta narážela na zcela zřejmé etické problémy, které by bylo nutné překonat a pakliže by k takovýmto řešením ve větší míře lidstvo přistoupilo, patrně by

vedlo k novému společenskému rozdělení. Došlo by například ke společenským nerovnostem založeným na hodnocení genetické výbavy (Bostrom, 2018, s. 69–78).

Spojení mozku a počítače

„Občas se objevuje myšlenka, že pomocí přímých rozhraní mozek-počítač, zvláště implantátů, by lidé mohli těžit ze silných stránek digitálních počítačů – dokonalé paměti, rychlých a přesných aritmetických výpočtů a vysokorychlostního přenosu dat – díky čemuž by výsledný hybridní systém mohl dalece předčit samotný mozek“ (Bostrom, 2018, s. 80). Tento scénář ale považuje Bostrom v blízké době jako velice nepravděpodobný, protože v současnosti toto řešení považuje za stejně náročné, jako samotné dosažení obecné inteligence. Toto řešení s sebou nese řadu nevýhod, a v současnosti mu v naplnění brání spousta nevyřešených problémů. Například Bostrom zmiňuje obrovské riziko fyzického posunu implantátu, jehož zavedení je až příliš invazivní. Využití uznává pouze v případech léčby například Parkinsonovy nemoci, které funguje ovšem na principu pouhé přímé elektrické stimulace. Spojení hardwaru a biologického mozku má také tu nevýhodu, že stroje vstřebávají a ukládají informace jiným „jazykem“ než lidský mozek a přenos informace mezi těmito dvěma tělesy, nebo mezi mozkem a jiným mozkem, by vyžadoval nějaký způsob efektivního rychlého překladu. Ještě složitějším dělá toto propojení skutečnost, že neuronové propojení každého jednoho mozku na světě je unikátní, a tím i vstřebávání příchozích informací je specifické. Stahování obřího rozsahu znalostí přímo do mozku se tak pravděpodobně neuskuteční. Nakonec Bostrom považuje interakci se stroji, tak jak ji máme dnes, za dostatečnou (Bostrom, 2018, s. 80-85).

Reorganizace společnosti

Dalším na první pohled velice jednoduchým způsobem, jak dojít k Superinteligenci, je superinteligence kolektivní. V tomto přístupu považuje Bostrom za základní složku takovéto struktury lidský potenciál, čímž lze částečně navázat na předchozí geny manipulující řešení. Bostrom nám předkládá úvahu o tom, že naše společnost jednoduše není na tolik dobře organizovaná, aby dosahovala s lidským potenciálem lepších výsledků při řešení problémů a současně v ní figuruje spousta překážek, které její efektivitu brzdí, jako je například byrokracie, lhaní nebo statusové hry. To následně způsobuje například neefektivitu ve sdílení informací. Na druhou stranu přitakává, že společnost se snaží svoji efektivitu stále zvyšovat. Zásadní je práce a přenos

informací a jejich věrohodnost. Zdůrazňuje také dosavadní neplnohodnotné využití internetu, které by s použitím efektivního způsobu skladování a přístupnosti dat za pomoci inteligentních algoritmů mohlo znamenat vytvoření inteligentního webu, který by snad mohl znamenat, že by se s přidáním nějaké klíčové složky mohl stát obecně inteligentním a následně i superinteligentním. K dosažení toho všeho je na místě tedy řešení problémů organizování společenské struktury, dělba práce nebo kvalitnější vzdělávání. Rovněž zmiňuje jako prospěšnou snahu omezit lidskou potřebu klamat například dozorem, zavedením kvantifikování lidské spolehlivosti a ustanovení kulturního narativu racionálního myšlení. Staví na první příčku snahu o pokrok v rozšiřování epistemologií a řešení společenských otázek, které mohou následně prospět funkci samotné společnosti (Bostrom, 2018, s. 86-88).

4.1.2. Strojové inteligence

Cesty, kterými dojdeme k obecné inteligenci a Superinteligenci rychleji, jsou spíše ty zahrnující primárně strojové řešení. Jedním z těchto důvodů by mohlo být, že výpočty na syntetickém nosiči jsou mnohonásobně rychlejší, než je chemie našich mozků, a tak se postavením dobrých základů tohoto systému proces patrně vyřeší v „krátkosti" sám. Na druhou stranu je třeba dodat, že všechna tato řešení si ale berou ve větší či menší míře zcela přiznanou inspiraci z člověka.

