Západočeská univerzita v Plzni

Fakulta filozofická

Diplomová práce

Funkcionalismus ve filosofii mysli a umělé inteligenci

Jan Korda

Plzeň 2015

Západočeská univerzita v Plzni

Fakulta filozofická

Katedra filozofie

Studijní program Humanitní studia

Studijní obor Teorie a filozofie komunikace

Diplomová práce

Funkcionalismus ve filosofii mysli a umělé inteligenci

Jan Korda

Vedoucí práce:

Mgr. Michal POLÁK, Ph.D.
Katedra filozofie
Fakulta filozofická Západočeské univerzity v Plzni

Plzeň 2015

Prohlašuji, že jsem práci zpracoval samostatně a použil jen uvedených pramenů a literatury.

Plzeň, duben 2005 ……

Tímto bych rád poděkoval Mgr. Michalu Polákovi, Ph.D. za ochotu vést mou práci stejně jako za kritické poznámky a přínosné rady při jejím psaní.

Obsah

Úvod ... 1

1. Problém mysli a těla ... 2

1.1. Přístupy ke zkoumání mysli a těla ... 3

1.2. Problematika zkoumání ... 5

2. Turingův test ... 10

3. Funkcionalismus Hilaryho Putnama ... 20

3.1. Povaha mentálních stavů ... 20

3.2. Putnamova kritika ... 24

4. Funkcionalismus Neda Blocka ... 28

4.1. Funkcionalistické přístupy ... 28

4.2. Blockova kritika ... 30

4.2.1. Homunkulární mozek a čínský národ ... 31

4.2.2. Argument chybějících kválií ... 33

4.2.3. Problematika vstupů ... 35

5. Rozšířená mysl ... 37

5.1. Koncepce Clarka a Chalmerse ... 37

5.2. Kritika ... 41

6. Umělá inteligence ... 47

6.1. Umělá inteligence jakožto vědní disciplína ... 47

6.2. Historie umělé inteligence ... 49

6.3. Přirozené a umělé ... 50

6.4. Dvě Paradigmata ... 52

6.5. Čínská komora Johna Searla ... 57

Závěr ... 67

Seznam použité literatury a pramenů ... 69

Resumé ... 73

Úvod

Funkcionalismus významně ovlivnil kognitivní vědy, konkrétně filosofii mysli a umělou inteligenci. Ač byl explicitně formulován ve 2. poloviny 20. století, jeho kořeny sahají mnohem hlouběji. Funkcionalismus získal mnohé stoupence, jelikož jeho hlavní teze o mnohonásobné realizovatelnosti mysli byla velice atraktivní. Funkcionalismus vzbudil velký zájem, ale i vlnu kritiky, a to jak na poli filosofie mysli, tak na poli umělé inteligence.

V české literatuře se funkcionalismus objevuje většinou v souvislosti s umělou inteligencí, z hlediska filosofie mysli o něm pojednává pouze Nosek (1997) a Polák (2013). Základním cílem této práce je představit funkcionalismus ve filosofii mysli, jeho hlavní myšlenky a východiska, a dále dvě oblasti bádání, které z funkcionalismu vycházejí. Nezbytnou součástí bude i kritika, a to jak samotného funkcionalismu, tak i zmíněných navazujících teorií.

Pokusím se zasadit funkcionalismus do širších souvislostí a ukázat na jeho vztah k jiným teoriím. První dvě kapitoly přiblíží problém mysli a těla, na který se funkcionalismus snaží odpovědět, a dále představí některé myšlenky Alana Turinga, jimiž byla inspirována první verze funkcionalismu. V druhé kapitole představím nejdříve myšlenky Hilaryho Putnama, jež byl první explicitní formulaci funkcionalismu ve filosofii mysli. Třetí kapitola je zaměřena na koncepci rozšířené mysli a čtvrtá, poslední, kapitola na umělou inteligenci.

Ve své práci budu argumentovat ve prospěch teze o mnohonásobné realizovatelnosti mysli, která tvoří jádro funkcionalismu. Čtvrtá kapitola má za cíl dokázat mnohonásobnou realizovatelnost alespoň některých kognitivních procesů, zatímco pátá kapitola se zaměřuje na realizaci mysli na jiném než biologickém substrátu, a představí dvě hlavní možnosti, jak toho dosáhnout. Vycházet budu hlavně z primární literatury, přičemž většina do češtiny nebyla přeložena. Výjimkou je první část první kapitoly, a některé podkapitoly poslední kapitoly, kdy jsem využil sekundární české literatury.

1. Problém mysli a těla

Na počátku úvah, jež stojí za celým tímto textem, bylo rozlišení těla a mysli[1]. Nelze říci, kdy k němu došlo, avšak pravděpodobně tomu bylo ještě „dříve, než se tímto rozlišováním začaly zabývat mýtus, náboženství, filosofie, nebo věda“ (Nosek, 1997, s. 8). Dá se tedy říci, že jde o filosofický problém s předfilosofickými počátky. Při jeho zkoumání se můžeme zaměřit na shody nebo na rozdíly mezi myslí a tělem. Od dob Descartových převažuje zaměření se na odlišnosti, a ne jinak je tomu i v současnosti. Nejsnáze uchopitelnou představou je pro nás kauzální vztah mezi tělem a myslí, ten však může vést jak od těla k mysli, tak opačně. Dostáváme se tedy k problému mysli a těla[2] (mind-body problem[3]), k otázce po jejich vzájemném vztahu, která je klíčová pro filosofii mysli. Na jedné straně máme mozek, hmotný, reálně existující a objektivně pozorovatelný orgán, v protikladu k němu nehmotnou mysl, která se vymyká empirickému zkoumání i jasnému a vševysvětlujícímu usouvztažnění s neurální aktivitou mozku. Na tento vztah lze nahlížet ze dvou rovin – „filosofické (metafyzické) a empirické. Ve filosofické rovině lze dále rozlišit problémy kauzální, ontologické, konceptuální a metodologické“ (Polák, 2013, s. 20).

Lidskou přirozeností je touha po poznání, co by tedy mohlo být významnějším cílem, než poznání sebe sama, naší mysli? S tou je spojena celá řada otázek jako: co je mysl, jak vzniká, je vlastní pouze člověku, a další. Dosud jsme ale nebyli schopni zodpovědět otázku, zdali mysl skutečně existuje, natož jak ji zkoumat a popsat.

[1] Nebudu se zde zabývat možnými důvody, proč k tomuto rozlišení došlo, zájemci nechť nahlédnou do (Nosek, 1997, s. 7-8).

[2] Pod pojmem tělo se většinou myslí „mozek a jeho specifické regiony, jejich koaktivace, neurotransmitery atd.“ (Havlík, 2013).

[3] Filosofické ustanovení problému těla a mysli je spojeno se vznikem filosofie mysli v polovině 20. století. Zásadním dílem pro filosofii mysli byl *Pojem mysli* G. Ryla. Filosofie mysli je součástí kognitivních věd, a přesahuje do metafyziky, etiky, epistemologie a filosofie vědy či jazyka. Kromě problému mysli a těla se filosofie mysli zabývá i vědomím, mentálními reprezentacemi a dalšími tématy (Jaworski, 2011, s. 1), a můžeme v ní rozlišit fenomenologický a analytický proud (Havel, 2001. s. 55). Block (1980, s. 6) zařazuje filosofii mysli pod filosofii psychologie.

1.1. Přístupy ke zkoumání mysli a těla

Vzhledem k obsáhlosti a povaze filosofie mysli není nikterak překvapivé množství a různorodost názorových proudů, jež se tímto problémem zabývají. Tyto přístupy lze rozčlenit na základě Bieriho trilematu, které lze shrnout následovně:

> 1) Oblast fyzických fenoménů je kauzálně uzavřená a je sama sebou vysvětlitelná.
> 2) Mentální fenomény nejsou fyzické fenomény.
> 3) Mentální fenomény jsou v oblasti fyzických fenoménů kauzálně účinné[4].
>
> Stavy a procesy mysli mohou pomoci vysvětlit fungování těla.

Teze 1) a 2) považuje Nosek za metafyzické, a jako takové v zásadě nedokazatelná či nevyvratitelná. Oproti tomu teze 3) je empirická a musí tedy být přinejmenším vyvratitelná. Pro zachování konsistence je třeba zvolit právě dvě z těchto tří tvrzení, přičemž přijetím 1) a 2) se hlásíme k epifenomenalismu či paralelismu, přijetím 2) a 3) k dualismu a konečně přijetím tezí 1) a 3) k materialismu či fyzikalismu, mezi které patří i funkcionalismus (Nosek, 1997, s. 9-10; Polák, 2013, s. 25-26).

Množství přístupů k problému mysli a těla nám poskytuje široké spektrum názorů, co vlastně mysl je. Obecně přijímané dělení je na monistické a dualistické teorie, avšak zařazení dílčích přístupů je již odlišné. Zde si dovolím vypůjčit stručnou charakteristiku, jak ji uvedl Nosek (1997, s. 13-14):

> A. Teze logického behaviorismu C. Hempela a G. Ryla: mysl je chování nebo dispozice k chování.
> B. Teze teorie psychofyzické identity U. T. Placeho, J. J. C. Smarta a H. Feigla: mysl není chování, mysl je část mozku.
> C. Teze funkcionalismus D. Armstronga, D. Lewise, H. Putnama, J. Fodora: mysl není chování ani mozek, ale druh kauzální funkce s fyzickou realizací, resp. program počítače s libovolnou fyzickou realizací.
> D. Teze epifenomenalismu K. Campbella a teorie supervenicence D. Davidsona aj. Kima: mysl není ani chování, ani mozek, ani funkce, ale kauzálně impotentní a anomální fenomenální kvalita, fyzicky fixovaná.

[4] Nosek (1997, s. 9) upozorňuje, že toto tvrzení nevypovídá nic o tom, jakého druhu jsou stavy a procesy mysli. Mohou být tělesné či netělesné; jevy, zdání nebo skutečnosti.

E. Teze eliminativního materialismu W. Quina, P. Feyerabenda, R. Rortyho, P. Churchlanda: mysl není ani fenomén, ani něco tělesného, ale zbytečně postulovaná entita nebo iluze.

F. Teze naturalismus J. Searla a T. Nagela: mysl je reálná fyzická subjektivní vlastnost mozku.

G. Teze interakcionismu K. Poppera aj. Ecclese: mysl je relativně samostatná subjektivní a reálná nefyzická entita, která je v interakci s mozkem.

Na první pohled je patrná převaha přístupů, jež považují mysl za určitým způsobem vázanou na fyzickou složku. Také je vidět, že ne všechny přístupy se vylučují se všemi ostatními a mohou mít (ba dokonce často mají) určité společné rysy (což vede k různým možnostem zařazení určitých přístupů). K popisu našich mentálních vlastností lze přistupovat následovně: „materialismus a fyzikalismus předpokládají, že fyzika může popsat všechny naše vlastnosti, dualismus vlastností předpokládá, že fyzika může popsat některé naše vlastnosti, [zatímco] dualismus substancí předpokládá, že fyzika nemůže popsat žádné naše vlastnosti“ (Jaworski, 2011, s. 9).

Navzdory veškeré kritice je stále zdrojem mnohých poznatků o mysli introspekce[5]. Ta je sice neomylná z hlediska zkoumání kvalitativních stavů, ovšem stejně tak je objektivně neověřitelná. Na druhou stranu, bez introspekce by naše poznání *pravděpodobně* nebylo úplné. Jinou možností je sledovat behaviorální či neurofyziologické projevy, jež umožňují oprostit se od mentalistických pojmů. Zde se dostáváme k roli úhlu pohledu, přesněji k protikladu hlediska první a třetí osoby. Z hlediska první osoby přistupuji k vlastním mentálním stavům a toto hledisko je spojováno s fenomenálním vědomím a kválii, zatímco z hlediska třetí osoby k mentálním stavům ostatních lidí a toto hledisko je spojováno s přístupovým[6] vědomím. Lidskou mysl lze také zkoumat pomocí umělých modelů – ty mohou být „matematické, počítačové, fyzikální nebo fyzické“ (Havel, 2001, s. 23). Tato metoda je úzce spojena s výzkumem umělé inteligence, a pro dokazování či vyvracení funkcionalistických hypotéz se jeví jako nejvhodnější.

[5] Havel (2001, s. 23) uvádí rozlišení mezi introspekcí a prožíváním, kdy „[i]ntrospekcí zkoumáme svou vlastní mysl a jednání jakoby *zvenku*, tj. sledujeme sebe z nějakého (nutně i časového) odstupu. Naproti tomu prožívání je to, co provází každou bdělou činnost, byť v různé intenzitě“.

[6] Havel (2001, s. 32) používá rozlišení na fenomenální a performační.

1.2. Problematika zkoumání

Při zkoumání vztahu těla a mysli se potýkáme s množstvím nejrůznějších obtíží, které nám naše pátrání znesnadňují, někdy dokonce i zcela znemožňují. Klíčové pojmy mysl a tělo jsou většinou chápány intuitivně, a odlišně v různých kontextech. Tato závislost na kontextu může způsobit, že „[o]dpověď buď nenajdeme, anebo nás neuspokojí, a ani nic nenasvědčuje tomu, že by se všechny disciplíny, které s tím mají co dělat, ve své odlišnosti vůbec kdy mohly na společné definici dohodnout“ (Havel, 2001, s. 22). Současné pojmy, které pramení z přirozené zkušenosti a lidové psychologie vyvolávají otázku, zda „mohou být funkcionalisticky interpretovány, a tím se stát i vědecky relevantními a využitelnými … [nebo jsou jen] vadně konstruované, a tedy vyloučitelné z jakéhokoli vědeckého zkoumání vůbec“ (Nosek, 1997, s. 30).

Stejně vágně vymezená je i inteligence. „[S]oučasní psychologové se nedokážou shodnout ani na tom, zdali vůbec existuje nějaká obecná inteligence, nebo se jedná o jev způsobený konglomerátem kooperujících a specializovaných modulů“ (Tvrdý, 2011, s. 95). Jako důležitou vlastnost inteligence uvádí Havel (2001, s. 20) „schopnost se vyvíjet – osvojovat si nové obecné schopnosti, které nebyly součástí původního vybavení“. Ryze behavioristická, avšak pro potřeby umělé inteligence vhodná, je Frenchova definice: „cokoli *jedná* postačujícím způsobem inteligentně, *je* inteligentní“ (cit. podle Tvrdý, 2011, s. 93). Naproti tomu popis ze sborníku *Umělá Inteligence 1*(Mařík, 1993, s. 15) v sobě skrývá jistou biologickou danost inteligence (z pohledu úvah, které budou následovat v podkapitole 6.3., by bylo vhodnější označení „přirozená inteligence“):

> Inteligence je vlastností některých živých organismů, která jim dává v přírodě mimořádné postavení. Vznikla a vyvíjela se v průběhu dlouhého vývoje. Dnes umožňuje některým živým organismům efektivně reagovat na složité projevy prostředí a aktivně je využívat ve svůj prospěch, k dosažení svých cílů.

Lze namítnout, že pro úspěšné poznání není zapotřebí (a mnohdy ani není možná) přesná definice zkoumaného. To samozřejmě platí u jevů dosud neobjevených, ale u jevů, jejichž poznáním (a popsáním) se zabýváme, jsou pojmy klíčové. Jedině s všeobecně přijímanými pojmy (jakkoliv budou nepřesné či nevhodné) je možné sjednocovat poznatky, formulovat a vyvracet hypotézy. Bez všeobecně přijímaných pojmů dospějeme ke konceptuálnímu chaosu, kdy budou autoři používat pojmy volně, každý dle svého (a

leckdy protikladně k autorům jiným). Obzvláště to platí u funkcionalismu, vzhledem k množství jeho přístupů[7].

Lycan (2003, s. 11) charakterizuje snahu funkcionalistů jako „vysvětlení mysli přírodními zákony, jako ontologickou (avšak samozřejmě ne typovou) redukci mentálního na fyzické či materiální“. Chomsky ale kritizuje uvedené použití slov fyzické či materiální, jelikož pojem „fyzický svět“ není striktně vymezený a může se měnit podle toho, jak bude postupovat naše poznání. Je možné, že učiníme vědecký pokrok, který bude vyžadovat zásadní změnu našich pojmů[8] tak, abychom byli schopni popsat dosud neobjevené a nepopsané entity a principy. Je tedy možné, že se problém mysli a těla „vyřeší způsobem podobným jakým se vyřešil pohyb nebeských těles, tedy uplatněním principů, které se zdály nesrozumitelné, či dokonce příčící se vědecké představivosti dřívějších generací“ (cit. podle Lycan, 2003, s. 12). Je tedy možné, že mentální nikdy nebude redukováno na „fyzické“ v současném smyslu, avšak není vyloučeno, že se pojem „fyzické“ rozšíří natolik, aby zahrnoval i mentální a tím umožnil jeho popsání v rámci přírodních věd. Chomsky dále tvrdí, že „[s] rozpadem tradiční teorie ‚hmoty‘, či ‚těla‘ se metafyzický dualismus stává neproveditelným; obdobně pojmy jako ‚fyzikalismus‘ či eliminativní materialismus‘ ztrácejí jasný význam – dokud nějaká nová představa ‚fyzického‘ nenahradí opuštěný karteziánský koncept“ (cit. podle Lycan, 2003, s. 12).

S rozvojem vědy a poznání se mění jak některé disciplíny (nejběžnější je prolínání disciplín, či naopak vydělování různých proudů z jedné disciplíny), tak jimi užívané pojmy. Můžeme definovat teorii z hlediska současných disciplín, ale není vůbec jisté (ba dokonce je to velmi nepravděpodobné), zda budou tyto disciplíny v nezměněné podobě existovat i v budoucnu. Tato předpokládaná změna v bádání s sebou nezbytně ponese i reformulaci dřívějších hypotéz tak, aby co nejlépe odpovídaly současnému stavu vědění.

Evoluci teorií mysli stručně charakterizuje Lycan (2003, s. 18) následujícím způsobem: „Z dobře známých důvodů, především kvůli interakčnímu problému, je karteziánský dualismus nepřípustný; behaviorismus je na vyšší úrovni, ale neadekvátní ve způsobech, které jsou důsledně a dramaticky překonány teorií identity; teorie identity je vynikající,

[7] Zde narážím především na odlišné pojetí „funkcionalismu“ u Neda Blocka a W. G. Lycana. Viz kap. 4.
[8] Tento fakt přímo koresponduje s Jaworskiho (2011, s. 26) popisem fyzické domény, jakožto „domény popsané a vysvětlené fyzikou“, do které spadá vše, co určí fyzika, a stejně tak má ty vlastnosti, které ji určí fyzika. Radikální změna fyziky jakožto vědního oboru by tedy znamenala neméně radikální změnu pojmu „fyzický“ a tím i, pravděpodobně stejně radikální, změnu chápání vztahu mysli a těla.

ale má jednu problémovou vadu, kterou napravuje funkcionalismus“. Lycan (2003, s. 17-18) dále cituje Chomského, dle kterého funkcionalismus pramení „ze strachu, že něco je špatně“, což reflektuje vznik funkcionalismu jakožto teorie mající za cíl nahradit behaviorismus a teorii identity, které se jevily jako „špatný“ přístup ke vztahu mysli a těla. Po funkcionalismu spatřily světlo světa další teorie, a o některých (například o emergentismu) můžeme říci, že reagují právě na jeho nedostatky.

Mezi další problémy, se kterými se zkoumání lidské mysli potýká, řadím myšlenkové experimenty. Jsem jejich zapřisáhlý odpůrce a význam pro mě mají pouze ověřitelná data. Myšlenkové experimenty považuji za samoúčelnou kritiku. Z teorií, kterými se ve své práci zabývám, žádná není postavena na myšlenkovém experimentu, avšak u všech jsou myšlenkové experimenty použity ke kritice. Ze samotné podstaty myšlenkového experimentu[9] vyplývá jejich velmi obtížná prokazatelnost i vyvratitelnost. Z tohoto důvodu nepovažuji myšlenkové experimenty za relevantní a troufám si tvrdit, že v seriózním akademickém diskurzu nemají místo. Bohužel jsou však až příliš častou pomůckou, kterou se autoři snaží podpořit své leckdy absurdní hypotézy, či naopak prostředkem, jak tyto leckdy absurdní hypotézy vyvrátit, avšak neméně absurdním způsobem.

Je nezbytné, abych se z tohoto postoje vyznal již na počátku, jelikož zkoumání mysli je s myšlenkovými experimenty nerozlučně spjato, a u funkcionalismu to neplatí o nic méně. Bude tedy nezbytné, abych uvedl a analyzoval některé důležité myšlenkové experimenty, které s funkcionalismem souvisí. Pokusím se omezit jen na obecně známé experimenty, které si vysloužily vícero dodatečných reakcí, či se považovaly za zcela, a bezpečně, překonávající nějakou teorii.

Ke všem experimentům budu přistupovat velice kriticky a pokusím se zhodnotit jejich relevantnost, přínos, a především jejich vztah k současnému stavu poznání (tedy i k případnému ověření či vyvrácení experimentu). Pokud to bude možné bez porušení

[9] „Zpravidla je v nich explicitně konstruován logický rozpor nalezený v argumentaci oponenta. Nejde však o důkaz praktické realizace obsahu myšlenkového experimentu, nýbrž pouze o stanovení logické možnosti této realizace. Prověření logické možnosti je považováno za nutnou, ale nikoli dostačující podmínku reálné existence daného experimentu. V myšlenkovém experimentu jde o to ukázat, že určitá situace, třebaže reálně zatím nenastává, je každopádně logicky možná, tj. v nějakém možném světě může taková konstrukce existovat logicky bezrozporným způsobem“ (Polák, 2013, s. 23). „Z pohledu logiky lze rozlišit myšlenkové experimenty *realizovatelné* (alespoň principiálně) a *sporné* (které vedou k paradoxům a slouží k vyvrácení nějaké hypotézy)“ (Havel, 2001, s. 34).

konzistence, budu ignorovat/redukovat na nezbytné minimum experimenty, v nichž jakýmkoliv způsobem figurují mimozemské bytosti (již teď musím čtenáře upozornit, že není možné se jim zcela vyhnout). Je pro mě velmi obtížně pochopitelné, jak často se k vyvrácení hypotéz týkajících se nás lidí používají stvoření, u nichž nemáme sebemenší jistotu, zda vůbec existují.

Na závěr mi přijde vhodné zmínit Chomského rozlišení na problémy a záhady. Problémy nám umožňují formulovat otázky tak, abychom nad nimi mohli seriózně bádat a pravděpodobně dosáhnout určitého pochopení. Oproti tomu záhady jsou pro nás neuchopitelné, pravděpodobně proto, že na ně nejsme vhodně vybaveni (dáno vrozenou strukturou naší mysli), takže pro nás mohou být neřešitelné (permanentně, nikoliv jen dočasně). Jako možná záhada se pak jeví naše svobodná vůle (Lycan, 2003, s. 22). Nechci zde pěstovat gnoseologický skepticismus, na druhou stranu je v povaze lidí přeceňovat své schopnosti, takže je možné, že se problém mysli a těla nakonec ukáže jako Chomského „záhada“ a všechny teorie cílící na jeho rozřešení jako marné. Podobně smýšlí i Ned Block (1978, s. 304), který problém mysli a těla označil za „problém, u kterého máme stěží jistotu, že má nějaké řešení“ a prohlásil, že „každé z navrhovaných řešení problému těla a mysli má závažné nedostatky, nedostatky, které považuji za fatální“. Jaworski (2011, s. 11) tento přístup označuje za „pesimismus mysli a těla“.

Celý diskurz o povaze mysli a jejího vztahu k tělu se nese v duchu descartesovského dualismu, který ovlivňuje jak naše představy, tak i metody zkoumání. Nepochybnou výhodou je srozumitelnost, intuitivně vnímáme určitý rozdíl mezi naším tělem a myslí, ovšem právě tato jednoduchost a srozumitelnost může zastřít náš úsudek a připravit nás o jiné možnosti, které si zasluhují naši pozornost. Není nikterak překvapivé, že se mnohé teorie odvrátily od dualismu a zaměřily se na materialismus (fyzikalismus), aby se tak uchýlily k objektivně měřitelným a popsatelným jevům hmotného světa, v protikladu k neznámé mysli. Ačkoliv byl Descartův substanční dualismus odmítnut, pozůstatky tohoto dělení jsou stále patrné, i monistické teorie operují s pojmy „mysl“ a „tělo“ a tuto distinkci vnímáme přirozeně (byť v odlišné podobě než Descartes a jiní myslitelé). Jaworski (2011, s. 24) upozorňuje, že „jedno z nejběžnějších nedorozumění – kterého se často dopouštějí i profesionální filosofové – je předpoklad, že mentální a fyzické jsou navzájem výlučné kategorie, že pojmenovat něco ‚mentální‘ znamená, že to není fyzické“ a naopak.

