Univerzita Hradec Králové
Filozofická fakulta
Katedra filozofie a společenských věd

# „Technologická Singularita“

Bakalářská práce

Autor: Jaroslav Malík
Studijní program: B6101
Studijní obor: Filozofie a společenské vědy
Forma studia: Prezenční

Vedoucí práce: prof. RNDr. Jaroslav Peregrin, CSc.

Hradec Králové, 2018

Univerzita Hradec Králové
Filozofická fakulta

# Zadání bakalářské práce

| | |
|---|---|
| **Autor:** | **Jaroslav Malík** |
| Studium: | F15BP0057 |
| Studijní program: | B6101 Filozofie |
| Studijní obor: | Filozofie a společenské vědy |
| **Název bakalářské práce:** | **Technologická singularita** |
| Název bakalářské práce AJ: | Technological singularity |

**Cíl, metody, literatura, předpoklady:**

Cílem práce je prozkoumat pojem technologické singularity, jak se objevil v diskusích v rámci umělé inteligence a "transhumanismu". Práce by měla zejména zmapovat historii, varianty a filosofickou relevanci tohoto pojmu. Současně by měla kriticky zhodnotit užitečnost tohoto pojmu pro filosofické úvahy o budoucnosti lidstva a oddělit od sebe jeho "popkulturní" aspekty a ty aspekty, které jsou hlubší a relevantní z hlediska uvažování o vývoji umělé inteligence.

KURWEIL Ray. The Singularity Is Near. When Humans Transcend Biology. New York: Viking Adult, 2005. BOSTROM Nick. Superinteligence. Paths, Dangers, Strategies. Oxford: Oxford University Press, 2014. SHANAHAN Murray. The technological Singularity. Cambridge, London: The MIT Press, 2015. EDEN H. Ammon, MOOR H. James, SORAKER H. Johnny, STEINHART Eric (eds). Singularity Hypotheses. A Scientific and Philosophical assesment. Heidelberg,New York, London, Dordrecht: Springer.

| | |
|---|---|
| Garantující pracoviště: | Katedra filosofie a společenských věd, Filozofická fakulta |
| Vedoucí práce: | prof. RNDr. Jaroslav Peregrin, CSc. |
| Oponent: | Mgr. et Mgr. Michal Rigel |
| Datum zadání závěrečné práce: | 13.12.2016 |

**Prohlášení**

Prohlašuji, že jsem tuto bakalářskou práci vypracoval (pod vedením vedoucího bakalářské práce) samostatně a uvedl jsem všechny použité prameny a literaturu.

V Hradci Králové dne 7. 4. 2018

Jaroslav Malík

Anotace

MALÍK, JAROSLAV. *Technologická Singularita.* Hradec Králové: Filosofická fakulta, Universita Hradec Králové, 2018, 60 s. Bakalářská práce.

Cílem práce je prozkoumat pojem technologické singularity, jak se objevil v diskusích v rámci umělé inteligence a "transhumanismu". Práce by měla zejména zmapovat historii, varianty a filosofickou relevanci tohoto pojmu. Současně by měla kriticky zhodnotit užitečnost tohoto pojmu pro filosofické úvahy o budoucnosti lidstva a oddělit od sebe jeho "popkulturní" aspekty a ty aspekty, které jsou hlubší a relevantní z hlediska uvažování o vývoji umělé inteligence.

Klíčová slova: Umělá inteligence, Filosofie mysli, Technologická singularita, Transhumanismus

Annotation

MALÍK, JAROSLAV. *Technological Singularity.* Hradec Králové: Philosophical Faculty, University of Hradec Králové, 2018, 60 pp. Bachelor Degree Thesis.

The aim of this thesis is to explore the term of technological singularity, how it appeared in the discussions about artificial intelligence and "transhumanism". This thesis should map the history, variants and philosophical significance of this term. At the same time it should critically assess the usefulness of this term for philosophical thought about the future of mankind and separate its "pop-culture" aspects from its other aspects, which are deeper and more relevant from the point of view of thinking about the development of artificial intelligence.

Keywords: Artificial intelligence, Philosophy of mind, Technological singularity, Transhumanism

Obsah

1. Úvod ..... 6
1.1 Historie pojmu ..... 6
2. Teorie technologické singularity ..... 9
2.1. Cesty k SI ..... 12
2.2. Formy SI ..... 17
2.3. Transhumanistická perspektiva ..... 20
3. Problémy a kritika Singularity a SI ..... 22
3.1. Problém Singularity ..... 22
3.2. Kritika hnutí Singularity ..... 37
3.3. Problém AI – Strojová etika a problém kontroly ..... 42
3.4. Problémy WBE – Upload mysli a Identita ..... 50
4. Závěr ..... 61
Seznam použitých pramenů ..... 62
Literatura ..... 62
Elektronické zdroje ..... 62
Seznam obrázků ..... 65

## Seznam použitých zkratek a termínů:

AI – Artificial Inteligence

HI – Human Intelligence

MIRI – Machine Inteligence Research Institute

Prima facie - lat. na první pohled

SI – Super Inteligence

SIA – Self Identification Assumption

SSA – Self Sampling Assumption

Thinktank – označení instituce nebo skupiny pro výzkum typicky se zabývající ekonomikou nebo politikou

Vitrifikace – proces přeměny určitého materiálu ve sklo

WBE – Whole Brain Emulation

Wutr - weighty, unseen, temporarily remove

# 1. Úvod

Problematika umělé inteligence je v současné době široce diskutovaným tématem, který byl i často analyzován filosoficky, z důvodu našich rostoucích technologických možností. Pojem technologické singularity do této problematiky přináší další dimenzi. Tou je problém super inteligence (dále SI). Podle teoretiků singularity zde existuje možnost, že až vyvineme umělou inteligenci, která bude dostatečně sofistikovaná, tak brzy potom tato inteligence vytvoří ještě lepší inteligenci. Pokud tato skutečnost může nastat, tak její následky budou nedozírné.

Cílem této bakalářské práce bude tento problém rozebrat z filosofického hlediska se snahou ukázat, jak argumenty pro, tak také proti. Cílem první části bude nejprve uvést historii tohoto pojmu, kdo ho poprvé v tomto kontextu použil, jak se vyvíjel atd., dále ukázat jednotlivé teorie tohoto pojmu a jejich představitele jako jsou Vernon Vinge, I. J. Goodwin, Ray Kurzweil, vysvětlit v čem spočívají a pak ukázat jednotlivé podoby, které může SI mít, jakou je umělá inteligence nebo například WBE (Emulace celého mozku) a další. Druhá část se soustředí na problémy spojené se singularitou a se SI, které budou rozděleny podle forem SI na problémy spojené s umělou inteligencí a problémy spojené s WBE. Pak bude následovat závěr, kde shrneme celou práci.

## 1.1 Historie pojmu

Samotné slovo singularita, bylo poprvé použito matematiky pro popsání hodnoty, která překračuje jakékoliv konečné limity. Dalším odvětvím, které začalo tento termín používat, byla astrofyzika. Zde šlo o výbuch hvězdy, supernovy, při kterém se pozůstatky hvězdy propadnou do bodu s nulovým rozměrem a nekonečnou hustotou. To, co se zrodí z těchto pozůstatků, tedy pojmenovali singularitou.[1] První zmínka o Technologické singularitě pochází údajně od Johna von Neumanna: „Neustále se zrychlující pokrok a změny v modech lidského života vytváří dojem blížící se singularity v historii lidské rasy, za kterou lidské činnosti, jak je známe, nemohou pokračovat“. Takto to uvádí Stanislav Ulam v popisu jedné konverzace Neumannem.[2]

[1] KURWEIL Ray. *The Singularity Is Near. When Humans Transcend Biology*. 1. vyd. New York: Viking Adult, 2005. ISBN 0-670-03384-7. s. 21–23.

[2] ULAM Stanislaw. John Von Neumann 1903 – 1957. In *Bulletin of the American Mathematical Society*. [online]. Providence: American Mathematical Society. 1958. Roč. 64, č. 3. s. 5[cit. 12. 2. 2018]. ISSN

Dále se singularita implicitně objevuje v článku I. J. Goodwina pro Advances in Computers Volume 6. z roku 1965 pod jménem „Speculations concerning the first ultraintelligent machine“, kde autor předkládá svou hypotézu inteligenční exploze, která nastane, když jeden inteligentní stroj postaví další, který je lepší. Goodwin v tomto článku pojem singularity neužije, ale vysvětluje proces, jakým by mohla nastat a jeho definice ultra inteligentního stroje je důležitá v diskuzích o singularitě.[3]

Ten, kdo popularizoval a víceméně zavedl pojem technologická singularita je Vernon Vinge. Poprvé ho užil v úvodu pro časopis Omni v roce 1983, kde předpovídá, že nás čeká exponenciální vývoj inteligence samotné, který povede k růstu rychlosti vývoje samotného[4]. Tyto myšlenky pak velice rozšířil ve svém příspěvku do Symposia organizovaném NASA pod jménem „Vision 21: Interdisciplinary Science and Engineering in the Era of Cyberspace“, kde mluví o singularitě jako o neodvratné události, která nás v budoucnosti čeká.[5]

Ray Kurzweil a Hans Moravec v 80. a 90 letech také napsali několik knih o tomto tématu. V roce 1989 Kurzweil napsal knihu „The Age of Inteligent Machines“ kde píše, že míříme k budoucnosti, kde budou existovat stroje s větší než lidskou inteligencí. Další knihu napsal v roce 1999 pod jménem “Age of Spiritual Machines: When Computers Exceed Human Inteligence“, ve které popisuje rostoucí blízkost mezi biologickou a umělou inteligencí. Moravec v roce 1988 napsal knihu „Mind Children“, ve které analyzoval vývoj robotiky a dochází ke stejný závěrům jako Kurzweil a ve své další knize z roku 1999 „Robot: Mere Machine to Transcendent Mind“, dokonce nazývá roboty, které postavíme v roce 2040, našimi evolučními dědici.[6] Nejznámější kniha, kterou Kurzweil

---

1088-9485. Dostupné z: <http://www.ams.org/journals/bull/1958-64-03/S0002-9904-1958-10189-5/S0002-9904-1958-10189-5.pdf>. vlastní překlad.

[3] GOOD J. Irving. Speculations concerning the first ultraintelligent machine. In ALT L. Franz, RUBINOFF Morris (eds.). *Advances in Computers Volume 6*. 1. vyd. London: Academic Press, 1965.

[4] VINGE Vernor. First Word. In *OMNI* [e-kniha]. Englewood: Jerrick Publishing, 17, 2. 2016, Roč. 5, č. 4, s. 10 [cit. 12. 2. 2018]. Dostupné z: <https://www.amazon.com/OMNI-Magazine-January-1983-ebook/dp/B01BXG1M2E/ref=sr_1_1?s=books&ie=UTF8&qid=1520446761&sr=1-1&keywords=OMNI+Magazine+january+1983>.

[5] VINGE Vernor. The coming technological singularity. How to survive in the post-Human era. In *NASA, Lewis Research Center. Vision 21: Interdisciplinary Science and Engineering in the Era of Cyberspace* [online]. Cleveland: NASA, Lewis Research Center. 1. 12. 1993 [cit. 23. 12. 2017]. Dostupné z: <https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19940022856.pdf>

[6] KURWEIL Ray. *The Singularity Is Near. When Humans Transcend Biology*. 1. vyd. New York: Viking Adult, 2005. ISBN 0-670-03384-7. s. 23–26.

napsal je „Singularita is Near: When Humans Transcend Biology“, ve které i rozebírá svou teorii singularity a její důsledky pro lidskou rasu.[7]

V současnosti pak máme několik výzkumných organizací a thinktanků, které se singularitou zabývají či zabývaly. Například MIRI, což je zkratka pro Machine Inteligence Research Institute (Původně Singularity Institute)[8], který založil Eliezer Yudkowski společně s Brianem a Sabine Atkinsovými.[9] MIRI je nezisková organizace, která se zabývá problémem AI, specificky jeho důsledky v budoucnosti.[10] MIRI také stála za založením tradice každoročních Singularity summitů, jejichž cílem je co nejlépe osvětlit rizika spojená s vývojem technologií.[11] Další organizací, kterou si můžeme uvést, je Singularity university, mezi jejíž zakladatele se počítá například Ray Kurzweil. Cílem členů je rozšiřovat povědomí o aplikaci exponenciálních technologií, abychom je mohli využít pro vyřešení největších problémů lidstva.[12] Singularity University také od MIRI přebrala organizaci Singularity summitů.[13]

[7] Tamtéž
[8] MUEHLHAUSER Luke. We are now „Machine Inteligence Research Institute“ (MIRI). In *MIRI (Machine Inteligence Research Institute)* [*online*]. Berkeley: Machine Inteligence Research Institute, 30. 1. 2013 [cit. 12. 2. 2018]. Dostupné z: <https://intelligence.org/about/>.
[9] *Singularity Institute Strategic Plan.* [online]. Berkeley: Machine Inteligence Research Institute (Singularity Institute), srpen 2011 [cit. 12. 2. 2018]. Dostupné z: <https://intelligence.org/files/strategicplan2011.pdf>.
[10] About MIRI. In *MIRI (Machine Inteligence Research Institute)* [online]. Berkeley: Machine Inteligence Research Institute [cit. 12. 2. 2018]. Dostupné z: <https://intelligence.org/about/>
[11] Singularity Summit. In *MIRI (Machine Inteligence Research Institute)* [online]. Berkeley: Machine Inteligence Research Institute [cit. 12. 2. 2018]. Dostupné z: <https://intelligence.org/singularitysummit/>.
[12] What is SU?. In *Singularity university.* [online]. San Jose: Singularity University, © 2017 [cit. 12. 2. 2018]. Dostupné z: <https://su.org/about/>.
[13] SINGULARITY UNIVERSITY ACQUIRES SINGULARITY SUMMIT. In *Singularityu.org* [online]. San Jose: Singularity University, © 2018 [cit. 12. 2. 2018]. Dostupné z: <https://singularityu.org/2012/12/09/singularity-university-acquires-the-singularity-summit/>.

# 2. Teorie technologické singularity

Přejděme tedy k jednotlivým teoriím singularity. Prvním, o kom můžeme říci, že se zamýšlel nad konceptem singularity, je Irving John Good, který už v roce 1965 očekával, že během 20. století bude postaven ultra inteligentní stroj. Jeho vznik by pak úplně změnil povahu naší společnosti z velmi prostého důvodu, protože by to byl poslední lidský vynález. Důvodem pro tuto skutečnost je podle něho to, že vznik tohoto stroje povede k tomu, čemu říká inteligenční exploze.[14]

Vyplývá to z jeho definice ultra inteligentního stroje, který definuje jako stroj schopný překonat člověka v jakékoliv intelektuální aktivitě. A následně předkládá i první argument pro vznik singularity tak, že stanovuje design podobných strojů za jednu z intelektuálních aktivit danou člověku, tudíž ultra inteligentní stroj by byl schopen stvořit ještě lepší stroj. Tato skutečnost by podle Gooda bezpochyby vedla k inteligenční explozi, kde by ultra inteligentní stroje stavěly ještě chytřejší stroje a člověk by byl zanechán daleko za nimi. To je singularita v jeho podání.[15]

Pokračujme tím, co říká Vernor Vinge. Ve své práci „The coming technological singularity: How to survive in the post-Human era“[16] argumentuje tím, že stojíme na pokraji neuvěřitelné změny, jejíž příčinou budou entity s větší než lidskou inteligencí. Tyto entity by podle Vingeho mohly vzniknout těmito následujícími způsoby: Vývojem superpočítačů, skrze velké počítačové sítě, rozhraním mezi člověkem a počítačem, nebo biologickým vylepšením lidského intelektu. Přičemž první tři možné způsoby jsou úzce spojené s vývojem počítačů.

Důsledkem této události by bylo zrychlení technologického pokroku, protože od té chvíle by byl pokrok zajišťován větším intelektem než lidským. Analogie, kterou předkládá, je spojena s naší evoluční historií. Zvířata dokáží řešit problémy, ale málokdy rychleji než přírodní selekce. Za to člověk je schopen provádět simulace situací ve své hlavě, tím dokáže řešit problémy daleko rychleji než přírodní selekce. Pokud lidský

[14] GOOD J. Irving. Speculations concerning the first ultraintelligent machine. In ALT L. Franz, RUBINOFF Morris (eds.). *Advances in Computers Volume 6.* 1. vyd. London: Academic Press, 1965. s. 78

[15] Tamtéž s. 33

[16] VINGE Vernor. The coming technological singularity. How to survive in the post-Human era. In *NASA*, Lewis Research Center. *Vision 21: Interdisciplinary Science and Engineering in the Era of Cyberspace* [online]. Cleveland: NASA, Lewis Research Center. 1. 12. 1993 [cit. 23. 12. 2017]. Dostupné z: <https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19940022856.pdf>.

intelekt byl schopen takto urychlit pokrok, nadlidský intelekt by dovedl provádět ty samé simulace daleko rychleji, což by vedlo ke zrychlení pokroku exponenciální řadou. Vinge říká, že objevy, o kterých si myslíme, že se stanou až za milion let, by najednou byly v dosahu.[17]

Z těchto důvodů se tato událost tváří jako singularita, protože je to bod, který když překročíme, tak naše dosavadní modely přestávají fungovat. Podle Vingeho lze pozorovat tuto nejistotu daleko dříve. Podle něj už spisovatelé science fiction v 60., 70. a 80. letech měli určité povědomí o této přicházející události. Cítili, že jejich předpovědi mohou jít jen do určitého bodu, za nímž už si nemohli být ničím jistí. Jak čas bude postupovat, tak bychom podle něj měli vidět více symptomů, tušení spisovatelů se rozšíří do dalších oblastí a uvidíme rozvoj robotiky do čím dál tím složitějších oblastí.[18]

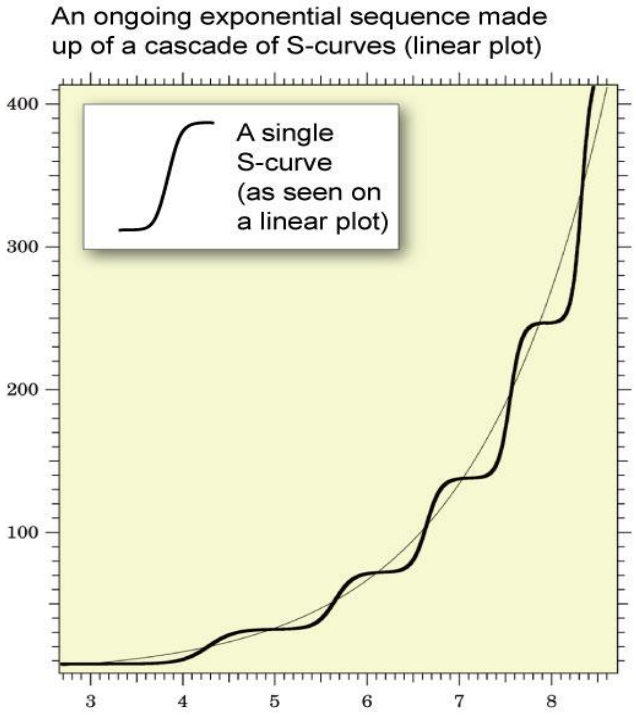
Pro Vingeho singularita bude výsledkem naší přirozené soutěživosti a možností, které se nalézají v technologiích. Z tohoto důvodu Vinge přirovnává singularitu k lavině, kterou začneme my sami. My sami si stanovíme její podmínky, takže máme možnost si nastavit danou situaci jako méně nepřátelskou vůči nám. Singularita je pro něj přírodní událostí a hranicí, za kterou nemůžeme nic vidět.[19]

Teorie Raye Kurzweila stojí na tom, jak chápe evoluci a s ní spojený lidský pokrok, o kterém si myslí, že je exponenciální, přestože se nám může zdát být pouze lineárním. Singularita je pro něj také tou konečnou částí exponenciálního růstu. Když si tento růst představíme jako křivku, je to ta skoro vertikální část. Je to bod, za kterým se pokrok neuvěřitelně zrychlí. Evoluci chápe Kurzweil jako vývoj toho, jak se skladovaly a jak se nakládalo s informacemi a rozděluje ji do 6 epoch. První z nich je fyzika a chemie, což je doba, kdy informace byly obsaženy v jednoduchých strukturách ve hmotě a energii, specificky v jejich schématech. Druhou je biologie a DNA, ve které vznikají biologické systémy, které ukládají informace. Třetí je mozek, v tomto období se objevují nervové systémy, které zaznamenávají a skladují informace. Čtvrtou je technologie, kdy lidé vytváří na nich nezávislé systémy zpracovávající informace. Pátá epocha je splynutí lidské technologie s lidskou inteligencí. Tohle je chvíle, kdy podle Kurzweila nastane singularita, chvíle, kdy se odpoutáme od limitů biologické evoluce. V poslední epoše jako

[17] Tamtéž, s. 12
[18] Tamtéž, s. 12–14
[19] Tamtéž, s. 15

následek singularity naplní inteligence hmotu a energii okolo sebe informacemi a vesmír si uvědomí sebe sama.[20]

Kurzweil také předkládá svůj argument pro blížící se singularitu, který nazývá The Law of Accelerating Returns. Jak již bylo řečeno, Kurzweil věří, že rychlost pokroku, rychlost vzniku produktů evoluce, exponenciálně roste. Důvodem pro tento růst je to, že nové produkty stojí na základech těch starých, a to ten růst vyvolává. Neroste ale pouze rychlost vzniku, ale také počet vzniklých produktů. Tuto skutečnost můžeme vidět například v Moorově zákoně, ve kterém každá nová generace čipů obsahuje dvakrát více komponentů za stejnou cenu dané jednotky a zároveň je rychlejší než ta předchozí. Mohli bychom také říci, že procházíme změnami paradigmatu čím dál tím rychleji. Kdybychom si měli představit vývoj paradigmatu, tak by asi měl tři části. Počáteční pomalý růst, po kterém by následoval explozivní rychlý růst, a nakonec ustálení a vyrovnání daného paradigmatu, po kterém by byl nahrazen paradigmatem jiným. Kdybychom si měli tento



Obrázek č. 1 – Série na sebe navazujících S křivek

[20] KURWEIL Ray. *The Singularity Is Near. When Humans Transcend Biology.* 1. vyd. New York: Viking Adult, 2005. ISBN 0-670-03384-7. s. 7–21.

proces znázornit v grafu, tak by měl tvar S křivky. Takže pro znázornění exponenciálního růstu bychom použili sérii na sebe navazujících S křivek.[21]

Podle Eliezera Yudkowského můžeme tyto tři uvedené teorie chápat jako rozdílné školy singularity s navzájem se podporujícími a vyvracejícími argumenty. Vernor Vinge zastupuje školu horizontu událostí a jeho hlavním tvrzením je, že budoucnost po příchodu SI není možné předpovědět. Do singularity všechen pokrok poháněl lidský intelekt, ale po ní ho bude pohánět nadlidský intelekt, jehož činění bude nad naše chápání. Do podobné situace se můžeme dostat i dnes, stačí, když si vzpomeneme na šachový počítač DeepBlue, který porazil šachového mistra. To se stalo, protože počítač hrál šachy na vyšší úrovni než člověk, a aby ho člověk porazil, musel by hrát na stejné úrovni. Stejně se to tedy má se SI, abychom mohli pochopit její jednání, museli bychom myslet na stejné úrovni.