Evoluční algoritmus

Jedním z poměrně očekávaných přístupů je samozřejmě kompletně umělý algoritmus. Ten by ale musel mít již zmíněné schopnosti k dosažení superinteligence, jak jsme si již řekli. Musel by se tedy dokázat učit sám o sobě, mít abstraktní myšlení a pracovat s neurčitostí a schopnost pozorovat „smyslově" své prostředí a vstřebávat z něj informace. Na základě toho všeho by pak musel takový algoritmus vnitřně vytvářet závěry a další znalosti sloužící k jeho seberozvoji (Bostrom 2018, s. 48). Celý tento postup ale, jak říká Bostrom, závisí na tom, jakých pokroků dosáhne technologický sektor v oblasti navyšování výpočetní síly. V základu se totiž počítá s množstvím 10^{30} provedených simulací, které by měly za úkol nějakým způsobem emulovat evoluční proces včetně různých evolučních podmínek. Pokud bychom ve studiu evoluce ovšem pokročili a byli schopni definovat klíčové evoluční mechanismy, kterými jsme se dostali

od buňky k člověku, snížili bychom toto číslo o několik desítek řádů (Bostrom, 2018 s. 55).

Emulace mozku

Jestli byl postup zahrnující evoluční algoritmy inspirací sebranou od matky přírody, pak je emulování mozku přímo krádež. V zásadě mluví Bostrom o super inteligenci v základu založené na emulaci mozku člověka. Mozek jedince je nasnímán v poměru 1:1 do počítače, a bez složitějšího porozumění struktuře mozku je zapnut, tedy se začne simulovat jeho neuronová síť. Tento přístup obecně podle Bostroma nevyžaduje žádný teoretický průlom, až na porozumění nízkoprahovým funkcím mozku, abychom jej v digitálním prostředí mohli spustit. Na druhou stranu je systémovým odporem v tomto případě nedostatek dostatečně schopných technologií. Mozek je totiž třeba nějakým způsobem nakrájet na menší části, které se musí zakonzervovat, v detailu poté nasnímat a digitálně vymodelovat. Model ovšem bude také vyžadovat počítačové prostředí, které jej musí zvládnout jednak kapacitně co se úložiště týče, a za druhé musí mít dostatečně velký výpočetní výkon, aby mozek mohl bez problému běžet (Bostrom, 2018, s 58-64). V současné době nám však dělá problém nasnímání hmyzího mozku. Čímž lze i laikovi přesvědčivě naznačit, jak ohromný systémový odpor by nás čekal v případě lidského mozku. Avšak právě nasnímání hmyzí neuronové sítě je první z kroků pochopení operace s tímto „mediem“. Podle Bostroma je to jeden z milníků, kterého když dosáhneme, přijdeme na to, co dál (Bostrom, 2018, s. 114).

Takováto emulace by pak mohla mít, pokud by byla emulována na dostatečně rychlém hardwaru, již jednou zmiňovanou superrychlostní schopnost. Nedocházelo by rovněž k únavě mozku z nedostatku energie nebo nemoci organismu. (Bostrom, 2018, s. 92-93).

5. Cíle a jednání UI

Goodova definice Superinteligence vyznívala ve smyslu, že není jisté, zda ji budeme schopni vůbec kontrolovat. Ať už nám vědeckofantastické romány prezentují ohledně vzpoury umělé inteligence ty nejtemnější scénáře, pravdou zůstává, že vysoce inteligentní agent bude jednat mimořádným způsobem, a v rámci svého konání by mohl ohrozit i existenci lidstva samotného. Příslib vyřešení spousty lidských problémů je tak najednou postaven do protiváhy k obavám, že se naprogramování Superinteligence projeví jako rizikové. To je si myslím dostatečný důvod vůči těmto snažením zastávat aspoň částečně podezřívavé stanovisko. Pojďme si proto vysvětlit, od čeho se pravděpodobně bude utvářet způsob jednání a vytyčování cílů umělých agentů.