U Descarta ještě zůstaneme, jelikož ve svém díle *Rozpravy o metodě* předznamenal vztah člověka a stroje a uvádí, jak je od sebe odlišit:

> A tu jsem se zvláště zastavil u důkazu, že kdyby existovaly takové stroje, jež by měly orgány a vnější vzhled opice nebo jiného nerozumného zvířete, měli bychom důvod se domnívat, že by byly ve všem stejné povahy jako tato zvířata; kdežto kdyby existovaly stroje, podobající se našim tělům a napodobující naše úkony potud, pokud by to mravně bylo možné, měli bychom vždy dva velice vážné důvody, abychom poznali, že proto ještě nejsou skutečnými lidmi. První důvod je, že by nikdy nemohly užívat slov ani jiných znaků, skládajíce je jako činíme my, abychom své myšlenky vyložili jiným. Neboť lze dobře chápat, že stroj může býti udělán tak, aby pronášel slova, ba dokonce aby pronášel některá ve spojení s tělesnými úkony, souvisejícími s nějakými změnami jeho orgánů: jako například když se ho dotkneme na určitém místě, aby se zeptal, co mu chceme říci, když na jiném místě, aby křičel, že ho to bolí, a podobně; nemůže však být udělán tak, aby slova různě sestavoval a takto odpovídal na vše, co se řekne v jeho přítomnosti, jak to i nejtupější lidé mohou činit. A druhý důvod je, že i kdyby vykonávaly určité věci stejně dobře nebo snad i lépe než kdokoli z nás, selhaly by nevyhnutelně v jiných, při nichž by vyšlo najevo, že nejednaly s vědomím, nýbrž toliko podle sestavení svých orgánů; neboť rozum je všestranný nástroj, kterého lze užívat ve všech možných případech, kdežto tyto orgány musí mít nějaké zvláštní uzpůsobení pro každý úkon jednotlivý, a proto je morálně nemožné, aby rozmanitost těchto orgánů v jednom stroji stačila přivést jej k tomu, aby jednal za všech okolností života stejně, jako jednáme my vlivem svého rozumu (Descartes, 1992, s. 41).

2. Turingův test

V duchu Descartova tvrzení, „že [stroje] by nikdy nemohly užívat slov ani jiných znaků, skládajíce je jako činíme my, abychom své myšlenky vyložili jiným“, se k této problematice staví o přibližně 400 let později matematik a logik Alan Turing. V roce 1950 představil Turing imitační hru (později známou jako Turingův test), jejímž účelem mělo být zodpovědět otázku, zda mohou stroje myslet. Turing se vyhnul definici myšlení a problémům spojených s vágností tohoto pojmu[10] a svou „hru“ koncipoval prakticky – máme zde tazatele, který nepřímo (například skrze textové rozhraní nebo prostředníka) komunikuje s mužem a ženou v jiné místnosti a snaží se odhalit, kdo z páru je muž a kdo žena. Žena má tazateli pomáhat, muž se ho snaží zmást. Nyní máme postavu muže vyměnit za stroj. Pokud by se tazatel mýlil stejně často, jako v původní hře, pak stroj dokáže věrohodně imitovat člověka a tím prokázat svou inteligenci. Výhodou testu je jeho realizovatelnost, nevýhodou pak ryze behavioristické pojetí mysli, které nijak nereflektuje vnitřní procesy, ani fenomenální vědomí. Představa myšlení, jak ji implikuje Turingův popis imitační hry předchází explicitní formulaci funkcionalismu ve filosofii mysli, ale v podstatě je s ním totožná.

Ve svém textu *Computing Machinery and Intelligence* zmiňuje Turing i některé námitky, které však nejsou mířené pouze proti testu samotnému, ale také proti Turingově stroji a funkcionalistickému pojetí mysli[11], které test implikuje. Z tohoto důvodu považuji za vhodné je zde zmínit (pro komplexnost uvedu veškeré námitky, ty méně relevantní pouze ve stručnosti).

1) „Teologická námitka – Myšlení je funkcí nesmrtelné lidské duše. Bůh dal nesmrtelnou duši každému muži a ženě, ale žádným jiným zvířatům či strojům. Proto žádná zvířata či stroje nemohou myslet“ (Turing, 2004, s. 449). Tento argument Turing zcela odmítá, především proto, že teologické argumenty se v minulosti ukázaly jako povětšinou chybné. Dále je tu

[10] Srovnej: „This should begin with definitions of the meaning of the terms "machine" and "think." The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words "machine" and "think" are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, "Can machines think?" is to be sought in a statistical survey such as a Gallup poll. But this is absurd“ (Turing, 2004, s. 441).

[11] Námitky 1) až 5) a dále 7) lze bez větších obměn aplikovat i na funkcionalismus, jak bude uveden v kap. 3.

spekulativní možnost, že by Bůh, ve své velikosti a všemohoucnosti, mohl rozhodnout učinit stroj či zvíře stejně moudrým, jako je člověk.

2) „Námitka hlavy v písku – Následky toho, že by stroje myslely, jsou příliš děsivé. Doufejme a věřme, že toho nejsou schopni“ (Turing, 2004, s. 450). Další argument, který staví člověka do neohroženě nadřazené role[12], a dle Turinga si stejně jako předchozí si nezasluhuje výraznější pozornost.

3) Matematická námitka. Zde se dostáváme k racionální námitce týkající se možností Turingova stroje. Existují totiž matematicko-logické důkazy svědčící o omezeních diskrétních stavových strojů. Za nejznámější Turing považuje Gödelův teorém[13]. Z hlediska imitační hry je ale Gödelův teorém nepodstatný, jelikož ani u lidí se nepředpokládá bezchybná znalost všeho, tudíž tuto schopnost nemusíme vyžadovat ani od počítače. Naopak se domnívám (a postupy chatbotů[14] to dokazují), že chybnou či žádnou odpověď považují účastníci imitační hry za více „lidskou“. Dále Turing (2004, s. 451) uvádí, že „otázky, které nemohou být zodpovězeny jedním strojem, mohou být uspokojivě zodpovězeny jiným“. Tvrdý (2011, s. 31) nabízí řešení tohoto problému po vzoru Tarského rozlišení objektového a metajazyka, které je možné vložit i do počítače – vybavit ho hierarchií jazyků (kterou už do jisté míry počítač má – přinejmenším je zde rozdíl mezi programovacím jazykem a samotným programem). Co není vyřešitelné v jedné soustavě je vyřešitelné v soustavě vyšší a tak dále ad infinitum. Tvrdý (2011, s. 34) se také domnívá, že „aplikace Gödelova teorému mimo oblast axiomatických disciplín, jako třeba v teorii mysli, je možná jen metaforicky, nikoli doslovně“.

Zde si dovolím odbočku a uvedu, že na počátku 60. let 20. století byla Gödelova věta prezentována jako nástroj „k vyvrácení mechanistické teze, tj. názoru, že lidskou mysl lze simulovat strojem“ (Havel, 2001, s. 28). Gödel sám uvedl, že „[b]uď je lidská mysl schopna odpovědět na více otázek z teorie čísel, než kterýkoliv stroj, anebo existují číselně teoretické otázky, na které lidská mysl odpovědět nedovede“ (cit. podle Havel,

[12] O tuto roli však člověk může během následujících přibližně 40 přijít, jak se domnívá Kevin Warwick ve své knize *March of the machines* (z roku 1999, vyšla i v českém překladu pod názvem *Úsvit robotů, soumrak lidstva*).

[13] Gödelův teorém dokazuje, že „v každém dostatečně silném logickém systému mohou být formulována tvrzení, jež nejsou v tomto systému dokazatelná, ani vyvratitelná, pokud samotný systém není nekonzistentní“ (Turing, 2004, s. 450).

[14] Chatbot, nebo též chatterbot je počítačový program vytvořený za účelem rozhovoru (většinou povrchního) s uživatelem.

2001, p. 28). I v tomto případě ale platí Turingova námitka, že co nezmůže jeden stroj, zmůže stroj jiný. Neplatí totiž mentalistická teze, kterou Havel (2001, s. 29) formuluje následovně: „Existuje zodpověditelná otázka, na kterou žádný program nedovede odpovědět“. Analogicky s hierarchií jazyků zmíněnou v předchozím odstavci lze formulovat hierarchii úrovní, kdy se Gödelův teorém vztahuje pouze na nejnižší úroveň, zatímco na nejvyšší úrovni mohou vznikat emergentní jevy (o těchto později v podkapitole 6.4.).

4) Argument z vědomí. Turing zde odkazuje na Geoffreyho Jeffersona, který tuto myšlenku vyjádřil již v roce 1949, tedy před formulací imitační hry. „Dokud nebude stroj schopen napsat sonet nebo složit koncert na základě svých myšlenek a emocí, nikoli náhodným výběrem symbolů, nemůžeme přistoupit na tvrzení, že stroj rovná se mozek. To znamená, že stroji nestačí sonet jen napsat, ale musí i vědět, že jej napsal. Žádný mechanismus nemůže cítit (nejen uměle dávat najevo, což je snadné) uspokojení z úspěchu nebo smutek ze spálené elektronky, nemůže být potěšen lichotkami, zarmoucen svými chybami, okouzlen opačným pohlavím, rozzloben nebo deprimován, když není schopen dosáhnout toho, co chce“ (Tvrdý, 2011, s. 35). Dle Turinga se jedná o nejlogičtější pohled, avšak poměrně solipsistický. Mysl ostatních lidí můžeme zpochybňovat podobně jako mysl počítačů, v čemž ale Turing nevidí žádný přínos. Argument samotný lze rozdělit na dvě části, přičemž dílčí část, která tvrdí, že stroje nejsou schopné vytvářet umělecká díla, jednak není oprávněná, a dále, v kontextu současného „umění“, působí přinejmenším úsměvně[15]. Naproti tomu požadavek na vědomí je zcela oprávněný a pro stroje problematický. Jelikož ale nemáme žádný prostředek, jak bezpečně určit, že ostatní lidé mají vědomí, nemůžeme to s jistotou tvrdit ani o strojích. Jako nejvhodnější se tedy jeví připsat strojům vědomí, pokud se chovají podobně inteligentně jako lidé.

5) Argument rozličných nedostatků. Tento argument připouští, že by stroj mohl dělat všechno možné, ale nebyl by schopný být X, kde X je např. laskavý, nápaditý, přátelský, iniciativní, zamilovaný (Turing, 2004, s. 453). Též bychom mohli říci, že stroj nikdy nebude mít kvalitativní stavy. Na druhou

[15] Narážím především na obraz *No. 5, 1948* Jacksona Pollocka, který je v současné době třetím nejdražším obrazem (Sedghi, 2015).

stranu všechny tyto stavy jsou poznatelné pouze na základě behaviorálních projevů, většina potom na základě verbálního chování, na které je zaměřena právě imitační hra. Turing (2004, s. 455) se dále domnívá, že „mnoho z těchto omezení je spojeno s velmi malou kapacitou paměti většiny strojů“ a předpokládá, že s jejím zvětšením bude možné stroje vybavit více lidským chováním.

6) Námitka Lady Lovelace[16]. Tato námitka vychází z informace o Babbageově analytickém stroji[17], předchůdci dnešních počítačů, o kterém Lady Lovelace prohlásila, že „nemá žádné ambice cokoliv vytvořit. Nicméně dokáže udělat vše, co víme jak mu nařídit“ (Turing, 2004, s. 455). Tuto námitku Turing modifikuje do podoby, že nás stroje nedokážou překvapit, což není pravda, ačkoliv toto překvapení většinou plyne z naší neznalosti. Dále je možná existence stroje, který by byl schopný sám něco originálního vytvořit, či se učit, ale v Turingovo době touto vlastností žádný stroj nedisponoval

7) Argument spojitosti nervového systému. Tento argument stojí na faktu, že nervový systém není diskrétní stavový stroj a že změna byť jediného neuronu může mít dalekosáhlé následky na celou nervovou síť. Zaprvé, tazatel v imitační hře nemá jak poznat, jakého druhu je „nervový systém“ (úmyslně v uvozovkách, jelikož tím metaforicky označuji i fungování počítače) účastníků, a zadruhé, analogový signál jsme schopni převést na digitální, lze tedy předpokládat, že podobně by bylo možné převést i náš nervový systém, čemuž napovídá zejména binární chování neuronů. Příkladem digitálního[18] uchování informací v lidském těle je DNA a RNA, jež obsahují informace „ve čtyřkové soustavě, pomocí bází cytosinu, guaninu, adeninu a thyminu, respektive uracilu“ (Tvrdý, 2011, s. 43).

8) Argument neformálního chování. Pravidla, kterými se stroj musí řídit, jsou v mezích formální logiky, což je omezení, které pro naše chování neplatí –

[16] Tvrdý (2001, s. 40) zmiňuje tzv. Lovelaceové test, jenž stroj splní, pokud vyprodukuje výstup V, který není výsledkem chyby, ale znovu proveditelného procesu, přičemž tvůrce stroje nedokáže tento proces vysvětlit. Tento požadavek připomíná popis emergentních jevů, či genetického algoritmu, a jako takový není zcela v rozporu s funkcionalismem.

[17] Pro představu, o jak odvážný projekt se jednalo, uvádím Tvrdého (2011, s. 38) popis: „mělo se jednat o kalkulátor schopný počítat logaritmické a goniometrické funkce, k čemuž sloužila paměť o kapacitě tisíce padesátimístných čísel. Vstup byl prováděn pomocí děrných štítků … Výstup měl být zajištěn nějakým typem ukazatele nebo tiskárny. Celý stroj včetně "procesoru" měl ryze mechanický design: funkci stroje zajišťovala nesmírně složitá soustava 96 ozubených kol, 24 hřídelí a nespočtu táhel, čepů, válců. Podle střízlivých odhadů mohl tento parou poháněný kolos vážit až dvě tuny“.

[18] Srovnej: „Life is just bytes and bytes and bytes of digital information“ (Dawkins, 1995, s. 19).

pokud se na semaforu rozsvítí červené i zelené světlo naráz, člověk není vázán pravidly a může se rozhodnout. Výhoda lidí spočívá v tom, že jsme schopni improvizovat. Na druhou stranu, svobodná vůle, počítačům odpíraná, může být pouhou iluzí pramenící z nedokonalosti našeho poznání. Pro počítače se jako nejhodnější jeví využití genetického algoritmu, který by mu umožnil upravovat svůj program na základě předchozích chyb, což by, po dostatečném množství tréninků, mohlo vést k použitelné imitaci lidské intuice.

9) Argument mimosmyslového vnímání. Tento argument, podobně jako první a druhý, nelze považovat za racionální námitku. Zabývá se totiž využitím mimosmyslového vnímání (telepatií, psychokinezí), jež by tazateli umožnila odhalit lidského účastníka imitační hry. Jelikož však mimosmyslové vnímání dosud nebylo vědecky podloženo, netřeba se jím více zabývat.

Turing (2004, s. 449) prohlásil, že „za přibližně padesát let bude možné naprogramovat počítač … aby hrál imitační hru natolik dobře, že průměrný tazatel nebude mít víc jak 70% šanci ho správně rozpoznat během pětiminutového rozhovoru“. Dále předpokládá, že otázka, zda mohou stroje myslet vymizí, jelikož „na konci století se použití slov a obecný názor vzdělaných změní natolik, že budeme mluvit o myslících strojích, aniž by to někdo zpochybňoval“ (Turing, 2004, s. 449).

Od roku 1991 se na principu imitační hry odehrává soutěž Loebner Prize, ve které se různí chatboti snaží během 25 minutového rozhovoru zmást alespoň polovinu ze 4 soudců. Za posledních 5 let se pouze jednomu chatbotovi podařilo zmást alespoň 1 soudce[19]. Oproti původnímu Turingovu testu je u Leobner prize navýšena doba rozhovoru pětinásobně, díky čemuž má tazatel delší dobu a více dat k rozhodnutí, zda komunikuje se strojem či člověkem. Osobně nepovažuji pět minut za dostatečnou dobu pro zhodnocení (ne)schopnosti používat jazyk, Loeber Prize se mi v tomto jeví jako vhodnější.

Historický zvrat přišel po takřka 65 letech od publikování Turingova testu (a 60 let od Turingovy smrti), kdy byla původní verze testu během události Turing Test 2014 poprvé překonána[20]. Tedy o přibližně 15 let později, než bylo původně předpovězeno. Úspěšným

[19] Vítězných chatbotem byl Timmy a jeho tvůrcem Bruce Wilcox (Hugh Gene Loebner, 2010).
[20] Loebnerova verze testu však dosud překonána nebyla.

chatbotem se stal Eugene[21], který simuluje 13letého ukrajinského chlapce. Eugenovi se podařilo zmást 10 z 30 soudců, čímž překonal Turingem danou hranici 30%[22] úspěšnosti. Aby se Eugene co nejméně dostával do úzkých, je naprogramován „změnit předmět rozhovoru když je to možné … pokládat otázky, řídit konverzaci, občas přihodit nějaký vtip“ (Markus 2014).

Je také důležité zmínit, že téma rozhovoru nebylo nikterak omezeno, přesto tento výsledek nelze brát za prokazatelný, a to zde dvou důvodů: Eugene simuloval chlapce, který nebyl rodilým mluvčím angličtiny, tudíž lze předpokládat určitou toleranci k jeho jazykovým nedostatkům. Dále, jeho věk byl 13 let, díky čemuž „dává perfektní smysl, že neví všechno“ (University of Reading, © 2014). Odpovědím 13letého chlapce neodepřeme určitou míru inteligence, avšak bezpochyby je nebudeme považovat za reprezentativní při jejím posuzování.

V obecnějších testech podobných tomuto se podařilo chatbotům zvítězit již dříve, s mnohem přesvědčivějšími výsledky. V roce 1991 bot PC Therapist, vytvořený Josephem Weintraubem, zmátl 5 z 10 soudců a v roce 2011 Cleverbot Rollo Carpentera zmátl 17 ze 30 soudců (Biever, 2014). Není tedy nikterak překvapivá zpráva z roku 2007 varující před „zločinným“ chatbotem CyberLover, který se na chatovacích webech snaží z uživatelů vylákat osobní informace. Tento chatbot by měl být schopný navázat vztah s až 10 lidmi během 30 minut (Rosi, 2007).

Turingův test je ve skutečnosti zaměřen pouze na schopnost stroje imitovat člověka, a to jen na jednu jeho specifickou schopnost, kterou je užívání přirozeného jazyka. Přestože myšlení je bezpochyby spojeno s řečí, nelze tvrdit, že používání řeči je jeho nutnou podmínkou. Hypotetickým konstruktem však zůstává člověk, který by nevyužíval žádnou formu jazyka – jak bychom poznali, že takový člověk myslí a není pouhým reaktivním agentem? Náš jazyk nemá v přírodě obdoby a označení nástroj myšlení považuji za trefné. Lze však prohlásit, že porozumění řeči (skutečné, ne jen pouhá imitace, která je postačující pro splnění Turingova testu) je postačující podmínkou inteligence. Tvrdý (2011, s. 7) dále shrnuje Quineův lingvistický behaviorismus, dle kterého „je třeba rezignovat na snahu o přímý, nezprostředkovaný přístup k mentálním stavům jiných lidí.

[21] Pro bližší informace o Eugenovi viz (Veselov, 2010) – tento text obsahuje také stručný popis jazykového modelu a je z něj dobře patrné, jaká část je pevně naprogramovaná, jaká závislá na proměnných a jak se analyzují jednotlivé výpovědi. V případě Eugena můžeme hovořit o ryze funkcionalistickém přístupu.
[22] V roce 2012 se Eugenovi podařilo zmást 29% soudců.

Za všech okolností jsme totiž odkázáni na vnější jazykové projevy ostatních, ze kterých můžeme pouze na základě analogie usuzovat, jaké jsou jejich myšlenky a pohnutky“. Kritik může namítnout, že mentální stavy jiných lidí můžeme pozorovat i z jejich neverbálních projevů – pokud jsou ale tyto identické, k jejich rozlišení nám mohou pomoci verbální projevy. Lidský pláč má různé příčiny, které se navenek projevují všechny stejně. Pokud člověk nevykazuje vnější poranění, může se stále jednat o vnitřní poranění, ale stejně tak se může jednat o projev smutku či jiných negativních emocí. V tomto případě nám (obyčejným lidem, ne odborníkům) pomohou právě verbální projevy dané osoby.

Existují různé projekty na nahrazení Turingova testu, jedním z nich je například vizuální Turingův test. V tomto testu jsou uživateli předkládány obrázky určité situace a máme vybrat „lidskou“ odpověď, ohledně umístění určitého objektu[23]. Příkladem otázky je: „Kde je Sharon? A) Za dodávkou; B) Nalevo od kanceláře“. Předmětem testu tak není jen schopnost přirozeného vyjadřování, ale také rozpoznávání předmětů. Troufám si tvrdit, že vzhledem k rostoucí úspěšnosti počítačového rozpoznávání předmětů bude i tento test brzy pokořen. Jako nejvhodnější se jeví komplexní testování různých lidských činností. Izolovaně jsou nás počítače schopné překonat v mnoha věcech (příkladem budiž IBM Deep Blue a hra v šach, nebo IBM Watson a soutěž Jeopardy[24]), avšak vždy se jedná o přístroj stvořený pečlivě na míru pouze tomuto konkrétnímu úkolu. Navzdory veškeré kritice je třeba se na Turingův test dívat jako na první jednoduše proveditelný test, který dosud nebyl nahrazen žádnou lepší alternativou. Dokud nebudeme schopni zodpovědět otázku, co vlastně znamená myslet, je hodnocení toho, jak lidsky/rozumně se stroje chovají tím nejlepším, co máme.

V poslední části svého textu *Computing Machinery and Intelligence* se Turing zabývá učením strojů. Turing (2004, s. 459) kapacitu lidského mozku odhaduje na 10^{10} až 10^{15} binárních čísel, přičemž pro úspěch v imitační hře by mělo postačovat 10^{9}, a technologie počátku 50. let dvacátého století umožňovaly 10^{7}. Turing vhodně rozdělil tři hlavní formanty mysli – počáteční stav mysli (při narození), vzdělání a další získané zkušenosti. Namísto simulace dospělé mysli máme simulovat dětskou mysl a tu vhodným vzděláváním (metodou odměn a trestů) přetvořit na dospělou. Z hlediska současné umělé

[23] Autor této diplomové práce měl úspěšnost testu 38%.
[24] V ČR se vysílala obdobná soutěž pod názvem Riskuj.

inteligence toto řešení skýtá přinejmenším tři výhody: 1) nižší nároky na výpočetní výkon, 2) nižší nároky na naprogramování (stejně jako nižší nároky na komplexnost znalostí programovaného), 3) více „lidský“ výsledek[25].

Dle původního předpokladu by se imitační hry mohl zúčastnit libovolným způsobem vytvořený stroj, jedinou výjimkou jsou „lidé narozeni běžným způsobem“ (Turing, 2004, s. 443). Avšak vzhledem k možnosti vytvořit člověka z jeho buňky, a tím obejít přirozený způsob početí, omezuje Turing svůj test pouze na digitální počítače. Dodává však své přání „povolit možnost, že by inženýr, či tým inženýrů mohl vytvořit stroj, který funguje, ale způsob jeho provozu nejsou konstruktéři schopni uspokojivě popsat, jelikož použily metody povětšinou experimentální“ (Turing 2004, s. 443). Z pohledu současné umělé inteligence Turing připustil použití nejen tradičního, ale i emergentistického paradigma (o obou více později v podkapitole 6.4.). Test samozřejmě není zaměřen na počítače Turingovy doby (jelikož počítače jak je známe nyní spatřily světlo světa až po Turingově smrti), jeho cílem je zjistit, zda by se vůbec nějakému představitelnému počítači podařilo ve hře zvítězit.