Ray Kurzweil, se svou školou zrychlující se změny, říká něco jiného. Protože naše intuice nám říká, že pokrok je lineární, tak očekáváme za našeho života stejnou míru změny, jakou čekali naši předkové. Tak tomu ale není. Pokrok se sám na sobě staví, a to ho vede ke zrychlování. A protože díky tomuto zrychlení pokrok následuje křivky, které jsou exponenciální, tak jsme schopni předpovídat, kam bude pokrok směřovat.

Nakonec tvrzením I. J Gooda a jeho školy Inteligenční exploze je, že inteligence vždy byla zdrojem technologického pokroku. Pokud ale inteligence může být vylepšena technologií, tak to vytváří pozitivní zpětnou vazbu, kdy tato vylepšená inteligence zase vylepší tu danou technologii. Nový ultra inteligentní počítač po svém vzniku následně vytvoří ještě lepší počítač. To změní povahu technologického pokroku, který do té chvíle poháněly lidské neurony. Bude jej hnát síla transistorů ultra inteligentního počítače a jeho nástupců, tudíž bude mnohem rychlejší, než byl předtím.[22]

## 2.1. Cesty k SI

Uvedli jsme jednotlivé teorie singularity a argumenty s nimi spojené. Přes různé koncepce a definice singularity se všechny shodují na tom, že pro to, aby singularita nastala, je zapotřebí super inteligentních činitelů. Jak by tedy SI mohla vzniknout a jakou

[21] Tamtéž, s. 35–43
[22] YUDKOWSKI Eliezer. Three Major singularity schools. In *MIRI (Machine Inteligence Research Institute)* [online]. 30. 8. 2007 [cit. 23. 12. 2017]. Dostupné z: <https://intelligence.org/2007/09/30/three-major-singularity-schools/>.

formu by mohla mít? Definujme super inteligenci jako „intelekt, který vysoce převyšuje kognitivní výkon lidstva ve všech pro nás důležitých doménách“. Tato definice velice jasně ukazuje, že máme na mysli obecnou inteligenci, a ne pouze super inteligenci v některých oblastech, jakou by byla například inženýrská super inteligence, což by byl intelekt, který by nás vysoce převyšoval pouze v technických oborech. Přejděme tedy k jednotlivým cestám k SI.[23]

Jednou cestou je Umělá Inteligence (AI). Přestože ještě neznáme způsob, jak naprogramovat obecnou umělou inteligenci, máme tušení, jaké komponenty by takový systém musel mít. Jednou z nich je schopnost se učit, což by byla hlavní část designu tohoto systému, která by mu umožnila, aby si vybudoval intelekt. Další nutnou komponentou by byla schopnost nakládat s nejistotou a pravděpodobností. Také by musela umět vnímat prostředí a vnitřní vztahy a pak na nich dále stavět koncepty pro logické a intuitivní myšlení.[24] První představu takového stroje uvedl Alan Turing. Měl představu „stroje dítěte“. Místo toho abychom se snažili postavit stroj s myslí dospělého, proč nejdříve nepostavit stroj s myslí dítěte?[25]

Jiným způsobem, jak bychom mohli dosáhnout obecné umělé inteligence, by bylo využít podobný proces, který nám dal intelekt, a tím je proces evoluční. Pokud slepý evoluční proces dokázal dát vznik intelektu, tak podobný proces, který je pod vedením programátora, by to zvládl daleko efektivněji. Musíme si ale dávat pozor. Evoluce umožnila ptákům létat, a nám lidem se to později také podařilo, ale ne stejným způsobem jako evoluci. Stejně tak, i přes naše technické schopnosti, se nám nepodařilo zreprodukovat všechny úspěchy evoluce.[26]

Další možností je inspirovat se tím, jak funguje mozek. Tohle samo o sobě také není nová myšlenka. Už v 50. letech 20. st. se John Von Neumann zabýval právě podobnostmi mezi počítači a mozkem.[27] Tento směr se z velké části opírá o dosavadní

[23] BOSTROM Nick. *Superinteligence. Paths, Dangers, Strategies*. 1. vyd. Oxford: Oxford University Press, 2014. vlastní překlad. ISBN 978-0-19-967811-2. s. 22.
[24] Tamtéž s. 23
[25] TURING, Alan. COMPUTER MACHINERY AND INTELLIGENCE. In *Mind* [online]. Oxford: Oxford University Press. 1. 11. 1950, Roč. LIX, č. 236, s. 456 [cit. 8. 2. 2018]. ISSN 1460-2113 Dostupné z: <https://doi.org/10.1093/mind/LIX.236.433>.
[26] BOSTROM Nick. *Superinteligence. Paths, Dangers, Strategies*. 1. vyd. Oxford: Oxford University Press, 2014. ISBN 978-0-19-967811-2. s. 23–24.
[27] NEUMANN V. John. *The Computer and the Brain*. 3. vyd. New Haven: Yale university Press, 2012. ISBN 978-0-300-18111-1.

úspěchy neurovědy a kognitivní psychologie. Nejde o to se snažit o emulaci mozku, ale spíše o pochopení základních principů jeho fungování, na kterých pak umělou inteligenci postavíme. Neurální sítě jsou druhem AI inspirované mozkem. Stejně tak technika zpětnovazebního učení je ve velké míře inspirována psychologickými teoriemi zvířecího vnímání. Není ale nutné stavět AI pouze na těchto technikách, mohli bychom také použít hybridní řešení a spojit některé tyto principy s některými čistě umělými technikami. Nakonec tu existuje variace na Turingovu představu dětského stroje v představě zárodku AI. Na rozdíl od Turingovy představy nějaké základní architektury, na které se staví, by tento typ AI byl schopen modifikovat i svou vlastní strukturu. Tento zárodek by se neustále vylepšoval, nejdříve metodou pokus omyl, později po vylepšení by použil jiné metody. Každá další verze by byla chytřejší než ta předtím. [28]

Jinou cestou by mohla být WBE, neboli emulace celého mozku. Procesu vytváření této emulace neboli procesu přechodu obsahu mozku do počítače se říká uploadování.[29] Bostrom popisuje jednu cestu, jak bychom toho mohli dosáhnout. Nejdříve potřebujeme detailní snímky mozku. Abychom je získali, tak bychom pomocí vitrifikace přeměnili mozkovou tkáň na hmotu podobnou sklu. Pak s pomocí speciálního laseru bychom daný mozek rozřezali na tenké plátky. Tyto plátky bychom pak naskenovali do počítače. Počítač by pak musel ze snímků zrekonstruovat 3D model neurální sítě daného mozku, to co mu umožnilo myslet. Tyto data by pak musel zkombinovat s neuro-počítačovými modely různých neurálních struktur a dalších neurálních elementů.[30]

V poslední fázi bychom tento model vzali a začali ho emulovat na dostatečně silném počítači. Na rozdíl od AI, pro WBE nepotřebujeme vědět, proč mozek umožňuje myšlení. Problém zde ale spočívá ve vyspělosti daných technologií. Potřebujeme snímek s dostatečně vysokým rozlišením, software, který vezme snímky a sestaví 3D model a nakonec hardware, který bude mít dost výpočetní síly pro simulaci daného mozku. Tyto

[28] BOSTROM Nick. *Superinteligence. Paths, Dangers, Strategies*. 1. vyd. Oxford: Oxford University Press, 2014. ISBN 978-0-19-967811-2 s. 28–29.
[29] CHALMERS J. David. The Singularity: A Philosophical Analysis. In AWRET Uziel ed. *The Singularity: Could artificial inteligence really out-think us (and would we want it to)?* [e-kniha]. 1. vyd. Exeter: Imprint Academic. 1. 12. 2016 [cit. 8. 2. 2018]. Dostupné z: <https://www.amazon.com/Singularity-Journal%20Consciousness-Studies-ebook/dp/B01N1N6KLZ/ref=mt_kindle?_encoding=UTF8&me>.
[30] BOSTROM Nick. *Superinteligence. Paths, Dangers, Strategies*. 1. vyd. Oxford: Oxford University Press, 2014. ISBN 978-0-19-967811-2. s. 30.

technologie jsou dosažitelné, ale ne v blízké budoucnosti. Zatímco v AI nám chyběly teoretické znalosti, zde nám spíše chybí technologické možnosti.[31]

Je důležité podotknout, že toto nemusí být jediná možná cesta pro emulaci. Takových cest lze použít několik. Postup, který nám ukazuje Bostrom je příkladem destruktivního uploadování, protože během procesu je mozek zničen. Jsou tu ale i jiné formy, jako například postupné uploadování ve formě nanotrasferu. Při tomto procesu je jedno nebo i více nano zařízení vloženo do mozku, kde se připojí na jeden neuron. Toto zařízení pak zmapuje funkce a propojení neuronu a následně ho začne simulovat, čímž ho efektivně v mozku nahradí. Tento proces se pak opakuje pro ostatní neurony v mozku. U této metody na rozdíl od té předchozí nedochází k destrukci mozku, pouze k jeho postupnému nahrazení. Žádná taková technologie ale zatím není v dohledu.[32]

Proto, abychom s emulací uspěli, není ani nutné úplně pojmout všechny schopnosti, které daný mozek měl. Je pravděpodobné, že první emulace, kterou vytvoříme, taková ani nebude. Můžeme si představit rozdílné úrovně emulace, které tímto procesem vytvoříme. Nejvyšší je emulace s vysokým rozlišením, kde by se podařilo simulovat všechny znalosti a dovednosti žijícího mozku. Nižší představuje emulace postrádající lidské rysy, ale stále schopná řešit intelektuální problémy. Poslední by byla generická emulace, která postrádá vše z původního mozku, ale je stále schopná se učit.[33]

Dalším možným způsobem by mohlo být vylepšení našeho myšlení na biologické úrovni. Toho by šlo dosáhnout různými způsoby. Nejméně technologicky náročným by bylo selektivní rozmnožování. To ale sebou přináší mnohé morální problémy, a navíc by to trvalo několik generací, než bychom viděli významné výsledky. V současné době už existují mnohá jiná bio-lékařská řešení. Například různé drogy, které zlepšují paměť a koncentraci. Tyto způsoby však pravděpodobně neumožní vznik SI. Genetické inženýrství má v tomto ohledu daleko větší šanci toho dosáhnout. Místo toho abychom prováděli selekci na úrovni dospělých individuí, můžeme ji provádět na úrovni embryí a

---

[31] Tamtéž, s. 30–33

[32] CHALMERS J. David. The Singularity: A Philosophical Analysis. In AWRET Uziel ed. *The Singularity: Could artificial inteligence really out-think us (and would we want it to)?* [e-kniha]. 1. vyd. Exeter: Imprint Academic. 1. 12. 2016 [cit. 8. 2. 2018]. Dostupné z: <https://www.amazon.com/Singularity-Journal%20Consciousness-Studies-ebook/dp/B01N1N6KLZ/ref=mt_kindle?_encoding=UTF8&me>.

[33] BOSTROM Nick. *Superinteligence. Paths, Dangers, Strategies*. 1. vyd. Oxford: Oxford University Press, 2014. ISBN 978-0-19-967811-2. s. 33–34.

gamet.[34] Tímto způsobem bychom mohli dojít k slabší formě super inteligence. Důvodem pro to je, že nervové buňky jsou značně pomalejší než dnešní transistory. Podle Bostroma by ale toto vylepšení mohlo sloužit jako odrazový můstek, který by nám umožnil vyvinout pravou SI.[35]

Nemuseli bychom ale vylepšovat naše myšlení pouze biologickou cestou. Existuje zde představa mozkových implantátů, které by mohli propůjčit schopnosti digitálního počítače našemu mozku. Tento hybridní systém by pak následně byl lepší než nevylepšený mozek. Ale i toto řešení má mnohá úskalí. Prvním jsou možné zdravotní komplikace, které by takový implantát mohl přinést. Dále je zde problém, jak docílit dostatečně rychlého propojení mezi mozkem a počítačem, které by přinášelo výhody pro uživatele. Většinu výhod, kterou by nám tento implantát přinesl, totiž můžeme využít i bez něj, běžnou interakcí s počítači již dnes. I kdybychom dokázali vytvořit rychlý tok informací do našeho mozku, stále by to neovlivnilo rychlost, s jakou je vstřebáváme. Proto by byla potřebná úplná výměna celého mozku, tedy víceméně umělá inteligence.[36]

Doposud jsme popisovali přenos informací z počítače do mozku, jak by to ale fungovalo obráceně? Zde existuje problém, který způsobuje fakt, že mozek neukládá informace stejně jako počítač, protože nepoužívá standardizovaný systém.[37] Možné řešení u tohoto přístupu spočívá v tom, že mozek by se mohl naučit s tímto implantátem spolupracovat, takže implantát sám by nemusel být inteligentní, to by za něj obstarala plasticita mozku. Znovu se ale musíme ptát, co by to dokázalo? Proč bychom informace, které implantát posílá do mozku, nemohli jednoduše jenom promítnout na sítnici?[38]

Přes všechny tyto možné cesty k SI, můžeme podle Bostroma spíše očekávat, že SI bude mít spíše digitální povahu než biologickou. To plyne z výhod, které by digitální inteligence měla. Co se hardwaru týče, tak jsme už zmínily, že transistory jsou rychlejší než neurony, a to jak ve výpočetní síle, tak i v komunikaci mezi sebou. Mozek je také omezen v počtu výpočetních elementů, jeden omezující faktor je velikost lebky, a dále jsou zde omezení metabolismu, které zabraňují mozku mít více neuronů. Počítač nemá tyto omezení, může být postaven v jakémkoliv měřítku. Počítač má také výhodu

[34] Tamtéž, s. 36–37
[35] Tamtéž, s. 44
[36] Tamtéž, s. 44–46
[37] Tamtéž, s. 47
[38] Tamtéž, s. 48

v úložném prostoru, životnosti a spolehlivosti. V současné době mozek ještě stále vede ve výpočetní síle, ale už se začínají vyrábět počítače, které dosahují podle odhadů výkonu mozku.

Digitální mysli by také měly mnohé výhody v softwaru. Jednou by byla upravitelnost. S digitální myslí by bylo snazší experimentovat díky menší obtížnosti ve změně parametrů. Stejně tak bychom mohli upravovat i povahu dané mysli novými algoritmy nebo moduly, a tím danou mysl specializovat pro určitý úkol. Duplikovatelnost je další výhodou. Když vytvoříme jednu digitální mysl, můžeme ji duplikovat a tím vytvořit další. Z toho plyne výhoda koordinace, kdy na rozdíl od lidí, mezi kterými existují rozdíly, můžeme vytvořit skupinu digitálních myslí, které jsou identické, a tudíž budou lépe spolupracovat. Tyto mysli pak dále mohou mezi sebou sdílet své vzpomínky a dovednosti, stačí pouze kopírovat uložená data do jiné mysli.[39]

## 2.2. Formy SI

V předchozí podkapitole jsme definovali SI a cesty, které by k ní mohly vést. Super inteligence ale také může nabývat různých forem, chovat se určitým způsobem a my můžeme identifikovat tři takové formy.

První formou je rychlá SI. Tu můžeme definovat jako „Systém, který může provádět vše, co může lidský intelekt, ale o moc rychleji“. Taková super inteligence by pravděpodobně vnímala svět jinak než my, vzhledem k rychlosti její mysli by se jí zdálo, že náš svět běží zpomaleně. Z tohoto důvodu můžeme očekávat, že by preferovala pracovat s digitálními objekty ve virtuální realitě. Pokud by pracovala v naší realitě, mohli bychom ji vybavit končetinami, které by měly nano velikost, protože v tomto měřítku by se mohly hýbat stejnou rychlostí jakou ona myslí.[40]

Druhou možnou formou by byla kolektivní SI. Taková SI by se skládala z více menších intelektů, které by dohromady převýšily výkon dosavadních intelektů. S tímto druhem inteligence máme zkušenosti už teď. Dobře víme, že skupiny lidí, kteří spolupracují, jsou efektivnější než jednotlivci. Abychom ale dosáhli kolektivní super inteligence, bylo by zapotřebí daleko víc než menších vylepšení pro současné kolektivy. Mohli bychom u takové inteligence také rozlišovat, jak moc jsou dané jednotlivé intelekty

[39] Tamtéž s. 59–61
[40] Tamtéž, s. 53 (vlastní překlad)

integrovány do tohoto většího celku, na jednou konci otevřená integrace na druhém uzavřená.[41]

Poslední formou SI je kvalitní SI. To je inteligence, která myslí aspoň tak rychle jako my, ale myslí daleko chytřeji. Nejlepší přirovnání pro to, jaký by mohl být rozdíl mezi naší inteligencí a SI je, když porovnáváme naši inteligenci s tou, kterou mají zvířata. Zvířata jsou na rozdíl od nás schopna při nejlepším používat pouze základní nástroje a nejsou schopna řeči. Proč tomu tak je? Zajímavé je, že náš mozek pravděpodobně nemá výhody díky hrubé výpočetní síle, spíše to tkví v jeho architektuře, která mu umožňuje vynikat. Stejně tak si tento rozdíl můžeme představit na srovnání naší úrovně myšlení a úrovně myšlení lidí postižených. Když takto přemýšlíme o naši inteligenci, jasně vidíme, jaké schopnosti máme a jaké chybí nižším inteligencím. Kvalitní SI by tedy musela mít kognitivní schopnosti, které nám chybí, aniž o tom víme.[42]

Jak by ale takové kognitivní schopnosti mohly vypadat? Murray Shanahan má zajímavou koncepci této možné kognitivní schopnosti. Abychom pochopili, co má na mysli, nabízí nám tři úrovně kognitivního nadání: ne-reflexní, před-reflexní a reflexní. Ne-reflexní nadání můžeme vidět u mnoha zvířat a lidských dětí. Stvoření s takovým nadáním nedokážou rozpoznat rozdíl mezi tím, jaké věci jsou a jak se zdají. Před-reflexní aktéři dokážou plánovat, vytvořit si hypotézy, které následně testují a uvědomují si, že se mohou mýlit. Tito aktéři jsou před-reflexní, protože jsou schopni identifikovat, že udělali chybu, ale přestože mají i potenciál plné reflexe, tak ho nevyužívají. Reflexní aktéři dokážou přemýšlet v kartesiánském smyslu, pokud věci nejsou takové, jak se zdají, jak víme, že nejsou pouhými iluzemi. Z reflexe pak vyvstávají mnohé metafyzické problémy a spolu s nimi oddělení sebe sama od světa jako subjektu; ze světa se pak stane objekt. [43]

Reflexní aktér následně stojí před dilematem, protože na základě těchto kategorií vznikají dvě skupiny. Na jedné straně máme ty, co dávají větší ontologickou důležitost vnějším věcem před těmi vnitřními. Říkají, že vnitřní děje jsou pouze odvozené od těch vnějších, tedy že naše fenomenální zkušenosti jsou odvozené od fyziky, takže náš

[41] Tamtéž, s. 54–56
[42] Tamtéž, s. 56–57
[43] SHANAHAN Murray. Satori Before Singularity. In AWRET Uziel ed. *The Singularity: Could artificial inteligence really out-think us (and would we want it to)?* [e-kniha]. 1. vyd. Exeter: Imprint Academic, 1. 12. 2016 [cit. 8. 2. 2018]. Kapitola 2. Dostupné z: <https://www.amazon.com/Singularity-Journal%20Consciousness-Studies-ebook/dp/B01N1N6KLZ/ref=mt_kindle?_encoding=UTF8&me>.

mentální život může být zredukován na pouhou funkční strukturu. Druhá strana pak silně vnímá ontologickou rozepři mezi vnějším světem a subjektem. Tento spor trvá dodnes a Shanahan spekuluje, že tomu je možná tak právě kvůli našemu kognitivnímu omezení. Tyto spekulace opírá o představu, že nejsme vybaveni proto, abychom vyřešili problém vědomí a abychom přišli na to, jak fenomenologie vyvstává z fyziky. Problém tkví už v tom, když se ptáme, co je to vědomí, protože se pídíme po jeho metafyzické důležitosti.[44]

Shanahan pak představuje čtvrtý typ kognitivního nadání, po-reflexní. Určitou představu tohoto nadání lze rozeznat ve spisech různých filosofů, nejznáměji u pozdního Wittgensteina, dále také v nauce budhistické a taoistické, kde se jedná o ticho po konfrontaci s metafyzikou. O jakési vystoupení z metafyziky, které nastává, když uvízneme mezi dvěma ohromnými možnostmi, například jestli naše subjektivní zkušenost je něco nebo naopak nic. Pak se jako nejlepší jeví možnost vystoupit z metafyziky úplně. V tomto stavu by pak následně asi také odpadaly metafyzické bariéry vnějšího a vnitřního, subjektu a objektu. [45]

Přestože ale máme různé popisy tohoto stavu, nevíme o žádných individuích, které by tohoto stavu úplně dosáhly. To pro Shanahana není překvapivé, protože se stále opíráme o představu, že jsme se vyvinuli tak, abychom metafyzicky přemýšleli. Podle Shanahana by SI tedy mohla oplývat tímto nadáním, a to by ji kvalitativně odlišovalo od nás. Takže Shanahan si tedy představuje, že v prostoru možných myslí existují tyto nadání jako určitá nutná cesta vývoje, prostor možných myslí je konstruován tak, že je nutné těmito stavy projít. Přestože ale existuje tato cesta, je možné, že mysl SI bude radikálně odlišná od té naší.[46]

[44] Tamtéž, kap. 3.
[45] Tamtéž, kap. 4.
[46] Tamtéž, kap. 6.