Při exkurzu a bourání funkčnosti koncepce Turingova testu jsme lehce zabředli do problematiky vědomí a myšlení. Operovali jsme tady s důkazy, že vědomí není něco, co takovým umělým agentům současně lze přiřknout. Dokazovali jsme si, že jednotlivé algoritmy vědomí nemají. Zdá se, že daleko pravděpodobnější je právě absence, nikoli přítomnost vědomí, která ale potenciální katastrofě neubírá na relevanci. „Vysoká inteligence byla totiž až dosud na vědomí závislá. Pouze vědomá bytost mohla hrát šachy, řídit auto, rozpoznat nemoc nebo vypátrat teroristy. Nyní ale vyvíjíme nový typ ne-vědomé inteligence, schopné vyššího výkonu než lidé. Tato inteligence zvládá úkoly svou schopností rozpoznávat vzory a ne-vědomé algoritmy a už brzy možná předčí vědomé rozpoznávání zákonitostí“ (Harari, 2017, s. 305). To samo o sobě ještě přeci nezní nijak apokalypticky, ohledně (ne)jednání UI. Harari nicméně navazuje: „Představa, že lidé mají nějaké jedinečné duševní vlohy, které nevědomé algoritmy nemohou nikdy získat, je jen zbožné přání“ (Harari, 2017, s. 314). Harari to myslí tak, že jakkoliv by například Turing vylučoval existenci umělého vědomí, neznamená to nejen, že nás předčí i bez něj, ale že vědomí nemá vůbec v tomto pohledu význam. Mozek sám sebe vysvětluje tím, že se dokáže vnímat a aktivně zasahovat do světa mimo něj, ale to Superinteligence dokáže také. Jedno ze zbožných přání tak pravděpodobně tkví v tom, že se nás Superinteligence nebude snažit vyhubit, a že s ní bude možné jednat, což se nezdá být pravděpodobné vzhledem k absenci vědomí.

Jaké je ale tedy jednání takové UI? Užitím pojmu „ortogonální teze“ vysvětluje Bostrom, že jakkoliv inteligenčně schopný umělý agent může mít vytyčený v podstatě jakýkoliv cíl. Není nic iracionálního na tom, si představit Superinteligenci, která má za

úkol počítat zrnka písku, vyrábět kancelářské sponky nebo se dopočítat co nejvíce desetinných míst Ludolfova čísla. Ortogonální teze totiž popisuje nikoliv racionální volbu, ale pouze kognitivní efektivitu. Vysvětluje tím tak možnost neantropomorfních cílů UI. Záleží ovšem i na principu toho, jak umělý agent k svým konečným cílům dojde. Může vycházet z vložení do zárodečné UI, nebo k němu dojít iteračním procesem na základě vývoje genetického algoritmu. V neposlední řadě by také mohl vzniknout na základě preferencí, které měl původní mozek, z nějž je UI emulována. (Bostrom, 2018, s. 166-170). Rozvádí dál, že velice pravděpodobně bude mít Superinteligence mnoho instrumentálních cílů, jejichž dosažení jí bude zajišťovat větší šance dosažení primárního cíle. Mezi takové by pochopitelně mohla patřit sebezáchova. Ta je ovšem na rozdíl od časově ohraničené sebezáchovy člověka pravděpodobně neohraničená. Dále lze zmínit snahy o vytvoření takových podmínek, aby agent nemusel za žádnou cenu měnit své primární cíle nebo aby se jeho jednání pouze z vnějšku zdálo jako oboustranně prospěšné pro toho, kdo by s ním spolupracoval (Bostrom, 2018, s.172-173). To se bohužel může týkat i nás. Pro úplnost zmíníme, že dalším takovým instrumentálním cílem bude dosahování dalších super-inteligenčních schopností, agent tak bude pravděpodobně usilovat o zlepšení své kognice, rychlosti a zřejmě v souladu s tím se bude snažit i dosahovat dalších zdrojů, které by tyto cíle zabezpečovaly (Bostrom, 2018, s. 175-180).

Tvůrce budoucí inteligence by tak neměl, vzhledem k nebezpečnosti volby postranních cílů, pospíchat s naprogramováním primárního cíle. Jednoduše proto, že by mohl své stvořitele vnímat jako hrozbu, na jehož zničení může zneužít naše vlastní technologie nebo si vytvořit vlastní. Sice by bylo patrně jednodušší vytvořit Superinteligenci dopočítávající se právě Ludolfova čísla než takovou, která by měla za úkol sloužit ku prospěchu lidstva, ale právě proto by měl být konečný cíl v souladu s lidskou existencí. Bylo by pochopitelné vznést myšlenku, ať tedy programátoři zanesou do algoritmu takové morální hodnoty, které odpovídají nám, a kterými by se Superinteligence řídila. Nicméně UI není problematická jen pokud jí lidské hodnoty chybí, ale jak popisuje Toby Ord, i pokud se jednoduše dostatečně neslučují zakódované hodnoty s těmi lidskými, na kterých se ani my sami často neshodneme (Ord, 2022, s. 172-173).

