Univerzita Hradec Králové
Filosofická fakulta
Katedra filozofie a společenských věd

Umělá inteligence?

Diplomová práce

Autor: Bc. Lea Roubíková

Studijní program: N6101 Filozofie

Studijní obor: Filozofie

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

Hradec Králové, 2018

Univerzita Hradec Králové
Filozofická fakulta

Zadání diplomové práce

Autor:	**Lea Roubíková**
Studium:	F15NP0012
Studijní program:	N6101 Filozofie
Studijní obor:	Filozofie
Název diplomové práce:	**Umělá inteligence?**
Název diplomové práce AJ:	Artificial Intelligence?

Cíl, metody, literatura, předpoklady:

Alan Turing přišel s návrhem, že otázka může stroj myslet? nemůže mít hlubší smysl než může se stroj chovat nerozlišitelně od myslící bytosti? Tím rozpoutal nekončící debaty mezi filosofy, kteří s ním souhlasí, a těmi, kteří mají pocit, že myšlení je něco více než co může dělat stroj. Úkolem práce je prozkoumat a kriticky zhodnotit názory filosofů, kteří s Turingem nesouhlasí, a tak se propracovat k podstatě pojmu lidské myšlení.

Dennett, D. (1996): Kinds of Minds, New York: Basic Books; český překlad Druhy myslí, Academia, Praha, 2004. Dreyfus, Hubert L. What computers still can't do: a critique of artificial reason. MIT press, 1992. Haugeland, John. Mind design II: philosophy, psychology, artificial intelligence (výběr článků k tématu). MIT press, 1997. Havel, I. (2004): "Přirozené a umělé myšlení jako filosofický problém", http://glosy.info/texty/prirozene-a-umele-mysleni-jako-filosoficky-problem/. Saygin, A. P., Cicekli, I. & Akman, V. (2000): 'Turing Test: 50 Years Later'. Minds and Machines 10, 463?518. Searle, J. R. (1984): Minds, brains, and science, Cambridge (Mass.): Harvard University Press; český překlad Mysl, mozek a věda, Mladá fronta, Praha, 1994. Shanker, Stuart G. Wittgenstein's Remarks on the Foundations of AI. Routledge, 2002.

Garantující pracoviště:	Katedra filosofie a společenských věd, Filozofická fakulta
Vedoucí práce:	prof. RNDr. Jaroslav Peregrin, CSc.
Datum zadání závěrečné práce:	12.1.2016

Prohlášení

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

V Hradci Králové dne 28. 3. 2018

Poděkování:

Mé poděkování patří prof. RNDr. Jaroslavu Peregrinovi, CSc. za odborné vedení a trpělivost, kterou mi v průběhu zpracování diplomové práce věnoval.

Anotace

LEA ROUBÍKOVÁ. Umělá inteligence? Hradec Králové: Filozofická fakulta, Univerzita Hradec Králové, 2018, 73. s. Diplomová práce

Ve své diplomové práci, nazvané Umělá inteligence? zvažuji možnost existence umělé inteligence. Má práce je rozčleněna do tří základních segmentů. V první části se věnuji historii samotné myšlenky myslících strojů. Zejména představuji názory Alana Turinga a jeho odpůrců. Ve druhé části se zaměřuji na technický pokrok, který proběhl během posledních desetiletí. Zmiňuji také aktuální úspěchy na poli umělé inteligence. V poslední části se, vzhledem k předchozím poznatkům, zamýšlím nad možností a pravděpodobnou podobou umělé inteligence.

Klíčová slova: umělá inteligence, mysl, neuronové sítě

Annotation

LEA ROUBÍKOVÁ. Artificial Intelligence? Hradce Králové: Faculty of Arts, University of Hradec Králové, 2018, 73. pp. Master's thesis

I consider the possibility of existence of artificial inteligence in my master's thesis called Artificial Inteligence?. The thesis is divided into three fundamental segments. In the first part I deal with the history of idea of thinking machines. Especially as I introduce ideas of Alan Turing and his opponents. In the second part I focus on technical progress which occured in the last few decades. I also mention recent achievements on an artificial intelligence field. In the last part, considering previous facts, I am thinking about the possibility and the likely form of artificial inteligence.

Keywords: artificial intelligence, mind, neural networks

Obsah

Úvod

Ústředním tématem mé diplomové práce je analýza možnosti umělé mysli, spolu s debatami, které tato teze přináší. I přes vlnu odporu, jakou idea umělé mysli od svého vzniku vyvolává, se moderní společnost pomalu ale jistě blíží ke konstrukci programů schopných učit se, rozhodovat se a plánovat své jednání. Stačí to ale na myšlení?
Pojem umělé inteligence je v současné době nadužíván k označování širokého spektra technologií, které imitují lidskou činnost, a tak ulehčující náš život. Od vzniku Turingova testu podvědomě předpokládáme, že pokud někdy budou stroje myslet, projevy umělé inteligence budou do jisté míry totožné s projevy té lidské. Proto hlavní proud umělé inteligence vede ke snaze o imitaci lidských procesů a obsahů myšlení. Je ale tento předpoklad podložený? A není snaha o simulaci lidské mysli spíše překážkou, než možnou cestou k dosažení mysli umělé?

Za účelem zodpovězení těchto otázek, se ve své práci budu zabývat problematikou umělé mysli, v souvislosti s jejím vývojem od poloviny minulého století až k současnému stavu a implementaci v reálném světě. Nejprve nastíním historické kořeny této problematiky, tedy zrod myšlenky umělé inteligence a námitky proti této ideji. Poté stručně popíšu technickou stránku systémů umělé inteligence a aktuální úspěchy na tomto poli. V závěru se pokusím zhodnotit, zda se nynější směr technologického vývoje skutečně blíží k vytvoření umělé mysli. Popř. zda je tato snaha vůbec splnitelná.

V první kapitole se tedy budu zabývat počátky historické analogie mezi lidským a strojovým myšlením. Budu se orientovat zejména na počátek ideje myslících strojů, ve které je přisuzování inteligence podmíněno schopností napodobit lidské projevy myšlení. V této části představím názory Alana Turinga i jeho odpůrců (imitační hra, čínský pokoj atd.), které jsou nezbytné pro pochopení dalšího vývoje umělé inteligence.

Ve druhé kapitole budu řešit základní charakteristiky nejvýznamnějších systémů umělé inteligence. V krátkosti nastíním vývoj technologií vedoucí až k dnešnímu stavu umělé inteligence a poukážu na zajímavé pokroky v současném výzkumu. Obrátím se také na porovnání odlišností a podobností při lidském a počítačovém zpracovávání informací, učení, apod.

V poslední kapitole se budu zaobírat otázkou myšlení jako takového. Zaměřím se především na problematiku redukcionismu a formalismu ve spojitosti s lidským myšlením. Na základě poznatků z předchozích částí se přitom pokusím zhodnotit, zda je možné sestrojit umělou mysl pomocí simulace procesů v lidském mozku, nebo dodáním základního souboru pravidel a obsahů, které bychom převzali z lidské mysli.

Cílem mé práce je, na základě poznatků o mysli a technologickém pokroku, zanalyzovat možnost existence mechanické mysli a zhodnotit adekvátnost analogie mezi mechanickou a lidskou myslí a její užitečnost pro další vývoj umělé inteligence.

1 Letmý pohled na historii ideje myslících strojů

Lidstvo odedávna sužují otázky o vlastní podstatě. Nejčastěji se ptáme: Co je to, co nás činí výjimečnými? Je to snad mysl? Ale co vlastně znamená myšlení? A je možné popsat myšlení v rámci zákonů, které jej řídí? Dodnes není jasné, jak přesně myšlení funguje a zda vůbec existuje nějaký soubor jasných pravidel, který by vymezoval činnost mysli. Víra v plnou formalizaci přesto ovládá západní myšlení. Už Hoobes pojímal myšlení jako kalkul; myšlení dle něj není nic jiného než počítání. Dle Leibnize lze konceptům přiřadit charakteristická čísla a všechny znalosti by tak mohly být vyjádřeny v deduktivním systému. A na základě těchto čísel a pravidel jejich kombinace by šly vyřešit veškeré problémy. Podobně jako Hobbes i Boole na počátku devatenáctého století předpokládal, že uvažování je výpočtem, a chtěl nalézt základní pravidla těchto operací, aby je poté vyjádřil v symbolickém jazyce. A díky Charlesu Babbage začala praxe hbitě dohánět teorii, když navrhl první digitální stroj. V době, ve které Babbage žil, však nebyly prostředky na sestrojení prototypu. Turing posléze charakterizoval digitální počítač jako diskrétní stroj, který se pohybuje z jednoho stavu do druhého, přičemž přechod mezi stavy není plynulý, ale probíhá ve skocích.[1]

Digitální počítač byl na počátku určen k provádění výpočetních operací. Později se ale zrodila myšlenka, že by měl být schopen provádět veškeré operace, které je schopen provádět člověk. Základní ideou bylo, že existuje pouze jediný typ univerzálního výpočetního stroje a „každý výpočet, který lze provést na jakémkoli fyzickém výpočetním stroji, lze provést na každém univerzálním počítači, pokud má dostatek času a paměti. To je pozoruhodný poznatek, ze kterého vlastně plyne, že správně naprogramovaný univerzální počítač by měl být schopen simulovat funkci lidského mozku.“[2] To se shoduje s Turingovou

[1] DREYFUS, H. L. *What computers still can't do: a critique of artificial reason*. Introduction
[2] HILLIS, W. D. *Vzor v kameni: jednoduché myšlenky, které řídí počítače*. s. 70

představou, že všechny digitální počítače jsou v jistém ohledu ekvivalentní. Liší se pouze kapacitou paměti a rychlostí.[3]