Za důležité pro úspěch považuje Turing následující tři faktory: vhodné *naprogramování*, dostatečná *rychlost* a adekvátní *paměť*. To že je digitální počítač elektrický a nervový systém také je podle Turinga (2004, s. 446) jen povrchní analogie a „máme se zaměřit na *matematické analogie funkcí*“. Digitální počítač sestává ze tří hlavních částí, kterými jsou: *paměť* (analogie k lidské paměti), *řídící jednotka* (zpracovává jednotlivé operace), *kontrola* (zabezpečuje, že se počítač řídí *tabulkou instrukcí*[26], která mu předepisuje, jak danou operaci vykonávat). Počítač nemusí být striktně deterministický, může být vybavený generátorem náhodných čísel (z pohledu uživatele je však obtížné říci, zda počítač skutečně obsahuje náhodné prvky, nebo je to výsledek určitého procesu[27]). Stejně tak není vyloučena potenciálně neomezená paměť, ačkoliv v daný okamžik bude k dispozici pouze její omezená část.

Počítače Turing řadí mezi „diskrétní stavové stroje“, jež se pohybují z jednoho určitého stavu do jiného, a tyto stavy jsou od sebe odlišeny, aby se nedaly zaměnit. Jedná se o teoretický konstrukt, Turing (2004, s. 446) přiznává, že v přesném slova smyslu

[25] Z pohledu stoupenců psychologického behaviorismu.
[26] Vytvoření tabulky instrukcí Turing nazývá (na)programováním.
[27] Neznalému uživateli by se počítač s částečně náhodných „chováním“ mohl jevit jako mající svobodnou vůli, ale od tohoto tvrzení se Turing distancuje.

„…neexistují takové stroje. Vše se pohybuje souvisle“. Nicméně pro naše potřeby můžeme některé stroje považovat za diskrétní stavové stroje, například kukačkové hodiny. Díky stavové tabulce stroje jsme pak schopni předpovědět stav stroje o libovolný počet operací později.

Představu budoucího počítače Turing nastínil již v *On Computable Numbers with an Application to the Entscheidungsproblem* roku 1936, v reakci na tzv. Entscheidungsproblem[28]. Turingův stroj (T-stroj)[29] se skládá z potenciálně nekonečné pásky (paměťového média), která je rozdělena na sekce, do kterých se zapisují *symboly* (symbolem může být i prázdná sekce). V daný okamžik může stroj číst symbol pouze z jedné sekce a při jeho načtení mu stavová tabulka (na základě načteného symbolu) zadá instrukce (ponechání symbolu / zapsání jiného symbolu; posun vpravo či vlevo. Chování automatického stroje je zcela determinováno jeho konfigurací, tj. současným stavem a právě čteným symbolem. Počítací stroj je pak automatický stroj, který používá pouze symboly 0 a 1 (Turing, 2004, s. 59-60). Zjednodušeně můžeme říci, že stavová tabulka udává instrukce ve formě: pokud stroj ve stavu S_1 obdrží vstup I_1, vyprodukuje výstup O_1 a přesune se do stavu S_2.

V závěru bych rád uvedl Turingovu (2004, s. 459) analogii mysli:

> Když uvažuje o funkcích mysli či mozku, nalezneme určité operace, které lze vysvětlit ryze mechanickými termíny. Řekneme si, že to neodpovídá skutečné mysli: je to určitá slupka, kterou musíme sloupnout, abychom našli pravou mysl. Ale co nám pak zbyde je další slupka, kterou musíme sloupnout, a tak dále. Dostaneme se tímto postupem někdy ke „skutečné“ mysli, nebo nakonec dojdeme ke slupce, pod kterou nic není? Ve druhém případě je celá mysl mechanická.

Přínos Turingových tezí na poli filosofie mysli a umělé inteligence shrnuje Georges Rey (2012, s. 89) následovně:

[28] „Ten spočíval ve snaze dokázat tvrzení, že ke každému matematickému výroku existuje ryze formální, mechanický proces, který by mu přiřadil pravdivostní hodnotu. Pokud by byl takový postup nalezen, znamenalo
by to, že matematika je naprosto soběstačný a úplný axiomatický systém nezávislý na logice či jiné teorii“ (Tvrdý, 2011, s. 14).

[29] Nejedná se o konkrétní fyzicky realizovaný stroj, pouze o abstraktní, ideální stroj (podobně jako např. geometrické obrazce).

Turingovu analýzu komputací bychom měli i nadále brát vážně při hledání vysvětlení očividně důmyslných způsobů jakými různé části naší mysli vytváří zjevně inteligentní chování. Turingův test by měl být zapomenut jako nic víc, než ukvapený návrh přespříliš ovlivněný pomýleným behaviorismem tehdejší doby.

3. Funkcionalismus Hilaryho Putnama

Ve své pozdější knize o sobě Hilary Putnam (1991, s. xi) prohlásil: „Je možné, že jsem byl prvním filosofem, který prosazoval tezi, že počítač je vhodným modelem mysli. Této doktríně jsem dal název ‚funkcionalismus', a pod tímto názvem se stala dominantním názorem – někteří tvrdí, že ortodoxním – v současné filosofii mysli". Následující podkapitola proto bude věnována tomuto, můžeme říci zakládajícímu, textu, kterým je *Povaha mentálních stavů*, napsaná roku 1967. Putnamovy myšlenky přímo vycházejí z Turingových, a to jak z jeho pojetí mysli, tak z Turingova stroje, který Putnam použil pro popis mysli.

3.1. Povaha mentálních stavů

Typické otázky, které si kladli doboví filosofové mysli shrnuje Putnam do tří následujících: „(1) Jak víme, že ostatní lidé cítí bolest? (2) Je bolest stav mozku? (3) Jaká je analýza pojmu *bolest*?" (Putnam 1975, s. 429), přičemž ve svém článku *Povaha mentálních stavů*[30] se zabývá otázkou číslo 2. Než přistoupí k samotnému zkoumání, vymezuje Putnam dva základní pojmy: „'vlastnost' jakožto zastřešující termín pro … mít bolest, mít určitý stav mozku, mít určité dispozice k chování … [a] termín ‚pojem' pro věci, které lze ztotožnit s třídou synonymních výrazů" (Putnam 1975, s. 429-430). Mezi pojmem a vlastností[31] je třeba důsledně rozlišovat (čímž Putnam oponuje Carnapovi, který zastával jejich rovnost), jelikož např. voda a H_2O mají stejné vlastnosti, ale jejich pojmy identické nejsou. Analogicky Putnam přistupuje k bolesti jakožto stavu mozku. To že jsem si vědom bolesti a zároveň si nemusím být vědom toho, že mám určitý stav mozku, neznamená, že nemohou být identické – mohou se shodovat vlastnostmi, avšak pojmově být odlišné. Aby spolu byly dvě vlastnosti totožné, nemusí být synonymní (synonymita jakožto mylný předpoklad totožnosti je způsobena nerozlišováním mezi vlastnostmi a pojmy).

Další důležitý rozdíl je mezi ve vyjádřeními „bolest je takový a takový stav mozku" a „bolest je korelátem takových a takových stavů mozku" (Putnam 1975, s. 432), jelikož

[30] Ned Block (1978, s. 303) se na toto téma vyjádřil poměrně skepticky: „Jakou máme jistotu, že existuje odpověď na otázku ‚Co jsou mentální stavy?'".

[31] Zjednodušeně si lze „vlastnost" představit jakožto predikát (jednomístný i vícemístný) a „pojem" lze ztotožnit se třídou synonymních výrazů (ačkoliv sám o sobě touto třídou není - příkladem jest číslo 2 ztotožnitelné se všemi párovými jevy) (Zarri, 2013).

první formulace nedává prostor k otázkám po povaze bolesti (jelikož ji explicitně označí za stav mozku), zatímco druhá ano. První formulace nám tedy poskytuje jasnou odpověď (ať už je pravdivá či nikoliv) na otázku, zatímco druhá pouze reformuluje naši původní otázku. Mylná je i představa, že aby bylo možné X zredukovat na Y, musí se obě nacházet ve stejném čase a prostoru. To by činilo hypotézu, že bolest ruky je stav mozku nepravdivou, jelikož ruka i mozek jsou v odlišných oblastech. Nikdo ovšem nezpochybňuje obraz v zrcadle jako odraz světla, i když se obraz jeví být za zrcadlem. Dále Putnam uvažuje nad smysluplností formulací: „1. Je zcela smysluplné (neporušuje žádné ‚pravidlo češtiny', nezahrnuje žádné ‚rozšíření použití [pojmu]') říci, že ‚bolest je stav mozku' … 2. Není smysluplné (zahrnuje ‚změnu významu' či ‚rozšíření použití' atd.) říci, že ‚bolest je stav mozku'" (Putnam 1975, s. 432). Zde Putnam nezaujímá žádné stanovisko, vyjadřuje skepsi nad představou „změny významu" a celou tuto otázku považuje za pseudoproblém. Jak bychom ale jinak měli zjistit, zda je bolest stavem mozku? Je třeba formulovat tvrzení ve formě „'bolest je A', kde ‚bolest' a ‚A' nejsou v žádném smyslu synonymní, a zkoumat, zdali existuje takovéto tvrzení, jež by bylo přijatelné jak empiricky, tak metodologicky" (Putnam, 1975, s. 433).

Putnam tvrdí, že aby byla hypotéza o stavu mozku pravdivá, bylo by třeba nalézt fyzicko-chemický stav mozku, který je možný u všech tvorů, schopných cítit bolest (lidí, savců, plazů, ale také případných dosud neobjevených mimozemských druhů) a zároveň tento stav musí být vyloučen u všech stvoření, které bolest necítí. Nejedná se o nic nemožného, Putnam uvádí srovnání člověka a chobotnice, kteří oba mají fyzicky v podstatě identické oko, ačkoliv jejich vývoj probíhal paralelně a daný orgán se v obou případech vyvinul z jiných buněk. Není tedy vyloučeno, „že paralelní evoluce napříč celým vesmírem by mohla *vždy* vést k *jednomu a tomu samému* ‚korelátu' bolesti" (Putnam, 1975, s. 436). Nejde však jen o bolest, ale o všechny psychické stavy. Stačí tedy nalézt jediný psychický stav, který bude u dvou stvoření totožný, ale bude mít odlišný korelát. Navzdory všeobecnému odmítání behaviorismu nám stále jako hlavní projevy duševních jevů (bolesti, žízně atd.) slouží chování daného organismu. Putnam se domnívá, že na základě podobnosti v chování můžeme usuzovat spíše na podobnost ve funkční organizaci, než na podobnost ve fyzické podobnosti. Dále můžeme předpokládat, že různé jevy budou mít stejnou či obdobnou pravděpodobnost přechodu z jednoho funkčního stavu do druhého (tak, jak bylo popsáno pro pravděpodobnostní automat), a že je větší

pravděpodobnost mezidruhové platnosti psychologických, spíše než neurofyziologických, zákonů.

Svoji hypotézu, že bolest není fyzický či chemický stav mozku (případně celého nervového systému) staví Putnam na základě, dle něj, přijatelnější hypotézy, kterou je, že „bolest, či stav bolesti, je funkční stav celého organismu“ (Putnam, 1975, s. 433). Nicméně sám uznává, že ověření této „hypotézy by byl stejně utopický úkol, jako … ověření hypotézy o mozkových stavech“ (Putnam, 1975, s. 433). Funkcionalismus tedy považuje Putnam za plausibilnější než teorii identity, avšak stejně obtížně ověřitelný.

K popisu své teze využívá Putnam tzv. pravděpodobnostní automat, což je obecná verze Turingova stroje[32], u kterého jsou přechody mezi stavy dány s určitou pravděpodobností, ne ryze deterministicky. Dále Putnam zavádí *Popis* systému, který vychází z možnosti realizovat libovolný systém pomocí různých pravděpodobnostních automatů. Popis systému S je libovolné pravdivé tvrzení o vztahu dílčích stavů systému (S_1, S_2 až S_n) mezi sebou navzájem a zároveň k vstupům a výstupům, přičemž pravděpodobnost přechodu mezi těmito stavy je určena *Strojovou tabulkou*. Tato strojová tabulka se nazývá *Funkční organizace* S ve vztahu k Popisu, a dílčí stavy S (s indexem) tvoří *Celkový stav* S (v daný čas) ve vztahu k Popisu. Důležité je, že znalost Celkového stavu zahrnuje znalost chování systému při různých vstupech, avšak nezahrnuje znalost fyzické realizace tohoto stavu – ten je specifikován pouze na základě Strojové tabulky.

Na základě výše uvedeného reformuluje Putnam (1975, s. 434) hypotézu, že „bolest je funkční stav organismu“ následovně:

> 1. Všechny organismy schopné cítit bolest jsou pravděpodobnostní automaty.
> 2. Každý organismus schopný cítit bolest je vybaven alespoň jedním Popisem určitého druhu (tj. být schopný cítit bolest *je* být vybaven odpovídajícím druhem Funkční organizace.)
> 3. Žádný organismus schopný cítit bolest nelze rozložit na části, které jsou samy o sobě vybaveny Popisem uvedeným v bodě (2).

[32] Zde bych rád upozornil na odlišnost přístupu těchto dvou myslitelů – zatímco T-stroj Turingovi sloužil pro popis umělé mysli, Putnam ho využil pro popis lidské mysli, bez zájmu o tu umělou. V roce 1988 Putnam vydal článek *Much Ado About Not Very Much,* který se k umělé inteligenci staví odmítavě.

4. Pro každý popis uvedený v bodě (2) existuje podmnožina senzorických vstupů takových, že organismus s tímto Popisem cítí bolest tehdy, a pouze tehdy, pokud nějaký z jeho senzorických vstupů je v této podmnožině.

Co se Funkční organizace týče, musí být stroj „schopný se učit ze zkušenosti“ (Putnam, 1975, s. 435), což, jak si ukážeme později, je vlastnost ne vždy zohledněná při vývoji silné umělé inteligence, a zároveň také vlastnost, která je, pro vývoj umělé inteligence (libovolné podoby) naprosto nezbytná (přinejmenším dle mého názoru).

Hypotéza, že bolest je stav mozku vylučuje jakoukoliv formu dualismu, zatímco hypotéza o funkčních stavech je s dualismem kompatibilní – Putnam uvádí, že pokud by existovalo něco obdařené tělem i duší, mohlo by to zároveň být pravděpodobnostní automat. Co se dualistického pohledu týče, uvádí Putnam, že „Celkové stavy a ‚vstupy‘ … nejsou samy o sobě ani fyzické, ani mentální“ (Putnam, 1975, s. 436). Je tedy evidentní, že Celkový stav do klasického dualistického pojetí nezapadá - Putnam nezmiňuje příklon ani k fyzickému ani k mentálnímu a já si také netroufám přiklonit ani k jedná variantě. Odlišně se mi ale jeví vstupy, které by mohly patřit do obou skupin. Za primární (*fyzické*; mající svůj původ vně systému) bychom označili vstupy z okolního světa (předpokládá se, že nezpochybňujeme jeho reálnou existenci), jako například vnímání okolní teploty, zatímco sekundární (*mentální*; mající svůj původ uvnitř systému) by pak byly vstupy na základě instrukcí (strojové tabulky). Nejedná se o ideální rozdělení, ale pokud bychom do jedné zařadili i Celkový stav – řekněme, že bychom ho označili jako *mentální*, pak bychom se zbavili oné třetí, obtížně zařaditelné skupiny, které celou hypotézu pouze zbytečně komplikuje. Nabízí se i reduktivní přístup, kdy bychom všechny kroky (od vstupů po výstupy) které pravděpodobnostní automat provádí, označili za ryze fyzické. Tím bychom se sice nebezpečně přiblížili hypotéze o stavech mozku, nicméně, i při označení funkcionálního stavu za fyzický netvrdíme, že je třeba tuto funkci realizovat právě na lidském mozku, a ne třeba na silikonovém čipu. Ačkoliv Putnam formuloval svou hypotézu v protikladu k hypotéze bolesti jako stavu mozku, není vyloučená jejich kombinace.

Na první pohled se na otázku po povaze mentálních stavů může zdát nejlepší behavioristická odpověď, tedy že se jedná o dispozici k určitému chování (a nikoliv o stav mozku či funkční stav). Tento pohled se zakládá na tom, jak poznáváme, že organismus je v určitém stavu. Na druhou stranu nám však nic neříká o vlastnostech

daného stavu. Můžeme si uvést příklad s hořením propanu – vidíme plamen, usuzujeme tedy, že daná látka hoří. To že probíhá chemická reakce $C_3H_8 + 5O_2 \rightarrow 3CO_2 + 4H_2O$ nám nemusí být známo. Projevy chování se nevylučují s hypotézou o stavech mozku, ani s funkčními stavy, jedná se pouze o viditelný projev určitého procesu, který však o povaze procesu nic nevypovídá. Chování a funkční stav mohou, ale nemusejí být identické, je možné jak stejné chování při odlišné funkci (není vyloučeno, aby odlišné funkce měly stejný výstup), tak odlišné chování při stejné funkci (dáno například stavbou organismu). Putnam (1975, s. 439) uvádí příklad s dvěma osobami, z nichž jedna by měla přetnutá vlákna přenášející informace o bolesti a druhá nikoliv, avšak úspěšně by potlačila veškeré projevy bolesti – první osoba by bolest necítila, zatímco druhá ano, jejich chování by ale bylo stejné.

Co se metodologické stránky týče, je dle Putnama stejná jako u ostatních reduktivních hypotéz. Zákony psychologie budou v tomto případě odvoditelné z Popisu pravděpodobnostního automatu. Funkční stav není pouhým korelátem určitého psychického jevu, nýbrž daný jev vysvětluje a ztotožněním funkčního stavu s daným jevem se zbavíme „neplodných a empiricky nesmyslných otázek“ (Putnam, 1975, s. 439-440).

Později shrnul Putnam (1991, s. 73) funkcionalismus následovně: „psychologické stavy (‚věřím, že p‘, ‚toužím po p‘, uvažuji, zda p‘, atd.) jsou jednoduše ‚komputační stavy‘ mozku. Správný způsob je uvažovat o mozku jako o digitálním počítači. Naše psychologie má být popsána jako software k tomuto počítači – jako jeho ‚funkční organizace‘“. Lycan (1987, s. 37) k tomu dodává: „Důležitá je funkce, nikoliv funkcionář; program, nikoliv prostředek realizace; software, ne hardware; role, ne její vykonavatel“.

3.2. Putnamova kritika

Hilary Putnam se proslavil nejen svými filosofickými názory (které kromě filosofie mysli zasahují i do filosofie jazyka, matematiky či vědy), ale i jejich poměrně častým střídáním[33]. Nejinak je tomu i v případě funkcionalismu, kdy se tento první proponent později stal odpůrcem. Jak sám vyjádřil v knize *Reprezentace a skutečnost*, „počítačová analogie, říkejme jí ‚komputační pohled na mysl‘ či ‚funkcionalismus‘, nebo jak chcete,

[33] Putnam přešel od positivismu k realismu, později k „internímu realismu“ a nakonec k „realismu s malým r“ (Ritchie, 2002).

nakonec nezodpověděla otázku, které jsme my filosofové (spolu s mnoha kognitivními vědci) chtěli zodpovědět, otázku ‚Jaká je povaha mentálních stavů'" (Putnam, 1991, s. xi).

Funkcionalismus reprezentoval přesvědčení, že „mentální stavy prostě nemohou být fyzicko-chemické stavy, ačkoliv emergují[34] z a supervenují na fyzicko-chemických stavech" (Putnam, 1991, s. xiii). Později však Putnam dospěl k názoru, že mentální stavy „nemohou být komputačními stavy, ani komputačními a fyzickými stavy (stavy, definovanými pomocí smíšeného slovníku vztahujícího se jak k fyzickým, tak komputačním parametrům), ačkoliv emergují z a mohou supervenovat nad komputačními stavy" (Putnam, 1991, s. xiii). Mentální stavy totiž mohou být součástí nejen fyzicky odlišných systémů, ale i komputačně odlišných systémů. „Mentální stavy nemohou být doslovně ‚programy', jelikož možné fyzické systémy mohou mít stejné mentální stavy, ale odlišné ‚programy'" (Putnam, 1991, s. xiv).

Jak již bylo zmíněno, Putnamův odklon od funkcionalismus vychází z jeho zjištění, že „význam není jen v hlavě" – odlišení pojmů a přesvědčení není možné bez reference k prostředí, ve kterém se subjekt nalézá. Tato externí povaha významu má za následek to, že „propoziční postoje … nejsou ‚stavy' lidského mozku a nervového systému uvažovaných v izolaci od sociálního a ne-lidského prostředí. A *fortiori,* nejsou ‚funkčními stavy'" (Putnam, 1991, s. 73). S touto kritikou nemohu souhlasit, a to ve více bodech, jak se pokusím uvést.

V první řadě, představa, že „význam není jen v hlavě" mi přijde nanejvýš absurdní, podobně jako substanční dualismus. Pokud není význam v hlavě, kde tedy je? Pluje si vzduchem kolem nás, nebo je snad nedílnou součástí každého předmětu? Troufám si tvrdit, že význam musí být v hlavě, nejen kvůli absurdnosti jiného (mimolidského) umístění, ale především proto, že význam je úzce spjat s jazykem a jazyk je výlučně lidský. Všichni si dokážeme představit Zemi bez lidí. V Putnamově pojetí by i tato Země bez lidí byla obdařena významem[35]. Stejně tak pokud bychom měli dostatečně vyspělé technologie, které by nám umožnily mezihvězdné cestování, a astronauti by se dostali na zcela neznámou planetu, byli by schopni ji pojmenovat a hovořit o ní, i kdyby byla

[34] Zde bych rád zdůraznil, že Putnam na mentální stavy nahlížel jako na emergentní jev ještě dříve, než se emergentismus stal přijímaným diskurzem v oblasti kognitivních věd (především pak umělé inteligence).
[35] Tento můj závěr Putnam explicitně ani implicitně nevyjadřuje, ale jak jsem již zmiňoval, pokud by význam nebyl jen v naší hlavě, musí být, alespoň jeho část, i někde v našem okolí.

diametrálně odlišná od Země. A tato schopnost hovořit o nepoznaném není způsobena „objevením“ významu, který by se v tomto nepoznaném skrýval, je to díky významům, které máme v sobě, které jsme schopni abstrahovat, modifikovat, zkrátka upravovat si je dle potřeby.

Nesdílím ani Putnamovo (1991, s. 76) přesvědčení, že by propoziční postoje nebyly redukovatelné na fyzikalistické či komputacionalistické predikáty. Naopak se domnívám, že tato redukce je nanejvýš jednoduchá – propoziční postoj „Jan si myslí, že tohle je nesmysl“ totiž není nic jiného, než že subjekt (Jan) má v paměti uloženo „‘tohle‘=‘nesmysl‘“. Jiné typy propozičních postojů pak lze definovat na základě dalších logických a matematických spojek, přičemž redukovaná forma je následující: „subjekt S má v paměti uloženo X ‚logická či matematická spojka‘ Y“ (případně negaci tohoto výroku). Jednoduchost tohoto popisu má výhodu z druhové nezávislosti, tímto způsobem lze popsat propoziční postoje pro všechny stvoření, známá i neznámá, která využívají libovolnou formu jazyka (jediným předpokladem je, aby měly paměť (což považuji za nutnost pro používání jazyka) a vztahovala se na ně pravidla logiky). Navzdory veškeré kritice, které funkcionalismus podrobil, uvažuje podobně i Ned Block – je možné ztotožnit „pamatuji si, že P, s ‚uložením‘ větné formy, která vyjadřuje propozici, že P“ (Block 1978, s. 306). Celou kategorii propozičních postojů považuji za redundantně postulovanou, mající nulový význam pro zkoumání lidské mysli (ba dokonce i při zkoumání jazyka). Důvody jejich odmítnutí jsou pak blízké těm, pro které je zavrhován Turingův test.

Dalším protifunkcionalistickým argumentem je dle Putnama odlišná funkční organizace, kterou lze najít i mezi členy téhož druhu. „Počet neuronů ve vašem mozku není zcela stejný, jako počet neuronů v mozku někoho jiného, a neurologové nám tvrdí, že žádné dva mozky nejsou ‚zapojeny‘ stejným způsobem. ‚Zapojení‘ jednotlivých mozků závisí na procesu dospívání a stimulacích z okolí“ (Putnam 1991, s. 82). Tento argument považuji za platný vůči typové teorii identity, nikoliv však vůči funkcionalismu (a ani vůči partikulární teorii identity). Svůj postoj nejlépe vyjádřím pomocí počítačové analogie – i na odlišném hardwaru lze spustit stejný software, se stejnými výsledky. Samozřejmě, každý software má určité minimální nároky a domnívám se, že totéž platí i pro mentální stavy. Lidoop nemá dostatečně rozvinutý mozek, aby mohl produkovat propoziční postoje. Všichni zdraví lidé ale mají, pro potřebu realizace mysli, stejně

vhodný mozek. Je tedy jedno, jestli máte víc neuronů než já, budeme schopni realizovat stejné funkce.