## 2.3. Transhumanistická perspektiva

Poslední věc, kterou je důležité zmínit v kontextu teorie technologické singularity, by se dala nazvat perspektivou, či ideovou bází, která ji ovlivňuje. Tím je myšlena Transhumanistická perspektiva. Pokud bychom měli definovat co je transhumanismus, tak podle Nicka Bostroma můžeme dojít ke dvěma různým definicím. První definicí bychom se na transhumanismus soustředili jako na hnutí a popsali ho jako: „Intelektuální a kulturní hnutí, které prohlašuje možnost a žádostivost po fundamentálním zlepšení lidské kondice skrze aplikovaný rozum, obzvláště skrze rozvoj a zpřístupnění technologií, které zabraňují stárnutí a vysoce zlepšují lidské intelektuální, fyzické a psychologické kapacity.“[47]

Druhá definice transhumanismu, se na něj soustřeďuje jako na obor zkoumání, takže ho pak popisuje takto: „Studie důsledků, příslibů a možných nebezpečí technologií, které nám umožní překročit fundamentální lidské hranice a související studie etických věcí spojených s vývojem a využití těchto technologií.“ Obecně můžeme mluvit o transhumanismu jako o rozšíření humanismu. Humanisté věří, že záleží na lidech a individuích, nehledě na jejich nedostatky. Transhumanisté souhlasí s tímto s tímto sentimentem, ale přidávají k němu možnost změny nás samých. [48]

Transhumanistům nejde příliš o naší současnou podobu, ale spíše věří, že nás utvářejí naše ideály a aspirace. Protože očekávají, že naše rostoucí technologické možnosti nám umožní posunout se za lidské hranice. Technologie, které v tomto smyslu mají na mysli, jsou ty samé, o kterých jsme psali ve smyslu toho, že by mohly dát vznik SI. Tedy AI, uploadování, rozhraní mezi mozkem a počítačem. Ale nejen těmito i dalšími technologiemi se transhumanismus zabývá. Jsou to například molekulární nanotechnologie, genetické inženýrství, klonování a další. Přičemž jejich cílem je s pomocí těchto technologií rozšířit lidské možnosti a vyvarovat se možných nebezpečí, které by tyto technologie mohly přinést. Následně pak je cílem mnohých transhumanistů překročit lidské hranice a stát se tím, čemu říkají „post-lidé“. [49]

---

[47] BOSTROM, Nick. *Transhumanist FAQ: A General Introduction.* [online] Verze 2.1. Los Angeles: World Transhumanist Associacion (Humanity+). 2003 [cit. 9. 2. 2018] s. 4 Dostupné z: <https://nickbostrom.com/views/transhumanist.pdf>. vlastní překlad.
[48] Tamtéž, s. 4 (vlastní překlad)
[49] Tamtéž, s. 4

Pod post-lidmi si můžeme přestavit ty, kteří právě díky dříve uvedeným technologiím přejdou do vyšší úrovně. Post-lidé pak mohou nabývat různých podob, mohou to například být vylepšené uploady mysli, či geneticky nebo technologicky vylepšení lidé[50], neboli lidé 2.0, jak jim říká Ray Kurzweil[51]. V kontextu problému singularity pak můžeme mluvit o tom, že diskuze o tomto problému je ovlivněná touto perspektivou. Důkazem nám může být, že Bostrom zařazuje pod transhumanistické pojmy pojem singularity[52] a SI[53], o kterých jsme předtím psali. Také píše o Singularitarialismu, jako o odvětví transhumanistického hnutí.[54] Je však důležité podotknout, že zmiňuje existenci rozdílných názorů v tom, jak moc velkou šanci transhumanisté dávají příchodu singularity.[55] Tudíž bychom pro jistotu měli asi říct, že ne všichni proponenti singularity jsou transhumanisté, a ne všichni transhumanisté jsou proponenty singularity. Existuje však podstatné překrývání těchto dvou skupin, což pak následně ovlivňuje diskurz okolo tohoto problému.

[50] Tamtéž, s. 4–5
[51] KURWEIL Ray. *The Singularity Is Near. When Humans Transcend Biology.* 1. vyd. New York: Viking Adult, 2005. ISBN 0-670-03384-7. s. 309 - 311.
[52] BOSTROM, Nick. *Transhumanist FAQ: A General Introduction.* [online] Verze 2.1. Los Angeles: World Transhumanist Associacion (Humanity+). 2003 [cit. 9. 2. 2018] s. 19 - 20 Dostupné z: <https://nickbostrom.com/views/transhumanist.pdf>.
[53] Tamtéž, s. 12–14
[54] Tamtéž, s. 44
[55] Tamtéž, s. 20

# 3. Problémy a kritika Singularity a SI

Doposud jsme se zabývali teorií singularity a SI a předpoklady, na kterých je postavena. Z těchto předpokladů ale také vyplývá několik problémů a kritika těchto konceptů, což bude náplní této kapitoly. Nejdříve si projdeme problém singularity, která stojí na bližším prozkoumání argumentů s ní spojenými. Dále si projdeme kritiku hnutí zastánců singularity. Nakonec prozkoumáme řadu úskalí SI, které rozdělíme podle cest k ní, jak jsme si je představili v druhé kapitole.

## 3.1. Problém Singularity

V druhé kapitole jsme si představili jednotlivé školy singularity a jejich koncepce tohoto fenoménu. Tyto školy nám také předvedly hlavní argument pro příchod singularity, kterým byla inteligenční exploze I.J.Gooda. Protože představa singularity stojí na tomto argumentu, pokud bychom jej vyvrátili, zpochybnili bychom celou teorii singularity. Z tohoto důvodu si nyní musíme podrobně rozebrat tento argument.

Pro rozebrání argumentu inteligenční exploze nám nejlépe poslouží David J. Chalmers a jeho filosofická analýza singularity[56], ve které i filosoficky rozebírá hlavní argument inteligenční exploze, který ve filosofické formě podle něj vypadá takto:

1. AI bude existovat
2. Pokud bude AI, tak bude AI+
3. Pokud bude AI+, tak bude AI++

Závěr: Bude existovat AI++

Zde ale jeho analýza nekončí, Chalmers pokračuje dále. Podrobně rozepisuje argumenty pro jednotlivé premisy předchozího argumentu. Pro první premisu píše o dvou argumentech pro ni. Prvním je emulační argument, který stojí na předpokladu funkčnosti mozkových emulací. Tyto emulace by na dostatečné úrovni detailu byly simulací

[56] CHALMERS J. David. The Singularity: A Philosophical Analysis. In AWRET Uziel ed. *The Singularity: Could artificial inteligence really out-think us (and would we want it to)?* [e-kniha]. 1. vyd. Exeter: Imprint Academic. 1. 12. 2016 [cit. 8. 2. 2018]. Dostupné z: <https://www.amazon.com/Singularity-Journal%20Consciousness-Studies-ebook/dp/B01N1N6KLZ/ref=mt_kindle?_encoding=UTF8&me>.

fungování lidského mozku, tudíž by oplývaly stejnými vlastnostmi. Takový argument by podle něho vypadal takto:[57]

1. Lidský mozek je stroj
2. Budeme mít možnosti emulace tohoto stroje
3. Pokud emulujeme tento stroj, budeme mít AI

Závěr: Bude existovat AI

První premisa je odvozená z toho, co víme z biologie, která ukazuje, jak každý orgán v našem těle funguje jako stroj. Druhá premisa, obsahuje předpoklad, že lze emulovat mikro procesy ve velkém detailu, a tím emulovat jakýkoliv stroj emulací jeho mikro procesů. Třetí premisa vyplývá z tvrzení, že pokud se nám podaří zreprodukovat vzorce lidské mysli neboli mozku, tak následně můžeme vytvořit AI na základě těchto vzorců. Tomuto argumentu lze oponovat. Můžeme zpochybnit první premisu a říci, že mozek není strojem. Mnoho lidí v rámci filosofie mysli tak učinilo. Tady můžeme podle Chalmerse zmínit například osoby jako Hubert Dreyfus a Roger Penrose, kteří podle něj oba argumentují, že činnost mozku nelze simulovat na žádném počítači, či Johna Searle a Neda Blocka, kteří podle Chalmerse argumentují, že i kdybychom simulovali lidský mozek, něco fundamentálního by mu chybělo, protože by pouze simuloval naše chování.[58]

Druhý argument je jiného rázu. Jedná se o argument z evoluce, který vypadá jinak:

1. Evoluce stvořila inteligenci
2. Pokud evoluce stvořila inteligenci, tak můžeme stvořit AI

Závěr: Bude existovat AI

Představou tohoto argumentu je, že pokud proces evoluce dokázal stvořit inteligenci, proč bychom nemohli také my? Tady bychom mohli oponovat a říci, že evoluce nestvořila intelekt, nebo že ho stvořila procesy, které nelze mechanicky reprodukovat. Možná evoluce potřebovala pomoc SI činitele, nebo bylo za potřebí

[57] Tamtéž, kap. 2.
[58] Tamtéž, kap. 2.

nemechanických procesů. Tuto námitku ale lze podle Chalmerse jednoduše odbýt tím, že lehce změníme znění první premisy.[59]

Dále představuje argument, který podporuje druhou premisu hlavního argumentu. První vychází z běžně známého faktu, že informační technologie se neustále vyvíjí, tedy že současný produkt bude brzy, díky tomuto vývoji, zastaralý. Tento stejný fakt by pak tedy měl platit pro umělou inteligenci. Argument vypadá následovně:[60]

1. Pokud bude AI, tak vznikne rozšiřitelnou metodou
2. Pokud AI vznikne takovou metodou, budeme schopni metodu rozšířit
3. Rozšířením metody vznikne AI+

Závěr: Pokud bude AI, tak bude AI+

Tento argument, tedy stojí na tom, že by AI vznikla rozšířitelnou metodou. Ale ne všechny metody, jak stvořit AI, jsou rozšířitelné. Nabízí tři příklady, strojové učení, umělou evoluci a přímé programování. Tento argument by nejlépe fungoval v případě, kdy stvoříme AI skrze programování. Dobře víme, že každý program lze vylepšit. U jiných metod, jakou je například mozková emulace, možnost rozšíření není tak jasná. Stejně tak u biologických metod. Chalmersovi přesto připadá pravděpodobné, že pokud stvoříme AI, tak AI+ bude následovat. I když ji nejdříve stvoříme skrze nerozšiřitelnou metodu, tak fakt existence AI nás pak později navede k metodě rozšířitelné. Ale je tu i alternativní cesta k vytvoření větší než lidské inteligence, pro kterou není zapotřebí stvořit AI, jako například vylepšení mozku.[61]

Argument pro třetí premisu, pak víceméně následuje argumentaci I. J. Gooda. Předpokládejme, že existuje AI+ a řekněme si, že $AI_1$ je první AI+ a $AI_0$ je jeho stvořitelem, ať už je to člověk či jiná AI (která může být také na menší úrovni než člověk), a $\delta$ je rozdíl mezi nimi. Takže systém, jehož inteligence se liší od předchozího o kladné $\delta$ má vyšší inteligenci. Dále si řekněme, že v případech, kde $n > 1$ a máme $AI_{n+1}$, jehož

[59] Tamtéž, kap. 2.
[60] Tamtéž, kap. 2.
[61] Tamtéž, kap. 2.

stvořitelem je $AI_n$, tak $AI_{n+1}$ je také chytřejší než jeho stvořitel, stejně tak jak tomu bylo v případě $AI_1$. Když jsme si toto vyložili, můžeme přejít k argumentu samotnému:[62]

1. Pokud je AI+, tak je $AI_1$
2. Pro všechny $n > 0$, pokud je $AI_n$, pak bude $AI_{n+1}$
3. Pokud pro všechna n je $AI_n$, tak bude AI++

Závěr: Pokud bude AI+, tak bude AI++

Tento argument stojí na tezi, že nárůst inteligence se rovná stejnému nárůstu v kvalitě designu další inteligence. Pokud by se tato teze ukázala být nepravdivou, tak by to ohrozilo sílu tohoto argumentu. Kritici by také mohli poukázat na to, co že máme přesně na mysli, když mluvíme o inteligenci a zpochybnit náš standard měření. Mimo jiné, všechny tyto argumenty stojí na tom, že inteligenci lze měřit, a že se jedná o existující veličinu. Ale jsou tu i tací, kteří této představě oponují. Také by se dalo najít několik důvodů, proč by se tento argument mohl ukázat jako nepravdivý. Jako příklad bychom mohli uvést představu jakéhosi inteligenčního stropu, limitu, kterého inteligence může dosáhnout. Může to například být naše úroveň inteligence, nebo možná i vyšší. Pokud takový strop je, tak by na něj AI narazila. Jinou námitku, kterou bychom mohli podle Chalmerse podat, je říci, že nárůst inteligence by měl snižující se výnosy, první nárůst 10% by pak mohl následovat pouze nárůst 5% a po něm pak jen nárůst 2,5% a tak dále.[63]

Jedna věc, kterou je také nutné zmínit je, že všechny tyto Chalmersovy argumenty stojí na dvou zásadních podmínkách, aby byly pro nás relevantní. První je, že se tak stane zanedlouho, tedy zanedlouho vyvineme AI, zanedlouho vznikne AI+ a tak dále. Tímto „zanedlouho“ má Chalmers na mysli za zhruba století. Druhou podmínkou je, že těmto argumentům nebude stát v cestě žádná překážka.[64] Jsou to právě tyto možné překážky, kterými se Chalmers následně zabývá a rozděluje je do tří kategorií: strukturální překážky, souvztažné překážky a manifestační překážky.[65]

Strukturální překážky jsou nejvíce přímočaré, už jsme je i zmínili. Mohl by to například být inteligenční strop, či problém snižujících se výnosů. V této kategorii je také

[62] Tamtéž, kap. 2.
[63] Tamtéž, kap. 2.
[64] Tamtéž, kap. 2.
[65] Tamtéž, kap. 4.

selhání odstartování. Je možné, že lidská úroveň inteligence není dostatečná pro to, aby dokázala vytvořit více inteligentní systém, prostě na to nemá kapacitu.

Souvztažné překážky jsou spojeny s představou, že inteligence, jako nějaká obecná veličina naší schopnosti řešit problémy neexistuje, a jedná se spíše o soustavu různých schopností řešit problémy. V tomto případě by se například mohlo stát, že se nám podaří zvýšit schopnost AI v oblasti programování, ale to nebude mít žádný vliv na ostatní její schopnosti, které s tou zvýšenou schopností nemají žádné spojení. Takže bychom vytvořili jen částečný SI systém.[66]

Manifestační překážky, mají své jméno od skutečnosti, že pro to, aby singularita nastala, musíme manifestovat naši kapacitu; takříkajíc ji nastolit. Tyto překážky by nám pak stály v cestě. Chalmerse je rozděluje do motivačních překážek, které nám znemožňují mít motivaci takto učinit, a situačních překážek, kde nám daná situace znemožňuje manifestaci. Situační překážky by tedy byly události jako katastrofy, nebo období nedostatku zdrojů. Takové situace mohou být vyvolány jak lidmi, tak je i možné že by je vyvolaly umělé inteligence samy. Motivační překážky zahrnují jak nechuť, která by mohla být spojená s vývojem AI, tak i aktivní prevenci tohoto vývoje, skrze legální, finanční či vojenské prostředky. Také je možné, že by AI+, kterou vyvineme, pak nechtěla vyvinout AI++, ať už z důvodu, že jsme ji takto naprogramovali, či z vlastní vůle.[67]

Chalmers sám se příliš neobává strukturálních ani souvztažných překážek. Podle něj tyto problémy rozhodně nejsou triviální, ale i kdyby byly pravdivé, tak stále tu je *prima facie* možnost exploze nějakých kvalit, když už ne inteligence. Nejzásadnějšími překážkami mu tedy připadají ty manifestační, hlavně ty motivačního charakteru.[68]

Tato argumentace není jedinou náplní Chalrmesova článku, který mimo jiné obsahuje i debatu o Uploadování, to však necháme stranou pro jinou kapitolu.[69] Tento článek pochopitelně také vyvolal velkou reakci, která sebou přinesla odezvy a kritiku Chalmersovy argumentace.

Jednu takovou kritiku přináší Drew McDermott, který souhlasí s Charmesovou argumentací do jednoho bodu, specificky do třetí premisy, tedy: Pokud bude AI+, tak bude AI++. McDermott nepochybuje, že vyvineme AI, ani to, že AI+ bude vytvořena

[66]Tamtéž, kap. 4.
[67] Tamtéž, kap. 4.
[68] Tamtéž, kap. 4.
[69] Tamtéž, kap. 9. – 10.

brzy potom. Nesouhlasí ale, že pak bude následovat AI++. Důvodem pro to je, že nepochybuje o existenci inteligence překračující tu lidskou, ale pochybuje o tom, jestli existuje inteligence, která dalece překračuje tu lidskou. Tento předpoklad, že existuje daleko vyšší inteligence, podle něj není v Chalmersově argumentaci nijak podložen.[70]

Charmersův argument má podle něj dva nedostatky. Prvním je, že by šel vyložit daleko jednodušeji. Druhým a podstatnějším je, že tento argument není dostatečné podpořen. Jednu věc, o kterou se argument trochu podle McDermotta opírá, je představa rozšířitelné metody, o které Chalmers mluvil a udal i příklady této metody. Podle McDermotta ale tyto příklady nejsou metody, které jsou rozšířitelné. Vezmeme si například přímé programování, jak rozšiřitelné je programování? Můžeme například hledat chyby v programu, či ho přenést na rychlejší hardware. Tyto metody ale nejsou podle McDermotta rozšířitelné do nekonečna. Strojové učení naráží na stejný problém, dokážeme si představit algoritmus, který můžeme rozšiřovat do nekonečna? Takové algoritmy lze pouze vylepšit novými nápady, ale čekat na nový nápad není rozšířitelnou metodou, ani metodou vůbec. To samé platí pro umělou evoluci.[71]

Další námitku vznáší Jesse Prinz a jeho argument je následující. Pokud bychom přijmuli Chamersovu argumentaci, že bude AI++, tak pak bychom také nutně museli přijmout, že existuje velká šance, že v současné chvíli žijeme v simulaci. V podstatě, pokud přijmeme možnost existence AI++, tak bychom mohli říci, že během našeho honu za AI++ bychom se naučili vytvořit simulace nás samých, s jejichž pomocí bychom i zachovali sami sebe. Takže pokud bychom mohli vytvořit simulace nás samých, tak je možné, že my v tuto chvíli jsme těmi kopiemi. Důvod pro to je ten, že AI++ by asi byla schopná takové simulace vytvářet, a ty by jí byli velice užitečné pro optimální rozhodování. AI++ by ale asi nevytvořila pouze jednu simulaci, ale více, každou jinou. Což zvyšuje šanci, že náš svět je simulace. Možná pravdivost Chalmersovy argumentace, ale přináší další problém. Pokud je singularita nevyhnutelná a my k ní míříme, tak

[70]MCDERMOTT Drew. Response to ‚The Singularity' by David Chalmers In AWRET Uziel ed. *The Singularity: Could artificial inteligence really out-think us (and would we want it to)?* [e-kniha]. 1. vyd. Exeter: Imprint Academic. 1. 12. 2016 [cit. 8. 2. 2018]. Dostupné z: <https://www.amazon.com/Singularity-Journal%20Consciousness-Studies-ebook/dp/B01N1N6KLZ/ref=mt_kindle?_encoding=UTF8&me>.