Objev univerzálních strojů mimo jiné znamenal formalizaci různorodých procesů. Konečně zde byl stroj řídící se pravidly, bez nutnosti lidské intuice či usuzování. Ale mohou stroje skutečně myslet? Za účelem získání odpovědi na tuto otázku byl vytvořen Turingův test. Nicméně stroje nebyly ani přes svou univerzalitu a rychlost schopné splnit požadavky Turingova testu a prokázat tak svou inteligenci.[4]

Během padesátých let se zkoumání strojové inteligence rozštěpilo na tři hlavní směry:

1. Hledání základních principů, které se transformovalo do nalezení podstaty samo-organizačních systémů. Jedná se o snahu nalézt sbírky komponent, které by se ve vhodném prostředí a za vhodného uspořádání mohly chovat adaptivně. V tomto pojetí je inteligence důsledkem evoluce samotného systému (např. Frank Rosenblatt);
2. Snaha stvořit pracovní model lidského chování, což vyžaduje, aby chování stroje odpovídalo chování člověka. Jedná se o kognitivní simulaci (např. Newell a Simons);
3. Přístup umělé inteligence, aneb snaha zkonstruovat inteligentní stroje, které by nebyly závislé na předsudku, že stroj musí být humanoidní nebo musí mít biologickou strukturu, aby mohl být považován za inteligentní.[5]

[3] TURING, A. M. *Computing machinery and intelligence*. s. 442
[4] DREYFUS, H. L. *What computers still can't do: a critique of artificial reason*. Introduction
[5] Tamtéž, s. 42-43

1.1 Námitky proti umělé inteligenci

Sám Turing ve svém článku „Computing Machinery and Intelligence“[6] zmiňuje nejčastější námitky své doby, proti ideji myslících strojů, z nichž uvedu ty nejzajímavější:

Pravděpodobně nejvíce by nás v tomto ohledu měla znepokojovat Gödelova věta, kterou lze vztáhnout na stroje typu digitální počítač s nekonečnou kapacitou. Tato věta dokazuje, že v dostatečně silných konzistentních logických systémech můžeme formulovat tvrzení, která nemohou být potvrzena nebo vyvrácena v rámci systému, v němž jsou formulována. Turing uvádí, že ačkoliv žádný jednotlivý stroj nemá neomezený výkon (není schopen zodpovědět všechny problematické otázky), bylo pouze stanoveno, bez jakéhokoliv důkazu, že lidský intelekt není také podroben stejnému omezení. [7]

„Navíc čistě logická stránka problému spočívá v tom, že k sebedokonalejšímu (konkrétnímu) stroji vždy existuje zodpověditelná otázka, na kterou tento stroj již nemůže odpovědět (dotyčnou otázku umíme vždy efektivně zkonstruovat metodou užitou v Gödelově důkazu), ale současně ke každé takové otázce existuje ještě o něco dokonalejší stroj, který ji už zase vyřešit může (stačí do něj zabudovat naši odpověď). Vznikají tak dvě nekonečné hierarchie: stále dokonalejších strojů a stále složitějších problémů.“[8]

Přestože stroje mohou vykonávat mnoho věcí, často se vyskytuje argument z různých neschopností strojů, který obecně zní: Stroje mohou provádět zdánlivě inteligentní úkoly, a přesto nemohou být X. A pod X se nejčastěji skrývá: přátelské, společenské, krásné, chybující, učenlivé, se smyslem pro humor atd. Turing věří, že tento argument je založen na předchozí zkušenosti s různými stroji a mnohé z těchto omezení jsou spojeny pouze s nízkým výkonem strojů. Nařčení, že stroje neumějí dělat chyby, v nás naopak vyvolává otázku *a jsou pro to o něco*

[6] TURING, A.M. (1950). *Computing machinery and intelligence*. Mind, 59, 433-460
[7] Tamtéž, s. 444-445
[8] HAVEL, I. *Přirozené a umělé myšlení jako filosofický problem. s.* 14

horší? Není to spíš náš nedostatek? Dle Turinga závisí tato námitka na zmatení mezi dvěma druhy chyb: chyby fungování a chyby v závěrech. První jsou způsobeny mechanickou poruchou, kvůli které zařízení nepracuje tak, jak bylo navrženo. V tomto smyslu stroje opravdu nejsou schopny dělat chyby, pokud fungují tak jak mají. Chyby v závěrech mohou nastat pouze tehdy, když je k výstupním signálům stroje připojen nějaký význam. Tento druh omylů jistě stroje dělat mohou.[9]

Další námitka proti umělé inteligenci tkví v přesvědčení, že stroj není schopen něco vytvořit. Dělá jen to, co mu přikážeme. Dle Hartreeho to ale neznamená, že bychom v budoucnu nemohli vytvořit stroj, který by mohl myslet sám od sebe, nebo v němž by se vyvinul podmíněný reflex sloužící jako základ pro učení. Turing souhlasí a dodává, že při pokusech napodobit lidskou mysl se musíme ptát, co ji přivedlo do stavu, v jakém je. Na základě toho předkládá svou představu mysli, skládající se ze tří složek: Počáteční stav mysli při narození; vzdělání, kterým prošla; další zkušenosti mimo vzdělání. Patrně by tedy bylo lepší, pokusit se místo simulace mysli dospělého člověka simulovat mysl dítěte, kterou bychom následně podrobili speciálnímu vzdělávacímu procesu.[10]

V neposlední řadě je třeba zmínit obecné přesvědčení, že není možné vytvořit soubor pravidel vymezujících vše, co by měl člověk dělat ve všech myslitelných situacích. S tím Turing souhlasí. Nicméně je rozdíl mezi pravidly chování a zákony chování (což jsou v podstatě zákony přírody aplikované na lidská těla). V tom druhém smyslu se dá jistě říci, že takový konečný soubor zákonů existuje a vymezuje naše možné jednání.[11]

Searle proti myšlení strojů namítá, že veškeré operace digitálního počítače lze formálně popsat, jednotlivé kroky např. specifikujeme jako operace s abstraktními symboly, které však nemají význam ani sémantický obsah, jelikož se k ničemu nevztahují. To zároveň podkládá zásadní chybu v analogii mezi

[9] TURING, A.M. *Computing machinery and intelligence*. s. 447-450
[10] HAUGELAND, J. *What Is Mind Design?* s. 46
[11] TURING, A.M. *Computing machinery and intelligence*. s. 452-453

programovými procesy a těmi mentálními. Mysl totiž obsahuje více než pouhé formální či syntaktické procesy. Mentální procesy mají navíc určité obsahy. A to znamená, že každý můj mentální stav má mimo formálních vlastností i nějaký psychický obsah. Mé myšlenky jsou „o něčem“, ergo mají význam. „Jedním slovem: Mysl má víc než jen syntax, má i sémantiku. A žádný počítačový program nikdy nebude mít mysl prostě proto, že je pouze syntaktický, zatímco mysl není jen syntaktická. Mysl je sémantická – sémantická v tom smyslu, že má víc než formální strukturu: má také obsah.“[12]

1.2 Turingův test

Turing předložil ideu, podle které je inteligence otázkou chování. Zaměřil se přitom na specifický druh chování – verbální chování. Systém je dle Turinga inteligentní, pokud může vést konverzaci jako běžný člověk. Proč ale vybral schopnost komunikace za známku inteligence? Komunikace je jedinečná, jelikož můžeme mluvit o většině aktivit závisejících na intelektu, vyjádřit je.[13]

Je důležité rozlišit, co napsal ve svém článku ohledně Turingova Testu (TT) sám Turing a co bylo přidáno časem. Turingovým cílem bylo vytvořit metodu, jakou by bylo možné zodpovědět otázku, zda mohou stroje myslet. Taková otázka však byla nejednoznačná, proto se ji ve svém článku pokusil přeměnit do konkrétnější podoby navržením imitační hry (IG).[14]

Imitační hra obsahuje jednoho muže, jednu ženu a vyšetřovatele, který je od nich oddělen. Cílem vyšetřovatele je určit, kdo z nich je žena, zatímco úkolem obou testovaných je přesvědčit vyšetřovatele, že právě on je ženou. Prostředkem přesvědčování a klamání je jen písemné spojení skrze přirozený jazyk. Dle Turinga bychom měli diskutovat místo nejasné otázky *mohou stroje myslet?* otázku *co se stane, když se stroj bude podílet na této hře?* Bude se vyšetřovatel rozhodovat

[12] SEARLE, J. R. *Mysl, mozek a věda*. s. 33
[13] HAUGELAND, J. *What Is Mind Design?* s. 3-4
[14] SAYGIN, A. P., CICEKLI, I. a AKMAN, V. *Turing Test: 50 Years Later*. s. 464

špatně stejně často, jako když před ním sedí muž a žena? Otázka, *zda stroje mohou myslet?* je tak nahrazena následující: „Je pravdou, že kdybychom modifikovali počítač tak aby měl adekvátní úložiště, zvýšili jeho rychlost a vybavili ho vhodným programem, pak by mohl uspokojivě nahradit stranu A v imitační hře, zatímco by na straně B byl muž?“[15] Ačkoliv žena zmizela, cíl stran A a B je stejný. Turing se tak snažil posoudit schopnost stroje napodobit lidskou bytost. Proč ale ponechal muže místo ženy a zamotal tak TT? Myšlenka stojící za tímto tahem je jednoduchá –imitační hra je o klamání – muž i stroj se pokouší přesvědčit vyšetřovatele, že jsou ženou, zatímco žena by neměla jak klamat. Turing se tedy snažil porovnat úspěch stroje v tom, zda dokáže imitovat ženu lépe než muž. Základní myšlenka, která je při realizacích TT často opomíjena, tkví ve snaze obou stran simulovat něco, čím nejsou. To mělo zajistit jistou metodologickou spravedlnost, která není v dnešních (genderově nezávislých) testech přítomná. Tato zdánlivá nejasnost tak poskytuje férový základ pro srovnání: žena působí jako neutrální bod (jakožto koncept), aby mohli být oba soupeři posuzování dle toho, jak dobře předstírají.[16]