Obecně považuji námitky proti teoriím mysli založené na analýze přirozené jazyka jako zcela nevhodné. Ano, dříve jsem uvedl, že jazyk je výlučně lidský a považuji jej za nástroj myšlení, jeho užívání je však pouze postačující, nikoliv nutnou podmínkou myšlení. Je třeba si uvědomit, že jazyk je arbitrární konstrukt, který v průběhu své evoluce nabývá na komplexnosti, složitosti a diversitě. Utopická, ale v principu reálná[36], je představa „dokonalého“[37] jazyka, který by byl logicky bezesporný, prostý víceznačných výrazů. Méně utopická je pak představa jednoho jazyka společného všem lidem.

[36] V odlišném kontextu uvedl Putnam (1991, s. 76): „Ale je to *v zásadě* proveditelné, a to jediné je pro filosofii podstatné“. Existujícím příkladem jazyků blížících se oné „dokonalosti“ je jazyk matematiky, logiky, programovací jazyky, atd.

[37] Dokonalost je samozřejmě subjektivní, narážím zde na staletou tradici evropských filosofů marně se snažících dosáhnout „dokonalého jazyka“.

4. Funkcionalismus Neda Blocka

Ned Block v době psaní svého článku *Potíže s funkcionalismem* přisuzuje funkcionalismu rostoucí oblibu a i dominantní pozici mezi teoriemi mysli. Jedním z možných důvodů je široké spektrum funkcionalistických přístupů, od snahy reformulovat logický behaviorismus, přes teorie o analogii mysli a stroje, pokusu aplikovat empirickou psychologii na filosofii mysli až po argumenty pro či proti teorii identity. Navzdory určitým společným rysům nesdílely tyto přístupy žádnou společnou doktrínu o povaze mysli.

4.1. Funkcionalistické přístupy

Vlastnost společná většině funkcionalistických přístupů je ta, že „každý typ mentálního stavu je stav sestávající z dispozice chovat se určitým způsobem *a mít určité mentální stavy*, za určitých senzorických vstupů a určitých mentálních stavů“ (Block, 1978, s. 262). Zde vidíme zřejmou podobu s behaviorismem, oproti kterému funkcionalismus přidává podmínku existence aktuálních mentálních stavů[38]. Díky této podmínce se přisuzování mentálních stavů u behaviorismu a funkcionalismu odlišuje - mohou existovat organismy, kterým behaviorismus přisoudí mentální stavy, zatímco funkcionalismus nikoliv a naopak. Společným rysem může být názor, že pojmy označující mentální stavy mohou být nahrazeny nementálními termíny[39], například za pomoci Turingova stroje. Nejjednodušší verze strojového funkcionalismus (jak jej představil Putnam) identifikuje mentální stavy se „souhrnem stavů Turingova stroje, jež jsou *implicitně* definovány pomocí strojové tabulky, která explicitně zmiňuje vstupy a výstupy, popsané nementalisticky“ (Block, 1978, s. 263). Block upozorňuje, že díky tomuto přístupu je funkcionalismus, stejně jako behaviorismus, vinen z liberalismu.

Různorodost funkcionalistických přístupů je patrná i v přístupu k teorii identity. Určitá verze funkcionalismu stojí v opozici k teorii identity, přesněji řečeno k její typové verzi[40]. To je nejpatrnější u strojového funkcionalismu, jak byl představen v přechozí kapitole,

[38] Nosek (1997, s. 123) tento rozdíl formuluje více ve funkcionalistickém duchu následovně: „Na rozdíl od behaviorismu, který přiřazuje pouze vnější výstupy mysli k vnějším vstupům mysli, vymezuje funkcionalismus mysl třemi funkcemi: vnější funkcí přiřazující vnitřní výstupy mysli k vnějším vstupům mysli, vnější funkcí přiřazující vnější výstupy mysli k vnitřním vstupům mysli a vnitřní funkcí přiřazující vnitřní vstupy a vnitřní výstupy mysli“.
[39] Avšak na rozdíl od behaviorismu, funkcionalismus obvykle přepokládá existenci mentálních stavů.
[40] Funkcionalismus nevyvrací partikulární teorii identity.

jelikož „pokud je bolest funkční stav, nemůže být, například, stav mozku, protože stvoření bez mozku mohou realizovat stejný Turingův stroj, jako stvoření s mozkem“ (Block, 1978, s. 264), resp. „je obtížné představit si, že by zde *mohla být* netriviální fyzikální vlastnost … společná pouze všem možným fyzickým realizacím daného stavu Turingova stroje“ (Block, 1978, s. 265). Tento názor zastával Hilary Putnam, když tvrdil, že „psychologické vlastnosti jsou funkčními charakteristikami a tedy, na základě logiky, nemohou být nikdy ztotožněny s libovolnou fyziologickou strukturou, u které by věda mohla prokázat, že s nimi koresponduje“ (Kalke, 1969, s. 83).

Block (1978, s. 265) ale odkazuje i na autory (konkrétně Lewis, Smart, Armstrong), kteří tvrdí opak, tedy pokud je funkcionalismus pravdivý, pak je pravdivá i teorie identity. Přesněji řečeno, tento přístup kombinující funkcionalismus s partikulární teorií stále stojí v opozici k typové verzi teorie identity. Jejich argument spočívá v popisu bolesti jakožto určité kauzální role a mozku jakožto majícímu tuto roli[41]. Abychom se vyhnuli problému s přisouzením bolesti stvořením bez mozku, je třeba zúžit danou vlastnost např. na vztahující se pouze ke konkrétnímu druhu stvoření, tj. lidská bolest je stav mozku a Marťanská bolest je stav něčeho jiného – čímž se ale tento přístup proviňuje šovinismem.

Block dále nabízí dělení na „Funkcionalismus“ a „Psychofunkcionalismus“[42]. Psychofunkcionalismus využívá hledisko empirické psychologie a mezi jeho zástupce patří Fodor, Putnam a Harman. Psychofunkcionalismus považuje funkcionalistickou analýzu za důležitou vědeckou hypotézu. Oproti tomu Funkcionalismus zastává hledisko apriorní psychologie, vychází z behaviorismu, zástupci jsou Smart, Armstrong, Lewis, Shoemaker, a funkční analýzu považuje za analýzu významů mentálních termínů. „[R]ozdíl mezi Funkcionalismem a Psychofunkcionalismem může být definován následovně: „Funkcionalismus ztotožňuje mentální stavy S s Ramseyho funkčním korelátem S s ohledem na teorie *lidové psychologie;* Psychofunkcionalismus ztotožňuje S s Ramseyho funkčním korelátem S s ohledem na *vědeckou* psychologickou teorii“ (Block, 1978, s. 265). Na otázku, zda Psychofunkcionalismus popisuje mentální stavy částečně také ve vztahu k *neurologickým* entitám, odpovídá Block (1978, s. 276), že „v

[41] Aby mohl platit funkcionalismus a zároveň partikulární teorie identity, je třeba zakázat možnost realizace abstraktního funkčního popisu nefyzickým předmětem (např. duší), což je jinak dle funkcionalismu logicky možné. Putnam (1975, s. 412) uvedl, že „z hlediska logiky je možné, aby lidská duše byla Turingovým strojem, či spíše konečným automatem“.

[42] Block tyto přístupy úmyslně pojmenovává s velkým počátečním písmenem, proto se budu tohoto stylu, v rámci této kapitoly, držet i já. Pokud se v této kapitole vyskytuje funkcionalismus s malým počátečním f, jedná se o obecný zastřešující termín, pod který spadá jak Funkcionalismus, tak Psychofunkcionalismus.

současnosti to tak nevypadá“, jelikož psychologie a neurofyziologie jsou od sebe ještě příliš vzdálené. Dále konstatuje, že „Psychofunkcionalismus je pouze výsledek aplikování přijatelné vědecké koncepce na mentální; Psychofunkcionalismus je pouhou doktrínou, že mentální stavy jsou ‚psychologické stavy' a je úkolem psychologie je charakterizovat“ (Block, 1978, s. 302).

V protikladu k Blockovo dělení na Funkcionalismus a Psychofunkcionalismus stojí dělení které představil W. G. Lycan. Ten za „Funkcionalismus“ označuje „*a posteriori* vědeckou spekulaci, že mentální stavy a události jsou funkční stavy a události v buď komputačním či teleologickém systémově-teoretickém smyslu slova ‚funkční'“ (Lycan, 2003, s. 24), což je podle něj původní význam slova a zároveň odpovídá „Psychofunkcionalismu“, jak jej představil Ned Block. Tento přístup stojí v protikladu k „apriorní Kauzální Teorii zdravého rozumu pocházející od Sellarse a rozvinutá Davidem Armstrongem a Davidem Lewisem, kterou Block zlomyslně neologizoval jako ‚Funkcionalismus'“ (Lycan, 2003, s. 24) a později se dočkala o něco trefnějšího názvu „Analytický funkcionalismus“. Vzhledem k pojmovým zmatkům považuji za vhodné uvést, že Funkcionalismus (Block) ≠ Psychofunkcionalismus (Block) a zároveň Funkcionalismus (Lycan) = Psychofunkcionalismus (Block).

4.2. Blockova kritika

Nyní přejdeme k Blockově kritice funkcionalismu. Tu rozděluje do několika částí, přičemž nejznámější je tzv. *čínský národ*, případně jeho modifikovaná verze *homunkulární mozek*. Obě verze této námitky jsou postavené na neschopnosti funkcionalismu zachytit a popsat kvalitativní mentální stavy. Další významným problémem funkcionalismu je jeho neschopnost vyhnout se zároveň liberalismu (ze kterého vinil behaviorismus) i šovinismu (ze kterého vinil teorii identity). Block se ale vyjadřuje i k jiným problémům funkcionalismu.

První problém, se kterým se strojový funkcionalismus potýká, je simultánnost lidských mentálních procesů. Náš mozek se může nacházet ve vícero stavech (např. věřím, že A; chci B), zatímco Turingův stroj pouze v jednom stavu. Tato námitka je podobná námitce *7) Argument spojitosti nervového systému* uvedenou u Turingova testu. Osobně tento nedostatek neshledávám přesvědčivým a přikláním se k Lycanovu řešení, které navrhuje realizaci pomocí vícero Turingových strojů. Jeden konkrétní Turingův stroj může být

realizován mnoha abstraktními stroji a naopak. Existovala by jedna hlavní funkce (hlavní Turingův stroj), jejíž vstupy a výstupy by měly formu množin, jež by byly dále zpracovávány podřadnými funkcemi, které by mohly být zpracovávány dalšími funkcemi atd., dle požadované specifičnosti. Block uvádí jako možnost ztotožnění každého stavu strojové tabulky nikoliv s jedním mentálním stavem, ale s jejich souhrnem, což je ale neuspokojivé, jelikož není umožněn vhodný popis vztahů mezi mentálními stavy. Není nikterak specifikováno, jak (podle jakých pravidel, jakým způsobem) by se měly tyto souhrny mentálních stavů přiřazovat, neuspokojivost tohoto popisu tedy není nikterak překvapivá. Nic to však nevypovídá o obecné nevhodnosti či neproveditelnosti takového přístupu. Samozřejmě by to znamenalo zeslabení původní teze, mysl by však stále zůstala popsatelná v rámci Turingova stroje.

4.2.1. Homunkulární mozek a čínský národ

Další problém funkcionalismu se Block snaží dokázat úvahou o robotovi s hlavou plnou homunkulů[43]. Tento experiment se konkrétně vztahuje na strojový funkcionalismu, který tvrdí, že „každý systém mající mentální stavy je popsán alespoň jednou tabulkou Turingova stroje určitého druhu, a každý mentální stav systému je totožný s jedním stavem strojové tabulky určeným strojovou tabulkou“ (Block, 1978, s. 278), a za vstupy slouží data ze smyslových orgánů a za výstupy pohyb končetin.

Představme si tělo, z vnějšku podobné našemu, v jehož hlavě ale není mozek, nýbrž skupinka malinkatých lidí (homunkulů), z nichž každý představuje jedno pole vhodné strojové tabulky, která by nás popisovala. Vstupy ze senzorických orgánů jsou napojeny na světla v lebce a pohyb končetin řídí páčkami. Každý človíček má na starost jednu konkrétní operaci, takže když například tělo zahlédne ženu, v hlavě se rozsvítí světlo „vstup-18“. Jelikož je na tabuli „stav-S“ človíček reprezentující pole S tedy zatáhne za páčku „výstup-66“ a tabuli přepíše na „stav-E“. Je zřejmé, že k provedení takto jednoduchého úkolu nemusí mít človíček vysokou úroveň inteligence, stejně jako nemusí mít ponětí o činnostech, které provádějí ostatní človíčkové. Každý je však skvěle vytrénovaný, aby bezchybně zvládal svůj úkol, takže tělo se chová jako bychom se chovali my. „Turingův stroj může být znázorněn jako konečná množina čtveřic – aktuální stav, aktuální vstup, následující stav, následující výstup. Každý človíček má úkol

[43] Block (1978, s. 322, pozn. 19) uvádí, že homunkuly je možné nahradit modelem umělého neuronu McCullougha a Pittse.

odpovídající jedné čtveřici“ (Block, 1978, s. 278), a díky jejich činnosti se systém chová jako my, je nám Funkčně ekvivalentní. Zde je třeba zdůraznit, že se jedná o Funkční, nikoliv Psychofunkční ekvivalenci, tudíž se na homunkulární systém nemusí vztahovat žádné psychologické, či neurologické teorie, jež se vztahují na nás. Jelikož existuje „velice přijatelný předpoklad, že mentální stavy jsou v doméně psychologie a/nebo neurofyziologie, či přinejmenším, že mentalita je zásadně závislá na psychologických a/nebo neurofyziologických procesech a strukturách“ (Block, 1978, s. 295), můžeme pochybovat o mentálních stavech entit Funkčně, ale ne již Psychofunkčně, ekvivalentních lidem.

Block si uvědomuje určitou absurditu této úvahy, proto navrhuje další, která je nomologicky přijatelnější. Jelikož je v mozku přibližně miliarda neuronů, poslouží nám v této úvaze Čínský národ, jehož představitelé konvertovali k funkcionalismu a rozhodli se světu dokázat, že lidskou mysl je možné realizovat i mimo mozek. Každého obyvatele vybaví vysílačkou, která bude spojená se všemi ostatními a také s tělem (jako v přechozím případě). Zjednodušeně můžeme říci, že skupinku homunkulů nahradíme obyvateli Číny, kteří ale budou provádět v podstatě stejnou činnost. I tato realizace je nám funkčně ekvivalentní, byť pouze po určitý čas. To ale není problém, jelikož zde máme určitý počáteční stav, a tabulku určující průběh funkce – Čínský národ by tak mohl být naší funkční ekvivalencí po řekněme čtyři hodiny, pak si dojít na oběd a znovu pokračovat.

Takovýto systém by samozřejmě byl mnohem náchylnější na nepříznivé vnější vlivy (povodeň, výbuch, ztráta spojení) než homunkulární. Musíme si ale uvědomit, že ani náš biologický mozek není nezranitelný, a poškození našeho mozku může také vést k chybným/chybějícím vstupům, které následně způsobí špatné fungování celku. Systém, u kterého bychom započítávali i vstupy pocházející z jeho chybně fungujících součástí by sice stále byl Turingův stroj, ale již by s námi nebyl funkčně ekvivalentní (pokud bychom sami netrpěli nějakou „poruchou“). Je třeba jasně stanovit, na jaké úrovni se má jednat o ekvivalenci. Zde Block (1978, s. 280) upozorňuje, že „funkcionalismus netvrdí, že každý mentální systém má strojovou tabulku, která opravňuje připsání mentality s ohledem na *všechny* specifikace vstupů a výstupů, ale spíše s ohledem na *některé* specifikace“.

Jiná námitka tvrdí, že by čínský systém pracoval příliš pomalu. Systém by tak nebyl schopný zachytit určité jevy či na ně reagovat. Dle Blocka nehraje čas roli, můžeme si

představit své mentální operace extrémně zpomalené, a stejně bychom si je neodepřeli. Jiným příkladem mohou být na první pohled nehybná stvoření, u kterých bychom viděli nějaké reakce pouze při použití časosběrných fotografií či velmi zrychleného videozáznamu. Časové hledisko je dle Blocka (1978, s. 281) „sprostě behavioristické“, a odvolává se k alespoň metafyzické možnosti realizace experimentu s homunkuly. Jelikož je však čínský národ konekcionistickým systémem, lze u něj rozlišit dva systémy o různé dynamice – rychlý aktivní a pomalý adaptační. „[T]en či onen systém lze snadno ztratit z očí pouhou změnou měřítka času“ (Havel, 2001, s. 40) – pokud bychom tedy přistoupili na Blockovo „bezčasové“ hledisko, mohla by naší pozornosti uniknout buď akce, nebo adaptace daného konekcionistického systému.

4.2.2. Argument chybějících kválií

Jak u experimentu s homunkulární hlavou, tak s čínským národem je zde pochybnost, zda mají vůbec nějaké mentální stavy, především pak kvália. S odkazem na Nagelův text *What Is it Like to Be a Bat?* chybí „jaké je to být tělo s hlavou plnou homunkulů“. Dle strojového funkcionalismu by měl být kvalitativní stav K totožný se strojovým stavem S_K, pokud ale chybí jakýkoliv kvalitativní stav, je zde pochybnost o K a tedy i o $K=S_K$, z čehož vyplývá pochybnost i o pravdivosti funkcionalismu[44]. Tato pochybnost, postavenou na Kripkeho teorii rigidních designátorů[45], se nazývá „arugment chybějících kválií“ (Block, 1978, s. 281). Block svůj argument zaměřuje proti typovému funkcionalismu, avšak nevylučuje tokenový funkcionalismus[46]. Je tedy možné, aby platil argument chybějících kválií a zároveň „*všechny* tokeny kvalitativních stavů byly tokeny funkčních stavů“ (Block, 1978, s. 288).

Ačkoliv je argument chybějících kválií přímo zaměřen na strojový funkcionalismu, dotýká se i jeho ostatních verzí. V případě ne-strojového funkcionalismus by se popis akcí, které by homunkulové prováděli, stal o něco složitějším, ale jádro experimentu by zůstalo stejné. Vzhledem k této zvýšené složitosti bychom museli homunkulům přiznat o

[44] Shoemaker se zpochybňuje logickou možnost chybějících kválií, jelikož je „logicky nemožné, aby dva systémy byly ve stejném funkčním stavu a jeden systém vlastnil kvalitativní stavy, zatímco druhý je postrádal“ (Block, 1978, s. 321, pozn. 15).
[45] Z té vyplývá, že „pokud F=G, neexistuje možný svět, a tím pádem nomologicky možný svět, ve kterém by F≠G; a proto neexistuje nomologicky možný svět, ve kterém by něco bylo v (nebo mělo) F a zároveň nebylo v (nebo nemělo) G“ (Block, 1978, s. 285).
[46] Čímž se odlišuje od Kripkeho, který se svou teorií možných světů snažil vyvrátit jak typovou tak partikulární teorii identity (Block, 1978, s. 288).

něco vyšší inteligenci, než v původní variantě. Block (1978, s. 285) se domnívá, že „čím více inteligence homunkulové využívají při naší simulaci, tím méně tíhneme k tomu, abychom simulaci připsali mentální vlastnosti, kterou simulují“.

Argument chybějících kválií však spočívá na intuici, že homunkulární hlava nemá kvália. Pokud bychom ale měli dát na intuici, proč bychom měli kvália přisoudit jiné hmotě, či mozku? Důvod je ten, že my jsme systém s mozkem, a že o sobě víme, že máme kvália. „Takže i když nemáme žádnou teorii kválií, která by vysvětlovala, jak je to *možné*, máme převažující důvod odmítnout libovolnou prima facie pochybnost o kváliích mozkového systému“ (Block, 1978, s. 293). Dalším rozdílem je naprogramovanost homunkulárního systému tak, aby nás napodoboval, zatímco my nic nenapodobujeme. Pokud křičíme či pláčeme, je to dáno našimi kvalitativními stavy, ale u homunkulů je to dáno jejich naprogramováním.

Jednou z možností, jak se může funkcionalismus vyvarovat Blockovy námitky s homunkulárním mozkem je ad hoc odmítnutí této možnosti, jako to udělal Putnam – systém nesmí být rozložitelný na části, které mají stejnou funkční charakteristiku jako celkový systém. Tuto podmínku Block (1978, s. 291) modifikuje tak, aby systém „neměl žádné části, které (1) samy vlastní daný druh funkční organizace a také (2) mají stěžejní roli v utváření funkční organizace celého systému“. Druhý bod kritéria má eliminovat problémy s např. těhotnou ženou či inteligentním parazitem, kdy sice část systému má stejnou funkční organizaci, avšak nikterak se nepodílí na konstituování celkové systémové funkční organizace. Osobně bych podmínku ještě reformuloval následovně: Systém nesmí mít žádné části, které (1) samy mají stejný druh funkční organizace jako systém, jehož jsou součástí, (2) mají aktivní roli v utváření funkční organizace celého systému, (3) mohou fungovat nezávisle na systému. Tím by se vyřešil i problém s homunkuly (nebo přinejmenším by dostal zcela nový rozměr) – pokud by homunkulové nemohli existovat mimo naši hlavu, podmínka je použitelná. Pokud by mohli existovat i mimo naši hlavu, podmínka je chybná, avšak tato schopnost homunkulů existovat i mimo naší lebku by znamenala mnohonásobnou realizovatelnost lidské mysli (homunkulové by se stejně tak mohli usadit v hlavě jiného člověka, či v hlavě vhodně postaveného robota), tedy jednu ze stěžejních myšlenek funkcionalismu.

4.2.3. Problematika vstupů

Posledním problémem, který Block uvádí, je neschopnost funkcionalismu vyhnout se jak liberalismu (v případě Funkcionalismu), tak šovinismu (v případě Psychofunkcionalismu). Tvrdí totiž, že „[k]aždý návrh na popis vstupů a výstupů … je vinen buď z liberalismus, nebo šovinismu“ (Block, 1978, s. 315). Jediným, hypotetickým, východiskem je utopická představa všesvětového Psychofunkcionalismu, který by se řídil psychologií platnou pro všechny stvoření ve vesmíru.

Příkladem liberalismus je dříve zmíněný Čínský národ či homunkulární mozek. Jedním z příkladů šovinismu je předpoklad, že vstupy a výstupy mohou být zadány popisem nervových impulsů, jelikož funkční popis nepřipouští u stvořeních bez neuronů, tedy ani u strojů. Jako jednu z možností, jak tomu předejít uvádí Block (1978, s. 314) „charakterizovat vstupy a výstupy *pouze* jako vstupy a výstupy“ – záleželo by pouze na zachování kauzálních vztahů, nikoliv na samotné povaze vstupů. Dle Blocka by však tento přístup vedl k přílišné liberálnosti[47]. Tento názor nesdílím, jelikož nijak nereflektuje potřebu mentálních stavů u obou srovnávaných entit (a tím se blíží spíše behavioristickému pojetí). Behavioristický přístup ke specifikaci vstupů a výstupů je taktéž šovinistický, jelikož odmítá mentální stavy např. ochrnutým lidem či mozkům v kádi. Vzhledem k tomu, že je druhově (lidsky) specifický, odmítá mentální stavy i stvořením s odlišnou tělesnou strukturou (která jednoduše neumožňuje lidský popis jejich chování). Je možné formulovat funkční popis pro každý druh zvlášť, čímž bychom se ale připravili o obecnou definici mentálních stavů platných pro všechny a funkcionalismus by se dostal na podobnou úroveň jako fyzikalismus.

