Superinteligence a problém kontroly: Skutečný problém nebo pseudo-problém?

Jaroslav Malík

Filozofická fakulta
Univerzita Hradec Králové
Rokitanského 62, 500 03 Hradec Králové
jaroslav.malik@uhk.cz

V tomto článku se zabývám konceptem SI (superinteligence) a s ní spojeným problémem kontroly. Podle určité skupiny teoretiků umělé inteligence stojíme na prahu události, která může radikálně změnit povahu technologického pokroku a lidské společnosti obecně. Touto událostí má být takzvaná technologická singularita, která je často spojována se vznikem první větší než lidské inteligence. Lidé jako Nick Bostrom varují před nebezpečím, které pro nás vznik SI znamená, a upozorňují, že musíme co nejdříve najít metody kontroly této inteligence. Podle Bostroma a dalších nebezpečí SI vyplývá z její povahy. Já na jedné straně zkoumám, jak SI může vzniknout, a na druhé straně posuzuji smysluplnost problému kontroly. Vznik SI vyplývá ze vzniku umělé inteligence. Proto podstatnou část textu věnuji argumentům pro vznik umělé inteligence. Ukazuji, jak se Bostrom a další nechávají unést jedním problematickým argumentem. Dále podrobuji jejich pozici klasické kritice umělé inteligence. Prezentuji, jak stále lpí na problematických domněnkách svých předchůdců. Zejména soustřeďuji svoji kritiku na tvrzení, že SI bude mít jeden konečný cíl, který bude interpretovat. Toto tvrzení označuji jako antitetické, a to představě, že SI bude obecnou inteligencí. V závěru ze své argumentace vyvozuji, že při formulaci problému kontroly teoretikové umělé inteligence zaměňují dva jiné různé problémy „kontroly“.

Klíčová slova: technologická singularita, superinteligence, umělá inteligence, emulace celého mozku, problém kontroly, tělesnost

1. Úvod

Představte si, tak jako Nick Bostrom, že jednou postavíme inteligentní stroj na výrobu sponek. Inteligentním míním to, že má v sobě zabudovanou funkční umělou inteligenci. Tomuto stroji dáme velice prostý úkol: „Vyrob tolik sponek, kolik jen dokážeš." Náš stroj na začátku plní svůj úkol více než dobře, díky učícím se procesům, které jsou v něm zabudovány, nejenže vyrábí sponky stále stejnou rychlostí, ale dokonce se postupně zlepšuje. Náš stroj se nejdříve pouze učí lépe provádět standardní proces výroby, ale netrvá to dlouho, než stroj začne nacházet inovace, které radikálně změní výrobu sponek. Nejdříve se jedná o pouze malé změny, možná přijde s novou technikou, nebo vynalezne nový přístroj pro splnění svého úkolu. Jeho objevy jsou však čím dál tím radikálnější, až nakonec přijde na „zázračnou" metodu, která našemu stroji umožní přeměnit cokoliv na sponky. V tuto chvíli zaregistrujeme první problémy. Náš stroj nejdříve začne přeměňovat vše ve své blízkosti, ale stejně jako optimalizoval své předchozí metody, se mu nakonec daří přeměňovat věci v širším prostoru. Lidé jsou samozřejmě zděšeni a pokusí se tento stroj v jeho snaze zastavit. Bohužel jejich snahy se ukážou být marné, náš nyní již superinteligentní stroj počítal s jejich reakcí, takže měl obranné prostředky. Dost brzy je celá planeta Země přeměněna na sponky a po ní následuje i celá sluneční soustava. Bostromův stroj se poté vydává dál, pokračovat ve svém úkolu po celé galaxii.[1]

Pravděpodobně se již každý z nás setkal s myšlenkou, že až se nám jednou podaří vytvořit umělou inteligenci (dále už pouze jako UI), tak nám dost brzy bude hrozit, že se UI vzbouří proti nám. Koneckonců jedná se o dost známé sci-fi klišé. Tyto myšlenky mají svoje kořeny už v samotných začátcích vývoje UI a pravděpodobně od první chvíle, kdy se někdo pokusil něco takového jako UI stvořit, musel začít uvažovat o takové myšlence. V oboru výzkumu UI se můžeme setkat s dlouhou historií techno-optimismu v souvislosti se vznikem UI. Lidé jako Marvin Minsky nebo Allen Newell už v 60. letech předpokládali, že UI je už nejen na obzoru, ale že nás bude schopná překonat.[2] Tato problematika

1 Bostrom (2003, 2012, 2014, s. 193).
2 Viz Minsky (1968) a Simon & Newell (1958).

je v současné době už široce diskutovaným tématem. Pojem technologické singularity do této problematiky však přináší nepříliš diskutovanou dimenzi. Tou je problém superinteligence. Podle teoretiků singularity (dále budu tuto skupinu označovat jako singularitáni) existuje možnost, že až vyvineme umělou inteligenci, která bude dostatečně sofistikovaná, tak brzy potom tato inteligence vytvoří ještě lepší intelekt. Pokud tato událost nastane, tak její následky budou podle singularitánů nedozírné. Varují před katastrofickými scénáři, ve kterých nás superinteligence (dále jako SI) zničí. Z toho pro nás vyplývají dvě otázky: Může skutečně vzniknout superinteligence? A pokud ano, vedlo by to k problémům?

Na úvod bude velmi důležité definovat, co míníme pojmy jako inteligence, umělá inteligence (UI) a superinteligence (SI). V tomto článku se budu držet definice inteligence, která je v rámci oboru UI obecně přijímána. Představa strojové inteligence konverguje k chápání inteligence jako optimálního plnění cílů napříč různými prostředími. Takže se budeme držet definice: „Inteligence je mírou agentovy schopnosti dosahovat cílů v širokém spektru různých prostředí.“[3] Pojem UI je obecně složité definovat. Russell a Norvig konstatují, že v historii UI můžeme identifikovat několik přístupů k jejímu chápání. Někteří preferují definovat inteligenci UI v měřítku lidského intelektu, jiní raději mluví o inteligenci jako o obecné racionalitě. Dále se někteří soustřeďují na to, aby se UI chovala inteligentně, a jiní dávají důraz i na to, aby UI byla schopná myšlení. To vytváří dvě dimenze, skrze které můžeme chápat přístupy k UI: myšlení/chování a lidské/racionální.[4] Singularitáni dle mého názoru UI chápou jako inteligenci, která se chová racionálně.[5] Proto v rámci tohoto textu budeme takto vnímat jejich výroky. SI budeme v textu chápat jako obecnou inteligenci, která je schopná dosahovat mnoha různých cílů. Bostrom SI definuje jako: „Jakýkoliv intelekt, který lidské kognitivní výkony dalece překonává prakticky ve

3 Legg (2008). Na této definici staví Bostrom (2014) a Muehlhauser & Helm (2012). Překlad vychází z Bostrom (2014, s. 405).

4 Russel & Norvig (2020, s. 1–4).

5 K této charakterizaci mě vedou Muehlhauser & Helm (2012), kteří popisují SI jako super optimalizátora, Yudkowski (2008), který chápe mysl jako optimalizační proces, Omohundro (2008, 2012), který se soustřeďuje na racionalitu SI, a Bostrom (2014, s. 47), který konstatuje, že SI nemusí být vědomá. Všichni také zastávají pozici, že SI bude inteligence odlišná od lidské.

všech relevantních oblastech.“[6] Z toho vyplývá, že budeme kompetence SI chápat v kontextu lidských kompetencí. Cokoliv člověk dokáže, SI zvládne také a lépe. Bostrom rozlišuje několik různých forem SI, které můžeme chápat jako odlišné systémy pro dosažení SI. Takže například rychlostní SI je inteligence na stejné úrovni jako ta naše, ale mnohem rychlejší. Tento typ se však chápe jako slabá forma SI. Další dva možné systémy jsou kolektivní a kvalitní SI. První dosahuje své úrovně intelektu skrze úzkou propojenost několika inteligencí a druhou můžeme chápat jako individuální intelekt, který převyšuje naše schopnosti s alespoň stejnou rychlostí jako náš mozek. Tyto různé SI také mohou vzniknout různými způsoby. Nehledě na typ, můžeme chápat všechny tyto SI jako součást hlavního problému.

Co je hlavní problém SI? Bostromův sponkový stroj je dobrý ilustrující příklad. V tomto myšlenkovém experimentu byl stroji dán velice prostý povel. Tento povel však skončil tím, že naše planeta byla přeměněna na hromadu sponek. Kde se stala chyba? Co zapříčinilo jednání tohoto stroje? Bostromův stroj a další možné SI musíme podle singularitánů chápat jako bytosti, které mají dvě zásadní vlastnosti. Zaprvé mají schopnost neuvěřitelně tvarovat realitu okolo sebe, a to jim umožňuje řešit problémy způsoby, které by člověka nenapadly. To vyplývá z vysokého intelektu SI, který jim teoreticky umožňuje nečekané objevy. Druhou jejich vlastností je to, co singularitáni nazývají „doslovnost“. Tím míní fakt, že jediné, co SI rozpoznává, jsou přesně specifikované hodnoty a pravidla. SI může naše výroky interpretovat pouze na základě těchto pravidel. To vyvolává potíže, protože je obtížné přesně definovat a specifikovat mnoho lidských hodnot. Kvůli této skutečnosti nemůžeme zaručit to, že SI bude naše pokyny interpretovat stejně jako člověk. Protože SI dost možná nebude vnímat jemné nuance našich výroků.[7] Při výkladu tohoto a dalších podobných příkladů se situace prezentuje jako problém kontroly. Proto musíme přemýšlet o tom, jak bychom zajistili kontrolu nad jednáním SI a tím se vyhnuli zkáze.

6 Bostrom (2014, s. 47). Podobné definice také lze najít v Bostrom (2003, 2006).

7 Muehlhauser & Helm (2012, s. 105–106). Doslovnost v originále jako Literalness, kterou definují následovně: „The [SI] recognizes only precise specifications of rules and values, acting in ways that violate what feel like ‚common sense‘ to humans, and in ways that fail to respect the subtlety of human values.“

Z tohoto důvodu bijí singularitáni na poplach. Pro ně tento problém vyplývá z povahy, kterou SI bude mít. Bostrom a další zastávají takzvanou ortogonální tezi, která říká: „Inteligence a konečné cíle jsou navzájem ortogonální: Víceméně jakákoliv inteligenční úroveň může v principu být spojená s víceméně jakýmikoliv konečnými cíli.“[8] Pro nás kritickým bude jejich tvrzení, že SI musíme chápat jako inteligenci, která bude mít jeden finální cíl. Tento fakt vyplývá z toho, že SI bude UI. Její cíl bude vycházet z jejího naprogramování. Bostrom a Omohundro v tomto bodě vychází z praktik vývoje UI. Když se vytváří program UI, je určeno, jaký problém bude řešit. Bostrom předpokládá, že SI bude výsledkem takového projektu UI. To, co však činí SI problematickou, není jenom tento samotný cíl, ale také hodnoty, které z něho vyplývají. Z tohoto faktu můžeme alespoň částečně předpovídat chování SI. Bostrom a další zastávají stanovisko, které nazývají tezí instrumentální konvergence, a ta říká:

> „Lze identifikovat několik instrumentálních hodnot, které jsou konvergentní v tom smyslu, že jejich uskutečnění by pro celou řadu konečných cílů a v celé řadě situací zvýšilo pravděpodobnost, že agent dosáhne svého cíle. Z toho plyne, že o jejich uskutečnění pravděpodobně bude usilovat široké spektrum inteligentních agentů nacházejících se v různých situacích.“[9]

Bostrom rozlišuje několik takových hodnot, jako je sebezáchova, integrita obsahu cíle, zdokonalování technologií, vylepšování kognice, získávání zdrojů. Nebezpečí SI neplyne pouze z faktu, že by si mohla špatně vyložit svůj konečný cíl, ale také z těchto konvergentních hodnot. SI by mohla mít skutečně neškodný cíl, ale tyto hodnoty by ji dostaly do konfliktu s námi.[10]

8 Bostrom (2014, s. 169). Viz i Bostrom (2012). Stejnou tezi zastává také Omohundro (2008, 2012, 2016).

9 Bostrom (2012 a 2014, s. 172). Stejnou tezi zastává Omohundro (2008, 2012, 2016).

10 Nejkritičtější z těchto hodnot, kterou musíme brát na vědomí, je integrita obsahu cíle. Podle Bostroma a Omohundry bude pro UI nejdůležitější splnění jejího cíle. Pokud by UI ztratila svůj cíl, tak by ho nemohla splnit, proto se bude snažit tento cíl zachovat. Z tohoto důvodu nemůžeme argumentovat, že SI by si mohla přeprogramovat svůj konečný cíl.

V tomto textu se budu zabývat problémem kontroly, přesněji se chci soustředit na domněnky, na kterých je vystavěn. Myslím, že pro zjednodušení bude nejlepší, pokud budeme problém kontroly (dále už jen jako PK) chápat jako formální argument:

(1) UI bude existovat.
(2) Pokud bude UI, tak vznikne SI (argument singularity)
(3) Protože SI je UI, bude mít jeden konečný cíl.
(4) Intelekt na jakékoliv úrovni, může mít jakýkoliv konečný cíl.
(5) SI může svůj konečný cíl interpretovat jinak, než jsme jej zamýšleli my.
(6) Z konečného cíle vyplývá několik konvergentních hodnot.
(7) Tyto hodnoty nemusí být v náš prospěch.
(8) Pokud SI špatně interpretuje svůj konečný cíl nebo získá pro nás škodlivé hodnoty, ztratíme nad ní kontrolu.

Závěr: Vznikem SI, nutně čelíme PK.

Tento článek se bude skládat ze dvou hlavních sekcí. V první se budu zabývat otázkou vzniku SI, specificky argumentem singularity Davida Chalmerse. PK může nastat pouze, pokud bude existovat SI, tudíž je důležité zhodnotit proces jejího vzniku. Budu zde posuzovat formalizaci argumentu singularity. Tento argument staví na tezích, ve kterých nalézám několik problémů. Také ukazuji, že způsoby vzniku SI mohou vést k různým inteligencím. To je podstatné pro konstituci PK, která je závislá na charakteru vzniklé inteligence. V druhé sekci zpochybním smysluplnost PK, protože singularitáni zakládají tento problém na pochybných domněnkách, které zdědili od svých předchůdců. Tento fakt je vede k formulování PK na základě konceptu SI, který je nekonzistentní s PK. Mým závěrem je, že pro singularitány existují pouze dvě cesty, kterými se mohou vydat, ale žádná z nich nemůže obsahovat PK v jeho současné formě.