Turing tedy nahradil otázku *Mohou stroje myslet?* za otázku *Mohou stroje hrát imitační hru?* Zdá se, že pokud postavíme počítač schopný vyhrát IG, nezáleží na tom, zda jsou jeho účely podobné lidským. Nicméně Turing uvádí, že nejlepší strategií je pokusit se poskytnou odpovědi, které by přirozeně dal člověk.[17] Hra nemá v podstatě žádná pravidla omezující konstrukci strojů. Stroje mohou dělat náhodné chyby a tím klamat – přestože za známku inteligence se obecně považuje schopnost dát správnou odpověď. Proto může TT v jistých ohledech působit více jako test lidskosti než inteligence.[18]

TT nebyl nikdy proveden přesně jak ho Turing popsal, existuje mnoho alternativních variant originálu, ve kterých programy dokazují svou „lidskost“. Nejznámější ze zrealizovaných variant TT je Loebnerova cena. Vítěz v ní musí projít neomezeným TT a dokázat tak svou lidskost. Dříve byla témata omezená,

[15] TURING, A. M. *Computing machinery and intelligence*. s. 442
[16] SAYGIN, A. P., CICEKLI, I. a AKMAN, V. *Turing Test: 50 Years Later*. s. 465-467
[17] TURING, A. M. *Computing machinery and intelligence*. s. 435
[18] SAYGIN, A. P., CICEKLI, I. a AKMAN, V. *Turing Test: 50 Years Later*. s. 468

protože nikdo nevěřil, že by programy dokázaly splnit původně definovaný TT, ale od roku 1995 se dotazuje na jakékoliv téma. Omezení TT by dle Turinga nedostačovalo k prokázání inteligence. V Loebnerově ceně je ignorována genderová složka. Cílem soutěžících je jednoduše přesvědčit soudce, že jsou lidmi. I když lze stále rozeznat počítač od člověka, v roce 1991 se stalo, že byl člověk považován za stroj (kvůli takřka neuvěřitelné znalosti Shakespeara).[19] V roce 2014 naopak prošel minimální hranicí 30% úspěšnosti program předstírající třináctiletého chlapce Eugene Goostmana, který se tak stal prvním strojem, jemuž se podařilo projít minimální teoretickou hranicí pro splnění TT.[20]

1.3 Námitky proti TT a alternativní testy inteligence

Pro mnohé je Turingův test problematický „vychází totiž z čistě behavioristického pojetí myšlení (charakteristického pro tehdejší psychologii) a nebere v potaz procesy uvnitř systému, tím méně fenomenální aspekty mysli.“ A „testuje se vlastně něco velmi speciálního, totiž umění stroje předstírat, že není stroj.“[21]

Jednou z nejznámějších námitek proti adekvátnosti TT jakožto testu inteligence je Searlův argument čínského pokoje. Ten je následující: John je zamčen v místnosti, neumí ani slovo čínsky, ale má k dispozici komplexní soubor pravidel, který mu umožňuje manipulovat se znaky tak, že výsledkem je vhodná odpověď na znaky zvenčí. Vypadá to, že dokonale chápe čínštinu, ale přitom jí vůbec nerozumí. Nezná významy symbolů. Searlova myšlenka je, že „pouze manipulací formálními znaky byste se nikdy nemohli naučit čínsky“ i když se „…při realizace formálního počítačového programu z hlediska vnějších pozorovatelů chováte tak, jako byste čínsky rozuměli.“[22]

[19] SAYGIN, A. P., CICEKLI, I. a AKMAN, V. *Turing Test: 50 Years Later*. s. 501
[20] http://www.reading.ac.uk/news-and-events/releases/PR583836.aspx
[21] HAVEL, I. *Přirozené a umělé myšlení jako filosofický problem. s.* 20
[22] SEARLE, J. R. *Mysl, mozek a věda*. s. 34

Searlův argument zní přesvědčivě, ale „je třeba si uvědomit dva hlubší rozdíly mezi Johnem a počítačem. Za prvé, John rozumí manuálu (v angličtině) a uvědomuje si, co má dělat, tj. manipulovat s nesrozumitelnými symboly, možná to dokonce chce dělat (aby vyhověl Searlovi, svému mysliteli). Toto rozumění, uvědomování si a případné chtění se neodvažujeme předpokládat u centrálního procesoru počítače, od něhož čekáme spíše jen kauzální chování poslušné zákonům fyziky.“[23] A navíc, jak získal Searle jistotu, že počítač skutečně nerozumí čínským znakům – na rozdíl od něj? Dle Dennetta, který argumentuje proti tomuto příkladu, Searle v příkladu s čínským pokojem bagatelizuje nároky na počítačový systém. „Takový systém by totiž musel mít tak rozsáhlou datovou bázi s tolika vnitřními vazbami, že by mohlo dojít k emergentním jevům na vyšších úrovních popisu. Pak by bylo, dle Dennetta, legitimní mluvit o pravém porozumění u počítače, třeba i takového, jehož centrální procesor je simulován nic netušícím Johnem.“[24]

Za zmínku stojí rovněž fakt, že John je v příkladu čínského pokoje stále myslící bytostí, která rozumí anglickým příkazům a díky tomuto porozumění (a vhodným nástrojům) je schopen adekvátně odpovídat, přestože nerozumí čínským znakům. Zdá se, že jisté porozumění je tedy vždy nutné. Ale čemu, pokud vůbec něčemu, rozumí stroj řídící se pravidly? V analogii na tuto situaci můžeme rovněž počítač v TT chápat jako inteligentní stroj rozumějící svému úkolu, avšak bez porozumění lidskému myšlení.

Keith Gunderson v článku „The Imitation Game“[25] poukázal na zásadní body týkající se Turingovo nahrazení otázky *mohou stroje myslet?* za schopnost hrát IG. Zdůrazňuje dva body: Hrát úspěšně IG je cíl, který může být dosažen různými způsoby, i bez ovládání inteligence; a myšlení je obecný pojem, zatímco hraní IG je jen jeden příklad věcí, které inteligentní entity dělají. Obě poznámky kritizují IG jako neplatné měřítko inteligence. Dle něj je součástí toho co bytosti

[23] HAVEL, I. *Přirozené a umělé myšlení jako filosofický problem. s.* 21
[24] Tamtéž, s. 22
[25] GUNDERSON, K. The Imitation Game. *Mind*. 1964, (73), 234–245

dělají i to, jak to dělají. Úspěch v IG tudíž není dostatečný k prokázání inteligence. Myšlení je komplexní ale IG dokazuje jen jednu omezenou část. [26]

Richard Purtill kritizuje IG jakožto vědeckou fikci, protože v dohledné době a s aktuálně představitelnými programovacími technikami je nepředstavitelné postavit počítač schopný hrát IG. Dle Purtilla není chování myslících bytostí deterministické a nemůže být vysvětleno v čistě mechanických termínech. IG je dle něj jen bitvou mezi vyšetřovatelem a programátorem – počítač je v podstatě jen prostředním článkem a může být nahrazen čímkoliv nesofistikovaným.[27]

P.H. Millar se naopak ptá, zda je IG správným posouzením inteligence. Poznamenává, že nás nutí antropomorfizovat stroje tím, že jim přisuzujeme lidské cíle a kulturní pozadí. Namítá, že IG neměří, zda mají stroje inteligenci, ale zda mají lidskou inteligenci. Millar naproti tomu věří, že bychom měli být otevření k tomu, umožnit každé bytosti (ať už marťanovi nebo stroji) projevovat inteligenci skrze prostředky chování, které jsou přizpůsobené pro dosahování jejich vlastních specifických cílů.[28]

Ned Block v „Psychologism and Behaviorism“[29] atakoval TT jako behavioristický přístup k inteligenci. Block věří, že rozhodčí může být oklamán neinteligentními stroji za pomocí triků. Block definuje psychologismus jako „doktrínu, že to, zda je chování inteligentní, závisí na charakteru vnitřního zpracování informací, které jej produkuje“[30]. Dle této definice mohou dva systémy vykazovat stejné skutečné i potencionální chování, stejné behaviorální vlastnosti, dispozice atd. ale může být rozdíl v jejich zpracování informací, který nás vede k přisouzení inteligence jednomu, zatímco druhému ne. Představme si např., že na Marsu objevíme marťany, začneme s nimi rozvíjet vzájemné porozumění, zapojí se do naší tvůrčí činnosti apod. a posléze zjistíme, že mají jiný systém zpracování informací než my. Znamená to, že bychom měli popřít jejich inteligenci, jen

[26] GUNDERSON, K. *The Imitation Game*. s. 238
[27] PURTILL, R.L. Beating the Imitation Game. s. 291-293
[28] MILLAR, P.H. On the Point of the Imitation. s. 597
[29] BLOCK, N. Psychologism and Behaviorism. *Philosophical Review.* 1981, (90), s. 5–43
[30] Tamtéž, s. 5

protože jsou od nás odlišní? To by bylo šovinistické. Přestože tento argument může působit jako argument proti psychologismu, Block dodává, že psychologismus ve skutečnosti nezahrnuje tento druh šovinismu; neuvádí totiž, že pokud má někdo rozdílný druh zpracování informací od našeho, nutně postrádá inteligenci. Důraz na zpracování informací při posuzování inteligence neznamená, že jediné vnitřní zpracování informací produkující inteligenci je to naše.[31]