Svůj text uzavírá Block (1978, s. 318) následujícím tvrzením: „Je velmi obtížné představit si jediný popis vstupů a výstupů, který se vztahuje na všechny mentální systémy, ale pouze na ně“. Funkcionalismus musí opustit původní cíl spočívající v popisu mentálního pomocí nementálních termínů, nebo se stát obětí buď šovinismu, nebo liberalismu. Osobně bych se přiklonil k „terestriálnímu šovinismu“, který by reflektoval naše přirozené prostředí a lze se o něm domnívat, že alespoň několik desítek let bude postačující. Oproti teorii identity však nabízí možnost realizace mysli i na nebiologickém substrátu a tím slouží jako přínosná metafora v umělé inteligenci. Stavovou tabulku je

[47] Block (1978, s. 55) uvádí příklad, kdy by s námi byl funkčně ekvivalentní určitý stát – jeho ekonomika totiž také má vstupy a výstupy.

navíc možné upravit tak, aby odpovídala danému druhu – je pravděpodobné, že bychom se těmito úpravami vzdali obecného popisu, ale není vyloučené, že bychom z takto odlišných tabulek byli naopak schopni abstrahovat popis dostatečně obecný, že by se vztahoval na všechny stvoření.

5. Rozšířená mysl

V této části se budu věnovat koncepci rozšířené mysli, která vychází z funkcionalistického předpokladu o mnohonásobné realizovatelnosti mysli. Rozšířenou mysl jsem zařadil ze dvou důvodů. Tím prvním je rozšíření tzv. nositelné elektroniky (např. chytré hodinky či brýle), která se může v následujících letech stát prostředkem rozšiřování našich kognitivních schopností (zde vidím největší potenciál v zařízeních pro rozšířenou realitu). Druhým důvodem je fakt, že text, který budu analyzovat, uvádí i externalizaci kognitivních procesů, které provádíme na denní bázi, aniž bychom si uvědomovali, že tak činíme. I tato kapitola bude rozdělena na dvě části, nejdříve představím danou koncepci a následně uvedu její kritiku.

Představu rozšířené mysli nastínil již zakladatel funkcionalismus Hilary Putnam. Ve své knize *Reprezentace a skutečnost* navrhl „rozšířit pojem komputačních stavů tak, aby zahrnovaly aspekty prostředí“ (Putnam, 1991, s. 74) a „charakterizovat referenci … jako funkční vztah mezi reprezentací používanou organismem a věcmi, které mohou být jak uvnitř, tak vně organismus“ (Putnam, 1991, s. 74). Jelikož však tato úvaha pochází z doby, kdy se Putnam distancoval od svých dřívějších (funkcionalistických) názorů, přistupuje k ní odmítavě, jedná se o pouhou iluzorní možnost, jak zachránit funkcionalismus. Nutno zdůraznit, že k rozšířené mysli přistupuje Putnam pouze z hlediska funkcionalismu, což kontrastuje s následujícím pojetím.

5.1. Koncepce Clarka a Chalmerse

Koncepce rozšířené mysli je spojena s textem Andyho Clarka a Davida Chalmerse *Rozšířená mysl* z roku 1998. Svůj text začínají Clark a Chalmers otázkou „[k]de končí mysl a začíná zbytek světa?“ (Clark a Chalmers, 1998) a uvádějí tři možné přístupy. První považuje za hranici naše tělo, tedy co je mimo tělo je i mimo mysl. Druhý přístup, vycházející z myšlenkového experimentu Dvojče Země[48] ve kterém Putnam tvrdí, že význam není jen v hlavě, a tento externalismus je pak analogicky aplikován i na lidskou

[48] Země Dvojče je téměř identická s naší Zemí, jen s tím rozdílem, že „voda“ je na Zemi Dvojčeti XYZ, nikoliv H_2O. Fyzikální vlastnosti mají ale XYZ a H_2O identické. Obyvatelé Země Dvojčete, ač velmi podobní nám pozemšťanům, když vysloví „voda“, neznamená to voda (v našem slova smyslu). Voda totiž znamená H_2O, zatímco „voda“ znamená XYZ, což je odlišná látka (můžeme říci voda-dvojče). Pozemšťan a pozemšťan-dvojče, mohou mít oba stejné mentální stavy, ale význam slova voda bude u obou odlišný. Z toho Putnam usuzuje, že význam nemůže být pouze v naší hlavě.

mysl. Clark a Chalmers navrhují třetí přístup tzv. aktivního externalismu, „založeného na aktivní roli okolního prostředí v řízení kognitivních procesů“ (Clark a Chalmers, 1998).

Pro snadné pochopení rozšíření kognice uvádějí příklad s člověkem, který sedí u počítače a má rozhodnout, zda jsou různě prostorově natočené dvojrozměrné objekty identické. Abychom se mohli rozhodnout, je třeba provést mentální rotaci prvního objektu a srovnat jej s druhým objektem (možnost M1). V tomto případě se bezpochyby jedná o kognitivní proces. Dále je tu ale možnost M2, že by za nás rotaci objektu obstaral počítač, po stisknutí určitého tlačítka (vzpomeňme na hru Tetris). A poslední možnost M3 je, že bychom byli vybaveni neurálním implantátem, který by za nás provedl rotaci předmětu jako počítač v předchozím případě. Clark a Chalmers se domnívají, že v druhém i třetím případě se jedná o stejnou míru kognice, jako v prvním případě. Poukazují na obecnou lidskou tendenci spoléhat na prostředky externalizace našich kognitivních procesů, mezi něž patří využití tužky a papíru k provedení složitých výpočtů, logaritmického pravítka, či přeskládání kostiček s písmenky ve hře Scrabble, abychom si snáze vybavili slova. „Ve všech těchto případech mozek samotný vykoná určité operace, zatímco jiné jsou svěřeny manipulacím s externím médiem“ (Clark a Chalmers, 1998). Výhodou externalizace může být jak usnadnění, tak i zrychlení daného procesu. Uveden je výzkum Kirsch a Maglio z roku 1994, při kterém bylo zjištěno, že mentální rotace tvaru o 90 stupňů zabere přibližně 1000 milisekund, zatímco fyzická rotace na zařízení pouhých 100 milisekund, plus 200 milisekund na stisk tlačítka (Clark a Chalmers, 1998).

Abychom mohli mluvit o externalizované kognici, musí být daná externalizace epistemicky věrohodná, tj. pokud při „nějakém úkolu funguje část světa jako proces, který, *kdybychom ho prováděli v hlavě*, bychom bezpochyby považovali za část kognitivního procesu, pak ta část světa *je* (to tvrdíme) součástí kognitivního procesu. Kognitivní procesy nejsou (pouze) v hlavě!“ (Clark a Chalmers, 1998). Tyto externí součásti kognitivního procesu nejen že fungují jako ryze interní kognitivní procesy, ale jejich odstranění vede k omezení naší kompetence, stejně jako by k ní vedlo odstranění části mozku – vzniká *spárovaný systém*.

Přístup Clarka a Chalmerse je aktivní externalismus, který je v protikladu s Putnamovým pasivním externalismem. V případě pasivního externalismu hrají stěžejní roli interní struktury a změna externích nemusí vést ke změnám chování. Naopak u aktivního externalismu mají externalizované části stejnou důležitost a jejich změna vede ke

změnám v chování. Např. význam vody je pro mě dán historicky, pokud bych se tedy nyní dostal na Zemi Dvojče, kde by nebyla „voda“ ale „XYZ“, moje myšlení o vodě by se okamžitě nezměnilo. Pokud bych však měl dříve zmíněný implantát, který by mi umožňoval počítačově provádět rotaci objektů v prostoru, v případě jeho odejmutí by se mé kognitivní schopnosti okamžitě změnily.

Autoři dále upozorňují, že ne všechny kognitivní procesy (např. zpracování jazyka a učení se dovednostem) jsou zároveň vědomé procesy, tudíž externalizace těchto procesů nijak nezpochybní interní podstatu vědomí. Dále je možné externalizaci zpochybnit pro snadnou „odpojitelnost“ externí a interní části kognitivního procesu. Zde je však rozdíl dán současným stavem techniky, je pravděpodobné, že dosáhneme bodu, kdy s námi budou externí prostředky spojeny stejně nedílně, jako naše končetiny – které také mohou být součástí externalizace, vzpomeňme např. na počítání na prstech.

Se snadnou „odpojitelností“ externích prostředků se pojí problém jejich spolehlivosti a důvěryhodnosti. Největší spolehlivost a důvěru samozřejmě mají procesy probíhající uvnitř našeho mozku. V případě externích prostředků je nezbytná jejich přítomnost vždy, když budou zapotřebí. Příkladem takového prostředku, který máme vždy spolehlivě k dispozici je jazyk (řeč).

Další otázkou je povaha mysli. Mentální stavy, jako přesvědčení, touhy a emoce se jeví být interní, závislé na mozku. Pokud však daný mentální stav závisí na paměti, kterou lze externalizovat, situace se začne komplikovat. Clark a Chalmers uvádějí příklad s Ottou, který trpí Alzheimerovou nemocí a svoji paměť externalizuje do zápisníku, který nosí stále s sebou. Otto a Inga se dozví o probíhající výstavě v muzeu a rozhodnou se ji navštívit. Inga je přesvědčená, že muzeum je v Ulici 53 ještě dříve, než si musela tuto informaci vybavit z paměti. Stejně tak Otto má jisté přesvědčení o poloze muzea aniž by se musel podívat do svého zápisníku. Stěžejní fakt je, že Otto na svůj zápisník spoléhá stejně, jako Inga na svou paměť a stejně jako Ingino přesvědčení nezmizí, když se stane nevědomým, nezmizí Ottovo přesvědčení, když odloží zápisník. „V obou případech je informace spolehlivě k dispozici když ji potřebují, dostupná vědomí a schopná řídit jednání stejným způsobem, jako to očekáváme od přesvědčení“ (Clark a Chalmers, 1998).

Nezbytnou vlastností Ottova zápisníku, je četnost a spolehlivost jeho využití. Pokud by Otto zápisník používal jen zřídka, nebo pokud by v něm v převažujícím procentu případů

nenašel potřebné informace, neplnil by zápisník přisuzovanou roli externalizované paměti. Můžeme namítnout, že Ottův zápisník není dostupný vždy (např. při sprchování), avšak autoři upozorňují na fakt, že ani naše paměť není vždy spolehlivě dostupná, což může být způsobeno třeba zvýšenou konzumací alkoholu. Stejně tak nehraje roli rychlost přístupu k informacím, pokud je přístup stále dostatečně jednoduchý a spolehlivý. Dle autorů nehraje roli ani způsob vybavování si dané informace – v případě Otty percepcí, v případě Ingy introspekcí. U Otty se může při vybavení objevit fenomenální zkušenost, ale to na podstatu přesvědčení nemá vliv.

Ottova „externí paměť" a Ingina „interní paměť" samozřejmě není identická. Podstatným rozdílem, jak uvádí Sprevak (2009, s. 4), je „negativní převod", který je běžný pro lidskou paměť (a to jak krátkodobou, tak dlouhodobou), avšak na externalizovanou paměť vliv nemá. Jedná se o problém, se kterým se potýkáme, když se máme naučit jinou variantu něčeho, co už známe. Příkladem může být chybně naučená slovíčka cizího jazyka – naučení se správného překladu nám bude činit větší obtíže, než naučení se novým slovům. Otto si jen přepíše informaci ve svém zápisníku a tyto problémy se ho týkat nebudou. Stejně tak bude mít na Inginu paměť vliv použití mnemotechnických pomůcek či křivka zapomínání, přičemž nic z toho se nebude vztahovat na Ottův zápisník.

Pro externalizaci přesvědčení hovoří i další variace na Putnamův experiment se Zemí Dvojčetem. Máme si představit Dvojče Ottu, který, stejně jako první Otto, trpí Alzheimerovou nemocí a má svůj zápisník. Do něj si však chybně zapíše, že muzeum je v Ulici 51. Přestože oba Ottové jsou fyzicky identičtí, jejich zápisník se liší a tím se liší i jejich přesvědčení. Zde se opět dostáváme k rozdílu mezi pasivním (Putnam) a aktivním (Clark a Chalmers) externalismem. V případě Putnamova experimentu mám já i můj dvojník na Zemi Dvojčeti odlišné přesvědčení, avšak naše chování je identické. V případě Otty a Otty Dvojčete mají externalizované vlastnosti aktivní dopad na chování, Otto a Otto Dvojče půjde každý do jiné ulice.

Clark a Chalmers (Clark a Chalmers, 1998) uvádějí čtyři kritéria pro připsání rozšířeného přesvědčení, přičemž první tři jsou stěžejní, čtvrté kritérium je do jisté míry sporné.

K1) Zdroj musí být spolehlivě k dispozici a běžně užíván.

K2) Informace obsažená ve zdroji je dostupná přímo a bez obtíží.

K3) Po obdržení informace ze zdroje je tato automaticky přijata.

K4) Informace musela být vědomě přijata někdy v minulosti a existovat jakožto následek tohoto přijetí.

S rozšířením představy, že lidský mozek funguje obdobně jako počítač a také s pokroky v neurobiologickém výzkumu bylo popsáno rozhraní člověk-stroj (brain-machine interface, zkráceně BMI), které souvisí i s koncepcí rozšířené mysli. Jedná se o uzavřené systémy, které využívají neurální aktivitu k řízení responzivních zařízení (např. počítač či robotická protéza). Koncepce těchto zařízení pochází již z šedesátých let dvacátého století a vycházela z behavioristických zkoumání. Jedno z prvních BMI zařízení pochází z devadesátých let a umožňovalo ovládat kurzor na počítačové obrazovce za využití encefalografu (EEG). V současnosti se krom zařízení založených na EEG využívají i systémy na bázi elektrokortikogramu (ECOG; invazivní EEG).

Výzkum společnosti Neurospace (Chang, Breshears a Rowland, 2013) zabývající se využitím responzivní neurostimulace u pacientů trpících epilepsií prokázal snížení záchvatů o 37.9% a navíc i zlepšení v řečových a paměťových úlohách. Jiným příkladem využití je hloubková mozková stimulace (DBS) pacientů trpících potížemi s pohybem. Stimulace je cílena na thalamus a basální ganglia a dlouhodobé studie ukazují přijetí DBS implantátů pacientem a trvalé zlepšení jeho stavu.

První BMI založené na bázi ECOG bylo stvořeno v roce 2004 a umožňovalo pacientům trpícím epilepsií ovládat kurzor na počítači. Během pokusů byla zaznamenána mozková aktivita pacientů při skutečném pohybu kurzoru a nalezeny skupiny frekvencí odpovídající pohybu v daném směru. Sledování těchto skupin frekvencí pak umožnilo ovládat kurzor pouze představou jeho pohybu. Tato technologie byla následně rozšířena k ovládání robotických protéz. Dále byl vyvinut i algoritmus, který umožňuje přesné ovládání prstů robotické ruky u pacientů s mozkovou mrtvicí (Chang, Breshears a Rowland, 2013).

5.2. Kritika

Kritizována je v první řadě identita možností při uvedené rotaci předmětů, jelikož M2 oproti M1 zahrnuje motorickou aktivitu a s ní spojené kognitivní procesy. Nezávisle na rychlosti a jednoduchosti M1 a M2 je mezi nimi rozdíl i v kauzální struktuře. Stejně tak vyhledávání v Ottově zápisníku je spojeno s motorickým a zrakovým systémem, na rozdíl od vyhledávání ve vnitřní paměti, které provádí Inga. V obou případech ale můžeme

mluvit o realizaci stejné funkce, která zpracovává stejné vstupy, podává stejné výstupy a liší se pouze ve zdroji (poskytovateli) těchto vstupů.

Další problém hypotézy rozšířené kognice (HRK) je její souvislost s funkcionalismem, díky čemuž přebírá i některé jeho problémy. Sprevak (2009, s. 1) tvrdí, že „[b]uď je HRK pravdivá, nebo je funkcionalismus nepravdivý“. Upozorňuje ale na fakt, že HRK vycházející z funkcionalismu je radikálnější, než původní hypotéza Clarka a Chalmerse, a to dokonce až takovým způsobem, že vytváří protipól k funkcionalismu. HRK Clarka a Chalmerse se jeví jako chybná, místo ní nastupuje radikální HRK, která je však také neudržitelná. Radikální HRK bude vysvětlena později, prozatím stačí uvést, že pramení z příliš liberálního funkcionalismu.

Funkcionalistická platnost HRK závisí na specifikaci funkčních rolí, přesněji řečeno na konkrétnosti jejich popisu. Lidské mentální stavy mají různé příčiny (vstupy) a účinky (výstupy), přičemž v popisu funkční role nejsou všechny zohledněny. Určité příčiny/účinky (např. snížená citlivost k méně závažným poraněním) totiž nejsou nutně průvodním jevem daného mentálního stavu (v tomto případě bolesti), a naopak určité příčiny/účinky je třeba zobecnit (bodnutí, zlomenina, spálenina = tělesné zranění). Pokud tedy bude funkční role specifikována příliš striktně, HRK nebude tato kritéria splňovat. Naproti tomu příliš benevolentní specifikace mohou vést k přisouzení mentálních stavů i stvořením, které jimi nedisponují.

Otázkou zůstává, jaké vlastnosti ignorovat a jaké zobecnit. Funkční model postavený na základě lidských mentálních jevů může vyloučit jiné živočišné druhy (včetně těch dosud neobjevených) i rozšířenou mysl/kognici. Těžko si dokážeme představit všechny způsoby, jak by mohla být realizována např. paměť tak, aby byla funkčně ekvivalentní s lidskou. Není vyloučeno, že existují stvoření, jejichž interní (můžeme se domnívat, že i biologická) paměť pracuje na stejném principu, jako Ottův zápisník.

Přikláním se k Sprevakovo skepticismu, že neexistuje ideální specifikace. I kdyby existovala, byla by ideální pouze pro určitý funkcionalistický přístup. Problematika nastavení specifikací se vynořuje jak u Funkcionalismu a Psychofunkcionalismu, tak i u teorie identity s funkčními stavy či strojového funkcionalismu. Obecně se dá říci, že u každé funkcionalistické teorie je třeba nastavit specifikace tak, aby bylo na jejich základě možno připsat mentální stavy i jiným formám života. Sprevak (2009, s. 5) zde operuje

s tzv. „Marťanskou intuicí[49]“, dle které „je možné, aby měla stvoření mentální stavy, i když tato stvoření mají odlišnou fyzickou a biologickou strukturu než my“. Pokud platí Marťanská intuice, pak platí i HRK.

Z hlediska funkcionalismu jsou problematická tři kritéria, která dle Clarka a Chalmerse musí rozšířená kognice splňovat. Pokud přistoupíme na rovnost externích a interních procesů, dostaneme se do nesnází, jelikož mohou existovat interní procesy, které nebudou splňovat kritéria, které Clark a Chalmers kladou na externí procesy. Buď jsou tato kritéria nesprávná, nebo je mezi externími a interními procesy rozdíl.

1) Existují interní kognitivní procesy, které nebudou ani spolehlivě k dispozici, ani běžně využívané. Příkladem může být změněný stav vědomí vlivem psychotropních látek, náhlá inspirace (umělecká či týkající se řešení určitého problému) či intuice.
2) Ani požadavek na snadno dostupné informace neplatí pro všechny interní procesy. Například nějaká traumatizující vzpomínka, kterou jsme vytlačili do nevědomí, může být dostupná pouze za pomoci psychoterapie. Jiným příkladem je student, který se naučil určitou věc, je schopný si ji bez obtíží vybavit, ale během zkoušky je natolik ve stresu, že si na tuto informaci nevzpomene.
3) „Některé lidské interní kognitivní procesy nemají své výstupy automaticky přijaté, např. představy, předpoklady, touhy“ (Sprevak, 2009, s. 11). Stejně tak si dokážeme představit kritický přístup a jistou nedůvěru k informacím poskytnutým těmito interními procesy.

Pokud mohou interní prostředky porušovat kritéria K1-K4, pak je mohou porušovat i externí prostředky, aby zůstal zachován princip rovnosti, který je nezbytný pro potvrzení HRK.

Představme si dříve zmíněného Ottu a jeho kamaráda Jürgena, který také trpí Alzheimerovou nemocí a též má svůj zápisník, který používá stejně jako Otto. Pokud by radikální HRK byla pravdivá, tak při setkání Otty a Jürgena by si tito dva přátelé mohli vyměnit své zápisníky a tím by si vyměnili svou paměť, přesvědčení a jejich mentální

[49] Tento „předpoklad“ se obvykle pojí s kvalitativními mentálními stavy, jako je bolest, avšak dle Sprevaka nic nebrání jeho použití s kognitivními stavy a procesy. Dále upozorňuje na to, že „Marťanská intuice není o tom, že by Marťan mohl mít mentální stavy ve všech ohledech identické s lidskými mentálními stavy“ (Sprevak, 2009, s. 5), vztahuje se pouze na daný typ lidského mentální stavu.

vlastnosti by se okamžitě změnily. Stejně tak bychom mohli Ottův a Jürgenův zápisník nahradit např. tabletem s neustálým a neomezeným přístupem k internetu, čímž by oba získali nezměrné množství nových přesvědčení.

Dostáváme se k otázce, zda existují podmínky, které by zajistily ekvivalenci interní a externí kognice, a zároveň nevedly k radikální HRK. Sprevak se domnívá, že je to velmi nepravděpodobné, a to ze dvou důvodů:

> 1) „Pro jakýkoliv případ lidského využití nástroje p k manipulaci s reprezentacemi si dokážeme představit Marťana, který je stejný jako my, až na to, že p je součástí jeho interních procesů. Zdá se být zcela v pořádku považovat p za jeden z Marťanových kognitivních procesů. … Pokud p splňuje podmínky možného interního kognitivního procesu, pak s ním nějaké dodatečné podmínky nemohou nakládat jako s ne-kognitivním – to by porušilo princip rovnosti[50]. Znamenalo by to říci, že pokud se p vyskytuje interně, bylo by kognitivní, ale pokud se p vyskytuje externě, v jinak funkčně identickém systému, je ne-kognitivní“ (Sprevak, 2009, s. 14.).
>
> 2) „Jediný způsob, jak se vyhnout radikální HRK je buď opustit princip rovnosti, nebo opustit tvrzení, že Marťané v těchto případech mají mentální stavy“ (Sprevak, 2009, s. 14). Jelikož je princip rovnosti pro koncepci rozšířené kognice stěžejní, je evidentní, že tato možnost by zabránila nejen radikální HRK, ale i umírněnějším alternativám. Situace s Marťany je ještě problematičtější, jeví se jako nutný ústupek k zachování umírněné HRK, avšak ústupek nepodložený, který dává vzniknout dalším otázkám[51].

Zdá se, že funkcionalismus, který stál u zrodu HRK, znamená i její konec, jelikož s sebou přináší pouze její radikální verzi. Jakou roli má princip rovnosti jsme si již ukázali, teď je třeba vysvětlit, z jakých funkcionalistických předpokladů pramení problémy s HRK. Dle Sprevaka jsou to následující dva:

> P1) Pokud je nám organismus dostatečně podobný na celkové funkční úrovni, pak je to kognitivní agent.

[50] Sprevak namísto Clarkova „parity principle“ používá „fair-treatment principle“, ale oba znamenají totéž. Pro jednoduchost oba překládám stejně.
[51] Podrobněji viz (Sprevak, 2009, s. 14-15).

P2) Pokud jsou činnosti kognitivního agenta významnou měrou řízeny procesem manipulace se symboly, pak je tento proces součástí jeho kognitivních procesů.

Odmítnutí P1 by znamenalo odmítnutí funkcionalismu jako takového a P2 je zapotřebí, abychom se vyvarovali funkcionalistického šovinismu, jak jej představil Ned Block. Odmítnutím P2 a obecně liberálního přístupu k přisuzování kognitivních procesů se sice přikloníme k jisté formě šovinistického funkcionalismu, ale HRK zůstane plausibilní a uchráněna své radikální verze.

Navzdory veškeré kritice nepovažuji hypotézu rozšířené mysli za zcela chybnou. Jako největší nedostatek a překážku vidím princip rovnosti, který je formulován příliš obecně a umožňuje tak poměrně snadné vyvrácení celé hypotézy, byť formou pro mě ne zcela relevantních argumentů. Řešením může být zúžení principu rovnost pouze na konkrétní druhy kognitivních procesů, např. paměť, senzorické vnímání atp. Tyto konkrétní případy by byly opatřeny vlastními kritérii (viz kritéria 1-4 uvedená Clarkem a Chalmersem). Pravděpodobně by nebylo možné takto postihnout celou lidskou mysl, ale některé její složky ano.