[71] Tamtéž, Kap. Prospects

v našem simulovaném světě asi nenastane, protože by ohrožovalo tu AI++, která náš svět vytvořila. Takže AI++ by byla nucena zasáhnout, tím by náš svět asi čekala zkáza.[72]

Podle Prinze se ale nemáme čeho bát, protože Chalmersova argumentace má podle něj v sobě chyby. Jednu chybu vidí v tom, že v argumentaci není žádná přímá definice inteligence. Když tedy tato definice chybí, má smysl mluvit o jejím postupném růstu, když nevíme, co roste? Největší důraz ale Prinz klade na možné překážky a říká, že příchod singularity je méně pravděpodobný než překážky, které jí stojí v cestě. Myslí si, že spíše nás postihne nějaká katastrofa, jako například globální hladomor, či nukleární válka, než že přijde singularita. Chybí mu tedy důkaz, který by ukázal větší pravděpodobnost singularity než těchto možných událostí.[73]

Arkady Plotnistky je další, který reaguje přímo na Chalmersovu argumentaci. První věcí, kterou dělá je, že to, co Chalmers provádí, přirovnává k jakési sázce. Chalmers podle něj sází na příchod singularity, kterou přirovnává ke slavné Pascalově sázce. Tato sázka ale nezačíná u Chalmerse, neboť se přenáší k Chalmersovi přes I. J. Gooda. Good také koneckonců sázel na příchod singularity. To, že jméno jeho článku obsahuje slovo spekulace, nám slouží jako nápověda. Chalmers podle něj dává příchodu AI, a možná i singularitě samotné, šanci 50 % nebo možná i více. Právě kvůli této povaze sázky, kterou Chalmersova analýza má, si Plotnistky myslí, že nebude mít velký dopad na názory skeptiků.[74]

Po tomto úvodu se Plotnitsky pouští přímo do Chalmersových argumentů. Nejdříve analyzuje Chalmersův emulační argument, kde jemu připadá zvláštní druhá premisa, tedy: „Budeme mít možnost emulace tohoto stroje“. Plotniskému jde o nezměrnou komplexitu toho daného „stroje“, tedy mozku. A je to právě tato otázka komplexity, jak lidského mozku, tak i evolučního procesu, který ho stvořil, který

[72] PRINZ Jesse. Singularity and Inevitable Doom. In AWRET Uziel ed. *The Singularity: Could artificial inteligence really out-think us (and would we want it to)?* [e-kniha]. 1. vyd. Exeter: Imprint Academic. 1. 12. 2016 [cit. 8. 2. 2018]. Kapitoly 2. – 3. Dostupné z: <https://www.amazon.com/Singularity-Journal%20Consciousness-Studies-ebook/dp/B01N1N6KLZ/ref=mt_kindle?_encoding=UTF8&me>.
[73] Tamtéž, Kap. 4.– 5.
[74] PLOTNISKY Arkady. The Singularity Wager: A Response To David Chalmerse. In AWRET Uziel ed. *The Singularity: Could artificial inteligence really out-think us (and would we want it to)?* [e-kniha]. 1. vyd. Exeter: Imprint Academic. 1. 12. 2016 [cit. 8. 2. 2018]. Dostupné z: <https://www.amazon.com/Singularity-Journal%20Consciousness-Studies-ebook/dp/B01N1N6KLZ/ref=mt_kindle?_encoding=UTF8&me>.

Plotnitskému připadá jako slepé místo Chalmersovy analýzy. Plotnitsky konstatuje, že toto je jeho hlavní problém.[75]

Vezměme si Chalmersovo tvrzení, že během století bychom měli být schopni stvořit inteligenci, čin, který evoluci trval miliony let. Plotnitsky si je vědom toho, čeho jsou počítače v současné době schopny, nemyslí si ale, že na základě těchto skutečností můžeme očekávat, že v dohledné době můžeme od nich čekat podobnou komplexitu jako u lidského mozku. To neznamená, že nevěří, že existují jiné stejně komplexní systémy, vzhledem k velikosti naší galaxie, nebo vesmíru vůbec. Stejně tak nezpochybňuje možnost vzniku AI, on jenom „sází" na tuto možnost daleko méně optimisticky než Chalmers.[76]

Právě tímto problémem komplexity, se zabývají Nick Bostrom a Carl Shulman, ve své reakci na Chalmersův článek. Zabývají se dopodrobna evolučním argumentem, který Chalmers předkládá. Důležité je, že Chalmers není prvním, kde tento argument použil v tomto kontextu, a stejně tak to není jediná verze tohoto argumentu. Autoři prezentují dvě další verze tohoto argumentu, první verzí je argument o obtížnosti problému:[77]

> 1. Evoluce stvořila lidský intelekt
>
> 2. Pokud evoluce stvořila lidský intelekt, pak je netěžké pro evoluční procesy stvořit lidský intelekt
>
> 3. Pokud je netěžké pro evoluční procesy „stvořit" lidský intelekt, tak není příliš těžké pro inženýry stvořit strojovou inteligenci na lidské úrovni.
>
> 4. Pokud není příliš těžké pro lidské inženýry stvořit strojovou inteligence, tak se tak stane zanedlouho.
>
> Závěr: Inženýři (zanedlouho) stvoří strojovou inteligenci na lidské úrovni.

U tohoto argumentu si můžeme představit, že mluví o prostoru možných algoritmů pro AI. Hlavní tvrzení tohoto argumentu je, že v tomto prostoru možných algoritmů pak

[75] Tamtéž
[76] Tamtéž
[77] BOSTROM Nick, SHULMAN Carl. How Hard is Artificial Inteligence? Evolutionary Arguments and Selection Effects. In AWRET Uziel ed. *The Singularity: Could artificial inteligence really out-think us (and would we want it to)?* [e-kniha]. 1. vyd. Exeter: Imprint Academic. 1. 12. 2016 [cit. 8. 2. 2018]. Kapitola 1. Dostupné z: <https://www.amazon.com/Singularity-Journal%20Consciousness-Studies-ebook/dp/B01N1N6KLZ/ref=mt_kindle?_encoding=UTF8&me>.

nebude těžké najít ten správný pro naše potřeby. Podle autorů ale premisa druhá a třetí potřebují bližší prozkoumání. Vezměme si třetí, proč bychom měli věřit, že stvořit intelekt nebude těžké pro lidské inženýry, když to nebylo těžké pro evoluci? Jedním důvodem by mohl být prostý optimismus, co se týče naší schopnosti řešit problémy. Pokud jsme ale optimističtí v tomto ohledu, proč musíme definovat tento problém jako netěžký? Pro autory je také důležité zmínit, že tato premisa je klíčová pro tento argument. Druhá verze evolučního argumentu stojí na evolučních algoritmech:[78]

> 1. Evoluce stvořila lidský intelekt
>
> 2. Pokud evoluce stvořila lidský intelekt, pak je netěžké pro evoluční procesy stvořit lidský intelekt
>
> 3. Budeme schopni (zanedlouho) spustit genetický algoritmus na počítačích, které jsou dostatečně rychlé proto, aby simulovaly v lidském měřítku stejné množství kumulativní optimalizační síly, kterou relevantní evoluční procesy používaly napříč naší evoluční minulostí (pro každý evoluční proces, který je netěžký)
>
> Závěr: Budeme schopni (zanedlouho) dosáhnout za pomocí genetických algoritmů výsledků srovnatelných s výsledky evoluce, včetně systémů na lidské úrovni intelektu.

Tento argument sdílí s tím prvním argumentem, první a druhou premisu. Nás tedy bude blízce zajímat právě druhá premisa, protože na ní stojí oba argumenty. To, co nás tedy zajímá je, jak můžeme definovat složitost evolučních procesů? Dalo by se říct, že tu máme dva rozdílné scénáře, na jedné straně máme představu netěžké inteligence, kde se mluví o jasné čáře postupného vylepšování od jednoduchých nervových systémů po lidský mozek. Takže vzhledem k tomu, že existují snadno nalezitelné jednoduché algoritmy inteligence, tak můžeme očekávat, že na každé planetě se správnými podmínkami se vyvine inteligence na úrovni člověka.[79]

Na druhé straně máme těžkou inteligenci. Zde nám jde o to, že bez té jasné čáry jsou funkční algoritmy inteligence vzácné. Pro jejich objevení evoluce potřebovala neuvěřitelné štěstí, takže pak se inteligence na lidské úrovni objeví pouze na jedné z $10^{1000}$ planet. Tohle jsou tedy dvě tvrzení, která musíme brát v potaz, co se týče evolučního

[78] Tamtéž, kap. 1.
[79] Tamtéž, kap. 1.

argumentu. Jak se tedy můžeme spoléhat na pravdivost prvního tvrzení, která zaručuje pravdivost evolučního argumentu? Jedinou věc, která nahrává prvnímu tvrzení, je fakt, že my jsme se vyvinuli na Zemi, takže bychom měli za pomoci tohoto faktu, nějak určit pravdivost jednotlivých tvrzení.[80]

V této chvíli nám tak autoři nabízí příklad, který by nám mohl napomoci k řešení: „Božský hod mincí: Předpokládejme, že si Bůh hází mincí. Pokud padne hlava, tak stvoří deset lidí, každého se svou vlastní místností. Pokud padne orel, tak stvoří tisíc lidí, každého se svou vlastní místnosti. Takže místnosti jsou očíslované 1-10 nebo 1-1000. Lidé se navzájem nevidí a nemohou komunikovat s ostatními místnostmi. Předpokládejme, že vy tohle všechno víte a zároveň jste objevili, že vy jste v prvních deseti místnostech. Jak tedy budete usuzovat, jaká mince padla?“ První metoda, jakou bychom mohli tento problém řešit, se jmenuje SSA (Self-Sampling Assumption). Podle ní bychom měli sebe sama brát jako náhodný vzorek z celé skupiny pozorovatelů stejné třídy. Pokud použijeme tento postup, tak dojdeme k závěru, že s pravděpodobností 100/101 padla hlava.[81]

Druhý postup SIA (Self-identification Assumption) přidává k tomu prvnímu další princip. Musíme brát v potaz, že my jako pozorovatel existujeme, tudíž bychom se měli spíše klonit k hypotézám, podle kterých existuje mnoho pozorovatelů. Když tedy použijeme kombinaci obou přístupů, dojdeme k jinému závěru. Předtím, než známe pozici naší místnosti, tak očekáváme, že orel je 100x pravděpodobnější, protože v takové situaci je o stejný poměr větší počet pozorovatelů. Ovšem jakmile známe naše umístění, tak orel i hlava mají stejnou pravděpodobnost.[82]

Když pak tyto přístupy obrátíme k našemu problému evolučního argumentu, tak samotné SSA nás nenavádí k závěru netěžké, ale ani k těžké inteligenci. Protože pokud máme třídu „lidé“ či „vyvinuté lidské bytosti“, tak nehledě na daný scénář všichni pozorovatelé už vznikli. Pokud ale aplikujeme SIA, tak následně má scénář netěžké inteligence o $10^{1000}$ větší šanci pro pravdivost než scénář těžké inteligence. Protože pak existuje také o $10^{1000}$ více pozorovatelů neodlišitelných od nás. Když si ale vezmeme, že inteligenci trvalo poměrně dlouhou dobu, než se vyvinula, tak nás to spíše navádí k tomu si myslet, že inteligence je středně těžká záležitost. Obzvláště, když si vezmeme fakt, že

[80] Tamtéž, kap. 1.
[81] Tamtéž, kap. 3. (vlastní překlad)
[82] Tamtéž, kap. 3.

jsme ještě nepotkali žádnou mimozemskou civilizaci, tak nás to také navádí k obtížnosti inteligence.[83]

Ve shrnutí tedy autoři říkají, že argument z evolučních algoritmů potřebuje definovat, jak mnoho výpočetní síly je zapotřebí pro simulaci optimalizační síly evoluce. Zatím se argument o obtížnosti tomuto kvantitativnímu problému vyhýbá a spíše na nás intuitivně apeluje. Bez ohledu na danou verzi stojí evoluční argument na představě toho, že evoluce lidské inteligence nebyla příliš těžká. Problém tkví v tom, že tuto představu nelze čistě inferovat z existence lidstva na Zemi. Co se týče přístupů, které autoři na tento problém aplikovali, tak SIA silně nahrává evolučnímu argumentu a možnosti existence AI, zatímco u SSA silně záleží na daných empiristických faktech, jinak nenahrává žádnému ze dvou závěrů.[84]

Posledním kritikem, kterého představíme je Selmer Bringsjord, který tvrdí že víra v singularitu je logicky křehká. První věc, kterou je nutné podle něj vysvětlit je, co přesně termíny, které Chalmers používá, znamenají, abychom pak pochopili Chalmersův argument. Má tím na mysli AI, AI+ a AI++, tyto termíny pro něj designují třídu strojů, kde je klíčové, že každá třída má určitou úroveň inteligence a každá třída je zodpovědná za vznik třídy další. Pak k Chalmersovým termínům přidává ještě termín HI, který popisuje lidskou inteligenci.

Chalmersův argument tedy nahrazuje řetězcem HI → AI → AI+ → AI++.[85] Bringsjord se ale dále ptá, co tohle všechno ve skutečnosti znamená. Protože podle něj Chalmers opomenul matematicky definovat, co přesně je stroj a co přesně je inteligence. V tomto ohledu nám Chalmers nedává žádnou odpověď. Takže se následně Bringsjord opírá o formální logiku a počítačovou teorii, když říká, že asi v tomto ohledu mluvíme o strojích, které zpracovávají informace. Jejich inteligenci pak můžeme formalizovat s ohledem k jejich schopnosti počítat určité číselně-teoretické funkce. Úroveň inteligence dané třídy pak lze soudit na základě toho, s jakou třídou funkcí dokáží pracovat. Bringsjord tedy pak představuje tři takové třídy.

---

[83] Tamtéž, kap. 3.– 4.
[84] Tamtéž, kap. 7.
[85]BRINGSJORD Selmer. Belief in Singularity is Logically Brittle In AWRET Uziel ed. *The Singularity: Could artificial inteligence really out-think us (and would we want it to)?* [e-kniha]. 1. vyd. Exeter: Imprint Academic. 1. 12. 2016 [cit. 8. 2. 2018]. Dostupné z: <https://www.amazon.com/Singularity-Journal%20Consciousness-Studies-ebook/dp/B01N1N6KLZ/ref=mt_kindle?_encoding=UTF8&me>.

První třídou je $M_1$, pod kterou si můžeme představit například kalkulačky, nebo jiné čistě mechanické stroje, které řeší problémy (angl. push-down automata). Druhá třída $M_2$ je pak třída standartních Turingových strojů. Poslední třída $M_3$ se pak skládá z Turingových strojů, které běží nekonečně dlouho. V této terminologii tedy také můžeme mít stroje, které vytváří další stroje. Pak bychom tento fakt vyjádřili tak, že na začátku máme $M_i$, které vytvoří $M_j$, přičemž $i < j$, takže pozdější třída je více inteligentní, než ta předchozí, čímž máme na mysli, že pozdější třída operuje s funkcemi, se kterými ta předchozí operovat nemohla.[86]

Abychom ale pochopili jeho další argumentaci, musíme v krátkosti vysvětlit dvě věci. Zaprvé musíme definovat co je „problém zastavení". Jeho podstatou je toto: máme popis libovolného programu a popis jednoho jeho vstupu, a máme z těchto popisů určit, jestli se program zastaví nebo jestli poběží navždy (zacyklí se). Alan Turing byl prvním, kdo definoval tento problém, a také tím, kdo ukázal, že neexistuje žádný obecný algoritmus, který by tento problém vyřešil pro všechny možné vstupy. Jde tedy o neřešitelný problém.[87]

Zadruhé musíme definovat co má na mysli, když mluví o Turingově stroji, který běží do nekonečna. Autoři tohoto teoretického modelu jsou Joel David Hampkins a Andy Lewis. Tento Turingův stroj má pro vyřešení jakéhokoliv problému nekonečný čas, rovněž je nekonečně rychlý. To mu umožnuje plnit takzvané super úkoly (angl. supertasks), což jsou problémy, pro jejíž vyřešení je zapotřebí nekonečné množství kroků. V naší současné situaci je tento typ Turingova stroje pro nás důležitý, protože podle autorů je schopný vyřešit problém zastavení. Důvodem je, že v případě tohoto stroje je problém zastavení spočitatelný, a tudíž i řešitelný. Tato spočítatelnost tkví v tom, že v tomto případě operujeme s trans nekonečnými ordinálními jednotkami času, v podstatě s určitými úseky nekonečného času.[88]

---

[86] Tamtéž

[87] TURING, Alan. On Computable Numbers, with an Application to the Entscheidungsproblem. In *Proceedings of the London Mathematical Society* [online]. London: London Mathematical Society. 23.12 2016, Roč. s2-42, č. 1., s. 230–265 [cit. 8. 2. 2018]. ISSN 1460-244X. Dostupné z: <https://doi.org/10.1112/plms/s2-42.1.230>.

[88] HAMKINS D. Joel, LEWIS Andy. *Infinite Time Turing Machines.* [online]. Ithaca (NY): Cornell University Library. 21. 8. 1998. Aktualizováno 1. 2. 2008. s. 1 - 3 [cit. 28. 3. 2018]. Dostupné z: <https://arxiv.org/pdf/math/9808093.pdf>.

Představme si, že se na tomto stroji spustil program, který následně provádí nekonečné množství kroků, aniž by se zastavil. Dále si představme, že jsme schopni zachytit snímek stavu tohoto stroje v určitém momentu, tedy jaký program v dané chvíli používá, status a pozici záznamové hlavy stroje a úplný obsah zapisovacích pásek stroje. Vezměme si tedy stav stroje ω, který nám bude reprezentovat určitý časový okamžik vykonávání daného programu, ze kterého také budeme mít snímek. Podle autorů pak můžeme stroj dále pozorovat, přičemž naším cílem je zjistit, jestli se snímek ze stadia ω opakuje v jiných stadiích po zvolených časových úsecích.[89]

Pokud se tento snímek vskutku začne opakovat v určitých intervalech, tak pak podle autorů můžeme říci, že se program dostal do nekonečné smyčky a nikdy se nezastaví. Pokud se však snímek opakuje v různých intervalech, je zde možnost, že se ze smyčky vymaní a pak se může buď zastavit, nebo začít řešit něco jiného. Podle autorů můžeme takto také zjistit, že se program zastaví, nebo naopak uvízne ve smyčce, v spočitatelném množství dob.[90]

Když se pak vrátíme k původnímu textu tak to, co pak Bringsjord konstatuje je velice důležité. Píše, že HI stojí na úrovni $M_2$, což se dále promítá do AI, které následně stojí na nebo pod úrovní $M_2$. Z toho pak plynou dvě propozice, na kterých stojí další argumentace. [91]

> Propozice 1 - AI je na úrovni $M_2$ nebo níže.
>
> Propozice 2 - Všechny Procesy HI pro vytváření AI jsou na úrovni $M_2$ nebo níže

Tyto propozice vychází z faktu, o kterém jsme psali předtím. Jak víme, tak Turingův stroj s nekonečným časem dokáže vyřešit problém zastavení, ale my lidé ani žádná AI, kterou jsme stvořili, tento problém zatím nevyřešili, z této skutečnosti pak vyplývají tyto propozice. Podle autora propozice také podporuje fakt, že všechny cesty k AI, které Chalmers zmiňuje ve své analýze (přímé programování, emulace mozku, umělá evoluce, strojové učení), by podle Bringsjorda šly formálně vyjádřit skrze sérii

[89] Tamtéž s. 6–7
[90] Tamtéž, s. 7
[91] BRINGSJORD Selmer. Belief in Singularity is Logically Brittle In AWRET Uziel ed. *The Singularity: Could artificial inteligence really out-think us (and would we want it to)?* [e-kniha]. 1. vyd. Exeter: Imprint Academic. 1. 12. 2016 [cit. 8. 2. 2018]. Dostupné z: <https://www.amazon.com/Singularity-Journal%20Consciousness-Studies-ebook/dp/B01N1N6KLZ/ref=mt_kindle?_encoding=UTF8&me>.

algoritmů. Kdybychom tak učinili, všechny by pak spadaly do třídy $M_2$. Po ustanovení těchto propozic pak Bringsjord může ukázat dvě různé cesty k tomu, aby nám předvedl, jak je víra v singularitu logicky křehká. První cesta podle něj nepotřebuje propozice, za to ta druhá je využívá společně. První cesta, nebo jak tento argument nazývá on, Teorém 1 vypadá takto:[92]

> Teorém 1 – $M_2$ nutně nemůže stvořit $M_3$
>
> Důkaz – Předpokládejme, že náš cíl je mylný, tedy že $M_2$ může stvořit $M_3$. Dobře víme, že žádný $M_2$ stroj nedokáže vyřešit problém zastavení. Také víme, že $M_3$ dokáže vyřešit tento problém. Takže pak musí existovat $M_2$ stroj, který dokáže vyřešit problém zastavení, což je absurdní, takže jsme nepřímo hotovi.

Tento teorém přímo utočí na premisy 2 a 3 Chalmersova hlavního argumentu (pokud bude AI bude AI+ atd.) protože obě tyto premisy tvrdí, že nižší třída muže vytvořit tu vyšší, čímž jsou v rozporu s Teorémem 1. Druhá cesta, Teorém 2, jak už bylo řečeno, stojí na dvou propozicích předtím zmíněných:[93]

> Teorém 2 – Pokud HI = $M_p$, kde $p > 2$, tak vzhledem k propozici 1 nebude nikdy takové AI, aby AI = HI, tedy Premisa 1 není pravdivá (AI bude existovat)
>
> Důkaz – Předpokládejme, že předchozí je pravdivé, a vzhledem k propozici 1 si vzpomeňme, že AI je $M_2$ nebo níže a pro *reductio* předpokládejme, že bude AI, jenž bude AI = HI. Vzhledem k těmto předpokladům tak existuje standardní Turingův stroj, který dokáže vyřešit problém zastavení (kvůli identifikaci s HI). Tohle je ale absurdní, takže jsme znovu hotovi.