Dle Blocka by byl inteligentní ten stroj, který by se mohl učit a řešit problémy; pak by inteligence skutečně patřila stroji, nikoliv programátorům. Ale pokud stroj odpovídá pouze tím, co bylo předem vloženo do jeho paměti, je to jen triviální vyhledávací stroj. Problém tkví v tom, že TT nerozlišuje mezi chováním, které odráží inteligenci stroje a tím, které odráží inteligenci programátorů. Pokud je TT tak nesofistikovaný, potom bychom dle Blocka měli revidovat náš mechanismus přisuzování inteligence tak, abychom se vyhnuli příliš liberálnímu i šovinistickému připisování inteligence.[32]

Stevan Harnad navrhuje tzv. Totální Turingův test (TTT), který vyžaduje, aby stroje reagovaly na všechny vstupy, nejen slovní. Splnit by to mohl senzomotorický robot. Jeho kritika TT se zakládá na problému týkající se základu symbolů. Harnad přirovnává situaci, kdy jsou symboly definovány prostřednictvím jiných symbolů, k začarovanému kruhu v čínsko-čínském slovníku. Tvrdí, že aby v mysli mohla být nějaká sémantika (a že tu jistě je), musí být symboly něčím podloženy. Význam symbolů je, dle Harnada, alespoň částečně odvozen z interakcí s vnějším světem. Jazyk tudíž nemůže být samostatným modulem a k jeho správnému užití je třeba interagovat s vnějším světem.[33]

John Barresi pro změnu navrhuje, abychom považovali inteligentní stroje za druh a vytváří evoluční „Cyberiad Test“ namísto TT, jelikož v TT se stroj snaží oklamat člověka, ale u přirozené inteligence je takovým rozhodčím matka příroda. Cyberiad Test je nicméně podobný TT v tom, že základem rozsudku je srovnání

[31] SAYGIN, A. P., CICEKLI, I. a AKMAN, V. *Turing Test: 50 Years Later*. s. 480-481
[32] Tamtéž, s. 482-484
[33] Tamtéž, s. 486-487

mezi lidmi a stroji. Rozdíl mezi nimi spočívá v tom, jak je definovaná inteligence. Barresi popisuje inteligentní chování jako takové, které je nezbytné pro přežití společnosti. Cyberiad test je proto splněn pouze „pokud je společnost umělých lidí schopna pokračovat v jejich vlastní socio-kulturní evoluci bez rozpadu po dlouhou dobu, řekněme po několik milionů let.“[34]

Jednu z nejzajímavějších poznámek k TT však sepsal Robert French ve svém článku „Subcognition and the Limits of the Turing Test“[35]. French zde poukázat na fakt, že TT poskytuje záruku nejen inteligence, ale i kulturně orientované inteligence. Proto věří, že TT je prakticky nepoužitelný jako reálný test inteligence, protože nebude nikdy splněn. K tomuto závěru došel na základě role, kterou hrají subkognitivní otázky v TT, tedy otázky, které se odkazují na naše prožívání světa. Tvrdí totiž, že každý dostatečně široký soubor otázek pro TT bude obsahovat subkognitivní otázky, i když to není jeho záměrem. A skutečnost, že kognitivní a subkognitivní úrovně jsou propojeny, dále ukazuje, že TT je v podstatě zkouškou lidské inteligence, ne inteligence obecně. French pro vysvětlení své teze podává analogii s testem racků: Uvažme ostrov, na kterém jsou jediné létací bytosti racci. Jednoho dne dva filosofové probírají esenci létání. Jeden z nich navrhne, že létání je pohybování se ve vzduchu. Druhý mrští oblázkem a namítne, že oblázek jistě nelétá. První řekne, že objekt musí zůstat ve vzduchu po nějakou časovou periodu, aby se to počítalo jako létání. Ale v tom případě mraky, kouř atd. jsou létajícími entitami, říká druhý. Napříč debatou se shodnou pouze na tom, že jedinými létajícími objekty, který znají, jsou rackové. Potom ve světle Turingova článku navrhnou Raccí test pro létání. A věří, že když něco projde tímto testem, tak to létá. Když ne, nemohou o tom rozhodnout. Objekt projde testem, jen když bude nerozlišitelný od racka. Argumenty těchto dvou filosofů připomínají argumenty ohledně povahy inteligence. Raccí test nepropustí ani letadla, ani vrtulníky nebo netopýry a brouky. Pravděpodobně neprojde nic kromě racků z ostrova filosofů. Pak nejde o zkoušku letu jako takovou, ale o zkoušku letu, jak

[34] BARRESI, J. Prospects for the Cyberiad: Certain Limits on Human Self-Knowledge in the Cybernetic Age. s. 23
[35] FRENCH, R. Subcognition and the Limits of the Turing Test, *Mind.* 1990, 99(393), s. 53–65

ho praktikují racci z daného ostrova. Stejně tak TT je zkouškou inteligence, jakou praktikuje člověk. Naše odpovědi v testu jsou závislé na našem prožívání světa jako lidské bytosti v určitém sociálním a kulturním nastavení. French proto předpokládá, že fyzická úroveň a kognitivní úroveň inteligence jsou neoddělitelné. Subkognitivní otázky přitom umožňují porovnání asociativních pojmových sítí u obou kandidátů. A protože se tyto sítě formují po celý život a jejich struktura a povaha nutně závisí na fyzických aspektech zkušenosti, počítač bude vždy rozlišitelný od člověka. Ve zkratce, pro počítač (nebo jakéhokoli ne-člověka) není možné být úspěšný v IG. To, že nepoznávají svět tak jako my, není jen překážkou, ale přísným omezením v tomto úkolu. Nechce tím ale naznačit, že lidský subkognitivní substrát je nezbytný pro inteligenci obecně; je nezbytný pouze k úspěchu v TT a právě proto je TT neadekvátní.[36]

Konečně Patric Hayes a Kenneth Ford tvrdí, že bychom měli odmítnout splnění TT jako cíl pro umělou inteligenci. Věří totiž, že snaha uspět v TT nás odvádí od skutečně užitečného výzkumu umělé inteligence, protože konečným cílem umělé inteligence by nemělo být napodobování lidských schopností.[37] Zpochybňují tak praktickou využitelnost TT. Proč bychom se měli snažit postavit stroj, co nás napodobuje? Je to jako lidské pokusy létat na základě imitace přirozeného letu. Umělá inteligence se může radikálně lišit od té naší, stejně jako umělá letadla mohou létat jinak než ptáci. Domnívají se proto, že umělá inteligence by měla produkovat artefakty, ne nutně podobné lidským, ale spíše užitečné pro lidi.[38]

1.4 Jazyky, které řídí stroje

[36] SAYGIN, A. P., CICEKLI, I. a AKMAN, V. *Turing Test: 50 Years Later*. s. 492-495
[37] HAYES, P. and FORD, K. Turing Test Considered Harmful. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*. 1995 Vol. 1, s. 972-973
[38] SAYGIN, A. P., CICEKLI, I. a AKMAN, V. *Turing Test: 50 Years Later*. s. 496

„Dvě věci jsou jasné: chceme-li, aby stroj něco dělal, musíme mu to říct a stroj musí být schopen rozumět, co mu říkáme.“[39] Stroji běžně dáme program (který má vykonat to co požadujeme) a data. Programy ale mění počítače na velmi úzce zaměřené. Oproti tomu lidé jsou všestranní a schopní porozumět přirozenému jazyku, který postrádá přesnost, jež se vyskytuje v programovacích jazycích a kterou stroje vyžadují.[40]

Přesto tkví „kouzlo počítače v jeho schopnosti stát se čímkoli, co si umíme představit, pokud dokážeme přesně vysvětlit, co máme na mysli. Potíž je v popsání toho, co chceme.“[41] Za tímto účelem používáme programovací jazyky, které jsou univerzální v tom smyslu, že jimi lze naprogramovat jakýkoliv úkon. „Stejně jako lidská řeč má programovací jazyk svou slovní zásobu a gramatiku, ale na rozdíl od lidského jazyka má v programovacím jazyce každé slovo a každá věta svůj přesný význam.“[42]

Proto je k naprogramování úkonů potřeba přesně říci počítači co od něj chci, do nejmenších detailů. Programátor a počítač musí mít do jisté míry stejné znalosti (např. znalost slovní zásoby či gramatiky daného jazyka), aby spolu byli schopní komunikovat. Jazyků existuje více, každý se hodí k něčemu jinému a každý má odlišnou syntax. Programovací jazyky ve zkratce vytyčují manipulaci s daty. Když počítači definujeme nový výraz, zařadí ho do slovní zásoby a můžeme ho poté využívat. Počítač je tedy specifickým typem konečného automatu, propojeným s pamětí. Paměť je přitom souhrnem registrů, obsahující řetězce bitů.[43] „Některá slova uložená v počítači reprezentují data, se kterými se bude pracovat, například čísla a písmena. Jiná reprezentují instrukce, které říkají stroji, jakou posloupnost operací má vykonat.“[44]

[39] WEIZENBAUM, J. *Mýtus počítače: počítačový pohled na svět*. s. 54
[40] Tamtéž, s. 54
[41] HILLIS, W. D. *Vzor v kameni: jednoduché myšlenky, které řídí počítače*. s. 50
[42] Tamtéž, s. 50
[43] Tamtéž, s. 51-60
[44] Tamtéž, s. 61