Nejjednodušší se mi pak jeví zeslabená verze principu rovnosti (i když by se již nejednalo o rovnost v původním smyslu, stále by se jednalo o funkční rovnost), kdy by se za kognitivní stavy daly z externích považovat pouze ty, které by splňovaly kritéria K1-K3 (případně jejich určitá varianta). Tím bychom předešli radikální HRK a umírněná HRK by zůstala plausibilní. Sprevak hovoří o JS- HRK[52], která vyžaduje kritéria K1-K3 spolu s funkčními podmínkami. V jeho pojetí je však JS- HRK stále podřízena nezměněnému principu rovnosti, takže opět vede radikální HRK[53].

Ať už přijmeme či odmítneme kritiku HRK, nelze ji upřít přístup k lidské mysli, který dává vzniknout zajímavým otázkám a vede nás k odlišnému pohledu na zkoumaný jev. HRK nám předkládá mysl v kontextu okolního světa, ne jako izolovanou a na okolí nezávislou, což razantně mění problém mysli a těla, ze kterého se stává problémy mysli, těla a světa. Pokud připustíme tuto reformulaci, bude třeba i nového přístupu ke

[52] Jointly sufficient HRK

[53] JS- HRK sama o sobě radikální HRK nepřipouští (nesplňuje její požadavky), ale radikální HRK je funkčně ekvivalentní s interními stavy, které požadavky JS- HRK splňují.

zkoumání, a „HRK, svědčící o problémech funkcionalismu, by mohla být nápomocná v utváření následnické teorie“ (Sprevak, 2009, s. 20).

6. Umělá inteligence

Tato kapitola se bude, jak již název napovídá, zabývat umělou inteligencí. Stejně jako hypotéza rozšířené kognice i umělá inteligence vychází z funkcionalismu. Kapitola bude rozdělena do čtyř částí, kdy první, úvodní, část má za cíl objasnit, čím se umělá inteligence jakožto vědní disciplína zabývá. Druhá část mapuje vývoj umělé inteligence (a jak uvidíme, tento vývoj je do jisté míry paralelní s vývojem teorií mysli), třetí část se zabývá problematikou přirozeného a umělého, čtvrtá část představí hlavní přístupy k realizaci umělé inteligence a poslední pátá část představí myšlenkový experiment Johna Searla.

6.1. Umělá inteligence jakožto vědní disciplína

Umělá inteligence (UI) má s filosofií mysli mnoho společného, nejvíce pak zájem o lidskou mysl. Shodný je i interdisciplinární charakter – setkává se zde „psychologie, neurologie, kybernetika, matematická logika, teorie rozhodování, informatika, teorie her, lingvistika atd." (Mařík, Štěpánková a Lažanský, 1993, s. 11) Umělá inteligence přitom není pouhou teorií, jaká by mohla být povaha naší mysli a jak by tato mohla fungovat, ale poskytuje nám i cenné možnosti modelování[54] lidské kognice, ať již jako celku, nebo pouze určitých částí. Shodně s filosofií mysli si klade otázky ontologické (které jsou oproti filosofii mysli doplněny o úvahy nad rozdílem přirozené - umělé), epistemologické a metodologické[55], ke kterým přidává ještě otázky futurologické a etické. V současnosti v UI převládá aplikační motivace, avšak podobnost s filosofií mysli je typická pro badatelskou motivaci UI – společně patří do rámce kognitivních věd[56]. Stěžejní je „předpoklad, že každý aspekt učení či jakýkoli jiný projev inteligence lze z principu detailně popsat a že lze vytvořit stroje k jejich simulaci" (Citováno dle Žáčková, 2014, s. 51). Havlík (2013, s. 35) shrnuje umělou inteligenci jako:

[54] Pod pojmem modelování myslím jak simulaci, tak duplikaci.

[55] Ontologické a epistemologické otázky, které si klade umělá inteligence můžeme ztotožnit s týmiž otázkami ve filosofii mysli. Metodologické otázky pak souvisí s různými přístupy (paradigmaty) v UI – o těch bude řeč později.

[56] Havel (2001, s. 19) jako předmět studia kognitivních věd uvádí „studium všech forem *lidské* inteligence, od vnímání a jednání (konání) až po řeč a myšlení". Dále dodává, že „s kognitivní vědou je zpravidla spojován názor tzv. ‚silné' umělé inteligence" (Havel, 2001, s. 19), jenž je v UI takřka identický s funkcionalismem. Stejně jako filosofie mysli se i kognitivní vědy vyznačují „obtížnost[í] vymezit vlastní předmět zájmu" (Havel, 2002, s. 22).

> [Č]lověkem stvořený (úmyslný) projekt, zabývající se vytvořením inteligence na nebiologické bázi (např. ve strojích, počítačích a robotech). K dosažení tohoto cíle je zde řada možností: nápodoba (Turingův test pro počítače, které se snaží chovat jako lidé); simulace či emulace (mozek může být simulován či emulován na nebiologické bázi); emergence (inteligence či vědomí je získáno jakožto emergentní vlastnost komplexního fyzického systému); evoluce (je možné ve strojích vyvíjet umělé mozky). Z hierarchického hlediska existují dva základní směry [paradigma]: začít s inteligencí na vysoké úrovni a reprodukovat ji přímo v počítači, nebo odzdola nahoru, tedy od věcí, které jsou velmi jednoduché části inteligence ke komplexnosti mozkové inteligence.

Jelikož dosud nebyla představena uspokojivá definice umělé inteligence (a domnívám se, že kvůli vágnosti obou pojmů tvořících název této disciplíny se jedná o cíl spíše utopické povahy) představíme si alespoň tři různé definice a ukážeme si jejich nedostatky.

1) Definice Marvina L. Minského z roku 1967: „Umělá inteligence je věda o vytváření strojů nebo systémů, které budou při řešení určitého úkolu užívat takového postupu, který – kdyby ho dělal člověk – bychom považovali za projev jeho inteligence“ (Mařík, 1993, s. 17).

Hlavní znaky této definice jsou shodné s Turingovým testem. Její výhodou definice je vyhnutí se otázce, co vlastně inteligence je. Problémem je pak behavioristické zaměření, které může vést k liberálnímu přisouzení inteligence.

2) Definice Richové z roku 1991: „Umělá inteligence se zabývá tím, jak počítačově řešit úlohy, které dnes zatím zvládají lidé lépe“ (Mařík, 1993, s. 17).

Tato definice je problematická ze dvou důvodů. Prvním z nich je závislost na stavu počítačů a lidské inteligenci – lze předpokládat, že s rostoucím výpočetním výkonem a pokroky v robotice budou lidé překonáváni ve stále větším počtu činností. Na druhou stranu je ale možné i to, že v určitých činnostech stroj člověka nikdy nepřekoná. Druhým problémem jsou úkoly, které není schopen řešit ani sám člověk – v tomto ohledu je implicitně zmíněná převaha člověka, od jehož schopností se umělá inteligence vyvíjí. I tato definice se vyhýbá definování inteligence.

3) Kotkova definice z roku 1983: „Umělá inteligence je vlastnost člověkem uměle vytvořených systémů vyznačujících se schopností rozpoznávat předměty, jevy a situace, analyzovat vztahy mezi nimi a tak vytvářet vnitřní modely světa, ve kterých tyto systémy existují, a na tomto základě pak přijímat účelná rozhodnutí, za pomoci schopností předvídat důsledky těchto rozhodnutí a objevovat nové zákonitosti mezi různými modely nebo jejich skupinami“ (Mařík, 1993, s. 18).

Tato definice je nejbližší lidské mysli a jejímu fungování. Stěžejní je interakce s okolním světem a schopnost učení. Systém splňující tuto definici by měl komplexní kognitivní aparát, který by mu umožňoval řešit rozličnou řadu úloh, včetně těch, se kterými si ani lidé neví rady. Osobně bych se nikterak nebránil takovému systému připsat inteligenci.

6.2. Historie umělé inteligence

Vznik umělé inteligence jakožto vědní disciplíny se datuje na rok 1956 (přibližně šest let po představení Turingovy imitační hry), kdy, pod vedením Johna McCarthyho, proběhla dartmouthská konference, na které „odborníci z matematiky, elektrotechniky, lingvistiky, neurologie, psychologie a filozofie … [diskutovali] domněnku, že ‚každé hledisko učení nebo jakýkoliv jiný příznak inteligence může být v principu tak přesně popsán, že může být vyvinut stroj, který ho simuluje‘“ (Mařík, 1993, s. 19). Tato domněnka předcházela funkcionalistickému přístupu k mysli (a nikoliv naopak, jak by se člověk mohl domnívat), se kterým byla umělá inteligence po následující desetiletí úzce spjata. Konference proběhla ve velmi optimistickém duchu, o čemž svědčí vize, že „v roce 1970 počítač bude velmistrem šachu, odhalí nové významné matematické teorémy, porozumí přirozenému jazyku a bude sloužit jako překladatel, bude schopen komponovat hudbu na úrovni klasiků“ (Mařík, 1993, s. 19-20). Nyní, o 45 let později se původní naivita jeví skoro až úsměvně a je otázkou, kdy se podaří naplnit některé z původních cílů (především pak porozumění přirozenému jazyku). Rozčarování z neuspokojivých výsledku na sebe dalo čekat přibližně deset let a tato skepse dalších deset let přetrvávala.

Zvrat přišel až v polovině 70. let, díky prokazatelnému úspěchu expertních systémů MYCIN[57] a PROSPEKTOR[58]. Tyto systémy simulují činnost „expertů“ a vyznačují se obsáhlou reprezentací znalostí („báze znalostí“) jejíž součástí jsou kromě explicitních

[57] Určen k diagnostice infekčních onemocnění krve.
[58] Určen k odhalování nerostných surovin.

informací i „znalosti nepřesné, neurčité, vágní“ (Mařík, 1993, s. 21). Jejich činnost odpovídá Minského definici UI a jedná se o příklad tradičního paradigma a ryze „umělé“ inteligence (o obojím později).

Další etapou je Projekt počítačů páté generace, který započal v roce 1981 a zkoumá „systémy zpracovávající znalostní nenumerické informace, založené na inovovaných teoriích a technologiích, které mohou poskytovat potřebné vyšší funkce, odstraňující omezení typická pro současné konvenční počítače“ (Mařík, 1993, s. 22). V tomto případě se již nejedná o jeden samostatný počítač, ale o jejich hierarchizovanou síť, jejíž hlavní předností bude schopnost učit se. Ani vize Projektu počítačů páté generace se nenaplnila tak, jak tvůrci původně zamýšleli. Na druhou stranu nastavila současný kurz směřování a představila koncepci genetických algoritmů a neuronových sítí.

6.3. Přirozené a umělé

Ještě než si představíme hlavní přístupy v UI, rád bych zmínil příznačnou filosofickou otázku týkající se umělé inteligence, kterou je rozlišení přirozeného a umělého. Tato otázka je důležitá i pro funkcionalismus ve filosofii mysli, jelikož pokud platí teze o mnohonásobné realizovatelnosti, pak by žádné „umělé myšlení“ nemělo existovat, veškeré kognitivní procesy funkčně popsatelné by byly na stejné úrovni jako naše přirozené. Pokud se nám jednoho dne skutečně podaří vytvořit mysl podobnou té naší, přinese to mnohé etické a právní problémy, obzvláště pokud budeme i nadále rozlišovat mezi přirozeným a umělým.

Havel (2001, s. 30) rozlišuje umělé „v *adverbiálním* smyslu, kdy jde jen o způsob, *jak došlo* k realizaci“ a „v *adjektivním* smyslu, kdy (navíc), jde o to, *jaká* realizace *je*, že k ní bylo totiž užito nepřirozeného materiálu či prostředí“, a dále uvádí tři znaky umělého:

1) „Existuje nějaká *přirozená* věc, logicky připouštějící duplikaci (v našem případě je to například lidské myšlení, vnímání, rozhodování apod.),
2) existuje *záměr* člověka (nebo týmu) vytvořit duplikát oné přirozené věci,
3) došlo k *provedení záměru*, čili proběhl intencionální proces vedoucí od záměru k *realizaci* (v našem případě byla dotyčná věc například implementována na počítači)“ (Havel, 2001, s. 31).

Musím se přiznat, že nesouhlasím s žádnou z těchto podmínek. Podmínka 1) totiž znemožňuje vytvoření něčeho skutečně originálního, resp. by tento výtvor označila za přirozený. Můžeme sice tvrdit, že nic skutečně nového nelze vytvořit, ale toto tvrzení vyžaduje značně metaforické pojetí světa. Stejně tak požadavek záměru by věci vzniklé omylem neoznačil jako umělé, a co je podstatnější, tento požadavek neumožňuje vytvořit umělé věci, pokud nemáme konstrukční projekt. V případě lidské mysli, u jejíhož poznání máme stále značné nedostatky je vytvoření takového projektu nereálné. Z hlediska UI pak v adjektivním smyslu umělá mysl nemůže být např. duplikovaný mozek, zatímco v adverbiálním smyslu umělá ano. Dle Havlových podmínek by funkcionalistická mysl realizovaná v počítači byla umělá, zatímco emergentní mysl nikoliv.

Otázkou zůstává, zda vůbec můžeme mluvit o umělé inteligenci (či obecně o umělé mysli). Většinu věcí a jevů rozdělujeme na přirozené a umělé na základě jejich původu – přirozené stvořila příroda, umělé člověk. Problém ale může nastat v případě inteligence (jejíž vymezení je samo o sobě problematické, jak jsme si ukázali dříve). Byť se jedná o ryze behavioristické hledisko (a tím se vracíme k Turingovo testu a problémům s ním spojeným), můžeme tvrdit, že „nelze mít falešnou inteligenci. Pokud se agent chová inteligentně, tak je inteligentní“ (Havlík, 2013, s. 16-17). To je ovšem velice kontroverzní tvrzení, jelikož jak jsme viděli v případě expertních systémů, tyto nedělají nic jiného, než že provádí úkony, na pohled inteligentní, které však imitují (pomineme-li odlišnou rychlost) úkony, které by dělal člověk.

Dostáváme se do situace, kdy bychom určitému programu (či robotovi) vykonávajícímu inteligentní chování mohli přisoudit inteligenci, pokud bychom nevěděli, že mu tento typ chování byl dán zvnějšku jeho programátorem (či konstruktérem). Abychom tedy mohli rozhodovat o přirozeném či umělém, musíme mít jisté znalosti o původu dané věci (či jevu). Příkladem mohou být různé druhy rostlin vzniklé šlechtěním – na první pohled je označíme za přirozené (a stran jejich složení tomu tak je), avšak jejich původ je umělý, čehož si ale člověk bez potřebných znalostí nemusí být vědom.

Problematiku přirozeného a umělého si troufám označit za další pseudoproblém, pramenící pouze z nedokonalosti našeho jazyka. Jelikož se jedná o vágní slova, skýtají v sobě neřešitelnost této dichotomie a přivádí nás k nepřínosným diskuzím. Jak přirozené tak umělé jsou pouhé nálepky, kterými označujeme věci a jevy ve světě, na základě naší libovůle. Pokud bychom umělé nahradili člověkem vytvořené, stalo by se jeho použití

méně problematickým. Dále je třeba se vyrovnat z různou mírou abstraktnosti, kterou můžeme na přirozené a umělé pohlížet. Příkladem budiž umělý les, který je ale tvořen přirozenými stromy. Ve své podstatě je většina věcí tvořena přírodními prvky, jsou ale známy i prvky (transurany s atomovým číslem vyšším jak 99), které se v přírodě nevyskytují, a můžeme je prohlásit za ryze umělé.

6.4. Dvě Paradigmata

Snahy o dosažení „umělé“ mysli vycházejí z funkcionalistických představ o jejím fungování, a především nápodoba lidské mysli je spojována s počátky umělé inteligence jako oboru. Zmíněné základní směry se nazývají *tradiční* a *konekcionistické*. „Tradiční paradigma v umělé inteligenci se v literatuře označuje různými přívlastky jako logicko-symbolické, symbolicko-reprezentační, algoritmické, komputacionalistické“ (Havel, 2001, s. 20). Jak již názvy napovídají, tradiční paradigma je zaměřeno na algoritmy a manipulaci se symboly, a je typické pro počítačový funkcionalismus neboli silnou umělou inteligenci. Silná UI vychází z přesvědčení, že „algoritmické procesy v počítači a kognitivní procesy v mysli jsou (opatrně řečeno) entity téhož řádu“ (Havel, 2001, s. 25). Konekcionistické paradigma vychází ze zjištění McCullocha a Pittse, že „jednotlivým neuronům lze přiřadit elementární logické proměnné a neuronovým obvodům funkce výrokové logiky“ (Havel, 2001, s. 26) – neurony jsou tedy chápány jako dvouhodnotové logické prvky a mozek jako rozsáhla množina těchto prvků.

Model lidské mysli může být dvojího druhu, věrný či metaforický. Věrný model se snaží o co nejlepší nápodobu originálu, a podle úspěšnosti (jak kvalitativní, tak kvantitativní) této nápodoby je hodnocena jeho věrnost. Metaforický model je naopak zaměřen na odlišnosti od originálu. Počítač je tedy metaforický model, který, pomocí programů, umožňuje tvorbu dalších modelů, metaforických i věrných. S modely dále souvisí základní úroveň analogie, kterou Havel definoval následovně: „úroveň popisu, ‚pod‘ níž již není rozhodující, zda si prvky modelu zachovávají či nezachovávají materiální, strukturní, nebo formální podobnost s odpovídajícími prvky modelované skutečnosti“ (Havel, 2001, s. 26). Tradiční paradigma klade základní úroveň analogie na rovinu manipulace se symboly, jedná se o přístup „shora“, zatímco konekcionistické paradigma přistupuje k lidské mysli „zdola“ a základní úroveň analogie je úroveň neuronů či ještě nižší.

Klasická UI je neschopná realizovat fenomenální vědomí – čím je jeho důležitost vyšší, tím více se bude umělá mysl odlišovat od přirozené. Pokud od fenomenální komponenty zcela odhlédneme, je možné vytvořit věrný model naší mysli. Ve snaze vytvořit i fenomenální složku „umělé“[59] mysli je silná umělá inteligence nahrazována genetickým programováním a konekcionismem[60] (emergentismem). Z hlediska problematiky „umělého“ můžeme genetické programování, na rozdíl od emergentismu, zpětně popsat jakožto funkci, i když se na první pohled nepodobá své ryze umělé původní verzi, a ani jejich tvůrce neví, jak se dospělo do současného stavu. Můžeme sledovat historii verzí funkce a v každém jednotlivém bodě jsme schopni ji popsat, zatímco u emergentismu se dostaneme do bodu, kdy nebudeme schopni popsat přechod z jednoho stavu do jiného.

Pojem emergence se objevil již na konci 19. století, kdy G. H. Lewes popsal emergenci jako „vznik něčeho, co nelze predikovat nebo vysvětlit z předchozího stavu věcí“[61] (Havel, 2001, s. 63), přičemž ale emergentní jevy[62] nemusejí být nutně komplexnější, než jevy, ze kterých vznikly. Lidskou „formu“ (stejně jako „formu“ ostatních stvoření) lze označit za emergentní (emergující nad slučováním, rozdělováním a diferenciací buněk). Dokonce můžeme takto uvažovat i o lidské mysli, jako „postupně emergující z interakcí mezi dítětem a jeho fyzickým a kulturním prostředím“ (McClelland, 2010, s. 761).

V umělé inteligenci se tímto směrem ubírá konekcionismus[63], který „je založen na myšlence, že na mnohé složité jevy, myšlení a inteligenci nevyjímaje, lze pohlížet jako na emergentní vlastnosti paralelních dějů v rozsáhlé síti třeba i jednoduchých a vzájemně si podobných *aktivních prvků* (formálních neuronů či procesorů), mezi nimiž existují

[59] Zde v uvozovkách, jelikož vytvoření takové mysli bychom pravděpodobně nenazývali umělou, jelikož by se od naší přirozené neodlišovala v žádném důležitém aspektu.
[60] Známý též jako neuronové sítě, paralelní distribuované procesy, konekcionistické systémy, spinová skla. (Havel, 2001, s. 38)
[61] Srovnej „In resultance, every result is either a sum or a difference of the cooperant forces. It is otherwise with emergence where there is a cooperation of things. The emergent is unlike his components insofar as these are incommensurable, and it cannot be reduced to their sum or their difference'' (Citováno dle McClelland, 2010, s. 752).
[62] Mnoho přírodních jevů lze označit za emergentní – např. přechody mezi skupenstvími, vlastnosti molekul i celých organismů, planety, sluneční soustavu ba i celý vesmír, mravenčí kolonie či města (McClelland, 2010, s. 753).
[63] Naproti tomu ve filosofii mysli může být konekcionismus považován za odnož mechanického funkcionalismu vycházející z představy, že „funkční relaci samu lze strukturovat prostřednictvím spojů, konekcí mezi jinými funkčními relacemi“ (Nosek, 1997, s. 129), tedy že Turingův stroj lze popsat jako spojení různých jiných T-strojů. Funkce „hlavního“ T1-stroje by byla množinou funkcí „dílčích“ T2-strojů, kdy výstupy T2 by tvořily vstupy T1 (celý proces může být samozřejmě více než dvouúrovňový). Nosek (1997, s. 127) dále uvádí, že v případě konekcionismu „funkcionalismus překračuje rámec svého vymezení, neboť přibírá pojem interakce … [konkrétně] interakce mezi elementy sítě“.

interakční vazby (*links*). V roli těchto prvků si můžeme představovat nejen neurony (či neuronové moduly), ale i umělé elektronické prvky anebo formální proměnné simulačních algoritmů, v širším pojetí dále i atomy (jejich spiny), molekuly, buňky, živé organismy, biologické druhy, lidi, národy, terminály internetu a kdoví co ještě.“ (Havel, 2001, s. 38) Základem konekcionismu jsou „dynamické systémy nezávislé na fyzické povaze substrátu, v němž jsou realizovány“ (Havel, 2001, s. 41), což je též hlavní myšlenka funkcionalismus, jakožto přístupu ve filosofii mysli[64].

Celá síť pak prochází dvěma módy: adaptačním a aktivním, kdy se v adaptačním módu modifikují interakční vazby mezi prvky, a do tohoto procesu může zasahovat „učitel“, ať už je jím programátor či přirozené prostředí. Dále je třeba upozornit na to, že lokální chování (tj. aktivita jednotlivých prvků) je „zpravidla velmi *jednoduché* (pravidla mají podobu jednoduchých matematických vztahů), *uniformní* (stejné vztahy platí pro všechny prvky a vazby) a především *srozumitelné*“ (Havel, 2001, s. 38-39), přičemž globální chování (systému jako celku), je z hlediska lokálního chování obtížně popsatelné.

Svým chováním i popisem je konekcionistický přístup nejvhodnější metaforou lidské mysli, jelikož neuronová tkáň je sama druhem konekcionistického systému. V umělém konekcionistickém systému lze najít „analogie s mentálními jevy, jako je například bistabilní vnímání, utkvělé myšlenky, váhání při rozhodování“ (Havel, 2001, s. 39) a další. Podobnost konekcionismu a lidské mysli je v pojetí paměti – distribuovaná paměť, založená na modifikaci vazebních vah, je podobná holografické teorii paměti, jež se ukázala jako vhodnější alternativa ke koncepci fyzických engramů (Havel, 2001, s. 40).

Konekcionistický systém je systémem kontrastů (lokální a globální, adapatační a aktivní), jež jsou nutnou bází pro vznik emergentních jevů. Těmi můžeme označit jevy kauzálně potentní, vznikající na vyšší[65] úrovni systému, které lze považovat za důsledek působení nižší úrovně systému, avšak nelze je na základě tohoto působení jasně popsat a předpovědět. Havel formuluje následující konekcionistickou tezi[66]: „Mentální stavy a

[64] Oproti tomu funkcionalismus v umělé inteligenci se s konekcionismem rozchází ve zvolené úrovni popisu (dříve zmíněný přístup shora a zdola).

[65] Rozlišení na nižší a vyšší úroveň je relativní a závislé na konkrétních úrovních, úroveň Y může být „vyšší“ úrovní pro Z, avšak nižší úrovní pro X. Z tohoto relativního dělení vyplývá i možnost řetězení emergentních jevů.

[66] Zde bych čtenáře odkázal zpět k myšlenkovému experimentu Neda Blocka s čínským národem. Ten by totiž byl také příkladem konekcionistického systému a existuje u něj reálná možnost emergence neočekávaných vlastností. Příkladem jest teorie davového chování, která ukazuje na taktéž emergentní jev

procesy lze pojmout jako emergentní jevy na některé vyšší úrovni dostatečně složitého konekcionistického systému“ (Havel, 2001, s. 41).