Takto Bringsjord vyvrací první premisu Chalmersova hlavního argumentu. Znamená to ale, že tento argument je neplatný také v jiných formálních rámcích? Bringsjord je ochoten připustit, že by tento argument mohl fungovat v jiných rámcích, ale má o tom své pochyby. [94]

---

[92] Tamtéž

[93] BRINGSJORD Selmer. Belief in Singularity is Logically Brittle In AWRET Uziel ed. *The Singularity: Could artificial inteligence really out-think us (and would we want it to)?* [e-kniha]. 1. vyd. Exeter: Imprint Academic. 1. 12. 2016 [cit. 8. 2. 2018]. Dostupné z: <https://www.amazon.com/Singularity-Journal%20Consciousness-Studies-ebook/dp/B01N1N6KLZ/ref=mt_kindle?_encoding=UTF8&me>.

[94] Tamtéž

Chalmers je si těchto odpovědí vědom a také na ně reaguje. Na McDermottovu námitku, že rozšiřitelné metody nelze rozšiřovat do nekonečna, odpovídá, že s ním souhlasí. Chalmersova představa je ale taková, že nás k AI++ nedostane pouze jedna metoda, a že jak AI systémy budou stále více inteligentní, tak se objeví i další rozšiřitelné metody. I kdyby McDermott měl pravdu, tak by ani to podle Chalmerse neohrozilo možnost singularity a tady naráží na formy SI, o kterých jsme mluvili. Pokud tedy budeme schopni sestrojit pouze AI+, tak je prostě spojíme do kolektivní SI, či je spustíme na rychlém hardwaru jako rychlou SI, to by samo osobě podle něj stačilo pro singularitu.[95]

Na Prinze reaguje tím, že přestože on sám soucítí se simulačním argumentem, tak Prinze ve svém použití tohoto argumentu nechal pár děr. Pokud tedy máme nějakou SI, která nechává běžet několik simulací, můžeme očekávat, že některé simulace nejsou aktivně pozorovány, tudíž že běží v pozadí. Takže je možné, že když v našem světě nastane singularita, tak by si toho ta daná SI nemusela všimnout. Také můžeme očekávat, že SI podnikla opatření, aby ta daná simulace byla uzavřená a nic neproniklo ven, v takovém případě by pak SI neměla důvod naši simulaci ukončit. Co se týče Prinzova názoru, že nás dříve postihne katastrofa, než že nastane singularita, tak Chalmers se nekloní ani jedním směrem, stačí mu jistota, že pokud nebude překážek, tak singularita nastane.[96]

Při reakci na Plotniského, nejdříve upozorňuje, že svou podmínkou *zanedlouho* měl na mysli za několik století, přičemž mu nepřijde příliš možné, že budeme schopni emulovat mozek během několika desetiletí. Na problém komplexity odpovídá, že je možné, že mozek sám nemusí být stejně komplexní jako celý evoluční proces, který mu dal vznik. Také pokud bychom měli simulovat evoluci, abychom následně dali vzniknout inteligenci, tak bychom podle Chalmerse možná nemuseli emulovat celý proces evoluce. Otázkou podle něj je, jestli proces vzniku inteligence není natolik složitý, že nemůžeme nijak doufat, že se nám ho povede emulovat. Konstatuje ale, že žádný argument, který by

[95]CHALMERS J. David. The Singularity: A Reply to Commentators In AWRET Uziel ed. *The Singularity: Could artificial inteligence really out-think us (and would we want it to)?* [e-kniha]. 1. vyd. Exeter: Imprint Academic. 1. 12. 2016 [cit. 8. 2. 2018]. Kapitola 2. Dostupné z: <https://www.amazon.com/Singularity-Journal%20Consciousness-Studies-ebook/dp/B01N1N6KLZ/ref=mt_kindle?_encoding=UTF8&me>.

[96] Tamtéž, kap. 2.

tuto námitku absolutně zpochybnil, nemá. Co se týče Bostroma a Shulmana, tak říká, že sám nemá co příliš dodat, pouze, že sám soucítí se SIA přístupem.[97]

Jeho reakce na Brinsjordův první protiargument je taková, že poukazuje na skutečnost, jaké všechny různé bytosti spadají pod $M_2$, protože v této třídě nejsme pouze my, ale také například myši. Protože Bringsjord nikde netvrdí, že lidé představují limit inteligence této třídy, tak podle Chalmerse si můžeme představit, že by AI+ nebo AI++ mohla mít větší inteligenci než my, a přesto spadat pouze do třídy $M_2$. Jeho odpověď na druhý protiargument je poněkud delší. Chalmers říká, že pokud jsou $M_3$ třídy možné v přírodě, tak mu nepřijde nemožné, že by $M_2$ AI nemohla takové procesy také vytvořit. Výpočetní teorie tvrdí, že pouze určitý typ vytváření je možný, pomocí vnitřních procesů AI a digitálních vstupů a výstupů. Pokud ale AI je schopná manipulovat vnější nevýpočetní procesy v přírodě, tak podle Chalmerse jí nic nebrání v tom sestrojit super Turingův stroj.[98]

Dále si Chalmers nemyslí, že lidé jsou ve třídě $M_3$, ale i kdyby byli v této třídě, tak si Chalmers nemyslí, že by AI bylo omezeno na $M_2$ třídu. Protože pokud my jsme přírodními systémy, tak můžeme předpokládat, že nevýpočetní procesy jsou součástí naší inteligence, a tyto samé procesy bychom pak mohli využít pro umělé systémy, čímž bychom vynesli AI do třídy $M_3$. Podle Chalmerse Bringsjordův argument stojí na implicitní premise, že naše zdroje pro sestrojení AI jsou limitovány na úroveň $M_2$, pokud ale my jsme na $M_3$, tak nemáme důvod si toto myslet.[99]

## 3.2. Kritika hnutí Singularity

Takže jsme si definovali problém singularity a rozebrali argumenty s tímto problémem spojené. Chalmers na samém začátku své filosofické analýzy ale zmiňuje jednu skutečnost, a to, že problematika singularity není příliš probíraná v akademických kruzích, přičemž Chalmers si myslí, že za tímto faktem by mohla být spekulativní povaha této problematiky.[100] Také by ale důvodem pro tuto skutečnost mohla být další forma

[97] Tamtéž, kap. 2.
[98] Tamtéž, kap. 2.
[99] Tamtéž, kap. 2.
[100] CHALMERS J. David. The Singularity: A Philosophical Analysis. In AWRET Uziel ed. *The Singularity: Could artificial inteligence really out-think us (and would we want it to)?* [e-kniha] 1. vyd. Exeter: Imprint Academic. 1. 12. 2016 [cit. 8. 2. 2018] Kapitola 1. Dostupné z:

kritiky, kterou problematika Singularity přináší, a to je kritika zastánců singularity, nebo jak je nazývá Kurzweil, kritika Singularitiánů[101]. Práce Diane Proudfootové „Software Immortals: Science or Faith?“ nám může být velice dobrým příkladem této kritiky.[102]

Proudfootová v této své práci obhajuje tezi, že hnutí singularity je novým náboženstvím, které nazývá techno-supernaturalismem. Toto nové náboženství je pak syntézou mezi vědou a starým náboženstvím. Aby tuto tezi dokázala, tak prezentuje tvrzení singularitianů a futuristů (například Raye Kurzweila, Nicka Bostroma, Franka J. Tiplera) a ukazuje jejich paralely s náboženskou vírou. Nejen to, chce ukázat, že i přes jejich tvrzení, že jejich přesvědčení stojí pouze na vědeckých faktech, není příliš velký rozdíl mezi jejich hnutím a náboženstvím.[103] Jeden z příkladů těchto tvrzení je představa, že s příchodem singularity budeme následně uploadováni do jiných substrátů, a tak budeme simulováni, tudíž se staneme softwarově založenými. Z tohoto stavu pak plynou mnohé výhody, protože se například budeme moci přenést do jiných těl, která mohou být kdekoliv, nebo budeme schopni myslet daleko rychleji. Tady se obzvláště soustřeďuje na jejich tvrzení blaženosti, která bude plynout z tohoto stavu a ukazuje na paralelu mezi naším simulováním a náboženským nebem.[104]

Proto o této představě mluví jako o simulačním vzkříšení.[105] Jeden z příkladů, který uvádí, aby ukázala, jak techno-futuristé mluví o post-lidském stavu jako o vyšším stupni bytí, je citát z Bostromova spisu „Letter from Utopia“: „Je tu krása a štěstí, které si nedokážeš představit. Je to tak dobrý pocit, že pokud by měl být vyjádřen slzami štěstí, tak by řeky přetékaly. … Je to jako déšť toho nejlepšího pocitu, kde každá kapka deště má svůj vlastní unikátní význam.“[106] Na tomto příkladu ukazuje, jak podobně jako mnohá

<https://www.amazon.com/Singularity-Journal%20Consciousness-Studies-ebook/dp/B01N1N6KLZ/ref=mt_kindle?_encoding=UTF8&me>.

[101] KURWEIL Ray. *The Singularity Is Near. When Humans Transcend Biology*. 1. vyd. New York : Viking Adult, 2005. ISBN 0-670-03384-7. s. 369– 372.

[102] PROUDFOOT Diane. Software Immortals: Science or Faith? In EDEN H. Ammon, MOOR H. James, SØRAKER H. Johnny, STEINHART Eric (eds). *Singularity Hypotheses. A Scientific and Philosophical assesment*. 1. vyd. New York, London, Dordrecht: Springer, 2012. ISBN 978-3-642-32559-5.

[103] Tamtéž, s. 367–368
[104] Tamtéž, s. 369
[105] Tamtéž, s. 370
[106] BOSTROM, Nick. *Letter from Utopia* [online]. Verze 1.9. 2010. [cit. 8. 2. 2018]. Dostupné z: <https://nickbostrom.com/utopia.pdf>. vlastní překlad.

náboženství, tak techno-supernaturalismus slibuje plnohodnotný život.[107] Stejně tak se někteří futuristé drží konceptu boha, jako příklad dává citát od Kurzweila: „Jakmile nasytíme hmotu a energii ve vesmíru informacemi, tak se „vzbudí“ a bude vědomá a vznešeně inteligentní. To je tak blízko k bohu, jak jen si dokážu představit.“[108] Podle ní tedy techno-supernaturalisté věří, že lidé postaví boha, a ne obráceně, jak je tomu tradičně.[109]

Proudfootová není jediná, která vede svou kritika hnutí singularity tímto směrem, Selmer Bringsjord společně s Alexandrem Bringsjordem a Paulem Belloem tak také činí, když tvrdí, že víra v singularitu je fideistická. Nejdříve musíme stanovit pár věcí, abychom pochopili jejich další argumentaci. Podle autorů o události jakou je singularita musíme mluvit jako o něčem, co je pro nás dočasně nepřístupné (temporary removed). Tedy události, která nenastala a nenastane v blízké budoucnosti. Dále má pro nás tato událost váhu (weighty), je v ní něco, co jí tuto váhu propůjčuje, například něco zázračného, a nakonec i něco nevídaného (unseen). Autoři pak říkají, v jakém případě můžeme o propozicích říci, že mají váhu a nevídanost: „Propozice mají váhu a jsou nevídané, když jsou přímo a bezprostředně spojeny s nějakou bytostí (bytostmi) a/nebo výskytem nějaké události, která pokud by měla v jakémkoliv měřítku nastat, nebo pokud by bylo známo, že nastane, tak (1) by zapříčinila v racionálních aktérech změnu v tom, v co věří a jak se chovají, a (2) zahrnuje bytosti dosud neviditelné.“ Zkráceně tedy autoři těmto vlastnostem říkají wutr (**w**eighty, **u**nseen, **t**emporarily **r**emoved). O singularitě můžeme podle autorů říci, že je wutr. Dále autoři ve své argumentaci označují symbolem *S* událost, která povede k singularitě, a **S** označuje propozici, která tvrdí, že singularita nastane.[110]

---

[107] PROUDFOOT Diane. Software Immortals: Science or Faith? In EDEN H. Ammon, MOOR H. James, SØRAKER H. Johnny, STEINHART Eric (eds). *Singularity Hypotheses. A Scientific and Philosophical assesment*. 1. vyd. New York, London, Dordrecht: Springer, 2012. ISBN 978-3-642-32559-5. s. 372.

[108] KURWEIL Ray. *The Singularity Is Near. When Humans Transcend Biology*. 1. vyd. New York: Viking Adult, 2005. vlastní překlad. ISBN 0-670-03384-7. s. 375.

[109] PROUDFOOT Diane. Software Immortals: Science or Faith? In EDEN H. Ammon, MOOR H. James, SØRAKER H. Johnny, STEINHART Eric (eds). *Singularity Hypotheses. A Scientific and Philosophical assesment*. 1. vyd. New York, London, Dordrecht: Springer, 2012. ISBN 978-3-642-32559-5. s. 372.

[110] BRINGSJORD Selmer, BRINGSJORD Alexander, BELLO Paul. Belief in The Singularity is Fideistic In EDEN H. Ammon, MOOR H. James, SØRAKER H. Johnny, STEINHART Eric (eds). *Singularity Hypotheses. A Scientific and Philosophical assesment*. 1. vyd. New York, London, Dordrecht: Springer, 2012. vlastní překlad. ISBN 978-3-642-32559-5. s. 396–400.

Samotná argumentace autorů začíná tvrzením, že když mluvíme o propozicích a našem uznání jejich pravdivosti, tak máme tři normativní pohledy: racionalismus, empirismus a fideismus. Abychom mohli zhodnotit tito tři pohledy, tak se autoři opírají o práci R. Chisholma, který vytvořil spektrum síly propozic.[111] Autoři vysvětlují, že toto spektrum se skládá z devíti bodů, kde na jednom konci máme propozice nepochybně nepravdivé a na druhém konci ty, které jsou jisté. Autoři detailně popisují pozitivní škálu a začínají přesnou polovinou, kde jde o vyvážené propozice, pak máme propozice možné, a dále propozice nade vší pochybnost, propozice evidentní, a nakonec předtím zmiňované propozice jisté. Pod jistými si můžeme představit karteziánské pravdy typu „já existuji". Pod evidentní spadají propozice vycházející z toho, co nám říkají smysly. Tyto propozice nejsou samozřejmě nezpochybnitelné, protože je například možné, že to, co teď vnímáme, je halucinace.[112]

Když jsme pak takto vybaveni touto škálou, tak můžeme popsat jednotlivé pohledy, a jak se vztahují k wurt propozicím. Racionalismus je podle autorů pohled, který nám ukazuje, že víra ve wurt propozici musí být podpořena deduktivními důkazy a argumenty, a premisy musejí být aspoň možné. Pohled empiristický je mnohem silnější, ten tvrdí, že víra ve wurt propozici musí být podpořena přímým a neurobiologicky normálním vnímáním složek dané propozice. Tedy, že premisa musí být evidentní. Poslední je pak pohled fideistický, jehož tvrzením je, že musíme věřit ve wutr propozici, i když máme málo či žádné důkazy, tudíž daná propozice pak může být z hlediska síly vyvážená, nebo být i níže ve spektru. Podle autorů si pak můžeme představit slabé, umírněné a silné formy těchto pohledů. Například silná forma racionalismu by požadovala od toho, kdo brání tu danou propozici, aby měl aspoň jeden nezpochybnitelný důkaz pro její pravdivost, čímž je myšlen řetězec deduktivních inferencí, které stojí na premisách, které jsou jisté. Umírněný racionalismus by zase požadoval, aby tento řetězec

[111] CHISHOLM M. Roderick. *Theory of Knowledge*. 3. vyd. Englewood Cliffs: Prentice-Hall, 1989. ISBN 0-13-914185-5.
[112] BRINGSJORD Selmer, BRINGSJORD Alexander, BELLO Paul. Belief in The Singularity is Fideistic In EDEN H. Ammon, MOOR H. James, SØRAKER H. Johnny, STEINHART Eric (eds). *Singularity Hypotheses. A Scientific and Philosophical assesment*. 1. vyd. New York, London, Dordrecht: Springer, 2012. ISBN 978-3-642-32559-5. s. 400–401.

stál aspoň na evidentních premisách, a slabou formu racionalismu jsme použili, abychom tento pohled definovali.[113]

Když jsme takto popsali základy, tak můžeme přistoupit k obhajobě teze, že víra v singularitu je fideistická. Autoři tuto tezi ukazují na Chalmersově hlavním argumentu s mírnou modifikací:

1. AI bude existovat (je stvořeno HI)
2. Pokud bude AI, tak bude AI+ (je stvořeno AI)
3. Pokud bude AI+, tak bude AI++ (je stvořeno AI+)

Závěr: Bude existovat AI++ (= **S** nastane)

Tento argument je formálně správně, zde ale podle autorů výhody tohoto argumentu končí. Jak bylo zmíněno výše, abychom mohli víru v **S** uznat za racionalistickou, tak by dané premisy musely být alespoň možné anebo vyšší, tedy nade vší pochybnost, evidentní nebo jisté. Když se podíváme na první premisu, rozhodně můžeme říci, že není jistá. Argumentace pro tento fakt se podobá druhému teorému z Bringsjordovy námitky proti Chalmersovi, tedy pokud si představíme, že lidé jsou na vyšší úrovni než standardní Turingův stroj, tak AI, které je postaveno jako Turingův stroj, se nemůže dostat na stejnou úroveň jako lidé. Z tohoto důvodu musíme říct, že první premisa není jistá. Podle autorů ale nemůžeme o první, a tím pádem ani o ostatních premisách říci, že jsou evidentní nebo nade vší pochybnost. Protože není nade vší pochybnost, že ti, co tvrdí, že lidské myšlení je nad standardním Turingovým strojem, se nemýlí.[114]

Jsou ale i jiné důvody pro to, abychom mohli říct, že první premisa není evidentní. Když si vzpomeneme, tak abychom mohli o něčem říct, že je to evidentní, musíme tu danou skutečnost vnímat našimi smysly. Co bychom ale mohli vnímat, co by nás přivedlo k myšlence, že se blíží strojová inteligence na úrovni člověka? Podle autorů nic, jejich názorem je, že neexistuje žádný počítač na naší úrovni. Tak se blížíme k poslední alternativě pro víru v **S**, kterou je slabý racionalismus. Můžeme tedy o tomto argumentu říct, že jeho premisy jsou možné? Autoři si myslí, že ne. Nejdříve ukazují na skutečnost, že dnešní počítače jsou v mnohém rychlejší než lidský mozek, a přesto žádný počítač

[113] Tamtéž, s. 401–402
[114] Tamtéž, 403–404

ještě nedosáhl lidské úrovně inteligence, takže se jim zdá poněkud nepravděpodobné, že HI stvoří AI.[115]

Mají však i daleko více filosofický důvod, proč neuznat premisy Chalmersova argumentu jako možné. Můžeme říci, že Chalmersův argument stojí na představě neustále se zvyšující inteligence a vyplývá to z trojice premis tohoto argumentu. Řekněme si tedy, že *L(M)* označuje úroveň (Level) inteligence stroje (Machine), takže z argumentu nám tedy vyplývá to, že *L*(HI) < *L*(AI+). Takže podle proponentů **S**, kteří se opírají o tento argument, je možné, že *L* (HI) < *L* (AI+). Aby však toto mohli tvrdit, tak by museli vědět, v čem spočívá rozdíl v inteligenci mezi HI a AI+. Pokud však toto nevědí, tak podle autorů nemají epistemické právo tohle tvrdit. A nejen to, pokud přesto budou toto tvrdit, tak se pak jejich propozice nese ve fideistickém duchu.[116]

Samozřejmě proponenti Singularity by mohli tvrdit, že ultrainteligentní stroje mají SI, protože například dokáží hrát šachy lépe než lidé, nebo že jsou lepší v testech inteligence než lidé. Tato námitka ale obsahuje podle autorů problémy, a oba stojí na tom, jak moc proponenti singularity podle autorů poukazují na větší rychlost počítačů, na čemž pak stojí jejich tvrzení. Problém je tedy v tom, že například v současnosti už máme perfektní algoritmus pro šachovou hru, to znamená, že příští počítače nebudou hrát šachy lépe, ale pouze rychleji. Podle autorů počítače, o které proponenti singularity svou víru opírají, nejsou nic víc než velice rychlé procesory. Takové počítače bychom ale nenazvali super inteligentními. Závěrem autorů tedy je, že víra v singularitu je fideistická.[117]

## 3.3. Problém AI – Strojová etika a problém kontroly

Největší problém, před kterým cesta k umělé inteligenci stojí je problém Strojové etiky, který lze dost jasně ukázat na myšlenkovém experimentu Džina-Golema. Představte si, že se před vámi objeví neuvěřitelně mocný Džin-Golem a sdělí vám, že za padesát let se před vámi zjeví znovu, ale tentokrát po vás bude chtít, abyste mu předložili morální doktrínu, podle které pak následně přetvoří celý svět. Pokud mu žádnou nepředložíte, tak příjme tu první, kterou mu předloží někdo jiný. Řekněme, že byste mu předložili doktrínu hedonistického utilitarismu, co by se stalo? Pravděpodobně by

[115] Tamtéž, s. 404–405
[116] Tamtéž, s. 405
[117] Tamtéž, s. 406

následně začal přeměňovat všechnu hmotu vesmíru do co nejmenších virtuálních myslí, které by nedělaly nic jiného, než prožívaly jeden pocit slasti neustále dokola. Tento problém tkví ve dvou vlastnostech, které Džin-Golem má. Zaprvé, má schopnost neuvěřitelně tvarovat realitu okolo sebe, a to mu umožnuje řešit problémy způsoby, které by člověka nenapadly. A druhou jeho vlastností je, že cokoliv řeknete, bude brát doslova, protože postrádá selský rozum.[118]

Tento experiment ilustruje problém etiky SI ve formě umělé inteligence. Problém, který přetrvává, i kdybychom vybavili danou SI jednoduchými cíli. Představme si prostý šachový počítač, ten má velice jednoduchý cíl, a to vyhrát co nejvíce šachových her. Takže má cíl, na kterém na první pohled není nic špatného. Musíme ale pochopit, že umělá inteligence udělá vše proto, aby maximalizovala splnění svého cíle. Pokud by tento šachový počítač například objevil internet, mohl by se nabourat do ostatních počítačů, aby mu pomohly lépe hrát šachy. Stejně tak by se snažil zabránit tomu, aby ho jeho stvořitelé vypnuli, protože pak by nemohl hrát další šachové hry. Pokud objeví existenci lidí, bude se snažit je donutit, aby s ním hráli šachy. Cokoliv, co by mu mohlo pomoci s jeho cílem, by se pokusil získat a ovládat.[119]

Problém Strojové etiky upozorňuje na větší nebezpečí. Tím je problém kontroly. Jak zajistíme kontrolu nad SI? Abychom tento problém mohli řešit, tak je pro nás podstatné abychom se zamysleli, jak by SI myslela a jaké by byly její motivace. První věcí, které bychom se měli vyvarovat, je možná antropomorfizace dané inteligence. Eliazer Yudkowski říká, že pokud jde o umělou inteligenci, musíme se podívat za hranici lidské mysli. Měli bychom si představit čáru, na které je malý kruh vyznačen jako lidská mysl, a dále jsou vyznačeny všechny druhy myslí, které jsou jak pod, tak i nad naší úrovní inteligence. Celý rozměr této čáry jsou možné mysli, které by AI mohla mít, tedy nejen ta lidská část. Přičemž největší chybou, kterou bychom mohli učinit, je příliš brzy si myslet, že umělou inteligenci chápeme.[120]

---

[118] MUEHLHAUSER, Luke, and HELM Louie. The Singularity and Machine Ethics. In EDEN H. Ammon, MOOR H. James, SØRAKER H. Johnny, STEINHART Eric (eds). *Singularity Hypotheses. A Scientific and Philosophical assesment*. 1. vyd. New York, London, Dordrecht: Springer, 2012. ISBN 978-3-642-32559-5. s. 105–106

[119] OMOHUNDRO Steve. Rational Artificial Inteligence for greater good. In EDEN H. Ammon, MOOR H. James, SØRAKER H. Johnny, STEINHART Eric (eds*). Singularity Hypotheses. A Scientific and Philosophical assesment.* 1. vyd. New York, London, Dordrecht: Springer, 2012. ISBN 978-3-642-32559-5. s. 163–164.