2. Argument Singularity

Prvním, o kom můžeme říct, že se teoreticky zamýšlel nad konceptem technologické singularity[11], je Irving J. Good, který už v roce 1965 očekával, že během 20. století bude postaven ultra inteligentní stroj. Podle Gooda, by jeho vznik úplně změnil povahu naší společnosti z velmi prostého důvodu, byl by to poslední lidský vynález. Faktem této skutečnosti je podle něho to, že vznik tohoto stroje povede k tomu, čemu říká inteligenční exploze.[12] Vyplývá to z jeho definice ultra inteligentního stroje, který definuje jako stroj schopný překonat člověka v jakékoliv intelektuální aktivitě. To vede k prvnímu argumentu pro vznik singularity. Protože Good stanovuje design podobných strojů jako jednu z intelektuálních aktivit daných člověku, byl by ultra inteligentní stroj schopen stvořit ještě lepší stroj. Stačí sestrojit jenom o trochu chytřejší stroj, než jsme my, aby se podle Gooda odstartovala inteligenční exploze. Po této události by ultra inteligentní stroje stavěly ještě chytřejší stroje a člověk by byl zanechán daleko za nimi. Podstatou je to, že jakmile se nám lidem podaří vytvořit umělý intelekt, tak přestane být přirozeným limitem pokroku. Po takovém objevu se intelekt stane už pouze inženýrskou výzvou, kde už jde jen o to vytvořit intelekt, který je lepší než ten předchozí. Toto tvrzení je obecně přijímáno mezi singularitány. Good samozřejmě nebyl jediným kdo formuloval teorii singularity[13], ale v tomto klíčovém bodě se jednotlivé teorie neliší.

11 Good nebyl prvním, koho napadlo, že by technologický pokrok mohl mířit k singularitě. Byl však prvním, kdo poskytl teorii takové singularity. První zmínka o technologické singularitě pochází údajně od Johna von Neumanna: „One conversation centered on the ever accelerating progress of technology and changes in the mode of human life, which gives the appearance of approaching some essential singularity in the history of the race beyond which human affairs, as we know them, could not continue.“ Viz Ulam (1958).

12 Good (1965, s. 78).

13 Viz Vinge (1993) a Kurzweil (2005). Vinge si myslí, že po singularitě nemůžeme nic předpovídat, Kurzweil naopak tvrdí, že vlastně můžeme dále předpovídat vývoj po singularitě, protože se jedná o přirozený výsledek pokroku. Naše modely pokroku budou fungovat nadále i po singularitě, jediné, co se změní, bude rychlost pokroku. Spor mezi jednotlivými teoriemi tkví v tom, kolik toho můžeme předpokládat o singularitě a vývoji po ní. Takže na jedné straně máme Vingea jako úplného skeptika a na straně druhé Kurzweila jako úplného optimistu. Good vytváří střední cestu mezi nimi, protože zastává názor, že inteligenční exploze bude dále pokračovat. Pokud bych měl sám sebe vložit někam na toto měřítko, byl bych mezi Goodem a Vingem. Všechny naše modely by se asi nestaly nepoužitelnými, ale myslím si, že obtížnost předvídatelnosti tkví v tom hádat, které modely vydrží. Ani Goodova inteligenční exploze není úplnou jistotou. Velice zde záleží na tom, jak by nové intelekty vznikaly.

Chalmers, který formuluje argument pro singularitu, staví na konceptu inteligenční exploze.[14] Na začátek je zapotřebí vysvětlit terminologii, kterou Chalmers používá.[15] Ten používá zkratku UI pro označení jakékoliv umělé inteligence na stejné úrovni jako lidská inteligence. UI+ označuje první ultra-inteligentní stroj, který převyšuje nejvyšší lidské hranice. UI++ pak označuje SI, která vytváří náš diskutovaný problém. Takže, pokud bude existovat UI, vyplývá z toho, že bude existovat UI+ a následně UI++. Proto argument singularity vypadá takto:

(1) UI bude existovat
(2) Pokud bude UI, tak bude UI+
(3) Pokud bude UI+, tak bude UI++
Závěr: Bude existovat UI++

I když je snadné toto formulovat, samo o sobě je to nic neříkající. Jsou zde otázky ohledně vzniku této UI a procesu, který by z ní měl UI+ a posléze UI++ učinit. Aby tyto otázky zodpověděl, nabízí Chalmers teze, které podpírají jednotlivé premisy tohoto argumentu. Těmito tezemi a problémy s nimi spojenými se budu zabývat v následujících podkapitolách. Na začátku se musíme ptát, jak by UI mohla vzniknout. Chalmers nabízí dvě cesty ke vzniku UI.

2.1. Emulační teze

První cestu k jejímu vzniku Chalmers opírá o předpoklad takzvané ECM (*emulace celého mozku*). Jedná se o teoretickou proceduru, která by nám umožnila emulovat živý mozek. Koncept je takový, že na základě detailních snímků mozku bychom mohli postavit model mozku. Díky tomuto modelu bychom následně mohli simulovat činnost mozku na počítači. Taková simulace by teoreticky měla mít ty samé vlastnosti jako živý mozek. Výhodou této metody je, že nepotřebujeme vědět, jak mozek produkuje inteligenci. Problém spočívá ve vyspělosti daných

14 Chalmers (2016a).

15 Chalmers v originálním znění používá anglické zkratky jako AI (Artificial Intelligence) a WBE (Whole Brain Emulation). Vzhledem k českému znění mého článku jsem jeho terminologii sjednotil s mojí.

technologií. Potřebujeme snímek s dostatečně vysokým rozlišením, software, který zpracuje snímky a sestaví 3D model a nakonec hardware, který bude mít dost výpočetní síly pro simulaci mozku. Tyto technologie jsou dosažitelné, ale ne v blízké budoucnosti.[16] Funkčnost této procedury závisí na myšlence, že mozek je ve své podstatě kauzálním mechanismem, a tak ho můžeme chápat jako stroj. Pokud je mozek stroj, tak ho můžeme zkopírovat a tato kopie by měla mít ty samé vlastnosti jako originál. Tuto myšlenku pak můžeme shrnout do Chalmersova argumentu z emulace:

(1) Lidský mozek je stroj
(2) Budeme mít možnost emulace tohoto stroje
(3) Pokud emulujeme tento stroj, budeme mít UI
Závěr: Bude existovat UI

Navrhoval bych toto tvrzení interpretovat jako prostý apel na funkcionalistické představy o fungování lidského mozku. To, co premisa (1) míní, je předpoklad, že mozek je ve své podstatě systém, jehož procesy se řídí určitými zákonitostmi. Je to systém obsahující části, které mezi sebou interagují na základě těchto pravidel. Díky tomuto faktu tyto části hrají roli v plnění celkové funkce tohoto systému. Klíčovým závěrem funkcionalismu je, že podstata, která umožňuje mozku plnit jeho funkci, nespočívá v materiálním charakteru jeho částí, ale právě v jejich daných rolích. Premisa (2) stojí na tezi, že lze emulovat procesy ve velkém detailu, a tím emulovat jakýkoliv stroj simulací jeho procesů. Takže pokud by se nám podařilo vytvořit umělou náhradu některé části mozku, která by byla schopná zastoupit stejnou roli jako přirozená část, tak by mozek měl dále plnit svou funkci úplně stejně. Premisa (3) vyplývá z tvrzení, že pokud se nám podaří zreprodukovat vzorce lidského mozku, tak následně můžeme vytvořit UI na jejich základě.

Díky této interpretaci by šlo toto tvrzení chápat i v širším smyslu. Možná není zapotřebí emulovat mozek tak do hloubky, jak se ECM snaží. Je zde otázka, jestli potřebujeme kopírovat všechny mikro procesy mozku, nebo jestli bychom si postačili pouze s makroelementy jeho

16 Bostrom & Sandberg (2008), Bostrom (2014, s. 58–68).

architektury. V rámci této cesty k UI můžeme mít celou škálu různých přístupů, které napodobují některé části mozku, ale radikálně se liší v jiných ohledech. Možná UI vznikne díky konekcionistické architektuře, protože inteligenci umožňuje paralelnost mozku. To, co musíme emulovat, je architektura mozku, která je složená z velkého množství jednoduchých jednotek. Koncept je takový, že to, co umožňuje kognici, je interakce mezi těmito body. Taková architektura je pak schopná provádět několik výpočtů najednou, díky jejich distribuci napříč celou neurální sítí. Stejně je to s reprezentací jednotlivých funkcí, která je distribuována mezi jednotlivými body neurální sítě. Jak moc jsou tyto reprezentace distribuovány, se však může lišit. Můžeme například mít systémy, kde jeden bod reprezentuje jeden koncept, nebo systémy, kde jeden koncept reprezentuje skupina bodů.[17] Také je tu otázka, jestli bychom se měli soustřeďovat pouze na mozek, nebo jestli není také zapotřebí emulovat celé tělo. To je představa vtělené a situované UI.[18] Naše myšlení je většinou zakotveno ve světě, málokdy se zabýváme čistě abstraktními objekty. Je dost možné, že inteligenci umožňuje interakce s prostředím, které na nás vyvíjí tlak. Takovou zkušenost však můžeme mít pouze, pokud je inteligence nějakým způsobem vtělená. Z tohoto důvodu mnoho autorů zastává tezi, že inteligenci umožňuje právě tělesnost.[19] Tuto tezi podporuje i sám Chalmers.[20] Dané tělo nemusí být nutně podobné tomu našemu, ale musí být vybaveno tak, aby přijímalo vjemy z vnějšího prostředí a bylo schopné se vyrovnat s komplexitou vnějšího světa. Takže je několik možností toho, jak můžeme emulovat a napodobovat lidskou inteligenci. Vzhledem k těmto dalším možnostem emulace lidské inteligence je otázkou, proč Chalmers svazuje tento argument tak úzce k funkčnosti ECM.[21] Nejpravděpodobnějším důvodem se zdá být to, že volí nejjistější cestu k úspěchu. Přesto je ale nutné zvážit i alternativy.

17 Sun (2014).

18 Beer (2014).

19 Viz Clark (1997), Pfeifer & Bongard (2006), Varela, Thompson & Rosch (2016).

20 Clark & Chalmers (1998). Chalmers v rámci textu i sám upozorňuje, že SI by mohla mít rozšířenou mysl. Vzhledem k argumentům, které rozvíjím později v textu, si nejsem jist, jak může Chalmers chápat rozšířenou mysl jenom jako jednu z možných SI myslí. Já zastávám pozici, že SI bude muset být rozšířenou myslí, aby byla skutečně obecnou inteligencí. Dle mého názoru Chalmers tak koná, aby nechal otevřený prostor pro svoje argumenty.

21 Bostrom & Sandberg (2008), Bostrom (2014, s. 58–68).

Emulačnímu argumentu lze samozřejmě oponovat. Jediné, co lze skutečně zpochybnit je premisa (1). Pokud bychom přijali, že mozek je stroj, je nejasné, jak bychom mohli oponovat možnosti jeho emulace. Chalmers sám zmiňuje filozofy, jako jsou Hubert Dreyfus a Roger Penrose, kteří podle něj tvrdili, že činnost mozku nelze simulovat na žádném počítači. Nebo John Searle a Ned Block, kteří argumentovali, že i kdybychom simulovali lidský mozek, něco fundamentálního by mu chybělo, protože by pouze simuloval naše chování. Když však Chalmers nahlíží jejich námitky, mám dojem, že poněkud zjednodušuje jejich postoje. Vezměme si Blocka, jeho myšlenka je taková, že pokud stroj jedná inteligentně, tak tato inteligence nemusí nutně být jeho. Může být inteligentní, protože mu tento intelekt propůjčil jeho programátor. V takovém případě jeho inteligence není jeho vlastní. Takže se zde nejedná ani tak o to, že by tento počítač simuloval naše chování, on jedná jenom díky našemu konání, nemá žádnou vlastní vůli.[22] Také si nemyslím, že by Dreyfus argumentoval proti tomu, že mozek může být strojem. Jeho kritika spíše mířila tím směrem, jakým druhem kauzálního mechanismu skutečně jsme. Více si však nechám pro druhou sekci, kde se budu blíže zabývat zejména argumenty Searlovými a Dreyfusovými.

2.2. Evoluční teze

Cesta první se opírala o fakt, že my lidé máme inteligenci. Snažila se o to postavit UI na jejím základě. Druhá cesta argumentuje z existence intelektu obecně a vychází z faktu, že vznikl určitým způsobem. Je to míněno tak, že náš lidský intelekt vznikl jako důsledek přírodního výběru. Pokud takto mohla vzniknout jedna inteligence, proč by nemohla vzniknout další? To pak Chalmers nazývá argumentem z evoluce:

(1) Evoluce stvořila inteligenci
(2) Pokud evoluce stvořila inteligenci, tak můžeme stvořit UI
Závěr: Bude existovat UI

22 Block (1981).

Prima facie se toto stanovisko zdá být velice prosté, je zde však pár skrytých interpretací. Bostrom a Shulman[23] ve své reakci ukazují, že můžeme tento argument číst dvěma různými způsoby. První interpretace říká to, že pokud slepý a nevědomý proces dokázal stvořit inteligenci, tak vědomý designer toho také může dosáhnout. Tradičně by následoval i výrok, že designer může tento úkol splnit rychleji a lépe, protože se o to vědomě snaží. Pokud by toto byla pravda, tak by UI mohla mít i formu, která se radikálně odlišuje od dosud známých forem inteligence. UI by mohla vzniknout jiným než evolučním procesem a nemusela by ani sdílet s tím spojené vlastnosti. Z této představy vyplývá jeden klasický argument pro vznik UI. Tento argument stojí na analogii mezi stvořením intelektu a vynálezem letadel. Nám lidem se podařilo létat, ovšem neudělali jsme to tím stejným způsobem jako ptáci, místo toho jsme postavili letadla. Možná stejným stylem sestrojíme UI, která sice bude inteligentní, ale jiným způsobem.[24] To přináší problém, který mám s touto interpretací evoluční teze. Protože tato myšlenka svádí k názoru, že prostor možných inteligencí je větší, než si myslíme. Musíme se ptát, jak moc je let ptáků odlišný od letu letadel. Jak letadla, tak i ptáci se musí řídit zákony aerodynamiky. To samé musí platit pro inteligenci. Přestože zatím nevíme, co umožňuje inteligenci, musí i v tomto prostoru existovat zákonitosti. Je dost možné, že pokud sestrojíme UI, tak bude výtvorem vědomého designera. Tento fakt nás ale nemůže svádět k tomu si myslet, že jakoukoliv metodou můžeme stvořit inteligenci. Pokud se necháme tímto argumentem příliš unést, mohli bychom postulovat inteligence, které nejsou možné. V příští sekci ukážu, jak singularitáni do této pasti padají.