V protikladu pak stojí teze symbolického paradigmatu Allena Newella: „K tomu, aby fyzický systém vykazoval obecnou inteligentní činnost, je nutnou a postačující podmínkou, aby to byl fyzický symbolový systém[67]“ (Havel, 2001, s. 43). Základní úroveň analogie symbolického paradigmatu stojí mnohem „výše“, než je tomu u konekcionistického paradigmatu. Symbolické paradigma dále předpokládá, že fenomenální vědomí lze popsat stejně jako přístupové, nebo že nemá vliv na projevy inteligence (což považuji za oprávněný předpoklad – pokud se budeme doslovně držet pojmů, teze symbolického paradigmatu tvrdí něco o „obecné inteligentní činnosti“, která je pouze dílčí částí „mentálních stavů a procesů“).

Na první pohled se obě teze jeví jako neslučitelné, ale nemusí tomu tak nutně být, jak upozorňuje McClelland: „Symboly a procesy které s nimi operují, jsou v současnosti často považovány za přibližný popis emergentních následků sub- či nesymbolických procesů, a celá škála konstruktů v kognitivních vědách může být považována za emergentní“ (McClelland, 2010, s. 751). Společnou vlastností obou paradigmat je pak tvorba vnitřní reprezentace světa – je tedy možné (šovinisticky) hodnotit podobnost této umělé reprezentace s naší přirozenou, rozlišovat správné a chybné výstupy. Jak symbolickému, tak emergentistickému paradigmatu se však dosud nepodařilo podat uspokojivé vysvětlení přirozené mysli, stejně jako se jim nepodařilo vytvoření té umělé.

Havel (2001, s. 52-53) dále uvádí srovnávací otázku „Jak poznáme, že kognitivní systém náležitě funguje?“, na kterou symbolické paradigma odpovídá: „Když symboly vhodně reprezentují určitý aspekt reálného světa a když zpracování informací vede k úspěšnému řešení zadané úlohy“, zatímco odpověď emergentismu je „Když uvidíme, že emergentní vlastnosti (a výsledná struktura) korespondují s nějakou specifickou kognitivní schopností – úspěšným řešením požadované úlohy“.

Emergentní jevy však nemusí vznikat pouze v hierarchizovaném konekcionistickém systému, ale též „na některé vyšší úrovni dostatečně složitého dynamického systému“

v dostatečně rozsáhlém systému. V přírodě lze na emergenci inteligentního chování usuzovat např. u mravenišť či úlů, kdy tyto celky vykazují vyšší míru inteligence (kterou nelze označit za pouhou sumu částí), než jejich jednotliví členové.

[67] Fyzický symbolový systém charakterizuje Havel (2001, s. 43) jako „zařízení, jehož fungování lze popsat v jazyce čistě kauzálních fyzikálních vztahů … a přitom operuje se symboly“.

(Havel, 2001, s. 41). Příkladem takového dynamického systému mohou být kauzální domény, kterými lze označit oblast s identifikovatelnými, srozumitelnými a koherentními kauzálními vztahy (vztahy příčinné souvislosti). Zobecněnou emergentistickou tezi pak formuluje Havel následovně: „mentální stavy a procesy lze pojmout jako emergentní jevy nad rozsáhlou množinou vzájemně vázaných kauzálních domén“ (Havel, 2001, s. 48). I kauzální domény mohou tvořit hierarchii, avšak v porovnání s konekcionistickými úrovněmi je od sebe nelze ostře oddělit, stejně jako nelze jasně určit, která doména leží „výše“ a která „níže“. Příkladem kauzálních domén v lidském mozku jsou „fyzikálně-chemické děje na synapsích a membránách, signální a logické závislosti na úrovni neuronů a funkční oblasti mozku jako celku“ (Havel, 2001, s. 47), což jsou kauzální domény relativně jasně vymezené a objektivně (empiricky) pozorovatelné. Kauzální doménou ale může být i doména mentálních stavů, která je uchopitelná pouze v pojmech lidové psychologie. Je evidentní, že kauzální domény mohou být vymezeny i volně a intuitivně. Stejně tak jejich hierarchie je pouze intuitivní, a dosud neexistují důkazy o správnosti této intuice – namísto jasně strukturované hierarchie se může jednat o funkční či kauzální paralelismus.

Konekcionistické intuitivní chápání „nahoře“ a „dole“, „úroveň“ či „hierarchie“ je podobné představám o fungování lidské mysli, může se tedy jevit jako nejvhodnější, na druhou stranu je podobně neuchopitelné (a neexaktní) a vzdaluje se funkcionalistické touze popsat mysl jasně a bez potřeby intuitivně chápaných pojmů. Stejně tak fyzikální emergentní jevy považuji za zřejmější, než (domnělé) emergentní jevy mysli. Dále se naskýtá otázka po vlastnosti (či stádiu vývoje), které umožní alespoň předpovědět (když už ne popsat) emergenci určité vlastnosti (v našem případě mysli).

Osobně se přikláním k analytickému pohledu na emergenci, který Romportl (2007, s. 92-93) charakterizuje následovně:

> Analytický pohled považuje reálné emergentní systémy za realizace složitých abstraktních nelineárních systémů (tj. abstraktních systémů popsaných netriviálními nelineárními diferenciálními či diferenčními rovnicemi), čili emergence je brána toliko za epifenomén, zdání v člověku vyvolané na základě neznalosti či neschopnosti zjistit nebo zajistit řešení příslušných rovnic a z toho plynoucí nemožnosti činit o systému a jeho chování jasné a ostré předpovědi a

závěry. Emergentní jevy vznikají díky neúplnému popisu jisté části reality – jde tak v zásadě pouze o „jazykový problém“.

Tento pohled totiž skýtá jistou naději, že dosáhneme znalostí, které nám umožní přesné popsání systému a jeho chování, a tím i vysvětlit příčiny a podmínky emergence (která se v této analýze rozplyne). Vymizení emergentního efektu nám umožní jednoznačný funkční popis systému, což v případě mysli znamená návrat k funkcionalismu. Existuje však i možnost, že by lidská mysl byla „speciálním druhem emergentní vlastnosti – vlastnosti, který nějak vyvstává z čistě fyzických a/nebo biologických procesů, ale způsob jejího vyvstávání je natolik komplexní, že věda nebude nikdy schopná přesně porozumět tomuto vyvstávání“ (McClelland, 2010, s. 764).

6.5. Čínská komora Johna Searla

V roce 1980 byl v časopisu *Behavioral and Brain Sciences* publikován článek *Minds, Brains, and Programs* amerického filosofa Johna Searla, ve kterém se vymezuje vůči paradigmatu silné umělé inteligence a předkládá svůj myšlenkový experiment s čínskou komorou, jímž se snaží dokázat nemožnost naprogramovaného počítačového realizování lidské mysli. Bezpochyby se jedná o dosti kontroverzní experiment, který má řadu odpůrců.

K pochopení cíle Searlova myšlenkového experimentu s čínskou komorou je nejprve třeba rozdělit umělou inteligenci na „silnou“ a „slabou“, kde pro slabou umělou inteligenci jsou počítače nástrojem pro studium mysli, zatímco pro silnou umělou inteligenci jsou počítače vybavené správným programem myslí samy o sobě. Searle se jasně vymezuje vůči silné umělé inteligenci, obzvláště vůči tvrzení, že správně naprogramovaný počítač by měl kognitivní stavy a že takovéto programy by vysvětlovaly lidskou kognici.

Svůj myšlenkový experiment s čínskou komorou Searle předkládá jako protiargument k Shankovu programu, který simuluje lidskou schopnost rozumět[68] příběhům. Čínská komora ale platí i pro „libovolnou simulaci lidských mentálních fenoménů pomocí Turingova stroje“ (Searle, 1980, s. 417). Máme si tedy představit Searla, který vůbec neumí čínsky, jak sedí v komoře, kde má čínské texty se znaky, kterým samozřejmě

[68] Na základě vstupních dat (textu) reprezentujících nějaký příběh je Shankův program schopen odpovědět na otázky týkající se tohoto příběhu.

nerozumí (a ani neví, že jsou čínské a ne třeba japonské) a k tomu anglický manuál, kterému jakožto rodilý mluvčí bez problémů rozumí. Tento manuál obsahuje pokyny, jaké symboly má přiřadit ke kterým. Searlovi je do komory vložen znak, kterému nerozumí, avšak na základě manuálu k němu přiřadí jiný znak, který pošle ven jako odpověď. Takto může „komunikovat" i s rodilým Číňanem, který zvenčí zajisté nabyde dojmu, že mu komora rozumí, když je schopná mu adekvátně odpovídat na vkládané texty. V této roli Searle ztvárnil počítač, který by se v daném úkolu choval identicky. Avšak Searle nerozumí čínštině, tudíž ani počítat provádějící stejný úkon nemůže rozumět. Jak Searle uvádí, on sám měl vstupy a výstupy nerozeznatelné od rodilého Číňana a přesto čínsky nerozuměl, proto také ani Shankův počítač nerozumí příběhům. Podstatné je, že Searle nerozumí čínštině, avšak rozumí angličtině, přestože při komunikaci v angličtině i v čínštině vykonává stejný program, tj. komputační operace s ryze formálně specifikovanými prvky. Z toho Searle (1980, s. 418) vyvozuje, že „dokud jsou programy definovány ve smyslu komputačních operací na ryze formálně definovaných prvcích … nemají žádné zajímavé spojení s porozuměním. Zcela jistě nejsou postačující podmínkou, ba dokonce ani netvoří důležitou součást porozumění".

V čem tedy dle Searla tkví rozdíl mezi jeho rozuměním angličtině a nerozumění čínštině, a proč nedokážeme schopnost rozumět vložit do počítače? Searle nevidí důvod, proč by nemělo být možné dát stroji schopnost rozumět (jelikož naše těla a mozky jsou přesně takovými stroji), ale není to podle něj možné u stroje, který provádí komputační procesy nad formálně definovanými prvky. Stěžejní je naše „biologická (tj. chemická a fyzická) struktura, a tato struktura je za jistých podmínek schopná kauzální produkce vnímání, jednání, porozumění, učení, a dalších intencionálních jevů" (Searle, 1980, s. 422). Žádný ryze formální model tedy nikdy nebude postačovat intencionalitě.

V závěru Searle metodou otázka – odpověď shrnuje některé filosofické myšlenky obsažené v argumentu čínské komory. Může stroj myslet? - Ano, jelikož my jsme přesně takový stroj. Může umělý, člověkem vyrobený stroj myslet? – Ano, pokud by měl lidský nervový systém. Pokud jsme schopni zcela duplikovat příčinu, můžeme duplikovat i následek. Může myslet digitální počítač? – Pokud digitálním počítačem myslíme cokoliv, co lze na určité úrovni popisu označit za instanci počítačového programu, pak ano, jelikož i my jsme instancí libovolného množství počítačových programů a myslíme. Avšak, pouhá instance vhodného programu není postačující podmínkou porozumění, jelikož

manipulace s formálními symboly postrádá intencionalitu, má pouze syntax a ne sémantiku. Je třeba si uvědomit, že rozdíl mezi programem a hardwarem není stejný jako rozdíl mezi mentálními operacemi a mozkem. Počítačový program může mít různé realizace, které postrádají intencionalitu – např. Weizenbaumem navrhnutý počítač z toaletního papíru a kamínků. Přitom je pro intencionalitu potřeba něčeho se stejnými kauzálními silami, jaké má mozek. Další odlišnosti jsou tyto: Programy jsou ryze formální, avšak intencionální stavy nikoliv; mentální stavy jsou produktem mozku, zatímco program není produktem počítače. Dle Searla se dopouštíme také té chyby, že stavíme do rovnosti simulaci a duplikaci. Počítačový program pouze simuluje lidské myšlení, což však neznamená, že by skutečně myslel.

Searle tvrdí, že myslet mohou pouze stroje, a to velmi specifický druh strojů, konktrétně mozky a stroje, jež mají stejnou kauzální sílu jako mozky. Stěžejní je pro něj intencionalita, bez které myšlení prostě nemůže existovat. A tato intencionalita je biologický fenomén, z čehož jasně vyplývá, že dle Searla je jakákoliv snaha o počítačovou realizaci (silné) umělé inteligence marné, jelikož v počítači máme pouze neživý hardware, na kterém běží nějaký program, zatímco potřebujeme cca 1400g čvachtavého[69] mozku.

V druhé části svého textu Searle zmiňuje šest námitek (včetně pracoviště odkud pocházejí) vůči argumentu s čínským pokojem a dodává k nim své připomínky.

1) Systémová námitka (Berkeley). Osoba v místnosti sice nerozumí čínsky, nicméně je pouhou součástí systému, má k dispozici pravidla, soubor dat (čínských znaků) a papír pro výpočty, a tento celek čínštině rozumí. Jako protiargument Searle jednoduše modifikuje svůj experiment tak, že se daná osoba vše naučí nazpaměť a výpočet provádí v hlavě, čímž by osoba obsáhla celý systém, avšak nic by se nezměnilo, stále by čínštině nerozuměla.

2) Námitka s robotem (Yale). V tomto případě si máme představit robota, který by měl senzory na vnímání svého okolí, a také např. ruce pro interakci s okolím. Robot by byl řízen počítačem a vykonával by lidské činnosti, jako je například chůze, pití, zatloukání hřebíků atd. Takovýto robot by měl mít skutečné porozumění a další mentální stavy.

[69] Zde můj názor zcela vystihuje Turingovo (2004, s. 477) prohlášení: "[N]ezajímá nás, že má mozek konzistenci studené ovesné kaše. Nechceme říkat 'Tento stroj je docela pevný, takže to není mozek, takže nemůže myslet'".

Searle nejdříve uvádí, že v tomto případě je již úkrok od silné UI, jelikož se připouští, že kognice není pouze otázkou formální manipulace se symboly, ale i určitého kauzálního vztahu s vnějším světem, avšak ani to není dostačující. Máme si představit, že by v robotově mozku byl opět Searle, který by manipuloval se symboly, avšak tentokrát by pocházely z robotových senzorů a byly by odesílány k robotovo pohybovým centrům. Searle by tedy opět manipuloval se symboly, kterým by nerozuměl a neměl by ponětí, co svým jednáním způsobuje.

Boden (1987, s. 11) uvádí, že „Searle-v-robotovi … vykonává funkce, které (dle komputačních teorií) vykonává lidský mozek. Avšak zatímco většina stoupenců komputačních teorií nepřipisuje mozku intencionalitu (a ti kteří ano … tak činí pouze omezeným způsobem), charakterizuje Searle sebe uvnitř robota jako obdařeného ryzí intencionalitou“. Searle tedy oponuje komputační teorii, avšak jeho argument se jí drží jen z části. Searlův popis robotova mozku (Searla-v-robotovi) „zahrnuje kategoriální chybu srovnatelnou s pohledem na mozek jakožto nositele inteligence, v protikladu ke kauzálním vztahům“ (Boden, 1987, s. 13).

3) Námitka se simulátorem mozku (Berkeley a M.I.T.). Nyní budeme simulovat skutečnou sekvenci neuronů pálících na synapsích v mozku rodilého Číňana, když rozumí čínsky vypráveným příběhům. Přístroj má čínské příběhy a otázky na ně jako vstup, simuluje formální strukturu skutečného čínského mozku zpracovávajícího tyto příběhy a jako výstup má čínskou odpověď. Pokud bychom odmítli, že takový přístroj rozumí čínštině, pak bychom museli odmítnout i to, že jí rozumí živý Číňan. Neboť jaký by mezi nimi byl rozdíl na úrovni synapsí? Searle v první řadě namítá, že dle silné umělé inteligence není potřeba vědět, jak pracuje mozek, abychom věděli, jak pracuje mysl. Pokud bychom pro potřeby UI museli vědět, jak funguje mozek, neobtěžovali bychom se s UI. Jako analogii k synapsím předkládá Searle muže ovládajícího vodovodní potrubí – muž obdrží neznámý symbol, v manuálu, kterému rozumí, najde, jakou trubkou má pustit vodu. Poté co otočí všemi správnými kohoutky, na konci potrubí vypluje správná odpověď. Tímto přirovnáním se nám Searle snaží dokázat, že ani na úrovní synapsí není porozumění (jelikož není ve vodovodních trubkách, které ale principiálně fungují stejně). „Problém mozkového simulátoru je ten, že simuluje špatnou vlastnost mozku. Dokud simuluje pouze formální strukturu sekvencí neuronů pálících na synapse, nesimuluje to,

co je na mozku skutečně důležité, konkrétně jeho kauzální vlastnosti, jeho schopnost produkovat intencionální stavy“ (Searle, 1980, s. 421).

Uznávám, že Odpověď mozkového simulátoru není obhajobou silné UI, vůči které se Searle tak zarytě vymezuje, nicméně, jeho odpověď mi nepřijde vůbec přesvědčivá a nedomnívám se, že intencionální stavy jsou zrovna ta stěžejní vlastnost mozku, která dává vzniknout myšlení. Searlovo přirovnání neuronů pálících na synapse k mužovi ovládajícímu vodovodní potrubí mi přijde nanejvýš absurdní. Vzhledem k nedostatečnému vědeckému popisu vztahu mozku a myšlení si mohu dovolit vyjádřit představu, v duchu Searlových námitek, že intencionální stavy nejsou nic víc, než odlišná teplota vody, která proudí daným potrubím.

S námitkou simulátoru mozku souvisí i úvaha Neda Blocka (1978, s. 310-318) o funkční ekvivalenci. Když vytvoříme svůj duplikát, shodný v každé buňce, tento s námi bude funkčně ekvivalentní. Pokud připustíme Putnamovo a Kripkeho argumenty (dle Blocka plausibilní) o významovém externalismu, bude mezi námi rozdíl v paměti a tím i v našich propozičních postojích. Já si například pamatuji, že jsem měl narozeniny, ale můj duplikát si to pamatovat nebude, protože je neměl. Příklad s pamětí mi nepřijde moc prokazatelný, ale tato problematika se vztahuje i na jiné jevy. Naše paměť je fyzicky uložena v mozku, tudíž by identický mozek měl mít i identickou paměť. Pokud bychom byli schopni přesně replikovat mozek, jednalo by se pouze o určitý statický obraz, bez probíhající dynamiky. Tu bychom museli dosadit dodatečně, což by mohlo změnit celý systém. Duplikace v reálném čase je velmi vzdálená našim možnostem a nevidím možnost, jak zcela přesně duplikovat nejen strukturu, ale i vazby a vztahy. Je tedy možné, že bychom duplikovali náš mozek, který by sice obsahoval všechny naše paměti, dojmy, pocity atd., ale nebyl by schopný k nim přistupovat, jelikož by v době duplikaci nebyly aktivně provázány s jinými částmi, tudíž bychom tento vztah neduplikovali, či bychom ho duplikovali špatně.

4) Kombinace námitek (Berkeley a Standford). Spojením tří předchozích odpovědí je s reálným světem interagující robot vybavený počítačem kopírujícím lidský mozek, včetně synapsí. Takovýto systém by měl být obdařen intencionalitou. Searle zcela souhlasí, že by bylo racionální přijmout hypotézu, že robot je obdařen intencionalitou, avšak nikterak to nepomáhá tvrzení silné UI. Přisouzení intencionality je však dáno robotovým vnějším chováním, které, dokud se nám nedostane jiného vysvětlení, se nám

jeví jako intencionální. Pokud bychom však věděli, že je robot řízen formálním programem, intencionalitu bychom mu nepřisoudili.

Námitku 3) považuji za nejvýznamnější a zároveň za nejspornější. Pokud je totiž Searlova kritika námitky 3) pravdivá, pak nemá šanci uspět ani námitka 4), ani žádná jiná kombinace námitek. Platnost původní (tj. nikoliv Searlovy reakce na ni) námitky 3) považuji za postačující (nikoliv nutnou) podmínku potvrzující funkcionalismus. Avšak její vyvrácení nemusí nutně znamenat nepravdivost funkcionalismu.

5) Námitka jiných myslí. „Jak poznáme, že ostatní lidé rozumí čínštině či čemukoliv jinému? Pouze dle jejich chování“ (Searle, 1980, s. 421). Pokud se tedy počítač chová jako člověk, také mu musíme přiznat kognitivní stavy. Searle (1980, s. 421-422) reaguje následovně: „Problém v této diskuzi není o tom, jak poznám, že ostatní lidé mají kognitivní stavy, ale spíše co jim přisuzuji, když jim přisuzuji kognitivní stavy“. Nemohou to být pouze komputační procesy a jejich výstupy, jelikož tyto mohou existovat i bez kognitivních stavů. Tato námitka nás vrací zpět k Turingovo testu a jeho kritice.

6) Námitka mnoha komor. Námitka, dle které je Searlův argument namířený pouze proti digitálním počítačům, tj. současnému stavu technologií. Předpokládá se, že jednoho dne postavíme zařízení mající kauzální procesy a bude umělou inteligencí. Tato námitka však dle Searle nemá význam, jelikož mění původní tezi silné UI, že „mentální procesy jsou komputační procesy nad formálně definovanými prvky“ (Searle, 1980, s. 422), proti které je argument čínské komory postaven (a na jinou hypotézu se však nevztahuje).

Dále si dovolím jednu velice zvrácenou úvahu, která však dle mého názoru dokládá chybnost Searlovy úvahy s čínským pokojem. Představme si, že by byl člověk ihned po narození uzavřen do místnosti, upoután tak, aby se nemohl hýbat, aby nic neviděl, nic neslyšel, nic necítil, zkráceně, byl zbaven veškerých podnětů z okolního prostředí. Jeho životní funkce by byly udržovány nějakým co nejméně vnímatelným způsobem. Pokud by tento člověk strávil řekněme dvacet let v takovéto izolaci, rozvinulo by se u něj myšlení, jaké známe u ostatních lidí? Osobně se domnívám, že nikoliv. Pravděpodobně by mu chyběla „sémantika“. Ta sémantika, která chybí počítači, a proto, dle Searla, nemůže mít lidské myšlení. I tato úvaha stojí v opozici k silné umělé inteligenci, avšak na problém nahlíží z opačné strany, než argument s čínskou komorou. Snažím se poukázat

na fakt, že pouhá struktura mozku dávající vzniknout intencionalitě není postačující podmínkou myšlení. Searle však může mít pravdu, že je to podmínka nutná.

Zajímavou poznámkou je to, že Searl v komoře nemanipuluje se symboly bez významu, ale s čínskými nápisy a tudíž je jeho aktivita taktéž smysluplná, ať už si to uvědomuje či nikoliv. Je tedy evidentní, že Searle v komoře oproti počítači má intencionalitu (vědomě manipuluje symboly, ač jim nerozumí), tudíž se nejedná o přesnou replikaci toho, co by v dané situaci dělal počítač. Na takto jednoduchém příkladu to pravděpodobně nebude hrát roli, nicméně, na výrazně složitějších a komplexních úlohách nemůžeme vyloučit emergenci určitého porozumění (Hauser, nedatováno).

Searle svůj argument považuje za jasné vyvrácení teze silné umělé inteligence, s čímž však, ani po jeho kritice kritiky nemohu souhlasit. Searlova jasná distinkce mezi syntaxí a sémantikou, kterou používá jako vysvětlení rozdílu mezi počítačem a člověkem mi přijde jako návod k vyvrácení samotného argumentu. Proč by syntax a sémantika nemohly být dvě úrovně myšlení, propojené, avšak nezávislé? Ano, pokud bychom toto připustili, stále by neplatila teze silné umělé inteligence, nicméně, na rozdíl od Searla, bychom připustili možnost, že počítače mohou myslet. V době formulování Searlova argumentu by byla následující představa bezpochyby přinejmenším velmi náročná, nicméně dnes, v době internetu a nepostihnutelného množství dat, mi přijde realizace víceúrovňové počítačové inteligence ne neproveditelná. Na jedné úrovni můžeme mít přesně tu syntax, kterou Searle demonstroval, a která k myšlení nestačí (s tímto předpokladem souhlasím), a na jiné (netroufám si tvrdit, zdali na vyšší, nižší, či stejné – nahlíželi bychom na ně jako na dvě kauzální domény) úrovni bychom měli sémantiku. Nevidím důvod, proč bychom nemohli dát Searlovi v komoře krom manuálu, jaký čínský symbol přiřadit k jakému, i vysvětlení, co jednotlivé symboly znamenají. Aby porozuměl pokynům, musí mít znalost nějakého jazyka. Navíc, lidé přeci fungují zcela stejně. Když s někým komunikuji, moje řeč (a stejně tak myšlení) je jednota syntaxe a sémantiky. Jedno bez druhého nefunguje.