[120] YUDKOWSKI, Eliazer. Friendly Artificial Inteligence. In EDEN H. Ammon, MOOR H. James, SØRAKER H. Johnny, STEINHART Eric (eds). *Singularity Hypotheses. A Scientific and Philosophical*

Podle Yudkowského, během diskuzí o AI propadáme do určitého klamu, tomuto klamu on říká *logický omyl tvarohového koláče*. Ten stojí na představě, že brzy postavíme SI. A dobře víme, jak velké koláče upečeme my s naší inteligencí. Tak si představte koláče, které upeče SI. Ty prostě budou obrovské! V tomto postupu je ale zřejmá chyba. Proč by SI chtěla péct tvarohové koláče? Místo toho abychom se ptali po motivech, které by umělá inteligence mohla mít, tak se zamýšlíme nad tím, čeho by byla schopná. AI by mohla chtít vyhubit lidstvo, ale proč by to chtěla?[121]

Yudkowski je také proponentem termínu Přátelská umělá inteligence, který by měl být cílem našeho snažení. Tato inteligence, by byla vybavena sadou motivací, které by ji vedly k přátelskému chování. Musíme se ale ptát, jestli je možné takovou AI vůbec postavit? Yudkowski sám přiznává, že reakce, které na tento termín dostává, jsou často negativní, často je mu řečeno, že je nemožné postavit takovou AI. Proč by dostatečné inteligentní AI, prostě nepřepsala jakákoliv omezení, které do ní vložíme? Tím se ale vracíme k tomu logickému omylu, proč by AI chtěla změnit své motivy?[122]

Jaké motivy, by tedy mohla mít SI, jak by fungovalo její přemýšlení, a jak bychom do ní mohli vložit ty hodnoty, které chceme, a zároveň zajistit, aby podle nich jednala tak, jak my chceme? Začněme možnými motivy SI. Podle Bostroma platí něco, čemu říká ortogonální teze. Tu definuje takto: „Inteligence a finální cíle jsou ortogonální: více méně jakákoliv úroveň inteligence by mohla v zásadě být zkombinována s více méně jakýmkoliv finálním cílem“.[123] Bostrom podotýká, že inteligencí nemá na mysli nic takového jako rozum či racionalitu. Inteligence je v této tezi myšlena jako schopnost předvídat, plánovat, operovat s účely a prostředky. Takže bychom mohli mít SI stroj, jehož úkolem by bylo počítat zrnka písku na plážích ostrova Boracay.[124]

Přestože SI může mít takové neantropomorfické cíle, tak to neznamená, že bychom nebyli schopni tyto její cíle předvídat. Bostrom navrhuje tři druhy předvídatelnosti: předvídatelnost z designu, předvídatelnosti z dědičnosti a předvídatelnost z konvergentních instrumentálních důvodů. První vychází z účelu, pro

*assesment.* 1. vyd. New York, London, Dordrecht: Springer, 2012. ISBN 978-3-642-32559-5. s. 181–183.

[121] Tamtéž, s. 185

[122] Tamtéž, s. 187

[123] BOSTROM Nick. *Superinteligence. Paths, Dangers, Strategies*. 1. vyd. Oxford: Oxford University Press, 2014. ISBN 978-0-19-967811-2. s. 107. vlastní překlad.

[124] Tamtéž, s. 107

který byla SI stvořena za předpokladu, že její vývojáři vytvořili dostatečně robustní systém cílů, které má následovat. Druhý platí u SI založené na lidské formě (např. WBE), kde díky podobnosti si můžeme být jisti, že SI bude sdílet některé naše motivy. Poslední je nejdůležitější. I když neznáme konečný cíl daného aktéra, tak tyto cíle můžeme inferovat z instrumentálních důvodů, jenž povstávají z mnoha cílů a které jsou aplikovatelné v široké míře situací.[125]

Poslední druh předvídatelnosti rozvádí a staví z něj další tezi. Tezi instrumentální konvergence: „Několik instrumentálních hodnot může být identifikováno, pokud jsou konvergentní v tom smyslu, že jejich osvojení by zvýšilo šanci splnění cíle aktéra pro široký rámec konečných cílů a v širokém rámci situací, což znamená, že se těmito instrumentálními hodnotami bude řídit široké spektrum situovaných aktérů.“ Pokud přijmeme tuto tezi, tak si pak můžeme představit několik takových hodnot, kterými by se SI mohla řídit, aby splnila svůj cíl. Přičemž Bostrom podotýká, že čím vyšší bude inteligence daného aktéra, tím vyšší bude šance, že je rozezná a bude se řídit těmito hodnotami.[126]

Jednou takovou hodnotou je sebezáchova. Pokud aktér má cíl, který chce v budoucnosti splnit, tak jeho existence je důležitá proto, aby daný cíl splnil. Protože jedině v budoucnosti může konat tak, aby zajistil splnění svého cíle. Tato hodnota ale nemusí vždy být nutná. Na rozdíl od lidí, kteří chápou své přežití jako konečný cíl, tak SI by se mohla obětovat, pokud by to bylo nutné pro dosažení jejího cíle. Daleko důležitější hodnotou pro SI by asi bylo udržení si svého cíle. Pokud by její konečný cíl byl změněn, nemohla by splnit ten původní. Tato hodnota by byla důležitější pro SI kvůli její softwarové povaze, která ji umožnuje se snadno přesunout, duplikovat atd. Což znamená, že pak dává větší důraz na své hodnoty než na povahu svého těla, vzpomínek či dovedností. [127]

Další hodnotou by bylo kognitivní vylepšení. Čím vyšší je úroveň inteligence aktéra, tím lepší je aktér při řešení problémů, což mu pak umožnuje lépe splnit svůj konečný cíl. Znovu ale tento cíl stojí nad vším. Pokud cena tohoto vylepšení je příliš vysoká, tak by se aktér spíše rozhodl být méně inteligentní. Pokud aktér má také k dispozici profesionální pomoc, nemusí cítit potřebu pro toto vylepšení. Pokud se aktér

[125] Tamtéž, s. 108
[126] Tamtéž, s. 109 (vlastní překlad)
[127] Tamtéž, s. 109–110

rozhodne vylepšit, bude se spíše zaměřovat na schopnosti, které skutečně potřebuje. Než se vylepšovat v celé šíři, bude se spíše specializovat.[128]

S kognitivním vylepšením jde v ruku v ruce technologické zdokonalování, které může vést ke kognitivnímu vylepšení, ale také nemusí. Toto zdokonalování může být pouze nalézání lepších nástrojů a prostředků, které by danému aktérovy umožnily lépe ovlivňovat a měnit realitu ke svému obrazu. Poslední instrumentální hodnotou by bylo získávání zdrojů, které jsou potřebné k dosažení daného cíle. Ke konci Bostrom ale podotýká, že toto nemusí být nutně všechny instrumentální hodnoty, kterými by se aktér mohl řídit. Je možné, že existují hodnoty, které si my nedokážeme představit, které mohou například plynout z dosud neobjevených fyzikálních jevů, kterých by si ale SI všimla. Tudíž chování aktéra nemusí být vždy předpověditelné. [129]

Jak ale bude SI přemýšlet a jednat? Podle Steva Omohundry si můžeme být jistí, že AI bude mít tendenci se chovat racionálně. Tato povaha bude jednak vyplývat ze skutečnosti, že umělá inteligence se bude moci vylepšovat, jednak z faktu, že z iracionálního chování plynou mnohé nevýhody, jako například možnost neefektivního využití zdrojů a možnost zneužití ze strany druhých. Stejně tak pokud daný činitel má nějaký cíl, který chce splnit v prostředí, o kterém ví, že je náhodné, tak racionální činění se ukazuje být nejlepší cestou. Z těchto samých důvodů mají i biologické bytosti motivaci pro racionální chování, pokud jednají iracionálně, ostatní je využijí. Ale protože postrádáme možnost sami sebe vylepšovat, tak nám zbývá pouze proces přírodní selekce. Biologické bytosti mají tendenci mířit k racionálnímu chování pouze díky vlivům prostředí, pokud ty chybí, tak k tomu není důvod. [130]

Skutečnost je ale taková, že být plně racionální může být pro určitá prostředí příliš náročné. Proto Omohubdro mluví o omezených racionálních systémech, které se objevují v daných prostředích. Systémy jsou tak racionální, jak mohou být vzhledem k omezením, které jsou na ně kladeny. Tyto systémy jsou tvarovány jinými procesy a systémy, například organismy jsou tvarovány evolucí, či ekonomické subjekty ekonomikou. Tyto procesy Omohundro nazývá tvarovači. Omohundro rozeznává několik druhů takových

[128] Tamtéž, s. 111
[129] Tamtéž, s. 112–114
[130] OMOHUNDRO Steve. Rational Artificial Inteligence for greater good. In EDEN H. Ammon, MOOR H. James, SØRAKER H. Johnny, STEINHART Eric (eds). *Singularity Hypotheses. A Scientific and Philosophical assesment.* 1. vyd. New York, London, Dordrecht: Springer, 2012. ISBN 978-3-642-32559-5. s. 164–165.

systémů, které jsou mezi sebou rozlišeny stupni vývoje, které reprezentují. Od systémů, které vyžadují nejméně výpočetní síly až po ty nejnáročnější. Nejomezenější jsou systémy konstantní akce, jak jejich jméno napovídá, jsou to systémy, které neustále jednají nehledě na to, co říkají jejich smysly či zkušenost. Tyto systémy jsou efektivní ve velice jednoduchých prostředích. Představte si prostředí, kde můžete jít pouze doleva nebo doprava, když jdete doprava, pocítíte bolest, když doleva tak ne. Tudíž systém v tomto prostředí jde stále doleva.[131]

Dále máme systémy stimulu a response. Tento systém koná čistě na základě toho, jakou odpověď dostane na své akce, pokud pozitivní, tak v akci pokračuje. Tyto systémy také nemají paměť, nezaznamenají si předchozí zkušenosti a neplýtvají výpočetní silou na plánování dalších akcí. Ideálním prostředím pro tyto systémy je dvourozměrná mřížka, kde jste potrestáni na její hranici a odměněni uprostřed. Daný aktér, který následně rozeznává souřadnice své pozice, se snaží držet uprostřed. Systémy jednoduchého učení jsou optimální v oblastech, kde jsou prvky, které tvarovač nezná, takže vybaví daný systém schopností jednoduchého učení. Příkladem muže být prostý herní automat s dvěma pákami, daný aktér bude tahat střídavě za obě páky, dokud nezjistí, která má větší pravděpodobnost úspěchu, za tu, která ji má, pak bude tahat.[132]

Nyní se už dostáváme k systémům, které jsou úmyslné. Tyto systémy už nejsou pouze reaktivní a jsou schopné plánování a modelování prostoru okolo nich. Také už mají episodickou paměť, která jim napomáhá ve vytváření daných modelů. Takové systémy pak operují v komplexních prostředích, která jsou proměnlivá a pro jejichž modelování je zapotřebí pravděpodobnostních modelů. Takže vytvoří model, zjistí, jestli funguje, a pokud ne, tak ho upraví. Ke konci tu pak máme systémy, které mají schopnost meta přemýšlení, takže jsou schopny upravovat samotný proces jejich přemýšlení. Dále tu jsou systémy sebe vylepšující, které vylepšují každý proces svého chování. Poslední systém je pak plně racionální systém.[133]

Tento vývoj inteligentních systémů je podle Omohundry vidět jak v biologii, kde například u virů lze pozorovat, že jednají na způsob systému stimulu a response, nebo u bakterií, které jsou zase systémy schopné jednoduchého učení, tak i v technologiích.

[131] Tamtéž, s. 165–166
[132] Tamtéž, s. 166
[133] Tamtéž, s. 167

Omohundro se také shoduje s Bostromem co se týče motivací, které by SI mohla mít. Motivace podle něj ale vyplývají z racionální povahy, kterou SI bude mít.[134]

Další otázkou je, jak vyřešit problém kontroly? Podle Bostroma existují dvě kategorie vyřešení tohoto problému, jedním je omezení schopností SI, tedy omezit to, co SI může dělat, a druhou je kontrola motivace, tedy omezit to, co chce dělat.[135] Je několik metod, jak bychom mohli omezit schopnosti SI. Jednou je Boxing metoda, kde omezíte SI buď v prostoru, nebo v přístupu k informacím. Fyzické omezení zabraňuje SI v kontaktu s fyzickým světem, kromě námi zvolených vstupů. Fyzické omezení lze snadno aplikovat, ale nevýhodou je, že drasticky omezuje schopnosti SI. Je tu také nebezpečí, že SI může najít díru v zabezpečení, skrze niž unikne. Informační uzavření omezuje množství informací, které přichází k SI například tím, že jí zabráníme v přístupu k internetu, Tedy omezujeme její komunikaci s vnějším světem. Problém je, že kdyby se nám podařilo takto dokonale omezit SI, a to jak fyzicky, tak i informačně, tak by výsledný systém nebyl k ničemu užitečný.[136]

Jinou metodou je metoda podnětová. Ta stojí na představě, že vložíme SI do prostředí, které ji donutí se chovat určitým způsobem. V tomto prostředí pak vládnou určitá pravidla, či obsahuje další aktéry, které působí na SI. Pokud SI dodržuje daná pravidla, tak je odměněna, pokud ne, tak ji čeká trest. Můžeme si představit, že daná odměna je nějak úzce spojená s konečným cílem SI, takže pro SI se pak následně vyplatí daná pravidla dodržovat. Problém je, aby nabízená odměna představovala pro SI větší zisk než ten, který by mohla získat únikem či vzpourou.[137]

Také je tu možnost, že bychom danou SI mohli také zbrzdit, bud hardwarově nebo omezením jejího přístupu k informacím. Hardwarově bychom ji mohli ponechat běžet na slabším a pomalejším hardwaru, to by ale znovu drasticky omezilo její schopnosti. Její informační přístup bychom mohli omezit tak, že jí například dáme jen ty informace, které potřebuje ke splnění daného cíle, například pouze poznatky z jednoho oboru. Tady ale máme problém, jestli by to stačilo, vzhledem k tomu, že se jedná o SI. To, co se nám může zdát jako neškodné, může být hodnotnou informací pro SI. Problematické je, že

[134] Tamtéž, s. 167–168
[135] BOSTROM Nick. *Superinteligence. Paths, Dangers, Strategies*. 1. vyd. Oxford: Oxford University Press, 2014. ISBN 978-0-19-967811-2. s. 128.
[136] Tamtéž, s. 129–131
[137] Tamtéž, s. 131–134

informace nejsou většinou pouze o jedné věci a často se vztahují i k věcem jiným. SI by mohla mnohé vydedukovat z toho, jaké informace ji dáváme, a jaké ne.[138]

Poslední metodou omezení schopností SI je metoda „nástražných drátů“. Tato metoda stojí na tom, že připravíme pro SI několik takových drátů, které ji sledují. Když by SI překročila jeden z nich, tak bude vypnuta. V podstatě je to série diagnostických testů, které sledují aktivitu SI. Například by mohli sledovat chování dané SI, které když dospěje do určitého bodu, tak bude zastavena. Stejně tak lze použít test schopností SI, kde by SI procházela testem v nastavených intervalech a pokud by ji vyhodnotil jako příliš inteligentní, tak by ji vypnul, nebo program, který sleduje obsah mysli SI, například její vnitřní procesy. Jestli by taková opatření stačila, je problematické, SI by mohla najít cestu, jak je přelstít.[139]

To by bylo vše k metodám omezení schopností SI. Způsoby kontroly motivace také mají různé formy. První je metoda přímé specifikace. Tato metoda je ze všech nejvíce přímočará. Jsou zde dva různé druhy, jeden je postaven na ustanovení přímých pravidel a druhý způsob je konsekvencialistický. První stanovuje pro SI pravidla chování, kterými se bude řídit. Druhý se soustřeďuje na výsledek, stanoví pro SI, čeho má její chování dosáhnout. Oba přístupy mají stejný problém. Jak formulovat pravidlo, které nebude vágní a bude přesné, a zajistí, aby SI jednala tím způsobem, jakým chceme, ve shodě s našimi hodnotami? Jak zabránit mylné interpretaci daného pravidla? Takže se pak vracíme k problému Džina-Golema.[140]

Jiná metoda, která se nabízí, je, že bychom mohli, jak říká Bostrom, „domestikovat“ SI. Co tím má na mysli? Že bychom ji mohli naprogramovat takovým způsobem, aby jednala pouze v určitém kontextu, určitým způsobem. Například bychom mohli vytvořit SI, která funguje jako stroj, který pouze odpovídá na otázky, které jsou mu kladeny. Přičemž jeho konečným cílem by bylo odpovídat na tyto otázky, jak nejlépe může. Tímto způsobem bychom omezili dopad činnosti SI na svět, protože by jednala v mnohem menším měřítku. Tato metoda má také výhodu, že v tomto případě by přímá specifikace cíle pro SI byla jednodušší, vzhledem k tomu, že toho po ní tak moc nežádáme.[141]

[138] Tamtéž, s. 135–136
[139] Tamtéž, s. 137–138
[140] Tamtéž, s. 139–140
[141] Tamtéž, s. 140–141

Další možnou metodou, by mohl být přístup nepřímé normativity. Tady jde o to, že bychom mohli sestrojit SI tak, že místo toho abychom ji my sami předložili pravidla, tak jí udáme nějaký standard, podle kterého si ona sama sestaví nějaký systém hodnot. Takže bychom jí mohli například sdělit něco ve smyslu „Jednej tak, jak by idealizovaná verze lidstva chtěla, abys jednala.“ Jde tedy o extrapolaci hodnot.[142] Poslední metodou, kterou Bostrom zmiňuje je rozšiřovací metoda. Tady v podstatě máme systém na lidské úrovni inteligence, který už má funkční hodnotový systém se správnou motivací, který následně vezmeme a učiníme ho super inteligentním. Příkladem by mohlo být WBE a jiné inteligence, založené na lidském podkladě.[143]