Druhá interpretace klade větší důraz na to, že náš intelekt vznikl díky evolučnímu procesu. Pokud takový proces dokázal stvořit intelekt, proč ho nenapodobit? Toto můžeme považovat za argument ve prospěch evolučních a genetických algoritmů, kterými vytvoříme proces podobný evoluci. Genetické algoritmy fungují tak, že generují omezené množství aktérů, kteří jsou vybaveni určitými schopnostmi. Této skupině je předložen cíl, a mají sami k řešení dojít. Jednotliví aktéři se chovají zcela náhodně, omezeni pouze danými parametry. Ze skupiny je vybrán ten

23 Bostrom & Shulman (2016).
24 Moravec (1976).

jedinec, který se splnění cíle přiblížil nejvíce. Na jeho základě je pak vygenerována další generace aktérů. Tento proces probíhá tak dlouho, dokud nevznikne aktér, který je schopen plnit daný cíl. Tato metoda je velice zdlouhavá, takže v praxi designer upravuje algoritmus tak, aby nezašel do slepé uličky. Tato interpretace ovšem stojí před problémem, jak náročné je simulovat evoluci. Bostrom a Shulman argumentují, že pokud bychom museli simulovat evoluční proces do každého detailu, tak bychom museli simulovat celé přírodní prostředí až do mikro-fyzikálních procesů. To by si vyžádalo výpočetní sílu, které bychom se nedočkali ani do konce století. Proto navrhují, že bychom měli simulovat pouze abstraktní prostředí.[25] Argumentoval bych, že tento fakt nás znovu musí vést k závěru, že výsledkem tohoto procesu musí být odlišný intelekt. Protože pokud takový intelekt vznikne v odlišném prostředí, mělo by se to promítnout na jeho struktuře.

Nehledě na interpretaci, tento argument stojí před problémem komplexnosti, který se promítá na několika úrovních. Obě interpretace spoléhají na to, že stvořit inteligenci není těžké. Otázkou je, jak skutečně složité je stvořit intelekt. Přírodě trvalo miliardy let, než vznikl intelekt na naší úrovni. Jak moc by mohl fakt, že nyní se úkolem zabývá skupina inteligentních designerů, zkrátit tuto dobu? Stejnou námitku lze aplikovat na snahu tento proces kopírovat. I když máme designera, který umělou evoluci navede na správnou cestu, mohli bychom čekat dost dlouhou dobu, než bychom se UI dočkali. To není příliš uspokojivý závěr pro inženýry UI a obzvláště ne pro singularitány. Proto oba tyto argumenty spoléhají na to, že stvořit inteligenci je ne-těžké. Protože v takovém případě to, co způsobilo dlouhou dobu vzniku inteligence, je slepost evolučního procesu.[26] Překážka komplexnosti je ale ještě silnější, kvůli faktu, že nevíme, jak se evoluci podařilo stvořit intelekt na naší úrovni. Co nás tak opravňuje tvrdit, že my budeme schopni stejného činu? Nebo v případě druhé interpretace je pochybné, jak bychom mohli genetický algoritmus správně navádět k jeho cíli. Z těchto důvodů tento argument nepovažuji za dostatečný důkaz toho, že by UI mohla vzniknout. Nelze pouze z existence věci usuzovat, že může vzniknout další. Musíme vědět,

25 Bostrom & Shulman (2016).
26 Tamtéž.

jak daná věc funguje, abychom ji mohli replikovat. Argument z emulace je lepší v tom ohledu, že nabízí návod, jak bychom UI mohli vytvořit, což evoluční cesta neobsahuje. První verze pouze argumentuje, že toho budeme schopni. Neposkytuje však žádné vysvětlení, jak toho dosáhneme. To má za efekt, že si díky tomuto argumentu můžeme představit široké spektrum intelektů, které bychom mohli stvořit různými způsoby. To nás však může svádět k tomu přehlížet, jakou skutečnou strukturu inteligence může mít. Druhá je rezignací na snahu zjistit, jak inteligence funguje. Tak pouze co nejlépe zkopírujeme proces jejího vzniku v naději, že tak dosáhneme našeho cíle. Proces vzniku intelektu je jen jedna část celého příběhu, druhá je skryta v jeho struktuře, kterou také musíme vzít v potaz.

2.3. Teze rozšíření

Nyní se tedy musíme ptát, jak můžeme od UI dojít k UI+. Jak bychom mohli stvořit vyšší než lidský intelekt? Pokud se nám podařilo vytvořit UI, museli jsme tak učinit určitou metodou. Tudíž abychom vylepšili naši UI, stačilo by nám, abychom rozšířili tu danou metodu. Tomu Chalmers říká teze rozšíření, která spoléhá na to, že UI vznikne rozšiřitelnou metodou.

(3) Pokud bude UI, tak vznikne rozšiřitelnou metodou
(4) Pokud UI vznikne touto metodou, budeme schopni ji rozšířit
(5) Rozšířením metody vznikne UI+
Závěr: Pokud bude UI, tak bude UI+

Proti tomuto argumentu můžeme namířit dvě námitky. (a) Nejjednodušší by bylo vyvrátit, že UI vznikne rozšiřitelnou metodou. Tuto námitku však Chalmers předpokládá, a proto říká, že i kdyby UI vznikla nerozšiřitelnou metodou, tak její vznik by nám poskytl vodítka k tomu, jak vytvořit UI rozšiřitelnou metodou. Toto se nezdá jako příliš jisté. Představme si, že se nám podaří stvořit UI skrze genetický algoritmus. Máme sice UI, ale nemáme tušení, proč je inteligentní. Kdybychom měli k dispozici zdrojový kód takové UI, jak bychom zjistili, které části jsou důležité? Vezměme

si například náš genetický kód. Přestože jsme schopni číst DNA, není snadné identifikovat, co je účelem jednotlivých genů. Tento argument může fungovat pouze, pokud UI vznikne cíleně, protože pak budeme vědět na čem dále pracovat. Ale například Bostrom zastává názor, že první UI nejpravděpodobněji vznikne omylem.[27] V takové situaci jsme na tom stejně jako v předchozím případě, protože nevíme, jak daná UI vznikla. Chalmers jako příklad rozšiřitelné metody zmiňuje přímé programování, které může zahrnovat činnosti jako ladění kódu. Co když však UI omylem vznikla právě kvůli chybě? Pak bychom během našeho procesu „rozšiřování", vlastně zničili danou UI. (b) Předpokládejme, že se nám skutečně podaří cíleně stvořit UI rozšiřitelnou metodou. Můžeme danou metodu rozšiřovat donekonečna? McDermott poukazuje na tento problém. Co se nám s ohlédnutím zpět může zdát jako hladká křivka technologického pokroku, bylo ve skutečnosti trnitou cestou, kterou se podařilo projít pouze s pomocí mnoha inovací. Naše představa může být taková, že postupujeme průběžným vylepšováním jedné metody, ale ve skutečnosti to nemusí být jedna metoda. Skutečný pokrok ve velké části závisí na objevování nových metod pro řešení problémů.[28] Chalmersova odpověď je, že pokud narazíme na strop jedné metody, můžeme se vždy obrátit k další. Dle jeho názoru s růstem inteligence daných systémů, budou tyto systémy objevovat další metody svého rozšíření.[29] To může být pravda, ale zůstává zde otázka, jak dlouho takové rozšiřování potrvá. Nemyslím si, že můžeme kategoricky odmítnout vznik UI+, ale myslím si, že bychom měli být skeptičtější ohledně data jejího příchodu.[30]

27 Bostrom (2014, s. 152–155).

28 McDermott (2016).

29 Chalmers (2016b).

30 Výše jsem citoval Simon & Newell (1958) a Minsky (1968), kteří jsou dobrým příkladem neoprávněného optimismu v oboru UI. Tento optimismus často zaslepuje výzkumníky UI vůči problémům spojeným s jejich tezemi, jak demonstroval Dreyfus (1992). Dreyfus (2012) poukazuje na celkový historický trend v oboru UI, kde se často začne prvním úspěšným krokem, který vzbudí tento optimismus, ale ke konci pokrok postupně pomine. Ten samý optimismus můžeme najít mezi singularitány. Stačí, když se podíváme na jejich předpovědi vzniku UI a příchodu singularity. Good (1965) předpovídal singularitu do roku 2000. Vinge (1992) nám pro příchod singularity nabízel rozmezí let mezi 2005 a 2030. Kurzweil (2005) předpokládá rok 2030 pro vznik první UI. Bostrom (2006) předpovídal UI v první čtvrtině tohoto století. Chalmers (2016a) je ve svém odhadu konzervativnější a s 50% jistotou odhaduje, že v rámci tohoto století vznikne první UI. Pro srovnání průzkum Müller & Bostrom (2016) ukazuje, že většina (přesněji 41 %) z dotazovaných výzkumníků UI předpovídá, že vznik první UI potrvá déle

2.4. Teze navýšení

Poslední premisu Chalmers zakládá na tezi navýšení, kterou opírá o matematickou indukci. Předpokládejme, že existuje UI+, a řekněme si, že UI_1 je první UI+ a UI_0 je jejím stvořitelem, ať už je to člověk či jiná UI (která může být také na menší úrovni než člověk), a δ je rozdíl mezi nimi. Takže systém, jehož inteligence se liší od předchozí o kladné δ, má vyšší inteligenci. Dále si řekněme, že v případech, kde n > 1 a máme UI_{n+1}, jejímž stvořitelem je UI_n, bude UI_{n+1} také chytřejší než její stvořitel, stejně tak, jak tomu bylo v případě UI_1. Z toho pak vyplývá teze navýšení:

(1) Pokud je UI+, tak je UI1
(2) Pro všechny n > 0, pokud je UIn, pak bude UIn+1
(3) Pokud pro všechna n je UIn, tak bude UI++
Závěr: Pokud bude UI+, tak bude UI++

Tomuto argumentu také můžeme oponovat různými způsoby. (a) Já bych protestoval proti tomu, jak dochází ke svému závěru. Když toto tvrzení posoudíme samostatně, musíme ho považovat za zvláštní, protože argumentuje *a priori*. Říká, že pokud máme jeden intelekt na určité úrovni, můžeme předpokládat, že by mohl existovat další intelekt na úrovni vyšší. Z toho můžeme odvodit, že existuje škála intelektů. To považuji za bizarní cestu, jak k tomuto závěru dospět. Samotný závěr není problematický, ale musíme k němu dojít *a posteriori*. Jakmile se setkáme s bohatým spektrem intelektů, můžeme začít přemýšlet o dalších možných intelektech. Možná byste mohli namítat, že zde je klíčové to, že se bavíme o UI. Takže pokud máme jeden uměle vytvořený intelekt, můžeme předpokládat, že by mohl vzniknout další. To však neřeší problém, na který jsem upozornil. Osamocená UI by se také musela nejdříve setkat s dalšími intelekty, aby mohla nad touto otázkou přemýšlet. Také byste mohli namítat, že Chalmers prezentuje tento argument v kontextu argumentů předchozích. Předchozí argumenty však byly schopné samostatného posouzení. (b) Tato teze je založena na představě, že nárůst

než 50 let. V průzkumu se jich také ptali, jak by s 10% / 50% / 90% jistotou odhadli rok vzniku UI. Median předpokládaného roku s 50% jistotou všech respondentů byl 2040. Průměrný předpokládaný rok s 50% jistotou byl 2081.

inteligence se rovná stejnému nárůstu v kvalitě designu další inteligence. Přičemž závisí na tom, že skok mezi jednotlivými inteligencemi bude vždy stejný. Tuto představu bychom mohli různými způsoby zpochybnit. Jeden možný protiargument by mohl být takový, že bychom mohli koncipovat inteligenční strop neboli limit, kterého inteligence může dosáhnout. Možná příliš vysoká inteligence je natolik náročná, že jí nemůžeme dosáhnout, nebo není dlouhodobě udržitelná. Pokud takový strop je, tak by na něj UI++ narazila. Jinou námitkou je říct, že nárůst inteligence by měl snižující výnosy, první nárůst o 10 % by pak mohl následovat pouze nárůstem o 5 % s postupnou ztrátou. Nejzávažnější námitka by mohla napadnout vztah mezi mírou inteligence a kvalitou designu. Možná vyšší inteligence není nutně lepším designerem.[31]

2.5. Celkové zhodnocení

Když se podíváme na argument jako celek, tak musíme říct, že nejvíce problematická je první premisa a teze s ní spojené. Premisa druhá a třetí sice mají své problémy, ty však nejsou nepřekonatelné. Jejich skutečné problémy se projevují v kontextu první premisy. Jak je asi očividné, abychom stáli před PK, musí nejprve vzniknout UI. Takže nejsilnější cesta, jak tento argument porazit, by bylo vyvrátit, že vznikne UI. Takovou argumentaci bych však chtěl v tomto textu ponechat stranou. Jedním z důvodů je, že už dost textů bylo věnováno takové argumentaci. Zadruhé, i kdyby UI byla potvrzena, tak PK stojí před dalšími překážkami, které by ho mohly vyvrátit. Protože se chci dále soustředit na smysluplnost tohoto problému, chci se na tyto překážky zaměřit. Jeden problém s argumentem singularity stojí na tom, jak by charakter dané UI mohl ovlivnit plauzibilitu následujících argumentů. Na jedné straně emulační teze argumentuje pro přístupy inspirované a založené na fungování lidského mozku. V úzkém smyslu argumentuje pro funkčnost ECM, v širokém smyslu ho vidím jako argument ve prospěch konekcionismu a vtělené inteligence. Zde jako výsledek můžeme očekávat inteligence, které budou podobné té naší. Na straně druhé, evoluční teze představuje dva opačné přístupy. Podle jednoho je inteligentní designer schopen

31 Chalmers (2016a).

stejného postupu jako slepá evoluce, možná i lepšího. Podle druhého by nám stačilo zkopírovat evoluční proces. V obou případech bych argumentoval, že by se výsledný intelekt lišil od toho našeho.