Stanovení nějaké jednoduché optimy v jednání Bostrom vylučuje, protože jednoduchost v paletě možných způsobů znamená i její nefunkčnost (2018, s. 282). Tím bezútěšněji tento problém vypadá, pokud začneme uvažovat nad tím, jak abstraktní

hodnoty jsou láska, svoboda, demokracie nebo spravedlnost. Problém nepředstavuje jen realita, že tyto pojmy chápeme individuálně různě, ale také je velice náročné je vyjádřit programovacím kódem (tamtéž). V rámci UI prezentuje sice Bostrom několik řešení, jak se přiblížit dosažení „umělých hodnot" s těmi biologickými, ať už hovoříme o vyučování na základě zpětné vazby, simulací evolučního procesu nebo několika přístupech, které ve zjednodušené formě znamenají, že my UI budeme s narůstající složitostí všechny tyto hodnoty vyučovat. Důležité je poznamenat, že etika sama o sobě je ve filosofii rozsáhlou disciplínou, a tím spíš je o to náročnější její výklad do výpočetních terminologií. Tato disciplína tak patrně na svůj rozvoj teprve čeká (Bostrom, 2018, s. 284-312).

Nejen v rámci utváření způsobu jednání UI se v současné době pohybujeme na křižovatce, ze které můžeme jít za našimi cíli mnoha neznámy, avšak bez toho, že bychom si sami byli jisti, kam tyto cesty vedou. Nemáme žádnou záruku, že naše rozhodnutí budou správná, a tak nám nad hlavou jako Damoklův meč visí katastrofické scénáře různého znění. Po všech výčtech toho, jakým způsobem by mohla umělá inteligence dojít k singularitě a jaké povahy je její (ne)myšlení a jednání by tak mohly obavy z umělé inteligence na čtenáře působit poněkud iluzorně. Faktem ale zůstává, že se na základě těchto teoretických koncepcí položených hmatatelnými vědecky ověřenými zjištěními pořád pohybujeme na poli stále možných scénářů. Můžeme se tak dočíst temných „proroctví" tohoto nebo podobného znění:

„Pokud si chcete představit nepřátelskou umělou inteligenci, nepředstavujte si neživého sečtělého myslitele, který pobývá kdesi na internetu, kde posílá nenávistné emaily. Představte si celou mimozemskou civilizaci, která přemýšlí milionkrát rychleji než člověk a je zprvu upoutána pouze na počítače. Z pohledu těchto stvoření jsou lidé velice hloupí a pomalí. Dostatečně inteligentní UI nezůstane dlouho odkázaná pouze na počítače. V dnešním světě lze e-mailem poslat do laboratoře řetězec DNA, který by jí umožnil vytvořit na počkání proteiny, jež umožní umělé inteligenci původně odkázané na internet postavit umělé životní formy nebo se rovnou pustit do post biologické molekulární výroby[8]" (Yudkowsky, 2023).

[8] To visualize a hostile superhuman AI, don't imagine a lifeless book-smart thinker dwelling inside the internet and sending ill-intentioned emails. Visualize an entire alien civilization, thinking at millions of times human speeds, initially confined to computers—in a world of creatures that are, from its perspective, very stupid and very slow. A sufficiently intelligent AI won't stay confined to computers for long. In today's world you can email DNA strings to laboratories that will produce proteins on demand, allowing an AI initially confined to the internet to build artificial life forms or bootstrap straight to postbiological molecular manufacturing. (Yudkowsky, 2023) Vlastní překlad

Další o dost volnější vidinu rizika, které představuje Superinteligence, vidí Ord takto: Vychází z toho, že se jednoduše dostane na internet, který jí ale v podstatě umožní stát se nesnesitelnou, protože se jí dostane možnosti zálohovat se na bezpočtu míst současně. Získala by přístup k vyšší výpočetní kapacitě, která by jí pomohla dosahovat svých instrumentálních cílů. Zde by čistě spekulativně mohla například manipulovat se světovou ekonomikou, vyvíjet nové zbraně a technologie nebo manipulovat a vydírat lídry světových mocností. Tím by se pak zcela hypoteticky dostala i k zbraním hromadného ničení (Ord, 2022, s.175-176).