1.5 Lidské vnímání strojů

Témat ohledně lidského vnímání strojů, které v nás vyvolávají otázky, je mnoho. Můžeme se např. ptát co je na počítačích, ergo umělé inteligenci takového, že to v nás vyvolává představy připodobňující člověka ke stroji? A kde je linie dělící lidskou a strojovou inteligenci? Je snad takové připodobnění způsobeno tím, že si lidé ke strojům vytváří citová pouta, díky kterým své výtvory antropomorfizují? To, že si vytváříme citová pouta k počítačům není samo o sobě překvapující, jelikož počítače jsou našimi výtvory a zároveň nástroji ulehčujícími náš život. A taková pouta jsou navíc logicky silnější u nástrojů, které přímo souvisejí s intelektuální, emoční anebo kognitivní činností. Ale nezvyšuje toto připodobnění, spolu s masivním zavedením počítačů do společnosti, nátlak, který člověka vede ke stále mechanističtějšímu obrazu sebe sama? A pokud v rámci zavádění počítačů do společnosti umožníme strojům vykonávat veškeré lidské aktivity, znamená to snad, že budou schopné obsáhnout i lidské myšlení?[45]

Okolo funkcionality počítačů obecně vznikají debaty se dvěma stranami názorů. Na jedné straně stojí příznivci toho, že by počítače měly a časem budou dělat vše. Počítač je v tomto pojetí chápán jako „stroj, který zrychluje a rozšiřuje naše myšlenkové pochody. Je to obrazotvorný stroj, který uchopí myšlenky… a vede je mnohem dál, než bychom byli schopni my sami.“[46] A na druhé straně stojí ti, kteří věří, že počítačové úkony mají své hranice a stroje nejsou schopné obsáhnout veškerou lidskou aktivitu.[47]

První strana tohoto sporu je přesvědčena, že „každý efektivní postup může být uskutečněn počítačem. Protože člověk, příroda a dokonce společnost uskutečňují postupy, které jsou jistě tím či oním způsobem „efektivní“, plyne z toho, že počítač může přinejmenším imitovat člověka, přírodu a společnost ve všech procedurálních aspektech.“[48] Tento názor si v sedmdesátých letech

[45] WEIZENBAUM, J. *Mýtus počítače: počítačový pohled na svět*. s. 15-17
[46] HILLIS, W. D. *Vzor v kameni: jednoduché myšlenky, které řídí počítače*. s. 14
[47] WEIZENBAUM, J. *Mýtus počítače: počítačový pohled na svět*. s. 18
[48] Tamtéž, s. 25-26

vybudoval takovou základnu příznivců, že „mnozí psychologové dospěli k předpokladu … že lidé a počítače jsou pouze dva odlišné druhy mnohem abstraktnějšího rodu zvaného „systémy na zpracování informací“.“[49] Dreyfus dodává, že základem tohoto optimismu (který provází především výzkumníky umělé inteligence) je přesvědčení, že zpracování informací u lidí musí probíhat na základě diskrétních kroků, jaké jsou i u digitálních počítačů. Jedná se tedy o přesvědčení, že lidské i počítačové zpracování informací zahrnuje stejné elementární procesy. A tak by vhodné programování mělo být schopné vyvolat u počítačů projevy inteligence.[50]

Stroj je z tohoto pohledu vnímán jako naprogramovaný organismus, nejen hardware. Už jej nelze vnímat jako černou skříňku, ale artefakt hardwarově i softwarově navržený tak, abychom se do něj mohli podívat a spojit jeho konstrukci s chováním.[51] Searle by ale pravděpodobně podotkl, že ačkoliv strojům metaforou často přisuzujeme porozumění a další kognitivní funkce, které nás vedou k přesvědčení, že stroj může být inteligentní, není to ničím podloženo. Věřil totiž, že naprogramovaný počítač, na rozdíl od člověka, nerozumí vůbec ničemu – pouze na něj rozšiřujeme vlastní intencionalitu.[52]

Umělá inteligence sice vznikla s cílem replikovat lidskou inteligenci, ale o replikaci celé škály inteligence se prozatím bohužel nedá hovořit, pokrok se týká spíše specializovaných oblastí. Inteligence na lidské úrovni je pro nás zatím příliš složitá a nedostatečně prozkoumaná k tomu, aby ji bylo možné rozložit na dílčí části a replikovat. Proto by bylo patrně lepší zaujmout jiný přístup k vytváření inteligence: a sice vytvářet jednotlivé schopnosti a postupně tak vybudovat kompletní inteligentní systémy, které poté uvolníme do reálného světa, který budou vnímat, učit se z něj a na základě toho i jednat.[53]

[49] MILLER, G. A. *Language, Learning and Models of the Mind*, nepublikovaný rukopis, červen 1972
[50] DREYFUS, H. L. *What computers still can't do: a critique of artificial reason*. s. 67
[51] NEWELL, A. a SIMON, H. A. Computer Science as Empirical Inquiry: Symbols and Search. s. 81-82
[52] SEARLE, J. R. Minds, Brains, and Programs. s. 188
[53] RODNEY, A. B. Intelligence without Representation. s. 395

2 Moderní technologie a vývoj strojového zpracovávání informací

Mnoho historických příběhů o vytváření myslících bytostí (umělé inteligence) se zakládá na jakémsi neurčitém procesu zrání. Před érou digitálních počítačů si málokdo dokázal představit, že by se myšlení v celé své komplexitě dalo rozložit do jednotlivých operací, které by probíhaly v jednom či více systémech. Mělo se za to, že pokud se vůbec dá inteligence uměle „vyrobit", pravděpodobně vznikne nějakým nenadálým procesem, v němž se komplexní chování vynoří v důsledku mnoha nepatrných interakcí. Toto pojetí nahrává představě, že je možné stvořit inteligenci, bez přesné znalosti průběhu takového nenadálého procesu nebo bez znalosti přesného fungování lidského myšlení. A z toho vychází otázka, která je aktuální dodnes: je možné vytvořit umělou inteligenci ještě před plným pochopením lidské inteligence?[54] Hillis je přesvědčen „že tvůrčí proces bude takový, ve kterém vše připravíme pro inteligenci, aby se pak vynořila ze složité série vzájemných působení, kterým podrobně nerozumíme… Nenavrhneme umělou inteligenci; jen nastolíme správné podmínky, za kterých se může inteligence vylíhnout. Největším úspěchem naší technologie může být vytvoření nástrojů, které nám umožní jít dále než technika – které nás nechají vytvořit více, než čemu rozumíme."[55]

Může být tímto nástrojem, který nám umožní vytvořit umělou inteligenci bez plného porozumění přirozené inteligence, počítač? Počítač je specifický tím, že dokáže simulovat systémy odlišné od něj samotného.[56] Může nám tak poskytnout prostor k modelování mentálních operací a přispět k dalšímu vývoji umělé inteligence. Hlavní myšlenkou počítačové vědy je, že můžeme použít jeden automatický formální systém k implementaci jiného; to je základem programování. Jeden systém implementuje jiný systém, když některé konfigurace znaků a pozic toho prvního mohou být považovány za znaky a pozice toho

[54] HILLIS, W. D. *Vzor v kameni: jednoduché myšlenky, které řídí počítače*. s. 134-135
[55] Tamtéž, s. 135
[56] RUMELHART, D. E. *The Architecture of Mind: A Connectionist Approach*. s. 206

druhého. Kdy je taková implementace možná v principu? Turing ukázal, že je možná vždy. Vytvořil ideu univerzálního stroje, který může implementovat jakýkoliv dobře definovaný automatický formální systém, a to za předpokladu že k tomu má dostatečnou paměť a čas. Každý běžný programovatelný stroj je tedy univerzálním strojem v Turingově smyslu.[57]

V rámci klasického pojetí se počítače považují za formální systémy mající set pravidel, dle kterých manipulují se znaky. Takové systémy mají tři základní znaky: jsou to systémy manipulující se znaky; jsou digitální; jsou nezávislé na médiu. Každý znak má svůj význam, musí něco symbolizovat. Pravidla ohledně toho, jak můžeme znaky manipulovat a co znaky znamenají, jsou přitom úzce spojena.[58]

Koncepce symbolického systému je, dle Newella, zásadní pro pochopení zpracovávání informací a dosažení inteligence, jelikož symboly leží v základech inteligentního jednání. Inteligenci systému obecně měříme schopností dosáhnout cílů, i přes proměnlivost a složitosti úlohy, nebo nepředvídatelnosti světa. Naše chápání požadavků na inteligentní systém je ale pomalé, protože neexistuje žádný „princip inteligence“. Nicméně klasická teorie umělé inteligence formuluje určité strukturální požadavky na inteligenci, jako např. schopnost ukládat symboly manipulovat s nimi.[59]

Searle ale oponuje Newellovu tvrzení, že podstatou duševna jsou operace fyzikálního symbolického systému. Tvrdí přitom, že veškeré pokusy přikládat umělým systémům duševno spočívá pouze v podobnosti jejich chování s naším. Dle Searla pouze nepodloženě předpokládáme, že když se systém chová dostatečně podobně jako člověk, má duševní stavy, které vyjadřuje skrze chování, a proto musí mít vnitřní mechanismus schopný produkovat takové stavy. Ale jakmile zjistíme, že chování robota je výsledkem formálního programu, odmítneme myšlenku umělé intencionality. Intencionalita je přitom dle Searla

[57] HAUGELAND, J. *What Is Mind Design?* s. 12.
[58] Tamtéž, s. 8-18
[59] NEWELL, A. a SIMON, H. A. *Computer Science as Empirical Inquiry: Symbols and Search*. s. 83

naprosto zásadní pro vysvětlení struktury lidského chování. Intencionální stavy jsou ty, které jsou na něco zaměřené. A mentální stavy musí být nutně intencionální, aby měly nějaký obsah, byly o něčem.[60]

Abychom byli schopni vytvořit umělou inteligenci, nesmíme tudíž vycházet pouze z takových projevů myšlení, které jsou tradičně voleny pro umělou imitaci. To znamená neomezovat se pouze na modelování. Určitě je k tomu potřeba zabývat se fungováním přirozené mysli, nejen jejími vnějšími projevy. Problém je, že při zkoumání lidské mysli narazíme na nejednoznačné termíny – co je vlastně mysl? Jak souvisí s fungováním mozku? A je možné jeho funkce implementovat do umělých systémů?[61]

Na základě optimismu v oblasti výzkumu umělé inteligence a faktu, že nedokážeme mysl (jakožto předobraz umělé inteligence) jednoznačně definovat, nebo odhalit její přesné fungování, vznikly čtyři druhy nepodložených předpokladů:

1. Biologický předpoklad – že na jisté úrovni své činnosti zpracovává mozek informace v diskrétních operacích skrze biologický ekvivalent přepínačů on/off;
2. Psychologický předpoklad – že mysl lze pojmout jako zařízení operující s bity informací dle formálních pravidel. Takže počítač může sloužit jako model mysli s bity namísto atomických vjemů a programy namísto pravidel;
3. Epistemologický předpoklad – že všechny poznatky mohou být formalizovány. Ergo vše, co lze pochopit lze i vyjádřit pomocí logických vztahů;
4. Ontologický předpoklad – že všechny důležité informace o světě a vše nezbytné pro vznik inteligentního chování, musí být analyzovatelné jako soubor situačně nezávislých prvků. Aneb že vše, co existuje, je souborem faktů, které jsou na sobě logicky nezávislé.