To vše nám poskytuje množství potencionálních intelektů. Otázkou je, jestli jsou všechny rozšiřitelné. Už jsem mluvil o problému rozšiřitelnosti inteligence vzniklé skrze genetické algoritmy, podobné problémy se vztahují i na další metody. Například ECM je metodou, která je obtížně rozšiřitelná. Nejvíce teoreticky možné je, že bychom mohli zvýšit rychlost přemýšlení dané simulace, skrze vylepšení jejího hardwaru, to by však mělo za výsledek pouze slabou formu SI. Problém závisí na tom, jak moc je architektura samotného mozku rozšiřitelná. Jak bylo zmíněno, tato metoda má výhodu v tom, že nemusíme podrobně vědět, jak mozek funguje. To ale ovlivňuje, nakolik budeme schopni jeho rozšíření. Bostrom a Sandberg[32] argumentují, že jakmile budeme mít simulaci mozku, zjistíme toho více o jeho architektuře. Protože taková simulace bude více otevřená k pozorování a experimentování. To se může ukázat jako pravdivé, není ale jisté, jestli by to vedlo k takovému skoku ve výzkumu, který si představují.[33] Stejně tak je nejisté, jestli je architektura mozku rozšiřitelná. Můžeme pouze spekulovat, jaký následek by například mělo více neuronů na celkovou architekturu. Mohlo by to vést k rozšíření, ale je i možné, že by celková architektura nebyla schopná takový větší výkon vydržet.[34] Některé metody se zdají být jako velice otevřené rozšíření, jiné se však jeví jako daleko uzavřenější.

Pro konstituci PK jsou důležité teze emulační a evoluční. V příští sekci ukážu, jak singularitáni zastávají pozici, že SI bude radikálně odlišná

32 Bostrom & Sandberg (2008).

33 Toto tvrdím kvůli faktu, že mnoho našich problémů nemusí záviset přímo na nedostatku poznatků. Vezměme si například problém vědomí v rámci filozofie mysli. Přestože toho víme o mozku stále více, stále tady panuje problém explanační propasti, na který upozorňovali Nagel (1974), Jackson (1982 a 1986) a Chalmers (1996). I pokud se tato propast ukáže jako iluzorní, tak demonstruje, že další poznatky nejsou vším. ECM nám může poskytnout přímý pohled do mozku, to ale neznamená, že nebudeme stát před konceptuálními problémy.

34 Colzato, Hommel & Beste (2021) upozorňují na jeden problém, na který bychom mohli narazit. Architektura mozku je postavená na soupeření neuronů. Jednotlivé neurony a i vetší neurální sítě soutěží o to, aby mohly zpracovávat a reprezentovat informace. To však znamená, že pokud bychom vylepšili pouze jednu část mozku, tak potencionální zisky by s sebou nesly ztráty v jiných oblastech. Například vylepšení neurální persistence (tj. naše schopnost se soustředit na jeden konkrétní problém) by bylo na úkor naší neurální flexibility (tj. naší schopnosti nacházet nová řešení).

od naší inteligence, a to vyvolává PK. Nejlépe tento fakt demonstruje Bostrom, který argumentuje, že můžeme identifikovat různé způsoby předpověditelnosti SI na základě její architektury. Pro nás je nejdůležitější předpověditelnost na základě zděděných motivací. Konceptem je, že pokud daná inteligence vznikla na základě lidské předlohy, mohla by jako důsledek zdědit některé lidské motivace a hodnoty. To by znamenalo, že na takovou inteligenci by se PK vztahoval v menší míře, než kdyby daná inteligence byla odlišná od té naší. Bostrom samozřejmě upozorňuje, že nevíme, do jaké míry by taková inteligence mohla zdědit naše motivace. ECM by je mohla ztratit během přesunu do počítače, nebo při svém dalším vylepšení.[35] Přesto je tady dobrá šance, že na takovou inteligenci by se PK vztahoval v menší míře, než kdyby daná inteligence byla radikálně odlišná od té naší. Proto se singularitáni musí držet evoluční teze, aby mohli tvrdit, že stojíme před PK. Jedna věc, která je k tomu opravňuje, je fakt, že inteligence podobné té naší se zdají být hůře rozšiřitelné.[36] Z toho vyvozují, že SI bude spíše odlišná od nás. Tímto krokem však na sebe berou problémy, které evoluční teze má a se kterými se nevypořádávají. V příští sekci demonstruji, jak je tento krok vede k tomu formulovat SI a koncept PK s ní spojený.

3. Problém kontroly a obecná inteligence

Vraťme se úplně na začátek, k Bostromově stroji na sponky. Vzpomeňme si na dvě vlastnosti, které SI stroj údajně bude mít. Zaprvé bude schopen díky svému neuvěřitelnému intelektu řešit problémy způsoby, které by člověka nenapadly. A druhou jeho vlastností je doslovnost. Je to hlavně tato vlastnost, která činí SI tak problematickou. V úvodu jsem zmiňoval,

35 Bostrom (2012 a 2014, s. 170).

36 Kromě ECM jsou zde i problémy s dalšími postupy, které jsem zmínil. Vtělená UI čelí faktu, že není zřejmé, jak bychom rozšiřovali tělesnost samotnou. Tělesnost spíše hraje roli podmínky inteligence. Tělo také vytváří různá omezení, pro vylepšování UI. Do robotického těla nemůžeme jen tak dále vkládat další hardware, aniž bychom ovlivnili jeho stabilitu. Neurální sítě se z těchto přístupů ukazují jako nejlépe rozšiřitelné. Ovšem i moderní neurální sítě schopné hlubokého učení stojí před několika překážkami. Nejzávažnější pro vylepšování je nedostatek transparentnosti. Často není jasné, proč neurální síť došla ke svému závěru. To velice komplikuje proces rozšiřování učících procesů. Pro přehled dalších problémů viz (Marcus, 2018). I kdybychom tyto potíže překonali, tak vzhledem k argumentům, které rozvíjím později v textu, si nemyslím, že můžeme skutečnou SI stavět pouze na základě neurální sítě.

že doslovnost SI vyplývá z faktu, jak obtížné je definovat mnoho lidských hodnot a konceptů. Například morální filozofie jako celek je toho dobrým příkladem. V jejím rámci stále vznikají nové morální systémy a stejně tak se u každého objeví protipříklad, který daný systém zpochybní.[37] Problémem je, že SI bude od nás přijímat pouze přímo specifikované povely a definované hodnoty. SI bude jednat pouze na jejich základě. Pokud špatně specifikujeme dané povely a definujeme naše hodnoty, tak to pro nás může mít neblahé následky. Proto singularitáni koncentrují tolik úsilí na to, aby našli takový konečný cíl pro SI, který by byl nezávadný.[38] Je zde ale zásadní otázka, jestli se máme o toto snažit. Proč je skutečně tak složité definovat naše hodnoty a co způsobuje doslovnost SI? Pokud bychom byli schopní SI vysvětlit přesně, co máme na mysli, tak by tento problém nemohl nastat. Proto PK je ve své podstatě problém interpretace. Jak zajistit, aby SI interpretovala naše povely tak, jak je míníme? S tímto problémem jsme se mohli setkat mnohokrát v historii UI. Yudkowsky předkládá jeden příklad, kdy MIT údajně vytvořil neurální síť, která měla být schopná rozeznat maskované tanky v lese. Po nějakém cvičení neurální sítě testy ukazovaly, že síť skutečně rozeznávala mezi fotkami, kde tanky byly a kde nebyly. Když však poslali síť armádě k dalšímu testování, brzy jim byla vrácena zpět. Důvodem bylo, že v jejich testování síť nebyla v objevování tanků o nic lepší než náhoda. Problém se ukázal být ve cvičebním vzorku, fotky s tanky byly udělány během zamračeného dne, zatímco na fotkách bez tanků bylo slunečno. To, co se jejich síť naučila rozeznávat, bylo to, jestli na fotce bylo zamračeno nebo slunečno, ne jestli tam byly či nebyly tanky.[39]

Takové potíže vyvolává jeden z největších problémů s UI, a tím je problém rámce. Tento problém se nejdříve objevil jako technický problém pro UI založenou na formální logice.[40] Jednalo se o problém toho, jak elegantně zachytit důsledky dané akce v rámci formální logiky. Tento

37 Muehlhauser & Helm (2012).

38 Viz Bostrom (2012 a 2014), Yudkowski (2001, 2004 a 2011), Muehlhauser & Helm (2012).

39 Dreyfus (1992) a Yudkowsky (2008), přestože se jedná o často citovaný příběh, Yudkowsky není schopen dohledat primární zdroj. Je tak dost možné, že se jedná o pouhou historku. Ale slouží dobře jako typický příklad dezinterpretace v oboru UI, proto ho zde cituji.

40 McCarthy & Hayes (1969).

technický problém ale dal vznik širšímu epistemologickému problému.[41] Dennett ho charakterizuje jako problém toho, jak má činitel svá přesvědčení aktualizovat vzhledem ke svému jednání ve světě. Představte si běžnou situaci, kdy si děláte kávu. Vezmete si hrnek, konvici a kávu. Konvici naplníte vodou a dáte ji ohřát. Vezmete kávu a nasypete ji do hrnku a zalijete ji ohřátou vodou. Nakonec byste mohli hrnek s sebou vzít do obývacího pokoje. Teď bych na vás ovšem měl otázku. Když jste si s sebou vzali hrnek, vzali jste s tím s sebou i kávu? Předpokládám, že byste odpověděli, že ano, je přece v tom hrnku. Kdybychom se zeptali UI programu, tak by nevěděl. Program by čelil problémům daleko dříve. Například, když zalejeme kávu horkou vodou, co se skutečně změní v dané situaci? Počítač může mít mnoho údajů, jako pozici hrnku, teplotu v dané místnosti, daný čas a další. Složité je pro něj určit, jaký z těchto údajů by měl aktualizovat po daném jednání, a přitom stále mít věrohodný obraz světa.

Problém rámce je kromě v oboru UI často diskutovaným problémem v komputační teorii mysli.[42] Pokud chápeme mentální stavy jako sérii propozic, tak se problém rámce stává problémem toho, jaké propozice máme změnit v daném kontextu. Tady se jako nejdůležitější ukazuje princip relevance. Daný kontext si žádá, abychom změnili pouze relevantní množinu našich přesvědčení. Takže pokud natřu můj dům modrou barvou, měl bych aktualizovat pouze ta přesvědčení, která se týkají mého domu. Je to právě kategorie relevance, která spojuje problém rámce s PK. PK poukazuje na užší situaci, kdy si SI musí aktualizovat svůj soubor přesvědčení na základě našich povelů. Relevance je důležitá, co se týče jejich interpretace. Stroj na sponky trpěl tím problémem, že nenahlížel na všechny pro nás relevantní důsledky svého cíle. Když mu jeho designer zadal jeho úkol, konal tak v nějakém kontextu. Designer předpokládal, že stroj bude jednat v určitých mezích. Například, že bude sponky vyrábět pouze z kovu, nebo že při jejich výrobě nezraní člověka. Designera ani nenapadlo se zamýšlet nad tím, že by jeho výrok šlo vyložit jinak. SI je tak nebezpečná z tohoto důvodu, že může špatně identifikovat kontext daného povelu. Z toho vyplývá, že si jej vyloží jinak

41 Dennet (1984).
42 Viz Fodor (1983).

než my. V tomto ohledu se ale SI vymyká definici, kterou jsme si stanovili v úvodu. Pokud SI má být obecnou inteligencí, tak by měla být schopná stejně jako my, správně identifikovat, jaká přesvědčení má mít v dané situaci. Otázkou je, proč se tento problém vztahuje i na SI. Co ji vede k tomu interpretovat naše výroky jinak?

3.1. Davidsonova teorie interpretace a SI

To nás nutí zamyslet se nad tím, co to znamená interpretovat výroky druhého. Co bychom museli vědět, abychom mohli někoho správně interpretovat? Myslím si, že nejlépe se můžeme na to dívat skrze Davidsonovu teorii interpretace. Ten ji založil na základě Tarského teorému, ale obrátil ho naruby. Zatímco Tarsky předpokládal správný překlad vět z objektového jazyka do metajazyka, aby mohl definovat pravdu, tak Davidson místo toho přijímá pravdu jako primitivní pojem, aby mohl vytvořit teorii překladu nebo interpretace. Takže vytváří teorii pravdy, která nestojí na konceptu překladu. Podle Davidsona pro každou větu *s* v objektovém jazyce musíme empiricky najít takovou větu *p* nám známého metajazyka, že *s* je pravdivá, právě když *p*. Sama o sobě by nám však taková koncepce T-vět nic o interpretaci neřekla, je nutné ji formálně a empiricky omezit. Proto se musíme zamyslet nad tím, co má interpret k dispozici. Jediné, na čem můžeme skutečně stavět, je fakt, že mnohé naše výroky také považujeme za pravdivé. Davidson demonstruje to, že naše interpretace závisí na povaze dané komunity mluvčích, kde danou skupinu můžeme identifikovat tím, jaká tvrzení považují za pravdivá a jaká naopak ne.[43] Při intepretaci druhého promítáme to, co bychom za daných okolností považovali za pravdivé do struktury vět mluvčího. To je podstata principu vstřícnosti, který nám napomáhá interpretovat druhé. Tohle se dále promítá v jeho teorii poznání, kterou můžeme chápat pouze v rámci totálního souboru našich přesvědčení. Význam našich výroků je úzce propojen s našimi přesvědčeními. Fakt, že jsme schopni interpretovat druhé a běžně správně používat jazyk, pak obecně potvrzuje pravdivost našich přesvědčení. Pokud by pravdivá nebyla, nemohli bychom si rozumět.[44]

43 Davidson (1984).
44 Davidson (2004a).

Jak máme chápat PK ve spojení s interpretací? Bostromův sponkový stroj, očividně svého designera interpretoval špatně. Pokud však má Davidson pravdu a vskutku interpretujeme druhé osoby takovým způsobem, tak by SI z definice měla být schopná stejného úkonu. Na SI by se ani neměl vztahovat problém rámce. Davidson se v jedné pasáži[45] zabývá představou vševědoucího činitele, který by měl pouze pravdivá přesvědčení. Pokud by takový činitel interpretoval nás jako omylné činitele, stejně by musel dojít k závěru, že omylný činitel se nemůže mýlit ve většině svých přesvědčení. To je situace analogická s SI. Jsou zde dva možné důvody, proč by SI dělala takové chyby. (1) SI je schopná správné interpretace daného úkolu, ale nemá motivaci se jí řídit. To znamená, že nás může velice dobře chápat, ale postrádá jakýkoliv popud k tomu jednat podle našich přání. (2) SI nemůže nahlížet relevantní fakta z toho důvodu, že se od nás radikálně odlišuje. Yudkowsky a Bostrom zastávají obě tato stanoviska jako klíčová pro PK. Říkají, že SI se bude od naší inteligence radikálně lišit, a tím pádem bude mít motivace odlišné od nás. Mým postojem je, že nemohou mít oba tyto názory zároveň.