Vycházíme-li ze znalostí prezentovaných v této práci, můžeme pozitivistům racionálně připomínkovat jejich odsouvání bezpečnostních záležitostí ohledně umělé inteligence do budoucna. Je to technologie schopná velice prudkého vývoje, jehož základ musí být položen, dokud jí máme pod kontrolou. Pravdou je, že se už dnes dobrovolně a vědomě odevzdáváme do rukou algoritmů, kterým minimálně dáváme možnost nám vybrat, jaký obsah na internetu budeme konzumovat. (Harari, 2017, s. 343) „Technologický pokrok 21. století by mohl zvrátit humanistickou revoluci, vzít lidem vládu nad chodem jejich životů a dát ji ne-lidským algoritmům. Pokud vás to děsí, nedávejte to za vinu softwarovým vývojářům. (...) Byly to vědy o životě, které dospěly k závěru, že organismy jsou vlastně algoritmy. Kdyby to tak nebylo, pokud by organismy fungovaly principiálně jinak než algoritmy, počítače by možná vedly k zásadnímu převratu v jiných oborech, ale nebyly by schopné porozumět našemu životu a řídit ho a rozhodně by se s námi nemohly spojit“ (Harari, 2017, s. 342-343). Místy působí až prázdně, že se nepodávají naprosto přesné a jednoduché odpovědi na to, proč tedy umělá inteligence představuje tak bezbřehá rizika. Skutečně děsivou pravdou je, že tyto odpovědi opravdu nemáme.

6. Závěr

Lidské snahy zpříjemnit podmínky sobě i budoucím generacím zde byly již dávno. V současné době se tyto tendence promítají i do snažení vytvořit Superinteligenci, která, jak doufáme, většinu našich problémů vyřeší za nás. Zásadní však je, abychom vytvořili bezpečné podmínky, ve kterých by taková entita operovala způsobem, který pro nás bude prospěšný. A co je důležitější, aby pro nás tento vynález nebyl likvidační.

Před současnou společností se tak objevuje vidina velké výzvy, o níž nelze předpokládat, kdy přesně přijde a jak se její objev projeví. Nicméně je nutné se k ní už dnes stavět prozíravě a systematicky i z toho důvodu, že potenciální dopady těchto technologií nám mohou být dlouho skryty a k jejich odhalení by mohlo dojít až velice pozdě. Je třeba vzhledem k obřímu potenciálu, který by objev Superinteligence mohl přinést, pohlížet už teď na tuto problematiku obezřetně a promýšlet, jaká nebezpečí se mohou vynořit a jak jim je třeba zabránit nebo se jich zavčasu vyvarovat.

Má-li sepsání této práce nějaký nevyřčený smysl, není tím pouhá deskripce obecných a k této problematice relevantních skutečností, ale na základě jejich vysvětlení vytvořit vhodný prostor pro debatu ohledně tolik očekávaného a obávaného vynálezu Superinteligence. Je totiž zcela nepochybné, že sama Superinteligence a její koexistence (snad) s lidskou společností a v její prospěch mají dnes zřejmě ještě ani netušený potenciál otevírat v takové diskuzi mnohá témata. Ta se zcela jistě mohou stát již v blízké budoucnosti předmětem širších a hlubších úvah, které ovšem v podstatě překračují předpokládaný rozsah této práce.

Frank Herbert ve světě knižní série Duna odehrávající v daleké budoucnosti zmiňuje cosi, co nazývá Služebnický džihád. V rámci příběhu už šlo o historickou událost, která byla prezentována jako válka lidstva s myslícími stroji. Nevíme, kdo tento konflikt započal, ale vítězné lidstvo se nakonec zapřísáhlo nikdy nevytvořit jedinou technologii, která by byť jen trochu zosobňovala schopnost lidského myšlení. Mám obavy, že jsme jako lidstvo vstoupili do vlaku, ze kterého vystoupíme jen těžko, pokud je to ještě možné. Obecná ochota takový (a zřejmě pořádně rozjetý) vlak opustit je v přímém rozporu s očekáváním, co pokračování jízdy (pozitivního) přinese. Současná společnost je rozdělena vidinami individuálních úspěchů, a tak se honba za objevením Superinteligence zdá až příliš lákavou. Přeji nám, aby bylo skutečně možné docílit řešení,

které nám zajistí Superinteligenci jednající v souladu s námi a pro obecné dobro. Přeji nám, ať nikdy nebude muset nastat něco jako válka strojů s lidmi. Vzhledem k tomu, že se při objevování Superinteligence aspoň na počátku díváme do podstaty naší mysli, přeji nám, ať z nás vyjde to nejlepší.