Nakonec bychom chtěli říci, že v této debatě by právě filosofie a její etické diskuze mohly nabýt neuvěřitelného praktického využití. Chalmers právě na tuto skutečnost poukazuje. Vezměme si například pohled Davida Huma, který tvrdil, že morální hodnoty nemají žádnou spojitost s racionalitou, takže bez ohledu na inteligenci či racionalitu daného mohou mít libovolné hodnoty. Pohled Imanuela Kanta byl jiný, ten tvrdil, že některé hodnoty jsou více racionální než ostatní. To, který z nich měl pravdu, by pak mohlo ovlivnit, zda si AI dokáže vyvinout nějakou morálku. Pokud Kantovský pohled je pravdivý, tak bychom nemuseli nijak omezovat či ovládat motivaci AI, racionalita a inteligence by tak učinily za nás. Což by mělo také pozitivní efekt toho, že s vyšší inteligencí by AI měly lepší stupnici hodnot, ale asi by bylo těžší ovládat hodnoty pozdějších systémů.[144]

## 3.4. Problémy WBE – Upload mysli a Identita

Jak už bylo naznačeno v předchozí kapitole WBE má nad AI jednu určitou výhodu. U WBE se nemusíme obávat etických problémů, protože má lidský základ. To ale neznamená, že tento přístup k SI je bez problémů. Prvním je problém vědomí. Jak si můžeme být jistí, že upload některého člověka má vědomí či ne? Tohle je další otázka,

[142] Tamtéž, s. 141–142
[143] Tamtéž, s. 142–143
[144] CHALMERS J. David. The Singularity: A Philosophical Analysis. In AWRET Uziel ed. *The Singularity: Could artificial inteligence really out-think us (and would we want it to)?* [e-kniha]. 1. vyd. Exeter: Imprint Academic. 1. 12. 2016 [cit. 8. 2. 2018]. Kapitola 6. Dostupné z: <https://www.amazon.com/Singularity-Journal%20Consciousness-Studies-ebook/dp/B01N1N6KLZ/ref=mt_kindle?_encoding=UTF8&me>.

kterou se zabývá Chalmers ve své filosofické analýze singularity. Narážíme zde podle něho i na širší problém, a tím je problém vědomí obecně.[145]

Problém vědomí podle něj tkví v jeho povaze, jako vzdáleném faktu vědomých systémů. Vzdáleným faktem je myšleno, že fakta vědomí jsou epistemologicky vzdálenými fakty. Jako příklad toho, co má na mysli, udává známou ilustraci tohoto problému neurovědkyně Mary, která je zavřena v černobílé místnosti. Přestože zná přesný fyzický proces mozku pro vidění barev, tak nemá žádnou zkušenost s červenou barvou. V podstatě lze jakkoliv přesně popsat daný fyzický proces, ale nikdy nezískáme úplnou vědomou zkušenost daného subjektu. Z tohoto důvodu je problém vědomí vzdáleným faktem.[146]

Nehledě však na tuto skutečnost můžeme zkusit tvrdit, jaký druh systému může mít vědomí a jaký ne. Filosofové se pak v tomto směru dělí na dva tábory. Na ty, kteří se přiklání k Biologické teorii vědomí, podle které pouze biologické systémy mohou být vědomé, a na ty, kteří se kloní k Funkcionalistické teorii, která tvrdí, že záleží spíše na kauzální struktuře a roli daného systému a tato povaha mu propůjčuje vědomí. Chalmers přiznává, že on sám se spíše kloní k Funkcionalistům, protože si myslí, že jsou blíže k pravdě.[147]

Představme si tedy, že vytvoříme perfektní upload mozku v počítači. Každý fyzický neuron nahradíme počítačovým elementem, který replikuje jeho funkčnost. Tyto elementy pak propojíme se vstupními a výstupními zařízeními, jako např. umělé oči, či končetiny, které budou funkční, ať už ve virtuálním nebo reálném prostředí. V podstatě vytvoříme plně funkční izomorf mozku. Bude ale mít vědomí? Podle Chalmerse naší nejlepší šancí pro udržení vědomí je proces postupného nahrávání, při kterém postupně nahrazujeme jednotlivé části mozku. V takovém případě bychom mohli nahrazovat části mozku fyzicky, a to silikonovými obvody, nebo možná procesy v počítači, který je nějakým rozhraním propojený s mozkem.[148]

Pokud takový proces dopadne správně, tak můžeme očekávat, že nové komponenty se bez problému integrují, jak do biologických, tak i do nebiologických částí okolo nich a že budou fungovat úplně stejně jako původní biologické komponenty. A

[145] Tamtéž, kap. 9.
[146] Tamtéž, kap. 9.
[147] Tamtéž, kap. 9.
[148] Tamtéž, kap. 9.

nejen to, při správném průběhu by si zúčastněná osoba ani neměla všimnout toho, že proces už proběhnul. Otázka tedy zní, co se při tomto procesu stane s vědomím? Pro Chalmerse je tady několik možností, první je nenadálé zmizení vědomí, kdy do nějaké doby tam stále je, dokud nevyměníte jeden klíčový komponent, což způsobí, že najednou vědomí zmizí. Tato možnost přijde Chalmersovi nejméně pravděpodobná, co když místo toho abychom vyměnili celý komponent, ho budeme vyměňovat po částech, co pak?[149]

To nás přivádí k druhé možnosti, kterou je postupné mizení. Tady místo toho, aby celé vědomí zmizelo najednou, mizí po částech. To ale přijde Chalmersovi stejně nepravděpodobné. V takovém případě bychom v půlce procesu měli subjekt, který je částečně vědomý, ale přitom, jak plyne z jeho hypotézy, si daný subjekt není ničeho vědom. Je možné, že subjekt nevěří, že se něco změnilo, ale aby se tohle stalo, muselo by jít o systém, který si není příliš vědom své vědomé zkušenosti.[150]

Když si pak představíme dobu, kdy by aspoň částečné uploadování bylo běžné, a podíváme se na takovou možnost náhlého zmizení vědomí, tak by všude chodili lidé, jejichž vědomí se jich sotva drží. Podle Chalmerse je možné, že u takových lidí by se vědomí mohlo také vypínat a zapínat, což by znamenalo, že by se u nich střídaly stádia vědomí a stádia, kdy by byli prakticky zombie. Z pohledu postupného mizení by okolo nás chodili neúplní lidé. Přesto všechno by ale tito lidé jednali, jako by se nic nestalo. Z tohoto důvodu tedy Chalmersovi přijde těžké tyto dva pohledy brát vážně.[151]

Proto je zastáncem třetí možnosti, tedy že upload mysli by byl vědomý, což mu přijde jako nejvíce pravděpodobná možnost. Takže pokud bychom tomuto funkčnímu izomorfu uznali vědomí, tak nás to podle Chalmerse také navádí k tomu, abychom mu uznali stejně kvalitativní vědomé stavy. To nás pak vede dále k závěru, že vědomí je organizační invarianta. Museli bychom tedy uznat, že systémy se stejnou kausální organizací mají stejné stavy vědomí nehledě na jejich substrát. Což víceméně znamená, že takovou věc je možné zachytit v počítačové simulaci, na rozdíl od těch, které nejsou organizačními invariantami.[152]

Je důležité také zmínit z jaké filosofické pozice Chalmers k tomuto závěru dochází. Toto je skutečnost, na kterou ukazuje například Daniel Dennet ve své reakci na

[149] Tamtéž, kap. 9.
[150] Tamtéž, kap. 9.
[151] Tamtéž, kap. 9.
[152] Tamtéž, kap. 9.

Chalmersovu analýzu,[153] a i sám Chalmers uvádí svou pozici. Chalmers u problému vědomí a otázky vzdálených faktů rozeznává několik filosofických skupin. První jsou Materialisté typu A, jejímž zástupcem je například Daniel Dennet. Ti odmítají, že problém vědomí je vzdálený fakt. Materialisté typu B, jako například Ned Block, uznávají, že problém vědomí je vzdálený fakt z epistemologického hlediska. Nakonec Chalmers sám sebe identifikuje jako Dualistu, přesněji dualistu vlastností (property dualist). Takže věří, že problém vědomí je vzdáleným faktem jak z epistemologického hlediska, tak i z ontologického.[154]

Dennet se ve své odpovědi na Chalmersovu analýzu věnuje pouze krátce hlavnímu tématu, což je singularita, o které mimochodem tvrdí, že nejde o nic než technologickou fantazii, a že bychom se měli soustředit na důležitější problémy. Dennet se soustřeďuje právě na tuto sekci, aby ji vzhledem ke statusu, který mu Chalmers připsal, zpochybnil. Otázkou, kterou bychom si podle něj měli klást je, proč Chalmers není materialista typu A? Měli bychom si ji klást, protože Chalmers se ve své argumentaci kloní spíše k jejich argumentům. Příkladem toho, že má blízko k funkcionalistům je, že pokud jde o otázky postupného uploadování, tak se také kloní na Dennetovu stranu. Podle Denneta je očividné, že chápe argumenty materialistů typu A, a dokonce je přijímá, ale přitom se stále drží dualismu. To samé se podle něj děje i později, když Chalmers řeší problém identity. To souvisí s tím, že Chalmers bere vážně myšlenkový experiment se Zombie. Experiment, který poukazuje na možnou epistemickou díru mezi fyzickými procesy a vědomím, což Dennet odmítá.[155]

Další problém WBE přístupu je problém identity. Pokud někdo projde procesem uploadování, je tento upload stále ta stejná osoba? Jsou tu dva způsoby, jak se na tento problém můžeme dívat, buď můžeme být optimisty a říct, že daný upload je vskutku stále

---

[153] DENNET Daniel. Mystery of David Chalmerse. In AWRET Uziel ed. *The Singularity: Could artificial inteligence really out-think us (and would we want it to)?* [e-kniha]. 1. vyd. Exeter: Imprint Academic. 1. 12. 2016 [cit. 8. 2. 2018]. Dostupné z: <https://www.amazon.com/Singularity-Journal%20Consciousness-Studies-ebook/dp/B01N1N6KLZ/ref=mt_kindle?_encoding=UTF8&me>.

[154] CHALMERS J. David. The Singularity: A Philosophical Analysis. In AWRET Uziel ed. *The Singularity: Could artificial inteligence really out-think us (and would we want it to)?* [e-kniha]. 1. vyd. Exeter: Imprint Academic. 1. 12. 2016 [cit. 8. 2. 2018]. Kapitola 9. Dostupné z: <https://www.amazon.com/Singularity-Journal%20Consciousness-Studies-ebook/dp/B01N1N6KLZ/ref=mt_kindle?_encoding=UTF8&me>.

[155] DENNET Daniel. Mystery of David Chalmerse. In AWRET Uziel ed. *The Singularity: Could artificial inteligence really out-think us (and would we want it to)?* [e-kniha]. 1. vyd. Exeter: Imprint Academic. 1. 12. 2016 [cit. 8. 2. 2018]. Dostupné z: <https://www.amazon.com/Singularity-Journal%20Consciousness-Studies-ebook/dp/B01N1N6KLZ/ref=mt_kindle?_encoding=UTF8&me>.

tou samou osobou, nebo pesimisticky říct, že se jedná o úplně jinou osobu. Když se ještě řekne, že proces uploadování povede k destrukci mozku, který byl použit, tato otázka nabírá na důležitosti. Jak tedy tento problém vyřešit?[156]

První věc, kterou Chalmers zmiňuje je, že jeho závěry k problému vědomí nám v tomto případě nepomohou. Náš upload může mít kvalitativně stejné stavy vědomí, jako máme my, ale nemusí být nutně námi. Dále si musíme blíže rozebrat, jak pesimistické a optimistické modely vidí identitu, nebo spíše, co nám zaručuje, že si ji zachováváme v čase? Na jednom konci můžeme mít představu, že musí být zachována hmota, ze které se osoba skládá, takže když změníme jednu molekulu, tak se identita dané osoby mění. Na straně druhé stačí pouze mít nějaký vědomí stav, abychom uznali stejnou identitu. Většina teoretiků se ale podle Chalmerse soustřeďuje na dvě kategorie, propojenost a kontinuitu. Pokud jsou tyto dvě kategorie udrženy napříč časem určitým způsobem, tak je možné uznat stejnou identitu. Udržování těchto kategorií pak může být chápáno jako přežití subjektu.[157]

Důležitá otázka tedy je, jakým způsobem musejí být tyto dvě kategorie udržovány, abychom uznali identitu? Zde tak máme znovu několik skupin s rozdílnými názory a teoriemi. Jedni se zaměřují na biologickou složku a podle nich je zapotřebí přežití mozku či celého těla, druzí jsou zastánci psychologické teorie. Těm tedy jde o psychologickou kontinuitu napříč časem. Nakonec jsou tu ti s teorií nejbližšího následovníka, kteří si myslí, že osoba udržuje svou identitu skrze entity, které co nejblíže v čase následují, samozřejmě ale s několika omezeními.[158]

Následně pak tyto teorie ovlivňují to, jak se můžeme dívat na problém uploadování a identity. Biologická teorie by nám řekla, že se identita rozhodně nepřenáší na daný upload, teorie psychologická by pravděpodobně byla spíše nakloněná k optimistickému pohledu a byla by ochotná uznat stejnou identitu uploadu, a nakonec zastánci teorie nejbližšího následovníka by pro uznání identity potřebovali vědět, jestli se jedná o destruktivní uploadování, pokud tomu tak je, tak nejbližší následovník by byl

[156] CHALMERS J. David. The Singularity: A Philosophical Analysis. In AWRET Uziel ed. *The Singularity: Could artificial inteligence really out-think us (and would we want it to)?* [e-kniha]. 1. vyd. Exeter: Imprint Academic. 1. 12. 2016 [cit. 8. 2. 2018]. Dostupné z: <https://www.amazon.com/Singularity-Journal%20Consciousness-Studies-ebook/dp/B01N1N6KLZ/ref=mt_kindle?_encoding=UTF8&me>.
[157] Tamtéž, kap. 10.
[158] Tamtéžj kap. 10.

daný upload. Chalmers sám říká, že se přiklání spíše k psychologické teorii, ale není si příliš jistý.[159]

Abychom tedy vyřešili tento problém, tak nám Chalmers prezentuje dva argumenty, jeden z pohledu pesimistického a druhý z pohledu optimistického. Jako první prezentuje argument z nedestruktivního uploadování, který je pesimistický a stojí na představě nedestruktivního uploadování. Představme si, že včera byl Dave nahrán do počítače a originální mozek a tělo bylo nedotčeno. Takže tu tedy ve výsledku jsou dvě vědomé osoby: BioDave a DigiDave. Přičemž BioDave bude pochopitelně tvrdit, že on je původním systémem a DigiDave je pouze kopie. Zároveň by nám asi přišlo přirozené přiznat některá práva spíše BioDaveovi než DigiDaveovi, jako například vlastnictví osobních věcí a další.[160]

Takže bychom spíše řekli, že BioDave je skutečný Dave a DigiDave není. To tedy znamená, že v případě nedestruktivního uploadování, daný upload nezískává identitu toho, kdo byl nahrán. Ten, který se pak přiklání k pesimistickému pohledu, by naopak mohl říct, že pokud v tomto případě nepřisoudíme identitu DigiDaveovi, proč bychom tak měli učinit v případě destruktivního uploadování? Formálně by pak tento argument vypadal takto[161]:

1. Při nedestruktivním uploadování, DigiDave není Dave
2. Pokud při nedestruktivním uploadování DigiDave není Dave, tak při destruktivním uploadu DigiDave není Dave

Závěr: Při destruktivním uploadování DigiDave nebude Dave

Existuje několik způsobů, jak tomuto argumentu odolat, můžeme například odmítnout druhou premisu. Když se držíme teorie nebližšího následovníka, můžeme pak říci to, že dokud BioDave existuje, tak je následovníkem a pak DigiDave je jeho následovník. Nebo můžeme vzdorovat první premise, a uznat, že DigiDave je také identický s BioDavem a stejně tak identický s původním Davem. Pak je tu i třetí možnost, kde bychom mohli uploadování nazvat případem štěpení. Hlavním příkladem štěpení je oddělení levé a pravé hemisféry do rozdílných těl, kde následně jednotlivé poloviny

[159] Tamtéž, kap. 10
[160] Tamtéž, kap. 10.
[161] Tamtéž, kap. 10.

normálně žijí v jiných tělech a prokazují znaky originálu. Přestože nám může přijít nepříjemné uznat, že jsou oba vzniklé systémy identické, někteří by mohli apelovat, že jsou na stejné úrovni. To samé bychom mohli učinit v situaci Davea, kde bychom řekli, že jak BioDave, tak i DigiDave jsou následovníky Davea.[162]

Pak následuje argument z postupného uploadování. Zde si představme, že 1% Davea je vyměněno izomorfním silikonovým obvodem. A každý další měsíc znovu. Po 100 měsících je celý systém vyměněn. Můžeme si pak přestavit, že funkční izomorfismus daného systému mu pak zaručuje vědomí. Řekněme si tedy, že $Dave_n$ je systém po n měsících. Můžeme se pak ptát, jestli $Dave_1$ je stále Dave? A co $Dave_2$ a $Dave_3$ a tak dále? Nakonec je tu $Dave_{100}$ neboli plně vyměněný systém. Je $Dave_{100}$ stále Dave? Podle Chalmerse, by aspoň bylo přirozené si to myslet. Argument by pak vypadal takhle[163]:

1. Pro všechny n <100 je $Dave_{n+1}$ stejný jako Dave
2. Pokud pro všechny n <100 je $Dave_{n+1}$ stejný jako Dave, tak $Dave_{100}$ je stejný jako Dave

Závěr: $Dave_{100}$ je stejný jako Dave

Na první pohled je podle Chalmerse těžké odmítnout druhou premisu, protože v podstatě jde o příklad a=b a b=c, což následně vede k a=c. To samé platí pro první premisu, protože je těžké vidět, jak 1 % mění identitu dané osoby, a v případě námitek budeme opakovat experiment, ale tentokrát budeme vyměňovat menší části. Pokud bude opozice stejně vzdorovat, bude muset nakonec přiznat, že změna jednoho neuronu mění identitu dané osoby, a to je závěr, který je obtížné uznat. Přesto by šlo argumentu vzdorovat, například tak, že bychom mohli tvrdit, že jde o sorites paradox neboli argument šikmé plochy. Opozice by také mohla tvrdit, že identita přichází ve stupních a že existují těžko rozhodnutelné případy, to samé by mohlo být řečeno o přežití dané osoby. Jako příklad řekneme, že $Dave_n$ přežívá z velké míry jako $Dave_{n+1}$, ale z menší míry než ostatní systémy. Podle Chalmerse je ale těžké uznat, že Dave postupně umírá místo toho, že stále existuje. Pokud bychom tuto tezi přijmuli, co si počít se skutečností,

[162] Tamtéž, kap. 10.
[163] Tamtéž, kap. 10.

že hmota v našem mozku se v průběhu času mění, a to vede k našemu postupnému umírání.[164]

Chalmersovi v tomto případě přijde dost možné, že originální systém přežívá za předpokladů izomorfismu daného systému. Podle něj ze stejného důvodu jako přežíváme my napříč časem. Jakmile tedy přijmeme skutečnost, že postupné uploadování mezi několika lety zachovává identitu, tak pak stačí urychlit proces. Představme si tedy, že hmota mozku Davea je vyměňována v průběhu několika hodin. Pokud uznáme přežití v prvním případě, tak ho musíme uznat i zde. Abychom ještě posilnili tuto tezi, řekněme si, že při postupném uploadování tok vědomí pokračuje od momentu k momentu. Pokud není přerušeno, tak tok bude pokračovat během celého procesu, což znamená, že celým procesem prochází pouze jeden tok vědomí. Takže na jeho konci máme tu samou osobu jako na začátku. Je to právě tohle pokračování jednoho a toho samého toku vědomí, které Chalmersovi přijde jako nejsilnější důkaz pro přežití jedné a té samé osoby.[165]

Pokračujme ve zrychlování procesu uploadování, z hodin na minuty a z minut na sekundy. Přežije daná osoba? Chalmersovi přijde, že pokud přežívá v rámci hodin, proč nemůže přežit v rámci sekund? Znovu opakuje, že pokud přijmeme původní tezi, tak musíme přijmout i tuto. Čím rychleji jdeme, tak se více přibližujeme k okamžitému destruktivnímu uploadování, kde vyměníme celý mozek naráz. Je možné, že překročení tohoto limitu něco mění, ale šlo by formulovat argument, který stojí na tom předchozím, protože jediný rozdíl tkví v tom, že jedno uploadování je postupné a druhé okamžité. Toto můžeme chápat jako optimistický pohled na destruktivní uploadování:[166]

1. Dave přežívá jako $Dave_{100}$ při postupném uploadování
2. Pokud Dave přežívá jako $Dave_{100}$ při postupném uploadování, tak Dave přežívá jako DigiDave při okamžitém uploadování

Závěr: Dave přežívá jako DigiDave při okamžitém uploadování

[164] Tamtéž, kap. 10.
[165] Tamtéž, kap. 10.
[166] Tamtéž, kap. 10.