3.2. Ontologická sázka

Pro jejich postoj je klíčový koncept prostoru myslí. Yudkowsky tento prostor explicitně vytváří. Dle jeho názoru, když se bavíme o inteligenci, často propadáme klamu, že víme, co tento pojem obsahuje. Tento klam se ukazuje během diskusí o UI. Když se o ní bavíme, máme tendenci ji antropomorfizovat. Soudíme-li její možný intelekt, konáme tak v kontextu lidského měřítka. Takový postup není problematický při interakci s dalšími lidmi. Tato perspektiva je pro nás přirozená, a proto když se bavíme o SI, představujeme si ji jako stereotypického genia. Lidské měřítko intelektu je však neadekvátní, když mluvíme o UI. Podle Yudkowského UI představuje prostor myslí obecně. Přesněji se dle jeho názoru musíme bavit o *prostoru optimalizačních procesů*. To odpovídá naší definici inteligence. Dle jeho názoru musíme odlišné mysli chápat právě takovým způsobem – jako systémy, které jsou schopné dosáhnout velice specifických cílů v celém prostoru možností. Nejde jen o to, že SI

45 Tamtéž, s. 144.

by měla větší intelekt než my, ale také o to, že její mysl by mohla mít radikálně odlišné cíle. Dle jeho názoru prostor možných myslí UI je daleko větší než naše lidské měřítko, a právě proto je tak neadekvátní. Tohle pak navazuje na problém motivace UI. Pro něj z toho jasně vyplývá, že tyto odlišné mysli by oplývaly jinými motivacemi než my, a tudíž by jinak interpretovaly naše výroky.[46]

Bostrom v tomto směru následuje Yudkowského. Nám se může zdát, že mezi dvěma lidmi je obrovský rozdíl. Pokud bychom se ale podívali jen na jejich mozky, museli bychom dojít k závěru, že se jedná o dvě inteligence stejného typu. Bostrom vytváří distinkci mezi námi jako evolučně vzniklými tvory, kteří mají určité potřeby, a UI, která bude mít jeden konečný cíl, vycházející z jejího naprogramování. To, co spojuje Bostroma a Yudkowského, je představa, že různé inteligence odlišují jejich cíle.[47] Bostrom také zastává podobnou představu nepřímo ve vztahu k PK. Když mluví o transhumanistických hodnotách, tak nám nabízí představu prostoru možného bytí. Tento prostor zakládá na rozdílu mezi našimi a zvířecími stavy. Tento rozdíl můžeme podle něj vidět jasně, když posoudíme rozdíly v modalitách jednotlivých druhů. Například je nám zřejmé, že psi asi vnímají svět jinak než my, tudíž jejich způsob bytí se v něčem radikálně odlišuje od toho našeho. Bostrom z tohoto rozdílu vyvozuje, že možná kromě těchto stavů mohou existovat další různé teoretické stavy bytí.[48]

Touto tezí se ale dostávají na šikmou plochu. Yudkowsky a Bostrom si nepředstavují jenom obecný prostor myslí. Aby PK byl takový, jak ho prezentují, musí postulovat velice určitý prostor odlišných myslí. Pokud by se ukázalo, že SI nemůže mít takovou mysl, jakou si představují, tak by PK padl. Konají tak něco, co nazývám ontologickou sázkou, kde sází na existenci něčeho, co nemohou potvrdit, aby vytvořili tento problém. Teisté dělají něco podobného. Ti nesází pouze na to, že může existovat vyšší bytost, ale zároveň předpokládají, že tato bytost bude mít právě ty schopnosti, které bychom od boha očekávali. Tato myšlenka by i šla aplikovat na celou problematiku UI. Já zde ovšem vnímám podstatný

46 Yudkowsky (2008).
47 Bostrom (2012 a 2014). Obecně zastává stejnou pozici v Bostrom (2003).
48 Bostrom (2005).

rozdíl mezi singularitány a obyčejnými inženýry UI. Inženýr UI pouze sází, že může existovat něco jako UI, která může být úplně stejná jako naše inteligence. Singularitáni potřebují nejen to, ale také, aby SI byla od ní radikálně odlišná a aby měla jeden konečný cíl. Takže se nabízí otázka: Dával by PK stále smysl, pokud bychom existenci tohoto prostoru zpochybnili? V kritice UI můžeme najít dvě různé strategie takového zpochybnění.

3.3. Searle a nevědomá inteligence

První strategií by bylo zpochybnit, že UI by vůbec mohla mít něco jako motivaci. John Searle formuluje argument proti UI, který míří tímto směrem. Terčem jeho kritiky je myšlenka, že inteligence je ve své podstatě pouze zpracovávání informací, které můžeme chápat jako manipulování s obecnými symboly na základě formálních pravidel. To je však podle Searla problematické, protože mysl je pro něj více než pouze formální struktura. Naše myšlenky nejsou pouhými symboly, také mají svůj obsah. Stavy naší mysli jsou intencionální stavy, vztahují se k věcem. Jazyk vychází z intencionality naší mysli. Proto můžeme mluvit o tom, že naše výroky mají význam. To ho vede k závěru, že mezi námi a počítačem je rozdíl v tom, že my neovládáme pouze syntax, ale i sémantiku. Naše myšlenky mají význam, kdežto počítač místo toho ovládá pouze holé symboly a jejich význam mu uniká. Searlovým závěrem, je, že pouhá manipulace s reprezentacemi nestačí pro skutečné vědomí.[49]

Na tyto myšlenky navazuje v recenzi na Bostromovu knihu.[50] V tomto textu staví distinkci mezi objekty, které existují nehledě na to, co si o nich myslíme, a těmi, které naopak závisejí na našich postojích. První kategorii označuje jako na *pozorovateli nezávislou*. Druhou máme naopak chápat jako na *pozorovateli závislou*. Tato kategorie je důležitá, protože mnohé věci, se kterými se ve světě setkáváme, jsou závislé na pozorovateli. To jsou věci jako instituce, peníze nebo státy. Tyto věci však závisí na vědomí, které je na *pozorovateli nezávislé*. Bez vědomí by tyto objekty vlastně neexistovaly. To se vztahuje na koncept vypočítávání

49 Searle (1980 a 1984).
50 Searle (2014).

ve vztahu s počítači. Původně termín „počítač“ označoval osoby, které se živily tím, že dělaly výpočty. Tyto osoby však byly postupně nahrazeny současnými počítači, protože zvládaly tu samou práci přesněji a rychleji. Je tu však jeden zásadní rozdíl v jejich aktivitě. Lidský „počítač“ chápal význam čísel, se kterými pracoval. Současný počítač pouze manipuluje čísly úplně stejně jako se symboly. Jediný důvod, proč to, co počítač dělá, také nazýváme vypočítáváním, je, že my to tak vidíme. Takže po vzoru Searlovy kategorizace, naše výpočty jsou na *pozorovateli nezávislé*, zatímco výpočty počítače jsou na *pozorovateli závislé*. Bez lidské interpretace výsledky výpočtů počítače nemají žádný význam. To samé pak můžeme aplikovat na SI. Její inteligence je také na *pozorovateli závislá*. Dle jeho názoru fundamentální rozdíl mezi námi a počítačem je, že nás lze psychologicky popsat. My lidé máme motivace k našemu jednání, ale měl by je počítač? Dle Searla ne. Podle jeho názoru počítač nejenže nemůže mít žádnou motivaci, ale ani není činitelem.[51] Počítač nemůže mít žádnou vlastní motivaci, protože podle Searla počítač ani nemůže mít vědomí. Pokud tedy nemá vědomí, nemůže mít motivaci, a tudíž nemá nic, co by ho motivovalo k tomu, aby pro nás byl hrozbou. Takže by ani SI nevznikla, protože by postrádala dostatečnou autonomii k tomu, aby započala inteligenční explozi.

Jeho argumentace není však příliš přesvědčivá. Mohli bychom ji uznat pro vyvrácení konceptu inteligence, jako pouhé manipulace s reprezentacemi. Ale jako argument proti možnosti vědomí digitálního počítače je nedostačující. Vezměme si jeho distinkci. Searle si je příliš jistý tím, že vědomí je na pozorovateli nezávislé. Může tomu tak být, pokud se bavíme o svém vlastním vědomí, ale co vědomí druhých? To úplně stejně záleží na naší interpretaci chování druhých.[52] Zde bych argumentoval,

51 Pro zjednodušení jsem se v hlavním textu rozhodl držet Searlova konceptu nevědomé inteligence. Nejsem si však jistý, nakolik můžeme spojovat jednatelství s vědomím. Searle zde zaměňuje dva různé způsoby, jak ho můžeme chápat. Musíme rozlišovat mezi minimálním jednatelstvím a morálním jednatelstvím. My lidé máme morální jednatelství z důvodů, které Searle popisuje. Minimální jednatelství se ale může vztahovat i na jednodušší činitele. Barandiaran, Di Paoloa Rohde (2009) stanovují tři podmínky, abychom mohli daný systém popsat jako činitele. 1. Systém musí být jedincem. 2. Systém musí hrát aktivní roli v interakci s prostředím, musí být zdrojem aktivity. 3. Aktivita daného systému se řídí nějakou normou či cílem daného systému, jeho chování není čistě arbitrární. Na základě těchto podmínek můžeme přiznat jednatelství i nevědomé UI.

52 Tady se opírám o Dennetta (1987). To, že druhé považujeme za vědomé, není ani tak spojeno

že jsme ve stejné situaci jako s počítačem a Searle nenabízí žádný argument, proč bychom se na to měli dívat jinak. Zajímavé je to, že přestože UI upírá vědomí, neupírá jí inteligenci. Searle je ochoten uznat, že UI může plnit složité úkoly, jako pilotovat letadlo nebo řídit automobil. Takže Searle přiznává, že může existovat něco jako nevědomá inteligence. Zde ale spočívá další problém s jeho argumentem.

Nevědomá inteligence by mohla být přesně taková, jak Searle říká. Neměla by žádnou svou autonomii, motivaci či přesvědčení. Místo toho bychom jí museli motivaci dodat. To však vůbec není nepodobné případu sponkového stroje. Jeho finální cíl také nebyl jeho vlastní, byl mu dán jeho designerem. Podstatou tohoto příběhu bylo to, jak se stroj choval, ne jak k tomuto chování došel. To vyplývá jak z definice inteligence obecně, tak i z definice UI, které jsem stanovil v úvodu. Muehlhauser a Helm na jejich základě argumentují proti Searlovi. Aby činitel mohl inteligentně postupovat ke svému cíli, není zapotřebí, aby měl vědomí.[53] Bostrom také zastává stejnou pozici.[54] Takže Searlův argument selhává při snaze vyvrátit možnost existence SI. Na druhou stranu ale jeho argument odhaluje jednu zásadní chybu ve formulaci PK, kterou si singularitáni neuvědomují. Jde o zásadní rozpor mezi jejich chápáním SI a konceptem PK. Tento rozpor vyplývá z jejich volby soustředit se na SI jako holý intelekt. Tím podceňují důležitost naší mysli a její intencionality pro naše schopnosti. Schopnost interpretace je toho dobrým příkladem.

Pokud SI může být nevědomou inteligencí, tak by to vysvětlilo, proč SI není schopná správně interpretovat naše výroky. Protože bez vědomí nemůže mít přesvědčení. Bez přesvědčení nemá co promítat do našich výroků. Ale znamená to, že nás ve skutečnosti SI neinterpretuje. To by vysvětlilo doslovnost SI. Zásadně se tím mění koncept SI a charakter PK. Výsledkem představy nevědomé inteligence nemůže být SI, jak ji singularitáni koncipují, protože taková inteligence nemá schopnost

s tím, že by měli na pozorovateli nezávislé vědomí. Spíše jde o to, že jejich chování nás vede k tomu interpretovat je jako vědomé, protože se tento postup ukazuje jako nejužitečnější. To je podstata Dennettova intencionálního postoje. Proto jsme ve stejné pozici, ať už interpretujeme chování dalšího člověka nebo počítače.

53 Muehlhauser & Helm (2012, s. 104).

54 Bostrom (2014, s. 47). V poznámce 88 stanovuje, že nemusí mít pravou intencionalitu (přestože si myslí, že ji asi mít bude).

interpretace. To znamená, že postrádá jednu velice relevantní lidskou schopnost. Tím se vymyká Bostromově definici SI. Možná ale jde o SI v jiném smyslu. Nazvěme si ji omezenou superinteligencí neboli OSI. V některých oblastech nás překračuje, v jiných nikoliv. Pokud SI má být obecnou inteligencí, tak jak ji singularitáni koncipují, tak bude muset mít vědomí. Singularitáni nám musí poskytnout nějaký obraz tohoto vědomí, pokud chtějí zastávat svoji pozici. Místo toho se této otázce vyhýbají.

Jak jsem ukázal v úvodu, PK stojí na představě, že SI bude interpretovat svůj konečný cíl. OSI ale nemůže interpretovat naše výroky, ze své povahy jako nevědomé inteligence. Pokud by tomu tak bylo, současná forma PK by musela padnout, ale v obecnějším smyslu by PK přežil. Jenom se už nemůže vztahovat na to, jak UI bude interpretovat svůj finální cíl, ale možná spíše na to, jakými motivacemi takovou UI vybavit. Což už je součástí projektu Yudkowského.[55] Je ale zásadní otázkou, jestli bychom takovou inteligenci mohli vybavit motivací, a proto se později k tomu vrátím. Kromě tohoto bodu se však zatím zdá, že PK může v tomto případě zůstat ve stejné podobě. Tento fakt je ovšem podezřelý. Pokud je PK víceméně aplikovatelný na koncept OSI, proč ho singularitáni spojují s konceptem SI? To je to hlavní, co si musíme z této podkapitoly odnést, protože se to ukáže jako důležité pro konstituci PK. V celku musíme soudit Searlův argument jako neadekvátní k tomu, aby vyvrátil existenci SI. Z toho plyne, že stále stojíme před PK. Jeho argument ovšem poukazuje na jeden důležitý fakt. Naše inteligence stojí na naší intencionalitě. Singularitáni si ji nemohou dovolit ignorovat. Prozatím však jejich argument víceméně platí. Jediné, co musí udělat, je uznat, že SI bude vědomá, aby byla obecným intelektem. Já jsem toho názoru, že tu jsou další rozdíly mezi SI a OSI, které musíme posoudit.