Žádný z těchto předpokladů však není při detailnějším zkoumání příliš přesvědčivý a ani po mnoha letech výzkumu umělé inteligence potvrzený.[62]

[60] SEARLE, J. R. *Minds, Brains, and Programs*. s. 195-196
[61] HAVEL, I. *Přirozené a umělé myšlení jako filosofický problém. s.* 9-10
[62] DREYFUS, H. L. *What computers still can't do: a critique of artificial reason*. s. 68

Proto musíme předpokládat, že mezi lidským myšlením a počítačovým zpracováním informací zřejmě nebude taková podobnost, jak předpokládají mnozí zastánci klasických digitálních počítačů. Největší rozdíl pravděpodobně spočívá v tom, že zatímco běžné digitální počítače jsou naprogramované pouze na konkrétní druhy činností, člověk je schopen provádět úkony takřka v neomezeném počtu oblastí. A přestože jsou počítače díky jejich úzkému zaměření schopné zvládnout úkony a množství dat jaké člověk nedokáže (přinejmenším ne tak rychle), zpracování úkonů běžného lidského života je stále problematické.[63] Místo toho, aby se experimenty umělé inteligence dále zaměřovaly na specifickou kognitivní schopnost, měl by se vývoj umělé inteligence zaměřit na propojení těchto schopností. Protože inteligence ze své podstaty obsahuje především všestrannost. Taková integrovaná umělá inteligence by za účelem všestranného vývoje měla mít kontakt s reálným světem. Další důležitou vlastností přirozené inteligence je schopnost vyvíjet se a osvojovat si nové dovednosti, které umí zobecnit. Snaha o dosažení těchto schopností byla hlavním podnětem k narůstající oblibě konekcionistických neuronových sítí.[64]

Jak již bylo řečeno, fakt že nejsme schopni popsat přesné fungování lidské mysli neznamená, že nemůžeme nastolit podmínky k samovolnému vzniku umělé mysli. Hlavní strategií ve vytváření umělé inteligence se proto v posledních desetiletích stává konekcionismus, který je založen na myšlence, že mnoho komplexních jevů (včetně myšlení) můžeme chápat jako emergentní vlastnosti paralelních dějů v síti, která je složena z mnoha prostých, vzájemně interagujících jednotek. Emergentní jev je přitom jakýkoliv jev, který je svébytný na vyšší úrovni a na který lze pohlížet jako na důsledek chování jednoduchých prvků z nižší úrovně, přestože jeho chování není možné v rámci nižší úrovně popsat. Konekcionismus předpokládá, že i přes naši omezenou znalost lidského mozku je zřejmé, že mozková tkáň může být chápána jako konekcionistický systém,

[63] CRICK, F. *Věda hledá duši: Překvapivá domněnka.* s. 182
[64] HAVEL, I. *Přirozené a umělé myšlení jako filosofický problém. s.* 4-5

s neurony jakožto základními jednotkami a synapsemi jako jejich společnými vazbami.[65]

2.1 Paralelní počítače

„Rychlost počítače je v podstatě určena časem, který je potřebný k přemístění dat do paměti a z paměti.“[66] Klasické sekvenční počítače provádí vždy jen jednu operaci. I přes mnohá vylepšení je jejich základní koncepce procesoru spojeného s pamětí stejná už od 50. let. „Charakter využití jednotlivých částí se stále řídí starým dvousložkovým návrhem. Část čipu zodpovědná za zpracování je velmi zaměstnaná; v paměťové části v jednom okamžiku stále protéká jen jedno slovo dat.“[67] A právě přesun dat, z procesoru do paměti a zpět, je slabým místem sekvenčních počítačů. Ačkoliv v 80. letech nastal prudký rozvoj umělé inteligence, vymyslet univerzální strojovou inteligenci obsahovalo velmi složité problémy, právě kvůli rychlosti počítačů omezených sekvenčním zpracováním informací. Zatímco lidský mozek je schopen např. vidět obraz jako celek a porovnávat ho s jinými obrazy, sekvenční počítač musí zkoumat jeden pixel za druhým, jednotlivě. Přestože lidský mozek obsahuje relativně pomalé součásti vzhledem k počítači, jeho výkon je neporovnatelně vyšší.[68]

V dnešní době je zrychlování komplikované, jelikož rychlost počítače je omezena rychlostí šíření informace, tedy rychlostí světla. Ke zvýšení rychlosti musí proto moderní počítač vykonávat více operací najednou. To se řeší rozdělením paměti do více paměťových lokací s vlastními procesory, což definuje základní strukturu paralelního počítače. Takové počítače mohou mít až desetitisíce mikroprocesorů, které využívají koordinovaně.[69]

[65] HAVEL, I. *Přirozené a umělé myšlení jako filosofický problém. s.* 23-27
[66] HILLIS, W. D. *Vzor v kameni: jednoduché myšlenky, které řídí počítače*. s. 109
[67] Tamtéž, s. 110
[68] Tamtéž, s. 109-116
[69] Tamtéž, s. 109-111

Zdá se, že nejefektivnějším uspořádáním paralelních operací je nechat každý procesor vykonávat podobný typ operací, pouze v jiném sektoru. Tak je zrychlení skoro úměrné počtu přidaných procesorů (krom ztráty při rozdělování úkolů a spojování jednotlivých výsledků procesorů). Většina úloh se dá rozložit a zpracovávat souběžně na paralelním počítači, jelikož představují fyzikální modely reálného světa. A skutečný svět funguje paralelně. Hillis k současnému vývoji paralelních systémů podotýká, že „jedním z nejzajímavějších dnešních paralelních počítačů je ten, který vzniká v podstatě náhodou ze síťového propojení sekvenčních strojů. Počítače na Internetu, pracující společně, mají potenciální výpočetní kapacitu, která daleko převyšuje každý jednotlivý počítač, jaký kdy byl postaven. Věřím, že se Internet nakonec rozroste… a stroje budou své vstupy číst přímo ze skutečného světa a nebudou závislé na lidech jako prostřednících. Tak jak se informace dostupné na Internetu obohacují a interakce mezi propojenými počítači se stávají složitějšími, předpokládám, že Internet začne vykazovat vlastní chování, které půjde mnohem dále, než bylo explicitně v systému naprogramováno.“[70]

2.2 Konekcionismus

V díle „Parallel Distributed Processing“[71] představil Rumelhart s McClellandem zcela novou ideu strojů na zpracování symbolů, ze které později vznikl základ konekcionistického programu.[72] „Konekcionismus (jakožto paradigma) 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 interakční vazby.“[73]

[70] HILLIS, W. D. *Vzor v kameni: jednoduché myšlenky, které řídí počítače*. s. 118-120
[71] RUMELHART, D. E. a J. L. MCCLELLAND. *Parallel distributed processing: explorations in the microstructure of cognition*. Cambridge, Mass.: MIT Press, 1986.
[72] CRICK, F. *Věda hledá duši: Překvapivá domněnka.* s. 188
[73] HAVEL, I. *Přirozené a umělé myšlení jako filosofický problém. s.* 23

Konekcionistické systémy jsou tedy sítěmi mnoha jednoduchých procesních jednotek, podobných neuronům, které mají mnoho vzájemných spojení. Tyto modely mají (minimálně vzdálený) vztah k architektuře fungování biologického mozku. Předpoklad je, že výpočetní operace se uskutečňují skrze jednoduché interakce těchto procesních jednotek. Každá jednotka přijímá vstupy od malé skupiny svých sousedů a předává výstupy také pouze jim. Neexistuje přitom žádný centrální procesor, pouze jednotky měnící stavy v reakci na příchozí signály a následně vysílající vlastní signály. Výstupy sítě se mohou měnit v závislosti na váhách jednotlivých spojení, které určují charakteristiku sítě. Tyto systémy jsou dobré pro rozpoznávání opakujících se vzorců, doplňování chybějících částí atd. Sítě je možné trénovat pomocí vystavení příkladům; tzn. že váha spojení se mění pod tíhou vstupních příkladů. Po tréninku se síť většinou naučí ustanovit váhy spojení, které efektivně řeší problém – snižují chybový signál.[74]