Pro Searla je intencionalita biologický fenomén, podobně jako je například fotosyntéza nebo trávení. Searle neodmítá, že by se látka mající stejné účinky jako chlorofyl mohla vyskytovat i jinde ve vesmíru a stejně tak neodmítá potenciální existenci látky, která by mohla být základem myšlení, pouze tvrdí, že počítačové čipy nejsou touto látkou. Zde vidím největší slabiny Searlovy teorie.

Zaprvé, v čem tkví stěžejní rozdíl díky kterému má mozek kauzální vztahy nutné pro myšlení a počítačový čip nikoliv? Odpovědí může být to, že lidský mozek je organický, zatímco čip anorganický. Na čem však spočívá toto dělení? Příroda nerozlišuje organické a anorganické, to člověk zavedl ono dělení (stejně v případě dříve diskutované dichotomie přirozené a umělé). V obou skupinách lze najít i prvky charakteristické pro skupinu opačnou. Lycan (2003, s. 15) upozorňuje na to, že „[s]tvoření s myslí jsou ve výsledku stvořeny ze stejných látek, jako běžné neživé předměty, a jejich vlastnosti jsou dány způsobem, jakým jsou tyto komponenty uspořádány, a jaký mají vztah k externím věcem“. Důraz na biologickou strukturu se tak přemění na důraz na uspořádání, či na přítomnost určitého prvku (resp. na jeho poměr ku ostatním prvkům konstituujícím daný organismus). Zadruhé, fotosyntéza je biochemický jev, který lze vyjádřit rovnicí 6 CO_2 + 12 H_2O → $C_6H_{12}O_6$ + 6 O_2 + 6 H_2O. Víme tedy jasně, za jakých podmínek co s čím reaguje a co vzniká. Pozornost si zaslouží především změna anorganického oxidu uhličitého na organické cukry. Pokud by tedy intencionalita měla být biologický fenomén, stejně jako zmíněná fotosyntéza, musela by být pozorovatelná, zapsatelná, změřitelná, a také potenciálně duplikovatelná. Naše vědomí sice může být vedlejším produktem biochemické reakce, nikterak to ovšem nebrání vytvoření umělé mysli. Silikonové počítače se samozřejmě v tomto případě jeví jako nevhodné, nicméně, určitá forma biologických počítačů (dosud neznámých), by mohla zastat takovou funkci, a to i za předpokladu, že by byly pouhou implementací určitého programu. Searle nám totiž neposkytuje jedinou plausibilní hypotézu, jak by měla být sémantika spojena s biologickou strukturou mozku. I pokud by tento vztah skutečně existoval, měl by být na základě výše zmíněného popsatelný (např. jako další vedlejší produkt biochemických reakcí probíhajících v mozku).

Může se zdát, že připuštěním nutnosti určitého biologického substrátu jsme opustili funkcionalistickou hypotézu, ale není tomu zcela tak. Domnívám se, že (případná!) potřeba biologického substrátu není s funkcionalistickou hypotézou zcela v rozporu. Je totiž možné, že naše mentální stavy jsou skutečně určitou funkcí, že naše mysl je analogická k počítačovému programu, avšak je výpočetně natolik náročná, že je třeba ji realizovat na počítačích zcela odlišných od těch, co máme nyní (např. na biologických počítačích). Mozek je tedy pouze dostatečně výkonný počítač, na jehož „architektuře“, pro potřeby realizace mysli, záleží pouze ve smyslu poskytnutí dostatečného výkonu). Již od nepaměti je známo, že určité materiály se na určitou činnost hodí lépe než jiné na

základě svých vlastností jako je hustota, tvrdost, atd. Není tedy vyloučeno, že v budoucnu objevíme, či budeme schopni syntetizovat látku takových vlastností, která ji předurčí jako nejvhodnější k funkcionalistické realizaci lidské mysli. Margaret A. Boden (1987, s. 8) uvádí, že kov a silikon jsou schopné realizovat určité funkce zrakové buňky, konkrétně např. mapování z 2D do 3D či rozpoznávání gradientu intensity. Nejedná se samozřejmě o nejvhodnější materiály k realizování zrakové funkce, ale pouze o materiály přijatelné (s určitými omezeními). Troufám si tvrdit, že aby měl lidský mozek natolik význačné vlastnosti, jaké mu Searle připisuje, muselo by se v lidském mozku nacházet „něco" (látka či způsob interakce) v přírodě zcela ojedinělého. Existence nějaké specifické a ojedinělé látky mi nepřijde moc pravděpodobná, zvláštní způsob interakce ale nevylučuji. Pokud by se však tento zvláštní způsob interakce týkal známých látek, které se přirozeně vyskytují, bylo by, alespoň teoreticky, možné tento způsob interakce duplikovat.

Boden dále dodává, že samotná intencionalita je filosoficky kontroverzní, a není vůbec jisté, zda bychom ji poznali, kdybychom na ní narazili. Navzdory stáří jejího článku *Únik z čínské komory* (1987) je tato obava stále aktuální. Obecně se intencionalita přisuzuje propozičním postojům, avšak pocitům a vjemům nikoliv, ohledně emocí však neplatí koncensus. Searle svým pojetím intencionality velmi volně navazuje na Brentana, nicméně, nejedná se o jedinou možnost, jak chápat daný jev (na rozdíl od fotosyntézy, kterou nelze chápat jinak). Intuitivně se nám sice šedá kůra mozková jeví jako přijatelnější varianta než silikonový čip, nicméně, upozorňuje Boden (1987, s. 8-9), „naše intuice se může změnit s vědeckými pokroky" a proto bychom měli k hypotézám na ni postavených přistupovat kriticky.

Nyní bych se vrátil k dříve zmíněnému článku Neda Blocka *Potíže s funkcionalismem* a zhodnotil Searlovu čínskou komoru z pohledu funkcionalismu nikoliv v umělé inteligenci, ale ve filosofii mysli. Block (1978, s. 263) v něm totiž rozdíl mezi behaviorismem a funkcionalismem charakterizuje následovně:

> Dle behaviorismu je nutné a postačující … aby byl systém charakterizován určitým souborem (pravděpodobně nekonečným) vztahů vstup-výstup; tedy dle behaviorismu systém touží po G pouze v případě, že pro něj platí určitý souboru podmínek ve formě ‚Obdrž I a vydej O'. Nicméně, dle funkcionalismu, systém může mít tyto vztahy mezi vstupy a výstupy a přesto netoužit po G; jelikož dle funkcionalismu to, zda systém touží po G, závisí na tom, zda má vnitřní stavy, jež

> mají určité kauzální vztahy k ostatním vnitřním stavům (a ke vstupům a výstupům). Jelikož behaviorismus nemá takový požadavek na ‚vnitřní stavy', mohou existovat systémy, kterým behaviorismus připíše mentální stavy a funkcionalismus nikoliv.

Čínská komora je tedy ryze behavioristické pojetí, nikoliv funkcionalistické, pokud se přikloníme k Blockově charakterizaci. Není pochyb o tom, že Searle uvnitř komory má vnitřní stavy, nicméně, tyto stavy bychom určitě nepřipsali počítači, kterého Searle-v-komoře imituje. Searle také rozumí anglicky psanému manuálu, což nás vede k otázce, zda počítač nějakým způsobem rozumí instrukcím, zdali chápe alespoň onu syntax. Pokud by totiž počítač rozuměl pokynům (tuto myšlenku zastává např. Boden), jak má vykonávat danou operaci (aniž by si uvědomoval, co vlastně vykonává), je zde bezpochyby možnost rozšířit toto rozumění i na obsah operace. Pokud však počítači upřeme i porozumění instrukcím, jedná se zde opět o ryze behavioristický přístup, kdy jsme počítač „vycvičili" k určitému druhu chování a celý experiment spíše připomíná trénování zvířat. O to zřejmější je pak nepodobnost mezi touž operací prováděnou Searlem a počítačem. Domnívám se, že jde o natolik významné odlišnosti mezi subjekty (tj. mezi Searlem-v-komoře a počítačem), že nemůžeme mluvit o relevantním experimentu, pokud se přikloníme k Searlově názoru, že počítač ničemu nerozumí, tj. ani svým instrukcím. Pokud se naopak přikloníme k názoru, že počítač svým instrukcím (svému programu) rozumí, pak lze Searlův experiment považovat za chybný.

V závěru bych chtěl říci, že Searlův argument čínské komory považuji za zajímavý, avšak mající pouze minimální faktický přínos. Oblast lidského mozku je dosud velmi neprobádaná, tudíž mi nepřijde jako přínosné zabývat se otázkou vztahu těla a mysli. Věřím, že technický pokrok, především v oblasti neurověd, nám poskytne cenné informace, které by mohly vést k nalezení správně odpovědi na povahu myšlení a tím i na jeho (ne)možnost realizace na jiném než biologickém substrátu. Do té doby by bylo nejlepší upustit od úvah a hypotéz a zaměřit se na řešení praktických problémů. Snaha o realizaci umělé inteligence, ať již slabé, či modifikované silné (která ve svém původním znění pravděpodobně nemůže být zrealizována, jak jsme si již ukázali Searlovým argumentem) může být naším cílem již teď, cílem, který budeme zpřesňovat s nově získanými poznatky, cílem, kterého, v to pevně věřím, jednoho dne dosáhneme.

Závěr

> Lidská mysl je biologicky daný systém s určitými možnostmi a limity… Skutečnost, že ‚přijatelné hypotézy' jsou tomuto specifickému biologickému systému k dispozici, svědčí o jeho schopnosti konstruovat bohaté a komplexní explanační teorie. Avšak ty stejné vlastnosti mysli, jež poskytují přijatelné hypotézy, mohou stejně tak dobře vyloučit ostatní úspěšné teorie jakožto nesrozumitelné lidem. Některé teorie prostě nemusí být mezi přijatelnými hypotézami určenými specifickými vlastnostmi mysli, jež nás uzpůsobuje k ‚představení si správné teorie určitého druhu', ačkoliv tyto teorie mohou být přístupné jinak uspořádané inteligenci (Lycan, 2003, s. 23).

Stejně tak si musíme uvědomit, že každý přístup má své výhody a své nevýhody. Na některé otázky poskytují lepší odpovědi dualistické teorie, na jiné naopak teorie materialistické. Dualismus, epifenomenalismus či emergentismus považuji za intuitivně nejvhodnější (nejsrozumitelnější), ale tyto přístupy nepřináší řešení (jehož požadavek je implicitně obsažen ve formulaci „*problém* mysli a těla"). Jediné skutečné řešení podle mého názoru může přinést pouze určitá forma materialismu. Ta jediná umožní popsat náš *problém* objektivně a ověřitelně. Všechny ostatní teorie nabízejí pouze hypotézy (které zatím nemáme jak verifikovat) či pseudořešení, která problém mysli a těla nahradí jiným problémem (například otázkou po podmínkách emergence určitého jevu).

Jsem přesvědčen o tom, že zkoumáním lidské mysli by se neměla zabývat filosofie, nýbrž věda. Důvod je prostý, a je jím samotná povaha filosofie, totiž ono hledání, které samo o sobě je cílem. Věda touží po dosažení konečného a jasně daného poznání zatímco „[f]ilosofie je … daleko více tázáním, než odpovídáním, spíše hledáním než nalézáním, otevíráním možností než jejich uzavíráním, pochybováním, než spoléháním na jistotu" (Nosek, 1997, s. 15), a jako taková nám nemůže přinést skutečný užitek, nedá v poklidu spočinout našemu duchu nad úspěšně dokonanou činností, naopak, bude nás neustále budit a vybízet k dalšímu pátrání, které bude stejně marné a zbytečné jako to předchozí.

Ať už je funkcionalismus pravdivý či nikoliv, je třeba na něj nahlížet jako na přínosnou metaforu fungování lidské mysli, která dala vzniknout počítačům, jak je známe dnes. I kdyby se nakonec funkcionalismus ukázal jako nepravdivá teorie pro popis vztahu těla a mysli, zásluhy na poli umělé inteligence mu nelze upřít (a nedokážu si představit jiný

přístup k problému těla a mysli, který by nám poskytl podobně plodné myšlenky). Ano, v současné době se funkcionalismus jeví jako překonaný, a to jak ve filosofii mysli, tak v umělé inteligenci, ale dosud nebyl přesvědčivě vyvrácen a jsem přesvědčen, že ani není v moci filosofie tak učinit. Jediným vážným protivníkem funkcionalismu, který mu může přinést buď porážku, nebo vítězství, je neurověda. Není vyloučeno, že neurologie nám poskytne materiální popis naší mysli, zatímco funkcionalismus funkční popis, avšak tyto popisy se nebudou vylučovat, naopak, budou se doplňovat.

Zatím mohu, možná dosti odvážně, tvrdit, že lidská mysl a mozek se k sobě skutečně mohou mít jako software k hardwaru, avšak software natolik složitý a komplexní, k hardwaru natolik výkonnému a postavenému na zcela neznámé architektuře, že zatím není v naší moci je popsat. Mnoho jevů se nám zprvu zdálo nepochopitelných a nepopsatelných, byly předmětem mnoha filosofických dohadů (vzpomeňme na řecké filosofy a jejich snahu objevit pralátku), avšak později se je podařilo vysvětlit exaktními metodami. Lidská mysl se však dosud vymyká materialistickým, a tedy i funkcionalistickým, popisům, které jako jediné jsou vědecky ověřitelné. I kdyby funkcionalismus byl neadekvátní, je alespoň postaven na ověřitelném předpokladu. I přes četnou kritiku považuji funkcionalismus za stále atraktivní koncept hodný dalšího zkoumání.

Seznam použité literatury a pramenů

BIEVER, Celeste. No Skynet: Turing test 'success' isn't all it seems. *NewScientist* [online]. 2014 [cit. 2015-04-11]. Dostupné z: http://www.newscientist.com/article/dn25692-no-skynet-turing-test-success-isnt-all-it-seems.html?full=true

BLOCK, Ned. Troubles with Functionalism. In: SAVAGE, Wade C. *Perception and cognition: Issues in the foundations of psychology*. Minneapolis: University of Minnesota Press, 1978, s. 261-325. ISBN 0816608415.

BLOCK, Ned. Introduction. In: BLOCK, Ned. *Readings in philosophy of psychology*. 2nd printing. Cambridge (MA): Harvard University Press, 1980, s. 1-10. ISBN 067474876x.

BODEN, Margaret A. *Escaping from the Chinese Room*. Sussex: The University of Sussex, 1987.

CLARK, Andy a David CHALMERS. The Extended Mind. In: *Cons.net* [online]. 1998 [cit. 2015-01-07]. Dostupné z: http://consc.net/papers/extended.html

DAWKINS, Richard. *River out of Eden: a Darwinian view of life*. 1st ed., 2nd impr. London: Weidenfeld, 1995. ISBN 02-978-1540-7.

DESCARTES, René. *Rozprava o metodě*. Praha: Svoboda, 1992. ISBN 80-205-0216-5.

HAUSER, Larry. Chinese Room Argument. In: *The internet encyclopedia of philosophy James Fieser, editor* [online]. [cit. 2014-12-07]. ISSN 2161-0002. Dostupné z: http://www.iep.utm.edu/chineser/

HAVEL, Ivan M. Přirozené a umělé myšlení jako filozofický problém. In: MAŘÍK, Vladimír, Olga ŠTĚPÁNKOVÁ a Jiří LAŽANSKÝ. *Umělá inteligence: 3. díl*. 1. vyd. Praha: Academia, 2001, s. 17-75. ISBN 80-200-0472-6.

HAVLÍK, Marek. Problém mysli a těla ve filosofii mysli. In: *E-learningová podpora mezioborové integrace výuky tématu vědomí na UP Olomouc* [online]. 2013 [cit. 2015-03-09]. Dostupné z: http://pfyziollfup.upol.cz/castwiki/?p=4073

HAVLÍK, Vladimír. Artificial or Natural Intelligence?. In: ROMPORTL, Jan, Pavel IRCING, Eva ŽÁČKOVÁ, Michal POLÁK a Radek SCHUSTER. *Beyond AI: Artificial Golem Intelligence*. Plzeň: Západočeská Univerzita, 2013, s. 15-27.

CHANG, Edward F., Jonathan BRESHEARS a Nathan C. ROWLAND. Neurosurgery and the dawning age of Brain-Machine Interfaces. *Surgical Neurology International* [online]. 2013, vol. 4, issue 2, s. 11-14 [cit. 2015-02-02]. DOI: 10.4103/2152-7806.109182. Dostupné z: http://www.surgicalneurologyint.com/text.asp?2013/4/2/11/109182

JAWORSKI, William. *Philosophy of mind: a comprehensive introduction*. Malden, MA: Wiley-Blackwell, 2011, x, 411 p. ISBN 14-443-3368-2.

KALKE, William. What is Wrong with Fodor and Putnam's Functionalism. *Noûs*. 1969, roč. 3, č. 1, s. 83-93.

LYCAN, William G. *Consciousness*. 1st MIT Press pbk. ed. Cambridge, Mass: MIT Press, 1987. ISBN 02-626-2096-0.

LYCAN, William G. Chomsky on the Mind-Body Problem. In: *Chomsky and His Critics*. Oxford, UK: Blackwell Publishing Ltd, 2003, s. 11-28. ISBN 9780470690024. DOI: 10.1002/9780470690024.ch1. Dostupné z: http://www.chomsky.info/onchomsky/20030401.pdf

MARKUS, Gary. What Comes After the Turing Test?. *The New Yorker* [online]. 2014 [cit. 2015-03-12]. Dostupné z: http://www.newyorker.com/tech/elements/what-comes-after-the-turing-test

MAŘÍK, Vladimír. Úvod. In: MAŘÍK, Vladimír, Olga ŠTĚPÁNKOVÁ a Jiří LAŽANSKÝ. *Umělá inteligence: 1. díl*. Vyd. 1. Praha: Academia, 1993, s. 15-32. ISBN 80-200-0496-3.

MCCLELLAND, James L. Emergence in Cognitive Science. *Topics in Cognitive Science*. 2010, vol. 2, issue 4, s. 751-770. DOI: 10.1111/j.1756-8765.2010.01116.x. Dostupné z: http://doi.wiley.com/10.1111/j.1756-8765.2010.01116.x

NOSEK, Jiří. *Mysl a tělo v analytické filosofii: úvod do teorií psychofyzického problému*. Vyd. 1. Praha: Filosofia, 1997, 202 p. ISBN 80-700-7091-9.

POLÁK, Michal. *Filosofie mysli*. Vyd. 1. V Praze: Triton, 2013, 259 s. ISBN 978-802-6103-134.

PUTNAM. *Representation and reality*. 1st pbk. ed. Cambridge, Mass: A Bradford Book, 1991, xv, 136 s. ISBN 978-026-2660-747.

PUTNAM, Hilary. *Mind, language and reality*. Cambridge: Cambridge University, 1975, xvii, 457 s. Philosophical papers (Cambridge University Press), vol. 2. ISBN 05-212-9551-3.

REY, Georges. The Turing thesis vs. the Turing test. *The Philosophers' Magazine* [online]. 2012, č. 57, s. 84-89 [cit. 2015-02-16]. DOI: 10.5840/tpm20125754. Dostupné z: http://www.pdcnet.org/oom/service?url_ver=Z39.88-2004

RITCHIE, Jack. Philosopher of the Month: June 2002 - Hilary Putnam. *Philosophers.co.uk* [online]. 2002 [cit. 2015-03-19]. Dostupné z: https://web.archive.org/web/20110709060653/http:/www.philosophers.co.uk/cafe/phil_jun2002.htm

ROMPORTL, Jan. Umělé myšlení a kauzální paradox emergentních systémů. In: MAŘÍK, Vladimír. *Umělá inteligence: 5. díl*. 1. vyd. Praha: Academia, 2007, s. 91-112. ISBN 978-80-200-1470-2.

ROSI, Sandra. Beware the CyberLover that Steals Personal Data. *PCWorld* [online]. 2007 [cit. 2015-03-13]. Dostupné z: http://www.pcworld.com/article/140507/article.html

SEARLE, John R. Minds, brains, and programs. *Behavioral and Brain Sciences*. 1980, vol. 3, issue 3, s. 417-457. DOI: 10.1017/S0140525X00005756. Dostupné z: http://www.journals.cambridge.org/abstract_S0140525X00005756

SEDGHI, Ami. The 10 most expensive paintings ever sold. *Theguardian* [online]. 2015 [cit. 2015-03-11]. Dostupné z: http://www.theguardian.com/news/datablog/2015/feb/10/the-10-most-expensive-paintings-ever-sold

TURING, Alan Mathison a Jack B. COPELAND. *The essential Turing: Seminal writings in computing, logic, philosophy, artificial intelligence, and artificial life, plus the secrets of Enigma*. New York: Oxford University Press, 2004, viii, 613 p. ISBN 01-982-5080-0.

TVRDÝ, Filip. *Turingův test: Filosofické aspekty umělé inteligence* [online]. Olomouc, 2011 [cit. 2015-02-06]. Dostupné z: http://www.kfil.upol.cz/doc/pgs/tvrdy/Disertacni_prace.pdf?lang=cz. Disertační práce. Univerzita Palackého v Olomouci, Filosofická fakulta, Katedra filosofie.

VADINSKÝ, Ondřej. Různé pohledy na otázku: Mohou stroje myslet?. *E-LOGOS: Electronic Journal for Philosophy* [online]. 2011, č. 4 [cit. 2014-12-10]. Dostupné z: http://nb.vse.cz/kfil/elogos/student/vadinsky11.pdf

VESELOV, Vladimir. Eugene Goostman the Bot. In: *Alicebot* [online]. 2010 [cit. 2015-03-12]. Dostupné z: http://www.alicebot.org/chatbots3/Eugene.pdf

ZARRI, Jason. Notes on Hilary Putnam on the Nature of Mental States. In: *Scholardarity* [online]. 12.1.2013 [cit. 2015-03-23]. Dostupné z: http://www.scholardarity.com/?page_id=2750

Loebner Prize 2010 Results. *Hugh Gene Loebner* [online]. 2010 [cit. 2015-03-12]. Dostupné z: http://loebner.net/Prizef/2010_Contest/results.html

Loebner Prize. *AIDB: The Society for the Study of Artificial Intelligence and Simulation of Behaviour* [online]. [cit. 2015-03-12]. Dostupné z: http://www.aisb.org.uk/events/loebner-prize

Turing Test Success Marks Milestone in Computing History. *University of Reading* [online]. 2014 [cit. 2015-03-12]. Dostupné z: http://www.reading.ac.uk/news-and-events/releases/PR583836.aspx

ŽÁČKOVÁ, Eva. *Člověk a nové technologie*. Vyd. 1. Plzeň: Západočeská univerzita v Plzni, 2014, 82 s. ISBN 978-80-261-0353-0.

Resumé

Functionalism is very influential theory in the cognitive sciences. The main purpose of my thesis is a comprehensive description of functionalism in the philosophy of mind as well as in the artificial intelligence. Especially the topic of functionalism in the philosophy of mind is not well-known in the Czech literature. It lacks description in wider context, with respect to its antecedent theory. Similarly, the concept of extended mind, which is based on the functionalism fundamental idea, attracts hardly any attention in Czech environment.

The thesis itself is divided into six chapters, all of them consists of two main parts – presentation of the topic itself and following criticism and/or problems of the topic. The first chapter introduces the mind body problem as an issue with which is functionalism closely connected and which caused its formation. The second chapter presents the work of Alan Turing, which can be said is the antecedent of functionalism. The third and fourth chapters both deal with one of the most important functionalism articles, *The Nature of Mental States* and *Troubles with Functionalism* by Hilary Putnam and Ned Block respectively. The fifth chapter describes the theory of extended mind, which I consider the second most interesting theory resulting from the functionalism. The most interesting is undoubtedly the artificial intelligence, which is being discussed in the last chapter.

I assert that human mind can be realized even on non-biological substrate and that the hypothesis of extended cognition demonstrate the possibility of external realization of at least some cognitive processes. Even if functionalism is wrong and human mind cannot be realized outside the brain, it provide us with powerful metaphor how to view human mind, metaphor, which was crucial to the development of the artificial intelligence.