3.5. Dreyfus a tělesnost

V minulé podkapitole jsem formuloval koncept OSI. Je otázkou, jaké další rozdíly by tu mohly být mezi OSI a SI. SI by měla být z definice obecnou inteligencí, stejně jako my. To nás vede k otázce, co utváří obec-

55 Yudkowski (2001, 2004, 2008 a 2011).

nou inteligenci a jestli se taková inteligence může od té naší lišit. Toto je druhá strategie zpochybnění prostoru myslí. Principem je ukázat, že prostor myslí má své hranice. Mohou existovat různé inteligence, ale jenom určité z nich jsou skutečně obecné. Nejlepším příkladem této strategie je kritika UI od Huberta Dreyfuse. Ten svůj argument cílil na výzkum UI v 60. letech, specificky na představy, na kterých stála SDUI[56] forma UI. Tehdejší programy UI byly postaveny na sérii explicitních pravidel, které pro ně vytvářely sérii možností. Takže tato pravidla ho například vybavila specifickými operátory, které určovaly, jak bude konat v určité situaci. Pravidla tak vytvářela rámec, v kterém se program pohyboval, aby mohl vyřešit daný problém. Tyto programy fungovaly na principu heuristických algoritmů. Heuristiky byly hrubé postupy, které program naváděly k tomu, aby co nejlépe vyřešil daný problém a vyhnul se slepé uličce.[57] Tehdejší výzkumníci si mysleli, že lidská mysl funguje na stejných principech. Ovšem to, co se neustále ukazovalo, byl fakt, že jejich programy mohly fungovat pouze ve velice omezených kontextech. Nehledě na to, jak dobrými pravidly daný program vybavili, vždy nastal kolaps. V obecnějších situacích počítač selhával, protože zaprvé nebyl schopen reprezentovat všechna potřebná data a zadruhé neuměl rozpoznávat, jaká data jsou pro danou situaci relevantní. Dreyfus argumentoval, že jejich problémy byly v domněnkách, na kterých jejich projekt spočíval. Na ně soustředil svou kritiku. Nás hlavně zajímají dvě následující domněnky.[58]

První je epistemologická domněnka, ta tvrdí, že znalost všeho ve světě lze formalizovat. Představou je, že budeme schopni formalizovat chování jako sérii úkonů, které se řídí pravidly. Takže podle této domněnky můžeme formalizovat jakékoliv chování jako soubor explicitních pravidel. Otázkou však je, jestli lze jakoukoliv kompetenci takto formalizovat. Nejlepší argument pro tuto tezi je postaven na základě úspěšné formalizace jazyka v rámci lingvistiky. Problém je v tom, že v tomto případě to, co bylo formalizováno, nebyla celá lidská kompetence, ale pouze naše schopnost ovládat gramatiku jazyka. Naše jazyková kompetence však také zahrnuje naši schopnost interpretace. Zde má počítač problém, může daný výrok

56 Stará dobrá umělá inteligence. V originále GOFAI (Good Old Artificial Intelligence).
57 Boden (2014).
58 Dreyfus (1992).

interpretovat pouze skrze předem daná pravidla. Člověk však má další možnost. My jsme schopni interpretovat i výroky, které se špatně řídí gramatickými pravidly. A i více než to. Někdy úmyslně porušujeme pravidla, abychom se lépe vyjádřili. Jestli můžeme porušit pravidla a stále být srozumitelnými, musíme se místo o pravidlech bavit o implicitním pochopení jazyka, které nám umožňuje ho používat. Rozdíl mezi člověkem a počítačem je v tom, že my máme druh praktické inteligence. Abychom tak mohli formalizovat užívání jazyka, museli bychom najít pravidla této praktiky. Tuto myšlenku ovšem Dreyfus zpochybňuje s pomocí Wittgensteinovské argumentace. Inteligence nemůže záviset na formálních pravidlech, protože pravidla propadají do nekonečného regresu. Abychom mohli mít nějaké pravidlo, tak to, jak ho čteme, se samo musí řídit nějakým pravidlem, ale i toto pravidlo musí být určováno pravidlem, a takhle bychom mohli pokračovat do nekonečna. Interpretace se musí zastavit na úrovni, kdy už je pochopení jednoznačné a není zapotřebí ho dále rozebírat.[59]

Zde je však spor, který stojí na druhé domněnce. Tou je ontologická domněnka, podle které lze vše analyzovat jako sérii nezávislých faktů, které můžeme chápat jako konkrétní reprezentace. Tento koncept umožňuje ignorovat argumenty proti epistemologické domněnce. Pokud bychom byli schopni vše takto reprezentovat, tak ani praktická inteligence by pro nás neměla být problémem. Tak ale stojíme před jinou překážkou. Jak máme všechna tato fakta počítači reprezentovat? Inženýři UI si představovali, že budeme schopni tato jednotlivá fakta lidské znalosti roztřídit do prostých kategorií. Problémem je, že tímto způsobem nelze kategorizovat ani jednoduché objekty. My lidé můžeme reprezentovat předměty ve světě, protože jsme schopni je chápat v rámci širšího kontextu. Židli například chápeme ve vztahu s jinými objekty, ale také s lidskou činností. Narážíme tak na problém, jak bychom mohli takové reprezentace vůbec oddělit od tohoto kontextu. Zde Dreyfus navazuje s fenomenologickou argumentací, aby tento koncept napadl. Proto, abychom mohli reprezentovat něco ve světě, museli bychom mít privilegovanou perspektivu, ve které bychom viděli svět o sobě. Problém je, že my žádnou takovou perspektivu nemáme. Svět, tak jak ho známe, už je re-

59 Tamtéž.

prezentací. Jediná reprezentace, kterou my můžeme vytvořit, je v rámci tohoto širšího kontextu. Nemůžeme vystoupit z rámce, ve kterém žijeme, abychom ho mohli popsat. Z tohoto důvodu jakékoliv modely, které můžeme UI poskytnout, nemohou fungovat. Ze stejného důvodu nemůžeme UI definovat naše hodnoty.[60]

Dreyfus používá několik příkladů, aby tuto představu vysvětlil. Představme si například melodii. Co ji vytváří? Jedná se o prostý soubor tonů? Problémem této myšlenky je, že tóny samy o sobě nic skutečně neznamenají, jsou pouhými abstrakcemi. Je to právě kontext dané melodie, který z tonů dělá to, čím jsou a z tohoto kontextu je nemůžeme jen tak vyjmout. Podstatou je, že vše, co vnímáme ve světě je nutně součástí takového rámce. Když požádáme někoho o sklenici vody a on nám místo toho podstrčí sklenici mléka, tak když se ho napijeme, nevíme, co máme zažívat. Nápoj, který v tu chvíli vypijeme, nechutná ani jako voda, ale ani jako mléko, místo toho je to určitý čistý požitek. Jsme z toho úplně překvapeni a naší první reakcí může být, že budeme chtít tu tekutinu vyplivnout. Možná si ale po chvíli můžeme uvědomit, že jsme se napili mléka a ne vody, a tím jsme si pak schopni daný požitek správně zařadit. Tohle ukazuje, nakolik naše inteligence závisí na tomto kontextu. Podle Dreyfuse jediný způsob, jak můžeme získat tento rámec, je pouze skrze tělesnost, která nám umožňuje úzce interagovat se světem. Chyba inženýrů UI byla v představě, že budeme moct simulovat inteligenci na počítači, který můžeme popsat pouze jako oddělenou a nevtělenou entitu. Jejich epistemologická domněnka závisela na této myšlence. Tím však přehlíželi roli, kterou tělo hraje při vzniku obecné inteligence. Například naše tělo neustále generuje potřeby, které můžeme naplnit pouze jednáním ve vnějším světě. Tyto potřeby pomáhají kategorizovat svět na věci, které jsou relevantní k uskutečnění těchto potřeb, a na věci, které k tomu relevantní nejsou. Tak nám tělesnost vytváří rámec, ve kterém musíme operovat. Potřeby nám také dávají motivaci k tomu, abychom kreativně řešili problémy. Takže nás tělo nutí k tomu, abychom jednali ve světě, a zároveň nám poskytuje prostředky k tomuto jednání.[61]

60 Tamtéž. Přesněji cituje argumentaci Heideggera a Merleau-Pontyho.
61 Dreyfus (1967 a 1992, s. 147–167).

3.6. Holistický argument

Dreyfusův argument je silnější než ten Searlův. V jednom smyslu to vypadá, že singularitáni tyto argumenty přijali. Už si nemyslí, že lidská mysl funguje na stejných principech jako počítač a přijali problém rámce, který z Dreyfusovy argumentace vyplývá. To jsme také mohli pozorovat, když nás Yudkowski varoval před antropomorfizací UI. V jiném smyslu se však těchto tezí stále drží. Pochopili, že lidská inteligence takhle fungovat nemůže, ale nemohla by tak fungovat inteligence jiná? K této představě je navádí evoluční teze, která jim umožňuje tvrdit, že taková inteligence může vzniknout. Dle mého názoru přestali problém rámce chápat jako chybu a místo toho ho chápou jako nedílnou vlastnost UI. Proto jejich koncept SI tímto problémem stále trpí. Z tohoto důvodu formulují PK na základě radikální odlišnosti SI od nás, protože chápou UI jako neoddělitelně odlišnou. To i vysvětluje, proč PK byl aplikovatelný na OSI. Stačí, když se podíváme na příklady možných SI, které nám nabízí.

Jednou teoretickou SI, kterou se Bostrom a další zabývají, je takzvaná věštírna.[62] Koncept věštírny popisuje druh SI, který je spíše nástrojem než činitelem. Máme si představit stroj, který má jako jediný cíl pravdivě a co nejlépe odpovídat na otázky, které mu klademe. Takže tomuto stroji bychom mohli předkládat různé problémy, které by řešil. Pokud mu nikdo nedává žádné otázky, tak je zcela inertní. S takovými věštírnami se už můžeme setkat i dnes, vlastně i obyčejná kalkulačka je druhem věštírny. Současné věštírny jsou však velice omezenými inteligencemi. Neumím si představit, jak bychom mohli takový stroj považovat za obecnou inteligenci. Věštírna stojí na principu omezení SI. Cílem je ohraničit schopnosti SI tak, abychom omezili případná rizika spojená se SI. Pokud ale obecný intelekt vyžaduje aktivitu ve světě, jak argumentuje Dreyfus, tak by věštírna nemohla být obecným intelektem. Přesto je Bostrom toho názoru, že bychom mohli mít všestrannou SI, která by byla věštírnou.[63] Bostrom může tuto pozici zastávat jedině, pokud přijímá domněnky, které Dreyfus kritizoval. Věštírna může být obecnou inteligencí, jenom pokud můžeme formalizovat jakoukoliv znalost. Totéž lze vztáhnout na

62 Bostrom (2014, s. 224–228), Armstrong, Bostrom & Sandberg (2012) a Armstrong (2013).
63 Bostrom (2014, s. 225).

Bostromův scénář vzniku SI, která vznikne v rámci náhodného projektu UI. Dle jeho názoru UI bude provádět strojové učení a najednou se z ní stane obecná inteligence, která posléze vylepší sebe sama na úroveň SI.[64] Mohla by však UI uzavřená na serveru dosáhnout obecné inteligence bez interakce s vnějším prostředím?

Napříč mnohými příklady možných SI formulují tyto inteligence na modelu současných počítačů. Tady však znovu dělají chybu, kterou objevil už Dreyfus. Mezi námi a počítači je zásadní rozdíl v tom, že my jsme vtělené inteligence, zatímco počítače nikoliv. Důvod, proč operujeme v rámci společného kontextu, je, že my jsme nuceni se světem neustále interagovat. Naše těla jsou uzpůsobena k tomu, aby nám umožňovala se vyrovnat se světem. Davidson má podobnou představu, kde naše přesvědčení vznikají díky triangulaci mezi námi, ostatními lidmi a světem. Kde poznatky o těchto kategoriích získáváme skrze naše smyslové orgány.[65] To, co Dreyfuse a Davidsona spojuje, je holistická představa intelektu, který je umožněn neustálou interakcí se světem. Scénář, který nám PK nabízí, je nesmyslný v tom, že stojí na představě SI, která je nevtělená a oddělená od světa.

Nejlépe se to ukazuje, když posoudíme tezi, že SI bude mít jeden konečný cíl. Inženýři UI používají užitkovou funkci pro modelování chování UI.[66] Z této myšlenky vyplývá představa konečného cíle SI. Principem je, že jednotlivé činy UI se hodnotí podle toho, jak moc jsou užitečné pro splnění cíle této UI. Postup při vytváření programu je takový, že si zvolíme, jak užitkovou funkci danému programu definujeme. Takže pro Bostromův sponkový stroj to chování, které maximalizovalo počet vyrobených sponek, bylo nejvíce hodnotné vzhledem k jeho užitkové funkci. Tato užitková funkce ale v programu figuruje jako model světa, který určuje, co je a není relevantní. Tento model nemusí být statický, například Hibbard se snaží vytvořit užitkovou funkci, kterou si počítač bude schopen upravovat vzhledem k okolním podmínkám. Ale i v tomto případě musíme danou funkci nejdříve definovat na základě omezených dat z prostředí. Takže například protože svět obsahuje lidské bytosti,

64 Tamtéž, s. 152–155.
65 Davidson (2004b).
66 Omohundro (2008), Dewey (2011) a Hibbard (2012).

musíme UI definovat různé lidské vlastnosti a koncepty, jako zdraví, bohatství, blaho a podobné.[67] To znamená, že UI stále operuje na základě modelu či rámce, který jí my předáme. Díky Dreyfusovi ovšem víme, že nemůžeme explicitně zachytit vše, co může být relevantní v jakékoliv situaci. Protože pokud se budeme snažit různé situace reprezentovat UI, tak velice brzy budeme čelit nekonečnému regresu. Z tohoto důvodu vzniká problém rámce, který vytváří PK.