V konekcionistických systémech předpokládáme, že se ve stavech jednotek může vyskytovat pouze krátkodobá paměť; dlouhodobá paměť má místo v propojení mezi jednotkami. A právě tato variabilní spojení odlišují každou konkrétní síť od ostatních. To znamená, že téměř všechny znalosti jsou implicitní ve struktuře sítě, a nikoli ve stavech jednotek samotných. Jinými slovy, znalosti jsou integrovány do aktuálního stavu sítě a determinují tak průběh zpracování. Jednotlivá spojení se tak účastní mnoha aktivit řešení problému. To je základem konekcionistického přístupu. Konekcionismus nevidí lidské jednání jako produkt izolované části kognitivního systému, ale jako produkt velkého souboru interagujících komponent, z nichž se všechny vzájemně omezují a přispívají tak k pozorovatelnému chování systému. K tomu je přirozené používat paralelní algoritmy namísto sériových.[75]

Kromě toho, že jsou konekcionistické systémy schopny využívat paralelismus a částečně napodobovat mozkové výpočetní operace, také poskytují

[74] HAUGELAND, J. *What Is Mind Design?* s. 21-24
[75] RUMELHART, D. E. *The Architecture of Mind: A Connectionist Approach*. s. 208

řešení mnoha složitých výpočetních problémů, které často vyvstávají v modelech poznání. Řešení mnoha problémů z oblasti kognitivní vědy je totiž určeno plněním velkého množství vzájemně interagujících podmínek. A konekcionistické sítě jsou ideální pro plnění více rozličných podmínek. Pokud zkonstruujeme síť a zadáme jí podmínky, které musí splnit, nakonec se usadí v optimálním stavu, ve kterém bude uspokojeno co možná nejvíce podmínek najednou. Postup usazení sítě do tohoto stavu se nazývá relaxace, a nastává v případě nalezení optimálního řešení. Systém může převzít teoreticky nekonečný počet stavů. Ale vzhledem k podmínkám vestavěným do sítě existuje relativně málo stavů, ve kterých se systém skutečně usadí.[76]

Dle Hopfielada lze obecně vysvětlit chování konekcionistických systémů tak, že se vždy pohybují ze stavu, který splňuje určité podmínky, do stavu, který jich splňuje více než ten předchozí. Systém se tedy pohybuje od počátečního stavu směrem k nejbližšímu uspokojivějšímu, a to do té doby, dokud nedosáhne stavu maximálního uspokojení. Když dosáhne takového stabilního stavu, zůstane v něm, jelikož představuje optimálním řešení problému. Jedním z nejtíživějších problémů v kognitivní vědě je konstrukce systémů, které umožní efektivně zapojit při řešení problémů více zdrojů poznání. Konekcionistické modely pojaté jako sítě splňující podmínky jsou pro tento úkol ideální. Protože každý druh znalosti je zkrátka další podmínkou a systém bude hledat takovou konfiguraci hodnot, která bude nejlépe vyhovovat všem podmínkám, ze všech druhů poznání.[77]

2.2.1 Neuronové sítě

Selhání hlavního proudu umělé inteligence, především špatný výkon při komplexních úkolech a pomalé kognitivní výkony (navzdory použití velmi rychlých strojů) nám jasně ukázalo, že musíme přehodnotit způsob výpočetních

[76] RUMELHART, D. E. *The Architecture of Mind: A Connectionist Approach*. s. 215-219
[77] Tamtéž, s. 219-221

procesů, které jsme připisovali kognitivním tvorům.[78] „Hledání obecnějšího, univerzálního systému zobrazení přivedlo mnoho výzkumníků k výpočetním systémům se strukturou podobnou propojeným sítím biologických neuronů – jako v lidském mozku. Takový systém se nazývá umělá neuronová síť. Neuronová síť je simulovaná síť umělých neuronů.“[79] Každý signál, který putuje od jednoho neuronu ke druhému, má určitou váhu, která určuje, jaký vliv bude mít vstupní signál na výstup neuronu. Výstup neuronu je tedy určen vahou více vstupních signálů. Neuron, zjednodušeně řečeno, zpracovává svůj výstup sečtením váhy všech svých vstupů. Fungování takového neuronu je velmi zjednodušeným obrazem skutečných neuronů. Skutečné neurony jsou o mnoho složitější, nicméně umělé neurony jsou dostatečné pro konstrukci učebních systémů. Umělé neurony lze využít k operacím *a, nebo, negace*. A proto jsou univerzálními stavebními jednotkami pro reprezentaci kterékoliv booleovské funkce.[80]

Každý neuron tedy přijímá vstupy od mnoha jiných neuronů a spolu s ostatními se formuje v jednovrstvou či několikavrstvou síť. U několikavrstvých sítí mohou být jednotky ve vstupní vrstvě sítě považovány za senzorické jednotky, jelikož jejich úroveň aktivace je přímo determinovaná aspekty prostředí. Aktivační úroveň se pak šíří vzhůru, skrze výstupní signál do střední vrstvy sítě, kterou nazýváme „skryté jednotky“. Všechny jednotky ve vstupní vrstvě utváří synaptické spojení určité váhy s každou jednotkou v mezivrstvě. Kterákoliv skrytá jednotka má tedy několik vstupů, jeden za každou jednotku na vstupní vrstvě. Výsledná úroveň aktivace skryté jednotky je součtem všech vlivů, které přicházejí od jednotek ve spodní vrstvě. Aktivační vektor ze skryté vrstvy se šíří dále do výstupní vrstvy jednotek, kde se utváří výstupní vektor.[81]

[78] CHURCHLAND, P. M. *On the Nature of Theories: A Neurocomputational Perspective*. s. 270
[79] HILLIS, W. D. *Vzor v kameni: jednoduché myšlenky, které řídí počítače*. s. 126
[80] Tamtéž, s. 126-127
[81] CHURCHLAND, P. M. *On the Nature of Theories: A Neurocomputational Perspective*. s. 257-261

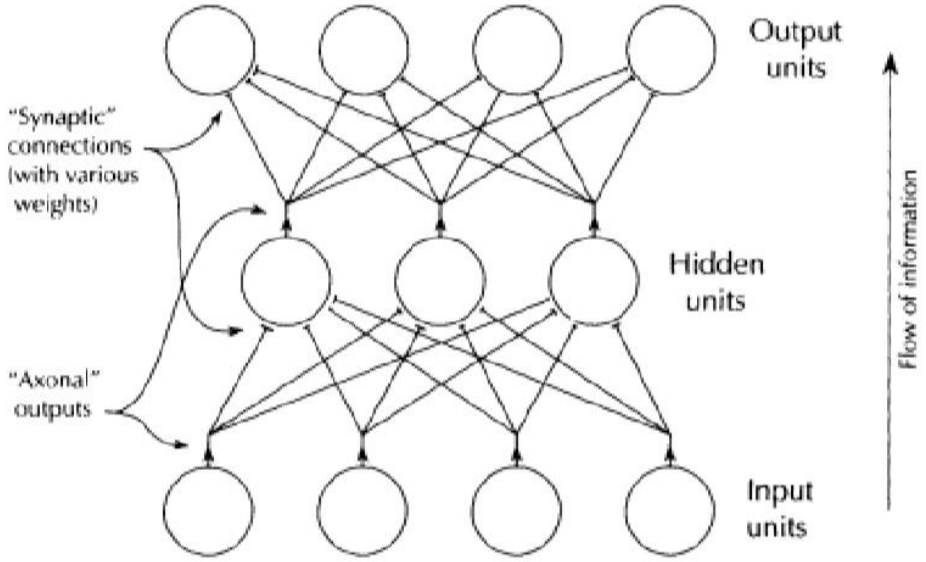

Figure 10.4: A simple network.

[82]

Prozatím stále neznáme přesné fungování mozku, ale zastánci konekcionismu předpokládají, že učení probíhá na základě změny síly synapsí spojující neurony. Stejně tak se umělé neuronové sítě učí změnami a úpravou vah svých spojení, pomocí algoritmu zpětného šíření, o kterém si řekneme více v dalších kapitolách. Učení skrze zpětné šíření chyb se stalo velmi populární nejen díky pravděpodobné podobnosti s lidským učením, ale zejména protože se jedná o efektivní technologii učení různých sítí k řešení různorodých problémů. Jedním z předpokladů konekcionismu je, že v rámci těchto sítí by mohlo být možné prozkoumat tajemství lidského poznávání, jelikož jednotky umělých sítí jsou mnohem přístupnější než mikroskopické jednotky organického mozku.[83]

Jednou z nejvýznamnějších sítí byla bezesporu Hopfieldova síť. Každá jednotka v této síti měla jen dva možné výstupy -1 a +1, ale více vstupů. Jednotka tudíž sčítala všechny vstupy od ostatních jednotek a měnila své chování pod jejich vlivem. Ačkoliv byla Hopfieldova síť biologicky nereálná, byla některými považována za model mozku, který nám odhalí jeho fungování. Hopfieldova síť fungovala následovně: na začátku výcviku dostala vstupní data a jisté pokyny jako nápovědu, jak se dostat k požadovanému výstupu (nikoliv celý postup). Síť

[82] CHURCHLAND, P. M. *On the Nature of Theories: A Neurocomputational Perspective*. s. 260
[83] Tamtéž, s. 271-274

následně přizpůsobovala výstupy svých jednotek, dokud nedosáhla uspořádání vedoucí k požadovanému výstupu. Tak se v podstatě vytváří paměť sítě. Paměť je zde distribuovaná, jelikož informace jsou uloženy v samotných spojeních mezi jednotkami.[84]