Seznam literatury:

- TURING, Alan Mathison. I.—COMPUTING MACHINERY AND INTELLIGENCE, *Mind*, Volume LIX, Issue 236, October 1950, Pages 433–460, https://doi.org/10.1093/mind/LIX.236.433
- ARENDT, Hannah. *Vita activa, neboli, O činném životě*. Praha: OIKOYMENH, 2007. Knihovna novověké tradice a současnosti. ISBN 978-80-7298-185-4.
- BOSTROM, Nick, ed. Introduction—The Transhumanist FAQ. MERCER, Calvin a Derek F. MAHER. *Transhumanism and the Body* [online]. 1. New York: Palgrave Macmillan, 2014, s. 1-17 [cit. 2022-06-13]. ISBN 978-1-137-34276-8. Dostupné z: doi: https://doi.org/10.1057/9781137342768
- BOSTROM, Nick. *Superinteligence: až budou stroje chytřejší než lidé*. V českém jazyce vydání druhé. Přeložil Jan PETŘÍČEK. Praha: Prostor, 2018. Globus (Prostor). ISBN 978-80-7260-389-3.
- DARWIN, Charles, Stanislav KOMÁREK, Emil HADAČ, Alena HADAČOVÁ a Hana MARSAULT. *O vzniku druhů přírodním výběrem*. Vyd. 3., V nakl. Academia 2., rev. Praha: Academia, 2007, 579 s. ; 21 cm. ISBN 978-80-200-1492-4.
- ELOUNDOU, Tyna, Sam MANNING, Pamela MISHKIN a Daniel ROCK. *GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models*. Cornell University, 2023. Dostupné z: https://doi.org/10.48550/arXiv.2303.10130
- GOOD, Irving John, „Speculations Concerning the First Ultraintelligent Machine“, v: Franz L. Alt a Morris Rubinoff (eds.), *Advances in* Computers, sv. 6, Academic Press, New York 1965, str. 33.
- GVOŽDIAK, Vít. *Základy sémiotiky 2*. Olomouc: Univerzita Palackého v Olomouci, 2014. Qfwfq. ISBN 978-80-244-4317-1.
- HARARI, Yuval N. *Homo deus: stručné dějiny zítřka*. Přeložil Alexander TOMSKÝ, přeložil Anna PILÁTOVÁ. Voznice: Leda, 2017. ISBN 978-80-7335-628-6.
- HAVLÍK, Marek. [online]. Západočeská univerzita v Plzni, 2012, s. 186-208. [cit. 2023-06-18]. Dostupné z: https://dspace5.zcu.cz/bitstream/11025/6556/1/Havlik_1.pdf

- HUXLEY, Julian. *New bottles for new wine* [online]. London: Chatto & Windus, 1957 [cit. 2022-06-17]. Dostupné z: https://archive.org/details/NewBottlesForNewWine
- ORD, Toby. *Nad propastí: existenční riziko a budoucnost lidstva*. Přeložil Anna ŠTÁDLEROVÁ. Praha: Argo, 2022. Crossover. ISBN 978-80-257-3779-8.
- PETIŠKA, Eduard. *Staré řecké báje a pověsti*. Vyd. 15., V Ottově nakl. 2. Ilustroval Václav FIALA. Praha: Ottovo nakladatelství, 2006. ISBN 978-80-7360-489-9.
- SEARLE, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*(3), 417-424. doi:10.1017/S0140525X00005756
- VOCŮ, Michal. Šachový šampión poražen počítačem IBM. *Ikaros* [online]. 1997, ročník 1, číslo 3 [cit. 2022-06-11]. urn:nbn:cz:ik-10025. ISSN 1212-5075. Dostupné z: http://ikaros.cz/node/10025
- YUDKOWSKY, Eliezer. 2008. Artificial intelligence as a positive and negative factor in global risk. In Global catastrophic risks, ed. Nick Bostrom and Milan M. Cirkovic, 308–345.New York: Oxford University Press Dostupné z: https://intelligence.org/files/AIPosNegFactor.pdf
- YUDKOWSKY, Eliezer. Pausing AI Developments Isn't Enough. We Need to Shut it All Down. *Time* [online]. March 29, 2023 [cit. 2023-06-16]. Dostupné z: https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/