Na další problémy také narazíme, když se podíváme na povahu lidských cílů. Přestože můžeme vyjmenovat mnoho lidských cílů, žádný z nich nemá privilegovanou pozici nad ostatními. Důvodem je, že žádný takový konečný cíl ani mít nemůžeme. Všechny naše cíle vždy soudíme v širším kontextu dalších cílů, a nejen těch našich ale i cílů ostatních. To samé musí platit pro SI. I kdyby jí kvůli její architektuře byl dán konečný cíl, mohla by ho začít interpretovat pouze, pokud by operovala v takovémto širším kontextu.[68] Tento rámec by potřebovala právě i díky ortogonální tezi. Pokud můžeme spojit jakýkoliv cíl s jakoukoliv inteligencí, tak musí znát vše, čeho by se její cíl mohl týkat. Z toho vyplývá fakt, že obecná inteligence si žádá kontext, v jehož rámci musí operovat. Jinak by nebyla schopná se s komplexním světem vyrovnat. Tohle se vztahuje i na motivace, které také nelze mít mimo tento širší rámec. A pokud si obecná inteligence žádá takový kontext, tak je PK seriózně zpochybněn. Ortogonální teze může být správná, co se týče vztahu cílů a inteligence obecně, ale SI se rozhodně nemůže řídit pouze jedním konečným cílem. Tohle však má i širší dopady na celkovou koncepci tohoto problému, protože SI by musela sdílet některé charakteristiky našeho intelektu. To by znamenalo, že SI by se od nás tolik lišit nemohla. Tudíž její motivace by nemohly být natolik odlišné od těch našich. Ovšem chci upozornit, že neříkám to, že SI pro nás nemůže být hrozbou. Nesouhlasím s tím, jak singularitáni chápou tuto hrozbu.

67 Hibbard (2012, s. 7–8).

68 Häggström (2018) a Totschnig (2017) nabízí podobné argumenty. Häggström jako já argumentuje z povahy lidských hodnot, proti konceptu finálního cíle SI, ale nevysvětluje, proč naše hodnoty mají takovou povahu. Podobně Totschnig konstatuje, že hodnoty SI budou propojené se světem, příliš však nevysvětluje, jak jsou s ním propojené. Já zahrnuji tělesnost jako vysvětlující pojem, abych vysvětlil jak povahu našich hodnot, tak i jejich propojenost se světem.

3.7. Dva problémy „kontroly“

Jsem toho názoru, že singularitáni při formulaci PK zaměňují dva různé problémy „kontroly“, které spolu příliš nesouvisí. První je postaven na myšlence, že se potýkáme s odlišnou inteligencí. To můžeme spojit s konceptem OSI. Taková inteligence může existovat, pokud je však tak odlišná od té naší, nemůžeme o ní mluvit stejným způsobem jako o naší inteligenci. Tím mám na mysli, že by o ní nešlo mluvit odkazováním se na její motivace, pokud by skutečně postrádala rámec nutný pro obecnou inteligenci. Jestliže takový problém vztáhneme k UI, tak tu máme problém ryze inženýrský nebo technologický.[69] Jak zajistit, aby naše UI dělala to, co má? Jak namířit danou UI ke správnému cíli? Bavíme se tu tak o tom, jak správně danou inteligenci ovládat. Druhý problém se týká obecných inteligencí. O nich můžeme říct, že mají motivace a přesvědčení. To znamená, že danou inteligenci chápeme jako osobu, která má morální zodpovědnost. Pokud vznikne SI, věřím, že to bude taková obecná inteligence, ale potom musíme daný problém chápat jiným způsobem. Tento problém je podobnější problémům etickým nebo politickým,[70] kde se potýkáme s individui, které mají větší moc než my. Jde o snahu tuto jejich větší moc dostat pod kontrolu tím, že na ně eticky apelujeme, aby jednaly v náš prospěch. Proto tento problém můžeme chápat v terminologii motivací, protože naším cílem je správně motivovat SI. Zde tak mluvíme o normativním motivování osob.

Oba tyto problémy jsou o kontrole, ale v jiném smyslu. První se typicky týká nástrojů, kde naším problémem je, jak je máme správně ovládat. Zde je problém rozšířen o to, že náš nástroj oplývá vlastní inteligencí. Druhý se naopak týká osob a jde spíše o kontrolu ve smyslu normativity, kde se daný jedinec řídí etickými pravidly, a tím kontrolujeme jeho motivaci. Tyto dva způsoby kontroly jsou však od sebe odlišné. Ovládat můžeme jen nástroje, normativně řídit pouze osoby. Status daného objektu či bytosti vyplývá z jeho charakteristik. Singularitáni tak mají tyto dvě možnosti, jak PK koncipovat. Buď jako problém ovládání, kde musí SI chápat jako OSI a vzdát se terminologie motivací. Nebo jako problém normativního

69 Viz Totschnig (2017).
70 Tamtéž.

řízení, pak se však musí vzdát konceptu radikálně odlišné inteligence a svůj problém vztáhnout blíže k běžným etickým problémům.

Pokud je prostor myslí omezen tak, jak tu bylo ukázáno, potom ontologická sázka singularitánů zřejmě nevyjde. Pokud chtějí dále mluvit o PK, musí si vybrat mezi dvěma problémy, které tu jsou. Tyto problémy jsou stejně závažné jako PK. Přestože jsem tu demonstroval, že Bostromův sponkový stroj asi není skutečnou SI, stále může být OSI. Je to právě jeho omezená inteligence, která z něho činí hrozbu. V takovém podání by ale tento stroj postrádal schopnost předvídání lidského chování. To by z něho mohlo učinit menší hrozbu. Problém normativního řízení je stejně důležitý. Jak jste si mohli všimnout, jediné, co jsem ukázal, bylo, že SI nebude mít příliš odlišné motivace od těch našich. To, co jsem nevyřešil, je problém toho, jestli SI bude mít motivaci k tomu se řídit našimi příkazy. Na druhou stranu ale fakt, že taková SI bude mít nám podobné motivace, ji činí daleko více předvídatelnou. Otázkou je, jestli budeme dávat takové SI příkazy. Jestli mám pravdu, a budeme nuceni SI v takovém případě považovat za osobu, tak se už nemůžeme bavit o příkazech. Místo toho budeme muset se SI vyjednávat stejně jako s člověkem. Můžeme tedy vidět, že problém normativního řízení operuje s jinou dynamikou než problém ovládání. Důležité je správně rozlišit před jakým problémem s danou UI stojíme.

4. Závěr

Závěrem se vraťme ke dvěma otázkám z úvodu. Může SI vzniknout? V první sekci jsem zkoumal otázku procesu jejího vzniku. SI je dle mého názoru možná, ale podstatné je, jakou povahu SI bude mít. Zastávám pozici, že pokud SI vznikne, bude to intelekt postavený na základě toho našeho. Jak jsem však již poznamenal, jsem otevřený možnosti, že se naší inteligencí necháme inspirovat nejen v úzkém, ale také v širokém smyslu. Jsem si však jist tím, že taková inteligence bude nutně vtělená. Obecně se přikláním k myšlence, že nejlepší cesta k UI povede skrze vtělenou kognici, protože tělesnost vytváří důležitý základ pro inteligenci. Nejsem úplně proti jakémukoliv reprezentacionalismu v rámci tělesnosti. Jsem toho názoru, že intelekt jako náš musí obsahovat i určitou reprezentační

složku, která možná je i v některých aspektech podobná SDUI, není však jádrem našeho intelektu. Samozřejmě je zde překážka rozšiřitelnosti, která je podle mého názoru obzvláště silná u biologicky inspirovaných systémů. Tuto překážku však nevidím jako nepřekonatelnou. Jenom ji chápu jako velice obtížnou, a myslím si, že její překonání potrvá déle, než jak by si singularitáni přáli. Také si nedovoluji, jakkoliv předpovídat, kdy by SI mohla vzniknout. Moje pozice je jenom taková, že nevidím důvod, proč bychom její existenci měli kategoricky odmítnout.

Pokud SI vznikne, budeme stát před problémy? Myslím si, že rozhodně budeme. Nesouhlasím však se singularitány v tom, jakého druhu tyto překážky budou. V druhé sekci jsem se zabýval jejich konceptem PK. Mám zásadní problémy s tím, jak je PK prezentován. Singularitáni trpí zásadním zmatením ohledně PK a konceptu SI. Jedním zdrojem tohoto zmatení jsou argumenty pro vznik UI. Singularitáni se nechali příliš ovlivnit evolučním argumentem. To jim umožňuje představu SI jako radikálně odlišné od nás. Druhým zdrojem tohoto zmatení je fakt, že zakládají svoji argumentaci na povaze současných počítačů. Singularitáni si představují SI, která je lepší než my ve všech ohledech, ale zároveň trpí stejnými omezeními jako současné počítače. Na jednu stranu je od nich rozumné, že staví na současném fungování UI, na druhou stranu však neberou v potaz to, jak moc by SI byla odlišná v porovnání se současnou UI. To se promítá do PK, který dle mého soudu neodpovídá jejich konceptu SI a spíše popisuje problémy spojené s OSI. To je vede k formulaci PK způsobem, který je neudržitelný. Samozřejmě toto zmatení není jediným zdrojem pochybností o PK. Je zde také mé stanovisko, že obecná inteligence si žádá obecný rámec, který může vzniknout pouze interakcí se světem. To vede ke konceptu holistického intelektu, který podkopává jak tezi konečného cíle, tak i problém interpretace ve spojení s PK. Osoba nemůže mít jeden konečný cíl, protože cíle můžeme chápat pouze v kontextu s ostatními cíli. To samé platí pro interpretaci, protože můžeme pochopit výroky druhých, jenom když do nich promítneme naše vlastní přesvědčení. Tato přesvědčení závisí na interakci se světem, který zahrnuje i ostatní osoby. Pokud obecný intelekt vyžaduje všechna tato kritéria, pak SI nutně nemůže být radikálně odlišná od nás, a tím pádem také nemůže mít příliš odlišné motivace. SI může stále být nebezpečná, tato hrozba však nestojí na velké ontologické propasti.

Pro mě je klíčový fakt, že z tělesnosti vyplývá holismus. Náš intelekt je na tomto holismu založen. To hrozbu SI nutně charakterizuje jinak než PK. Tato povaha obecné inteligence, výrazně omezuje prostor možných myslí. Pokud SI bude obecným intelektem, tak bude muset být vtělená. Z této tělesnosti budou vyplývat hodnoty této SI. Otázkou do budoucna je, nakolik by takové hodnoty byly podobné našim. Přestože jsem stanovil tělesnost jako podmínku, i v rámci tohoto omezení můžeme najít dost různorodosti.[71] Možné rozdíly mezi tělesností SI a naší by se mohly promítnout jako rozdíl v motivacích. Nejsem ale toho názoru, že tento rozdíl by mohl být stejně velký, jako rozdíl, který koncipují singularitáni. Osobně jsem v tomto ohledu celkem optimistický, díky poznatkům z etologie. Možná nikdy nepoznáme, co to znamená být netopýrem.[72] Ale samotný fakt, že jsme byli schopni objevit, že netopýr vnímá svět jinak, je významné, protože nám to umožňuje lépe chápat jeho chování. Díky znalosti echolokace lépe odhalíme jeho motivace. To samé je aplikovatelné na SI, pokud bude od nás odlišná. Proč tohle tvrdím? Protože tělesnost netopýra vyplývá z jeho životních podmínek. Echolokaci jsme objevili, protože jsme hledali odpověď na otázku, jak může netopýr navigovat svůj pohyb v temných prostorách, ve kterých žije. Tato otázka přímo vyplývala z prostředí, které netopýr obývá. Tělesnost nás váže s naším prostředím, pomáhá nám se s ním vyrovnat. To znamená, že naše hodnoty vyplývají z našeho prostředí. Jaké kroky netopýr podnikne, vychází z jeho okolí. Prostředí skrze tělesnost vyvíjí tlak na naše hodnoty. Já argumentuji, že to samé musí platit pro SI. Svět vytváří rámec, ve kterém dané hodnoty mohou být, pokud se takové hodnoty dostanou s prostředím do konfliktu, tak nemohou dále existovat. Možná lepší je říci, že bytosti s takovými hodnotami dlouho ve světě nepřežijí. Pokud moje argumenty platí, tak SI bude muset být

71 V textu jsem se hlavně zaměřil na vztah mezi tělesností a obecnou inteligencí. Mohli bychom ovšem dále spekulovat, jakou roli dále hraje tělesnost pro inteligenci na lidské úrovni. V tomto bodě si myslím, že prostor je ještě užší než prostor obecných inteligencí. Pro to, aby vznikla inteligence na stejné úrovni jako my, bych argumentoval, že tělesnost musí obsahovat určité elementy. Například bych argumentoval, že je zapotřebí tělesné ústrojí podobné našim rukám. Vždy jsem souhlasil s myšlenkou, že pro vývoj naší inteligence byly důležité, protože nám umožnily interagovat s prostředím způsoby, které dále rozvíjely náš intelekt. To by mohla být jedna podmínka pro inteligenci na lidské úrovni.

72 Viz Nagel (1974).

vtělená. Z toho vyplývá, že její hodnoty budou vycházet z toho samého prostředí, jaké obýváme my. To znamená, že její hodnoty se nebudou příliš lišit od těch našich.

Mohli bychom ale skutečně pochopit chování SI? Nehráli bychom my roli netopýra v případě SI? Můj názor je takový, že na rozdíl od netopýra my máme určitou výhodu, při interpretování SI. Jak konstatuje Dennett,[73] my jsme bytosti, které jsou schopny kompetence s pochopením. Nejenom, že jsme schopni něčeho dosáhnout, také jsme schopni pochopit, jak jsme toho dosáhli. Dennett v této knize také zaujímá postoj, že UI bude blíže zvířatům v tom smyslu, že bude postrádat sebeuvědomění. V tomto bodě s ním nesouhlasím, jednoho dne UI dosáhne sebeuvědomění, a možná i přesáhne naše schopnosti a stane se SI. Tohle však není tak nepodobné mnoha situacím, ve kterých se můžeme ocitnout už dnes. Často se setkáváme s lidmi, kteří překračují naše schopnosti. Geniové existují, a přesto jsme schopni s nimi sdílet svět. Jsme schopni s nimi vycházet. Jak jsme toho schopni, když mnohdy nemůžeme úplně chápat jejich činy? To je pro mě daleko důležitější otázka, než otázka toho, jestli SI budeme schopni pochopit. Dle mého názoru, jsme toho schopni právě proto, že jsme schopni si uvědomovat sami sebe. Tím pádem jsme schopni určit co je pro nás důležité, co je našimi hodnotami. To nám umožňuje mezi sebou vyjednávat. Nejsem toho názoru, že musíme SI úplně chápat. Stačí pouze to, že bude schopná nám dát najevo své hodnoty. To samé my musíme být schopni formulovat pro SI. Netvrdím, že proces tohoto vyjednávání bude snadný, ale musíme tento problém takovýmto způsobem nejdříve uchopit, abychom ho mohli začít řešit.