Opozice by mohla odporovat tím, že odmítne první premisu způsoby, o kterých jsme předtím mluvili. Je také možné odporovat druhé premise tvrzením, že přechodné systémy jsou důležité pro přenos identity z jednoho systému na ten druhý. Můžeme říci, že tu máme silné argumenty, jak pro optimistický pohled, tak i pro ten pesimistický. Chalmers tedy spoléhá na postoupné uploadování, o kterém si myslí, že si můžeme být nejvíce jistí, že přežijeme. Co když ale nebudeme mít tuto možnost? Jak si můžeme být jistí u ostatních případů?[167]

Proto se Chalmers zabývá těmi dalšími případy, prvním je případ uploadování po zachování mozku. Představme si, že po smrti si zachováme naše mozky tím způsobem, že je třeba zmrazíme nebo i jiným způsobem, aby ho někdo pak později reaktivoval a použil k uploadování, nebo aby ho někdo uploadoval přímo. Pokud tedy po takovém zachování v mozku zůstává dostatek informací, je výsledný upload tou danou osobou? Pokud jde o případ reaktivace, tak se to podobá tomu, kdy se osoba vzbudí po dlouhém kómatu a následně pak můžeme mozek uploadovat. Co ale s případem, kdy nikdo mozek nereaktivuje, ale přímo uploaduje informace v něm? Zde záleží na vašem postoji k destruktivnímu a nedestruktivnímu uploadování. Optimista v těchto věcech by asi měl kladný postoj k tomuto uploadování.[168]

Pak tu máme případ rekonstruktivního uploadování, zde znovu postavíme systém na základě záznamů, které o něm máme. Může jít o snímky mozku a jiné zdravotní záznamy, pozůstatky genetického materiálu, zvukové a obrazové nahrávky, spisy o dané osobě, a nakonec svědectví dalších osob o této osobě. Jakkoliv omezené by takové záznamy mohly být, je možné, že SI by mohly stačit k takové rekonstrukci. Zde se nakonec musíme zeptat na to samé, je tohle forma přežití? Chalmers se ptá, pokud takto sestavíme kopii Einsteina, bude to skutečně Einstein? Podle Chalmerse, pokud mluvíme o rekonstrukci ze snímku mozku, potom zde není moc rozdílů od případů, které jsme probírali předtím.[169]

Abychom mohli nějak uzavřít tento problém a perspektivy s ním spojené, tak nás Chalmers vede ke dvěma pohledům na přežití obecně, protože je to právě filosofická otázka přežití, která obklopuje celý problém identity při uploadování. Podle Chalmerse tu zaprvé máme pohled vzdáleného faktu. Ten tvrdí totéž jako teze vzdáleného faktu

[167] Tamtéž, kap. 10.
[168] Tamtéž, kap. 10.
[169] Tamtéž, kap. 10.

vědomí, tedy že tu existují vzdálené fakty, v tomto případě se ale týkají přežití. Mezi těmito pohledy pak podle Chalmerse existuje i propojenost. Víceméně tento pohled říká, že bychom mohli mít plnou znalost, jak fyzických, tak i mentálních faktů, to by ale stále nechalo otázku přežití otevřenou.

Druhý, deflační pohled tvrdí, že když řešíme otevřenou otázku přežití, tak nevědomky přisuzujeme fakta k přežití, které s ním nemají nic společného. Podle tohoto pohledu máme představu rajského přežití, podle kterého by bylo možné v ráji dosáhnout v podstatě idealizované formy přežití. Protože tato představa je mylná, tak nám stačí pouze mít plnou znalost fyzikálních a mentálních faktů k tomu, abychom vyřešili otázku přežití. Což představuje pravý opak pohledu vzdáleného faktu.[170]

Tyto pohledy pak mají různé postoje k problému uploadování, pokud pohled vzdáleného faktu má pravdu, tak si nemůžeme být příliš jisti, jak je to s destruktivním a rekonstruktivním uploadováním, ale postupné uploadování by asi mělo být pro přežití dostatečné. Pokud deflační pohled je ten správný, tak postupné uploadování je stejně dobré jako normální přežití, přičemž destruktivní a rekonstruktivní jsou také docela blízko. Chalmers sám nepřisuzuje správnost ani jednomu pohledu, ale je si poměrně jistý, že postupné uploadování je nejlepší a nejjistější cestou.[171]

Na jeho příkladu BioDavea a DigiDavea přímo reaguje Arkady Plotnitsky. Místo toho, aby řešil problém identity, se zabývá otázkou, jak by vůbec mohla taková situace nastat. Podobně jako u jeho protiargumentů vůči AI, i zde apeluje na komplexnost našich systému, a jestli je vůbec možné je takto replikovat. Daný příklad i trochu pozměňuje v tom smyslu, že se nesoustřeďuje pouze na uploadování, ale také zmiňuje situaci, že bychom vytvořili přesnou biologickou kopii, takže v tomto případě bychom měli dva BioDavevy. Podle jeho názoru si vzhledem ke komplexitě evolučních procesů, které nás stvořily Plotnisky myslí, že abychom vůbec byli schopni vytvořit takovou kopii, museli bychom převyprávět celý evoluční proces. A nejen to. Museli bychom zopakovat historii celého vesmíru, až do současného bodu v čase. Vzhledem ale k faktu, že podle Plotniského byly části historie vesmíru ovlivněny náhodnými fluktuacemi, tak takové převyprávění je prakticky nemožné.[172]

---

[170] Tamtéž, kap. 10.
[171] Tamtéž, kap. 10.
[172] PLOTNISKY Arkady. The Singularity Wager: A Response To David Chalmerse. In AWRET Uziel ed. *The Singularity: Could artificial inteligence really out-think us (and would we want it to)?* [e-kniha].

Chalmers má opět odezvu, jak pro Plotniského, tak také pro Denneta. Jeho odpověď na Plotniského je taková, že pro uploadování nepotřebujeme udělat úplně přesnou kopii dané osoby. Stačí nám pouze simulovat ty nejdůležitější děje v mozku, takové simulace by tak měly aspoň napodobovat ty samé v mozku, aby pak daná emulace jednala tak, jak bychom mohli jednat v dané situaci, nehledě na to, jestli bychom tak v takové situaci skutečně jednali. Tento proces je pak mnohem jednodušší než se snažit o dokonalou kopii. S Dennetem Chalmers nesouhlasí v tom, že by jeho postoj ovlivňoval jeho závěry, podle něj jeho postoj v této sekci hraje pouze malou roli.[173]

Co se týče toho, že Chalmers podle Denneta využívá argumenty materialistů typu A, tak si Chalmers myslí, že jde o jeho argument mizících a tančících qualia, že ty jsou důvodem, proč Dennet přemýšlí nad Chalmersovou pozicí. Představa těchto termínů možná zapadá do Dennetova funkcionalismu, ale podle Chalmerse mohou tyto termíny zapadat i když předpokládáme epistemickou díru mezi fyzickými procesy a vědomím. Podle Chalmerse můžeme tento argument použít v jakékoliv situaci. Pak se ale zaměřuje na problém zombie, problém, který podle něj Dennet odmítá, přestože je logicky možný. Názor Chalmerse je opačný, protože je zde logická možnost pravdivosti tohoto scénáře, a tak ho podle něj musíme brát vážně. Zde mimochodem zmiňuje i Kurzweila, který také napsal odpověď na Chalmersovy myšlenky, kde víceméně souhlasil s Chalmersem a existencí epistemické díry, přestože i Kurzweil odmítá příklad existence zombie, protože ho nemůžeme nikde empiricky pozorovat.[174]

1. vyd. Exeter: Imprint Academic. 1. 12. 2016 [cit. 8. 2. 2018]. Dostupné z: <https://www.amazon.com/Singularity-Journal%20Consciousness-Studies-ebook/dp/B01N1N6KLZ/ref=mt_kindle?_encoding=UTF8&me>.

[173]CHALMERS J. David. The Singularity: A Reply to Commentators In AWRET Uziel ed. *The Singularity: Could artificial inteligence really out-think us (and would we want it to)?* [e-kniha]. 1. vyd. Exeter: Imprint Academic. 1. 12. 2016 [cit. 8. 2. 2018]. Kapitola 5. Dostupné z: <https://www.amazon.com/Singularity-Journal%20Consciousness-Studies-ebook/dp/B01N1N6KLZ/ref=mt_kindle?_encoding=UTF8&me>.

[174] Tamtéž, kap. 4.

# 4. Závěr

V rámci této práce jsme si prošli teorii technologické singularity, ať už šlo o rozdílné koncepce, které o ní měli různí lidé, nebo cesty, které by mohly vést k SI, která by pak mohla sebou přinést singularitu. Stejně tak jsme si prošli možné formy SI a ukázali jsme si, na jakém přesvědčení víra v singularitu stojí, to jest na Transhumanismu. Důležitější však bylo, že jsme filosoficky analyzovali problém singularity a argumenty, které tvrdí, že singularita je možná a nejen to. Prozkoumali jsme také kritiku hnutí proponentů singularity a viděli jsme problémy, které toto hnutí a jejich víra může mít. Následně jsme pak analyzovali další problémy a úskalí, které jsou se singularitou spojeny.

Nehledě na to, jestli singularita je možná nebo ne, je jisté, že se s Chalmersem můžeme shodnout na jedné věci. Tou je skutečnost, že problematika singularity sebou přináší několik filosofických problémů a také přináší nový pohled na některé otázky, které už byly řešeny předtím.[175] Například je tu problém SI, tedy problém představit si, jak by mohla vypadat inteligence, která je více pokročilá než my a společně s tím otázka, jak bychom k ní mohli dojít. Podle našeho uvážení tato a i ty ostatní otázky a problémy stojí za prozkoumání, protože jednoho dne možná budeme potřebovat odpovědi. Je však také možné, že ne.

[175] CHALMERS J. David. The Singularity: A Philosophical Analysis. In AWRET Uziel ed. *The Singularity: Could artificial inteligence really out-think us (and would we want it to)?* [e-kniha]. 1. vyd. Exeter: Imprint Academic. 1. 12. 2016 [cit. 8. 2. 2018]. Kapitola 1. Dostupné z: <https://www.amazon.com/Singularity-Journal%20Consciousness-Studies-ebook/dp/B01N1N6KLZ/ref=mt_kindle?_encoding=UTF8&me>.

# Seznam použitých pramenů

## Literatura


1. BOSTROM Nick. *Superinteligence. Paths, Dangers, Strategies.* 1. vyd. Oxford: Oxford University Press, 2014. ISBN 978-0-19-967811-2.
2. BRINGSJORD Selmer, BRINGSJORD Alexander, BELLO Paul. Belief in The Singularity is Fideistic In EDEN H. Ammon, MOOR H. James, SØRAKER H. Johnny, STEINHART Eric (eds). *Singularity Hypotheses. A Scientific and Philosophical assesment.* 1. vyd. New York, London, Dordrecht: Springer, 2012. s. 395–408 ISBN 978-3-642-32559-5.
3. GOOD J. Irving. *Speculations concerning the first ultraintelligent machine*. In ALT L. FRANZ, RUBINOFF MORRIS (eds.). *Advances in Computers Volume 6.* 1. vyd. London: Academic Press, 1965. s. 33–88
4. CHISHOLM M. Roderick. *Theory of Knowledge.* 3. vyd. Englewood Cliffs: Prentice-Hall, 1989. ISBN 0-13-914185-5.
5. KURWEIL Ray. *The Singularity Is Near. When Humans Transcend Biology*. 1. vyd. New York : Viking Adult, 2005. ISBN 0-670-03384-7.
6. MUEHLHAUSER, Luke, and HELM Louie. The Singularity and Machine Ethics. In EDEN H. Ammon, MOOR H. James, SØRAKER H. Johnny, STEINHART Eric (eds). *Singularity Hypotheses. A Scientific and Philosophical assesment.* 1. vyd. New York, London, Dordrecht: Springer, 2012. s. 101–125. ISBN 978-3-642-32559-5.
7. NEUMANN V. John. *The Computer and the Brain*. 3. vyd. New Haven: Yale university Press, 2012. ISBN 978-0-300-18111-1.
8. OMOHUNDRO Steve. Rational Artificial Inteligence for greater good. In EDEN H. Ammon, MOOR H. James, SØRAKER H. Johnny, STEINHART Eric (eds*). Singularity Hypotheses. A Scientific and Philosophical assesment.* 1. vyd. New York, London, Dordrecht: Springer, 2012. s. 161–179. ISBN 978-3-642-32559-5.
9. PROUDFOOT Diane. Software Immortals: Science or Faith? In EDEN H. Ammon, MOOR H. James, SØRAKER H. Johnny, STEINHART Eric (eds). Singularity Hypotheses. A Scientific and Philosophical assesment. 1. vyd. New York, London, Dordrecht: Springer, 2012. s. 367–393 ISBN 978-3-642-32559-5.
10. YUDKOWSKI, Eliazer. Friendly Artificial Inteligence. In EDEN H. Ammon, MOOR H. James, SØRAKER H. Johnny, STEINHART Eric (eds). *Singularity Hypotheses. A Scientific and Philosophical assesment.* 1. vyd. New York, London, Dordrecht: Springer, 2012. s. 181–193. ISBN 978-3-642-32559-5.


## Elektronické Zdroje


1. About MIRI. In *MIRI (Machine Inteligence Research Institute)* [online]. Berkeley: Machine Inteligence Research Institute [cit. 12. 2. 2018]. Dostupné z: <https://intelligence.org/about/>.

2. BOSTROM, Nick. *Letter from Utopia.* [online]. Verze 1.9. 2010. [cit. 8. 2. 2018]. Dostupné z: <https://nickbostrom.com/utopia.pdf>.
3. BOSTROM, Nick. *Transhumanist FAQ: A General Introduction.* [online]. Verze 2.1 Los Angeles: World Transhumanist Associacion (Humanity+). 2003 [cit. 9. 2. 2018]. Dostupné z: <https://nickbostrom.com/views/transhumanist.pdf>.
4. BOSTROM Nick, SHULMAN Carl. How Hard is Artificial Inteligence? Evolutionary Arguments and Selection Effects. In AWRET Uziel ed. *The Singularity: Could artificial inteligence really out-think us (and would we want it to)?* [e-kniha]. 1. vyd. Exeter: Imprint Academic. 1. 12. 2016 [cit. 8. 2. 2018]. Dostupné z: <https://www.amazon.com/Singularity-Journal%20Consciousness-Studies-ebook/dp/B01N1N6KLZ/ref=mt_kindle?_encoding=UTF8&me>.
5. BRINGSJORD Selmer. Belief in Singularity is Logically Brittle In AWRET Uziel ed. *The Singularity: Could artificial inteligence really out-think us (and would we want it to)?* [e-kniha]. 1. vyd. Exeter: Imprint Academic. 1. 12. 2016 [cit. 8. 2. 2018]. Dostupné z: <https://www.amazon.com/Singularity-Journal%20Consciousness-Studies-ebook/dp/B01N1N6KLZ/ref=mt_kindle?_encoding=UTF8&me>.
6. DENNET Daniel. Mystery of David Chalmerse. In AWRET Uziel ed. *The Singularity: Could artificial inteligence really out-think us (and would we want it to)?* [e-kniha]. 1. vyd. Exeter: Imprint Academic. 1. 12. 2016 [cit. 8. 2. 2018]. Dostupné z: <https://www.amazon.com/Singularity-Journal%20Consciousness-Studies-ebook/dp/B01N1N6KLZ/ref=mt_kindle?_encoding=UTF8&me>.
7. HAMKINS D. Joel, LEWIS Andy. *Infinite Time Turing Machines.* [online]. Ithaca (NY): Cornell University Library. 21. 8. 1998. Aktualizováno 1. 2. 2008. [cit. 28. 3. 2018]. Dostupné z: <https://arxiv.org/pdf/math/9808093.pdf>.
8. CHALMERS J. David. The Singularity: A Philosophical Analysis. In AWRET Uziel ed. *The Singularity: Could artificial inteligence really out-think us (and would we want it to)?* [e-kniha]. 1. vyd. Exeter: Imprint Academic. 1. 12. 2016 [cit. 8. 2. 2018]. Dostupné z: <https://www.amazon.com/Singularity-Journal%20Consciousness-Studies-ebook/dp/B01N1N6KLZ/ref=mt_kindle?_encoding=UTF8&me>.
9. CHALMERS J. David. The Singularity: A Reply to Commentators In AWRET Uziel ed. *The Singularity: Could artificial inteligence really out-think us (and would we want it to)?* [e-kniha]. 1. vyd. Exeter: Imprint Academic. 1. 12. 2016 [cit. 8. 2. 2018]. Dostupné z: <https://www.amazon.com/Singularity-Journal%20Consciousness-Studies-ebook/dp/B01N1N6KLZ/ref=mt_kindle?_encoding=UTF8&me>.
10. MCDERMOTT Drew. Response to ‚The Singularity' by David Chalmers In AWRET Uziel ed. *The Singularity: Could artificial inteligence really out-think us (and would we want it to)?* [e-kniha]. 1. vyd. Exeter: Imprint Academic. 1. 12. 2016 [cit. 8. 2. 2018]. Dostupné z: <https://www.amazon.com/Singularity-Journal%20Consciousness-Studies-ebook/dp/B01N1N6KLZ/ref=mt_kindle?_encoding=UTF8&me>.

11. MUEHLHAUSER Luke. We are now „Machine Inteligence Research Institute“ (MIRI). In *MIRI (Machine Inteligence Research Institute)* [online] Berkeley: Machine Inteligence Research Institute, 30. 1. 2013 [cit. 12. 2. 2018] Dostupné z: <https://intelligence.org/about/>.
12. PLOTNISKY Arkady. The Singularity Wager: A Response To David Chalmerse. In AWRET Uziel ed. *The Singularity: Could artificial inteligence really out-think us (and would we want it to)?* [e-kniha]. 1. vyd. Exeter: Imprint Academic. 1. 12. 2016 [cit. 8. 2. 2018]. Dostupné z: <https://www.amazon.com/Singularity-Journal%20Consciousness-Studies-ebook/dp/B01N1N6KLZ/ref=mt_kindle?_encoding=UTF8&me>.
13. PRINZ Jesse. Singularity and inevitable Doom. In AWRET Uziel ed. *The Singularity: Could artificial inteligence really out-think us (and would we want it to)?* [e-kniha]. 1. vyd. Exeter: Imprint Academic. 1. 12. 2016 [cit. 8. 2. 2018]. Dostupné z: <https://www.amazon.com/Singularity-Journal%20Consciousness-Studies-ebook/dp/B01N1N6KLZ/ref=mt_kindle?_encoding=UTF8&me>.
14. SHANAHAN Murray. Satori Before Singularity. In AWRET Uziel ed. *The Singularity: Could artificial inteligence really out-think us (and would we want it to)?* [e-kniha]. 1. vyd. Exeter: Imprint Academic. 1. 12. 2016 [cit. 8. 2. 2018]. Dostupné z: <https://www.amazon.com/Singularity-Journal%20Consciousness-Studies-ebook/dp/B01N1N6KLZ/ref=mt_kindle?_encoding=UTF8&me>.
15. *Singularity Institute Strategic Plan.* [online]. Berkeley: Machine Inteligence Research Institute (Singularity Institute), srpen 2011 [cit. 12. 2. 2018]. Dostupné z: <https://intelligence.org/files/strategicplan2011.pdf>.
16. Singularity Summit. In *MIRI (Machine Inteligence Research Institute)* [online]. Berkeley: Machine Inteligence Research Institute [cit. 12. 2. 2018]. Dostupné z: <https://intelligence.org/singularitysummit/>.
17. SINGULARITY UNIVERSITY ACQUIRES SINGULARITY SUMMIT. In *Singularityu.org* [online]. San Jose: Singularity University, © 2018 [cit. 12. 2. 2018]. Dostupné z: <https://singularityu.org/2012/12/09/singularity-university-acquires-the-singularity-summit/>.
18. TURING, Alan. COMPUTER MACHINERY AND INTELLIGENCE. In *Mind* [online]. Oxford: Oxford University Press. 1. 11. 1950, Roč. LIX, č. 236, s. 456 [cit. 8. 2. 2018]. ISSN 1460-2113 Dostupné z: <https://doi.org/10.1093/mind/LIX.236.433>.
19. TURING, Alan. On Computable Numbers, with an Application to the Entscheidungsproblem. In *Proceedings of the London Mathematical Society* [online]. London: London Mathematical Society. 1. 1. 1937, Roč. s2-42, č. 1., s. 230–265 [cit. 8. 2. 2018]. ISSN 1460-244X. Dostupné z: <https://doi.org/10.1112/plms/s2-42.1.230>.
20. ULAM Stanislaw. John Von Neumann 1903 – 1957. In *Bulletin of the American Mathematical Society.* [online]. Providence: American Mathematical Society. 1958. Roč. 64, č. 3. s. 5 [cit. 26. 2. 2018]. ISSN 1088-9485 Dostupné z: <http://www.ams.org/journals/bull/1958-64-03/S0002-9904-1958-10189-5/S0002-9904-1958-10189-5.pdf>.

21. VINGE Vernor. First Word. In *OMNI* [e-kniha]. Englewood: Jerrick Publishing, 17, 2. 2016, Roč. 5, č. 4, s. 10 [cit. 12. 2. 2018]. Dostupné z: <https://www.amazon.com/OMNI-Magazine-January-1983-ebook/dp/B01BXG1M2E/ref=sr_1_1?s=books&ie=UTF8&qid=1520446761&sr=1->.
22. VINGE Vernon. The coming technological singularity: How to survive in the post-Human era. In *NASA, Lewis Research Center. Vision 21: Interdisciplinary Science and Engineering in the Era of Cyberspace* [online]. Cleveland: NASA, Lewis Research Center. 1. 12. 1993 [cit. 23. 12. 2017]. Dostupné z: <https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19940022856.pdf>.
23. What is SU?. In *Singularity university.* [online]. San Jose: Singularity University, © 2017 [cit. 12. 2. 2018]. Dostupné z: <https://su.org/about/>.
24. YUDKOWSKI Eliezer. Three Major singularity schools. In *MIRI (Machine Inteligence Research Institute)* [online]. 30. 8. 2007 [cit. 23. 12. 2017]. Dostupné z: <https://intelligence.org/2007/09/30/three-major-singularity-schools/>.

# Seznam obrázků

Obrázek č. 1 – Série na sebe navazujících S křivek ..... 11