Literatura

Armstrong, S., Sandberg, A. & Bostrom, N. (2012): „Thinking inside the box: controlling and using an oracle AI.“ *Minds and Machines* 22 (4): 299–324.

Armstrong, S. (2013): „Risks and Mitigation Strategies for Oracle AI.“ In *Philosophy and Theory of Artificial Intelligence. Studies in Applied Philosophy, Epistemology and Rational Ethics vol. 5*. ed. V. Müller, Springer, Berlin, Heidelberg, s. 335–347.

73 Dennett (2018).

Barandiaran, X. E., Di Paolo, E. A. & Rohde, M. (2009): „Defining agency: individuality, normativity, asymmetry, and spatio-temporality in action." *Adapt Behav* 17 (5): 367–386.

Beer, R. D. (2014): „Dynamical systems and embedded cognition." In *The Cambridge Handbook of Artificial Intelligence*, eds. K. Frankish & W. Ramsey, Cambridge University Press, Cambridge, 2014, s. 128–148.

Block, N. (1981): „Psychologism and behaviorism." *Philosophical Review* 90: 5–43.

Boden, M. A. (2014): „GOFAI." In *The Cambridge Handbook of Artificial Intelligence*, eds. K. Frankish & W. Ramsey, Cambridge University Press, Cambridge, 2014, s. 89–107.

Bostrom, N. (2003): „Ethical Issues in Advanced Artificial Intelligence." In *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, eds. I. Smit et al. International Institute for Advanced Studies in Systems Research and Cybernetics, Tecumseh, s. 12–17. Dostupné z: https://nickbostrom.com/ethics/ai.html.

Bostrom, N. (2006): „How long before superinteligence?" *Linguistic and Philosophical Investigations* 5 (1): 11–30. Dostupné z: https://www.nickbostrom.com/superintelligence.html.

Bostrom, N. (2014): *Superinteligence, Paths, Dangers, Strategies*. Oxford University Press, Oxford; český překlad (J. Petříček) *Superinteligence: Až budou stroje chytřejší než lidé*, Prostor, Praha, 2018.

Bostrom, N. (2012): „The superintelligent will: motivation and instrumental rationality in advanced artificial agents." *Minds and Machines* 22: 71–85. Dostupné z: https://www.nickbostrom.com/superintelligentwill.pdf.

Bostrom, N. (2005): „Transhumanist Values." *Review of Contemporary Philosophy* 4 (1–2): 87–101. Dostupné z: https://www.nickbostrom.com/tra/values.html.

Bostrom, N. & Shulman, C. (2016): „How Hard is Artificial Inteligence? Evolutionary Arguments and Selection Effects." In *The Singularity: Could artificial inteligence really out-think us (and*

would we want it to)? [e-kniha], ed. U. Awret, Imprint Academic, Exeter, 2016 [cit. 15. 8. 2021]. Dostupné z: https://www.amazon.com/Singularity-Journal%20Consciousness-Studies-ebook/dp/B01N1N6KLZ/ref=mt_kindle?_encoding=UTF8&me.

Clark, A. (1997): *Being There: Putting Brain, Body and World Together Again*. MIT Press, Cambridge, MA.

Clark, A. & Chalmers, D. J. (1998): „The extended mind." *Analysis* 58 (1): 7–19.

Colzato, L. S., Hommel, B. & Beste, C. (2021): „The Downsides of Cognitive Enhancement." *The Neuroscientist* 27 (4): 322–330.

Chalmers, D. J. (1996): *The Conscious Mind*. Oxford University Press, Oxford, New York.

Chalmers, D. J. (2016a): „The Singularity: A Philosophical Analysis." In *The Singularity: Could artificial inteligence really out-think us (and would we want it to)?* [e-kniha], ed. U. Awret, Imprint Academic, Exeter, 2016 [cit. 15. 8. 2021]. Dostupné z: https://www.amazon.com/Singularity-Journal%20Consciousness-Studies-ebook/dp/B01N1N6KLZ/ref=mt_kindle?_encoding=UTF8&me.

Chalmers, D. J. (2016b): „The Singularity: A Reply to Commentators." In *The Singularity: Could artificial inteligence really out-think us (and would we want it to)?* [e-kniha], ed. U. Awret, Imprint Academic, Exeter, 2016 [cit. 15. 8. 2021]. Dostupné z: https://www.amazon.com/Singularity-Journal%20Consciousness-Studies-ebook/dp/B01N1N6KLZ/ref=mt_kindle?_encoding=UTF8&me.

Davidson, D. (1984): „Radical Interpretation." In *Inquiries into Truth and Interpretation*, D. Davidson, Oxford University Press, Oxford, 1984, s. 125–139.

Davidson, D. (2004a): „Koherenční teorie pravdy a poznání." In *Subjektivita, Intersubjektivita, Objektivita*. D. Davidson; český překlad (J. Kolář a T. Marvan) Filosofia, Praha, 2004, s. 129–151.

Davidson, D. (2004b): „Tři druhy poznání." In *Subjektivita, Intersubjektivita, Objektivita*. D. Davidson; český překlad (J. Kolář a T. Marvan) Filosofia, Praha, 2004, s. 194–209.

Dennett, D. C. (1984): „Cognitive wheels: The frame problem of AI.“ In *Minds, Machines and Evolution: Philosophical Studies*, ed. C. Hookway, Cambridge University Press, Cambridge, s. 129–150.

Dennett, D. C. (1987): *The Intentional Stance*. MIT Press, Cambridge, MA.

Dennett, D. C. (2018): *From Bacteria to Bach and Back: The evolution of Minds*. Penguin Books, London.

Dewey, D. (2011): „Learning what to value.“ In *Artificial General Intelligence: Proceedings of 4th International Conference*, eds. J. Schmidhuber, K. R. Thórisson & M. Looks, Springer, Berlin, s. 309–314. Dostupné z: https://intelligence.org/files/LearningValue.pdf.

Dreyfus, H. L. (2012): „A History of First Step Fallacies.“ *Minds and Machines* 22 (2): 87–99.

Dreyfus, H. L. (1992): *What Computers Still Can't Do*. MIT Press, Massachusetts.

Dreyfus, H. L. (1967): „Why Computers Must Have Bodies in Order to Be Intelligent.“ *The Review of Metaphysics* 21 (1): 13–32.

Fodor, J. A. (1983): *The Modularity of Mind*. MIT Press, Cambridge, MA.

Good, J. I. (1965): „Speculations concerning the first ultraintelligent machine.“ In *Advances in Computers Volume 6*, eds. F. Alt & M. Rubinoff, Academic Press Inc., New York, s. 33–88.

Häggström, O. (2018): „Challenges to the Omohundro-Bostrom framework for AI motivations.“ *Foresight* 21 (1): 153–166. Dostupné z: http://www.math.chalmers.se/~olleh/ChallengesOBframeworkDeanonymized.pdf.

Hibbard, B. (2012): „Model-based utility functions.“ *Journal of Artificial General Intelligence* 3 (1): 1–24. arXiv:1111.3934, Dostupné z: https://arxiv.org/ftp/arxiv/papers/1111/1111.3934.pdf.

Jackson, F. (1982): „Epiphenomenal Qualia.“ *Philosophical Quarterly* 32 (127): 127–132.

Jackson, F. (1986): „What Mary Didn't Know.“ *Journal of Philosophy* 83 (5): 291–295.

Kurzweil, R. (2005): *The Singularity Is Near. When Humans Transcend Biology*. Viking Adult, New York.

Legg, S. (2008): *Machine Super Intelligence*. Univerzita Lugano, Lugano. Doktorská dizertace.

Marcus, G. (2018): „Deep learning: A critical appraisal.“ In *ArXiv* [online]. 2. 1. 2018 [cit. 15. 8. 2021]. arXiv:1801.00631. Dostupné z: https://arxiv.org/abs/1801.00631.

McCarthy, J. & Hayes, P. J. (1969): „Some Philosophical Problems from the Standpoint of Artificial Intelligence.“ In *Machine Intelligence 4*, eds. D. Michie & B. Meltzer, Edinburgh University Press, Edinburgh, s. 463–502.

McDermott, D. (2016): „Response to ‚The Singularity‘ by David Chalmers.“ In *The Singularity: Could artificial inteligence really out-think us (and would we want it to)?* [e-kniha], ed. U. Awret, Imprint Academic, Exeter, 2016 [cit. 15. 8. 2021]. Dostupné z: https://www.amazon.com/Singularity-Journal%20Consciousness-Studies-ebook/dp/B01N1N6KLZ/ref=mt_kindle?_encoding=UTF8&me.

Minsky, M. (1968): „Machines Are More Than They Seem.“ *Science Journal* 4 (10): 3–43.

Moravec, H. P. (1976): *The Role of Raw Power in Intelligence*. [online]. Nepublikovaný spis. 12. 5. 1976 [cit. 15. 8. 2021]. Dostupné z: https://frc.ri.cmu.edu/~hpm/project.archive/general.articles/1975/Raw.Power.html.

Muehlhauser, L. & Helm, L. (2012): „The Singularity and Machine Ethics.“ In *Singularity Hypotheses. A Scientific and Philosophical assesment*, eds. H. A. Eden et al., Springer, Heidelberg, 2012, s. 101–125.

Müller, V. C. & Bostrom, N. (2016): „Future progress in artificial intelligence: A survey of expert opinion.“ In *Fundamental Issues of Artificial Intelligence*, ed. V. C. Müller, Synthese Library, Berlin, s. 553–571.

Nagel, T. (1974): „What Is It Like to be a Bat?“ *Philosophical Review* 83 (4): 435–450.

Omohundro, S. M. (2016): „Autonomous technology and the greater human good.“ In *Risks of artificial intelligence*, ed. V. C. Muller, CRC Press, Boca Raton, s. 9–27.

Omohundro, S. M. (2012): „Rational artificial intelligence for the greater good.“ In *Singularity Hypotheses. A Scientific and Philosophical assesment*, eds. H. A. Eden et al., Springer, Heidelberg, 2012, s. 161–176.

Omohundro, S. M. (2008): „The basic AI drives.“ In *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, eds. P. Wang, B. Goertzel & S. Franklin, IOS, Amsterdam, s. 483–492.

Pfeifer, R. & Bongard, J. (2006): *How the Body Shapes the Way We Think: A New View of Intelligence*. MIT Press, Cambridge, MA.

Russel, S. J. & Norvig, P. (2020): *Artificial Intelligence: A Modern Approach*. Pearson, Boston.

Sandberg, A. & Bostrom, N. (2008): *Whole Brain Emulation: A Roadmap* [online]. Future of Humanity Institute, University of Oxford, 2008 [cit. 15. 8. 2021]. Dostupné z: https://www.fhi.ox.ac.uk/brain-emulation-roadmap-report.pdf.

Searle, J. R. (1980): „Minds, brains and programs.“ *The Behavioral and Brain Sciences* 3 (3): 417–424.

Searle, J. R. (1984): *Minds, Brains and Science*. Harvard University Press, Cambridge, Massachusetts; český překlad (M. Nekula) *Mysl, mozek a věda*, Mladá fronta, Praha, 1994.

Searle, J. R. (2014): „What Your Computer Can’t Know.“ *The New York Review of Books* 9. října 2014.

Simon, H. A. & Newell A. (1958): „Heuristic Problem Solving: The Next Advance in Operations Research.“ *Operations Research* 6 (1): 1–10.

Sun, R. (2014): „Connectionism and neural networks.“ In *The Cambridge Handbook of Artificial Intelligence*, eds. K. Frankish & W. Ramsey, Cambridge University Press, Cambridge, 2014, s. 108–127.

Totschnig, W. (2019): „The problem of superintelligence: Political, not technological.“ *AI & Society* 34 (4): 907–920.

Ulam, S. (1958): „John Von Neumann 1903–1957.“ *Bulletin of the American Mathematical Society* 64 (3): 1–49.

Varela, F. J., Thompson, E. & Rosch, E. (2016): *The Embodied Mind*. MIT Press, Cambridge, MA.

Vinge, V. (1993): „The coming technological singularity. How to survive in the post-Human era.“ *Vision 21: Interdisciplinary Science and Engineering in the Era of Cyberspace* [online]. NASA, 1. 12. 1993 [cit. 15. 8. 2021]. Dostupné z: https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19940022856.pdf.

Yudkowsky, E. (2008): „Artificial Intelligence as a positive and negative factor in global risk.“ In *Global Catastrophic Risks*, eds. N. Bostrom & M. Cirkovic, Oxford University Press, New York, s. 308–345.

Yudkowsky, E. (2001): *Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures* [online]. The Singularity Institute, San Francisco, CA. [cit. 15. 8. 2021]. Dostupné z: https://intelligence.org/files/CFAI.pdf.

Yudkowsky, E. (2004) : *Coherent Extrapolated Volition* [online]. The Singularity Institute, San Francisco, CA. [cit. 15. 8. 2021]. Dostupné z: https://intelligence.org/files/CEV.pdf.

Yudkowsky, E. (2011): „Complex value systems in friendly AI.“ In Schmidhuber, T. & Looks, M. (2011): 388–393. Dostupné z: https://intelligence.org/files/ComplexValues.pdf.

Abstract

Superintelligence and the control problem: Real problem or pseudo--problem?

In this paper, I deal with the concept of SI (superintelligence) and the control problem. According to a group of AI theorists, we will soon experience an event that can change technological progress and human society. This event is the technological singularity associated with the emergence of the first greater than human intelligence. People like Nick Bostrom stress the SI's dangers and urge us to find methods to control this intelligence. According to Bostrom and others, the threat of SI stems from its nature. This article considers how SI can be created and judges the logic of the control problem. SI is possible only if we can create AI. For this reason, a section of the text concentrates on the arguments for the creation of AI. It is shown how Bostrom and others base their thesis on one problematic

argument and assumptions of their predecessors. Their position is subjected to the classical critique of artificial intelligence. I primarily focus my criticism on the claim that SI will have one final goal, which it will interpret. This statement is antithetical to the idea that SI will be a general intelligence. I conclude that the control problem confuses two other different “control” problems.

Key words: technological singularity, superintelligence, artificial intelligence, whole brain emulation, control problem, embodiment

Malík, J. (2021): „Superinteligence a problém kontroly: Skutečný problém nebo pseudo-problém?“ *Filosofie dnes* 13 (2): 73–118. Dostupné z www.filosofiednes.ff.uhk.cz.