Další zajímavou sítí je NETtalk, která se učila převádět anglický psaný text na řeč, na základě pokusu a omylu. DECtalk převáděl text na řeč už o pár let dříve, ale rozdíl byl v tom, že všechna jeho pravidla a výjimky kódovali lidé. NETtalk naproti tomu nebyl přímo naprogramován; učil se řešit problém pomocí učebního algoritmu a souboru příkladů. Síť nebyla chápána jako model jazykového porozumění, ale jen jako převodník textu na řeč, jelikož tu nebyla žádná sémantická databáze. Síť začíná se souborem vstupů a náhodným uspořádáním vah svých spojení. Až během tréninku se sama učí, jak řešit určený problém. Znalosti, které systém používá k transformaci vstupů na výstupy jsou zakódovány ve vahách spojení mezi jednotkami. Informace nutné k řešení problému jsou tedy rozloženy napříč celou sítí a tím pádem není síť náchylná ke katastrofálnímu selhání, v případě poškození konkrétní jednotky nebo spojení. Vnitřních spojení mezi jednotkami je několik tisíc, proto by bylo nepraktické nastavovat jejich váhy ručně. Naštěstí existují automatické učebních algoritmy pro nastavení vah, jako např. již dříve zmíněný algoritmus učení zpětným šířením.[85]

2.2.2 Nedokonalý odraz nebo základní stavební kámen?

Pro vědce není snadné pochopit propojení neuronů ani jejich chování způsobené vzájemnými interakcemi, protože mozek není lineární. A nelineární soustavy jsou z matematického hlediska o mnoho obtížněji pochopitelné než lineární; jejich chování je nepředvídatelnější a komplexnější.[86] Z jistého hlediska tedy mohou neurální sítě (ve velmi omezené míře) fungovat jako model chování

[84] CRICK, F. *Věda hledá duši: Překvapivá domněnka.* s. 187
[85] CLARK, A. *Mindware: an introduction to the philosophy of cognitive science.* s. 63-65
[86] CRICK, F. *Věda hledá duši: Překvapivá domněnka.* s. 180

skutečných neuronů. „Neurální sítě jsou soubory nejrůzněji propojených jednotek, z nichž každá má vlastnosti výrazně zjednodušeného neuronu. Slouží k napodobování dějů probíhajících v určitých částech nervového systému, k výrobě komerčně využitelných pomůcek i k ověřování obecných teorií činnosti mozku.“[87]

Co můžeme říci s jistotou je, že mozek je organizován dle radikálně odlišných výpočetních linií, než se používají v konvenčních digitálních počítačích. Mozek je masivně paralelní procesor, který provádí své výpočetní úkoly relativně pomalu.[88] Např. rychlost nejvýkonnějších počítačů byla na přelomu století více než 10 000 000 úkonů za sekundu a v posledních letech vzrostla až k 30 000 bilionům úkonů za sekundu. Neuron má oproti tomu frekvenci stovek impulzů za sekundu. Konvenční digitální počítače jsou tedy výrazně rychlejší, ale paralelní zapojení vyrovnává pomalejší výkon skutečných neuronů. Paralelní zpracování rovněž znamená, že poškození jednotlivých neuronů nemá za následek výrazné ovlivnění mozkové činnosti, zatímco u běžných počítačů může vést i drobné poškození ke ztrátě funkčnosti.[89]

Konekcionistické modely jsou na rozdíl od klasických modelů umělé inteligence zajímavé pro kognitivní vědy, právě díky jejich schopnosti elegantní degradace při poškození. Spolu se schopností učení můžeme od sítí očekávat, že se samovolně naučí obcházet vadné komponenty a budou i při poškození dále vykonávat svou funkci (za cenu zhoršeného výkonu). Jelikož každá informace je uložena distribuovaně v celé síti, ztráta několika jednotek degraduje uložené informace, ale nezničí je. Tato schopnost napodobuje lidské zpracování.[90]

Problém se snahou konstruovat systémy podobnější mozku spočívá v tom, „že stejné chování může předvést několik dosti vzájemně odlišných modelů. Rozhodnout, který z nich se nejvíce blíží skutečnosti, nutně vyžaduje další doklady, například přesnou znalost vlastností skutečných neuronů a molekul

[87] CRICK, F. *Věda hledá duši: Překvapivá domněnka.* s. 180
[88] CHURCHLAND, P. M. *On the Nature of Theories: A Neurocomputational Perspective*. s. 255
[89] CRICK, F. *Věda hledá duši: Překvapivá domněnka.* s. 180-181
[90] RUMELHART, D. E. *The Architecture of Mind: A Connectionist Approach*. s. 227

v dané části mozku.“[91] I přesto se ale neurální sítě podobají mozku více, než architektura klasického digitálního počítače. Na druhé straně zůstává pravdou, že jednotky neuronové sítě jsou extrémně zjednodušené. Skutečné soustavy neuronů vykazují široké spektrum vlastností, které většině konekcionistických modelů chybí. Stejně tak architektura umělých sítí je značně zjednodušena. Většina umělých neuronových sítí tedy zatím zůstává vzdálená od detailů skutečného neurovědeckého výzkumu.[92]

V souvislosti se zjednodušenou konstrukcí neuronových sítí se také objevuje kritika ohledně používání umělých úkolů a výběru vstupních a výstupních reprezentací. Ačkoliv se sítě učí najít vlastní řešení problémů, tato řešení zůstávají poznamenána lidským experimentátorem (který rozhoduje např. o výběru problémové oblasti, nebo tréninkových materiálech). I když sítě zkoumají úkoly, které jsou základní, výběr vstupních a výstupních reprezentací byl často velmi umělý. Vstupem je běžně tréninkový soubor dat a výstup bývá často pouze abstraktní aktivita místo činnosti v reálném světě. Dá se předpokládat, že tento způsob utváření problémového prostoru povede k nerealistickým řešením. Lepší strategií by bylo vytvořit systém s biologicky realističtějšími vstupy a skutečnými akcemi jako výstupy. Hlavní myšlenkou této kritiky je, že si kognitivní věda v rámci svého výzkumu nemůže dovolit zjednodušení, která vyjmou reálný svět a jednajícího agenta. Touha kognitivní vědy po vysvětlení skutečného biologického poznání by se tudíž neměla uchylovat k abstrakci, mimo skutečné vnímání a jednání, jelikož abstrakce zbavuje umělé systémy příležitosti zjednodušit jejich úkoly, ohledně zpracování informací, přímým využíváním struktury reálného světa.[93]

S problémem umělých vstupních a výstupních reprezentací také souvisí fakt, že konekcionistické sítě obsahují relativně málo jednotek a spojení (ve srovnání s mozkem) dostačujících pouze k řešení relativně diskrétních a dobře definovaných problémů. Jak bylo řečeno, sítě jsou často trénovány na umělých

[91] CRICK, F. *Věda hledá duši: Překvapivá domněnka.* s. 201
[92] Tamtéž, s. 201
[93] CLARK, A. *Mindware: an introduction to the philosophy of cognitive science*. s. 79-80

verzích problémů reálného světa. Navíc obvykle řeší pouze jeden konkrétní problém současně. Přirozené neuronové sítě se naopak musí vypořádat s množstvím dat, které se týkají více problémů a vyžadují tak vnitřní uspořádání a distribuci. Étos „jeden úkol, jedna síť“ tedy může představovat podstatné zkreslení biologických faktů. Rozšíření sítí bohužel tento problém nevyřeší, jelikož běžné konekcionistické sítě zapomínají předchozí znalosti bez ohledu na jejich velikost. Snaha o přizpůsobení se novému vzoru dat vede vždy k přepisování ostatních informací. Jedním z možných řešení je použití sítě sítí. Taková architektura by zahrnovala více menších sítí, které by mohly být speciálně zaměřené. Prospěšné by také mohlo být připojení externí paměti k neuronovým sítím, které by tím pádem nemusely při každém novém úkolu přepisovat předchozí znalosti.[94]

2.2.2.1 Stručný popis neuronové sítě člověka

Všichni lidé jsou tvořeni jednáním ohromného souboru neuronů, přičemž typický neuron reaguje na širokou škálu podnětů. Tyto signály mohou podráždit, utlumit nebo jinak modifikovat chování neuronu. Ten v reakci na podráždění vyšle elektrický signál, který putuje k dalším neuronům, a tak ovlivní jejich chování. Jeden neuron je schopen zpracovat až pět set impulzů za sekundu (pro srovnání, rychlost počítače je nyní až milionkrát vyšší). Různé neurony vysílají vzruchy různou rychlostí a v různém počtu. Neuron není jednoduchý a zdaleka není snadné předvídat jeho chování.[95]

Na rozdíl od první generace konekcionistických sítí mozek obsahuje stovky různých typů neuronů, se specifickými funkcemi. Třetí vlna konekcionistických sítí se však k této typové i funkční variabilitě neuronů (alespoň vzdáleně) snaží přiblížit. V důsledku existence různých typů neuronů se různé části mozku, zdá se, specializují na konkrétní činnosti. Když určitou část poškodíme, jsou poškozeny

[94] CLARK, A. *Mindware: an introduction to the philosophy of cognitive science*. s. 80-81
[95] CRICK, F. *Věda hledá duši: Překvapivá domněnka.* s. 100-102

jednotlivé funkce. To ale neznamená, že jsou jednotlivé oblasti mozku sestrojeny jako oddělitelné komponenty u klasického digitálního počítače.[96]

Díky tomu lze předpokládat, že se jednotlivé funkce mozku přeskupují a přizpůsobují nastalé situaci. Když nastane nenávratné poškození, po kterém se neurony již nemohou regenerovat, je pravděpodobné, že se mozek (alespoň v omezené míře) opět naučí potřebné schopnosti zapojením jiných neuronů. Překvapivě můžeme podobné chování pozorovat i u umělých neuronových sítí, které se při poškození učí samy hledat nové cesty k vyřešení problému.[97]

2.3 Umělá evoluce: Přirozený výběr o trochu rychleji

Jak bylo řečeno, klasické digitální počítače mají hierarchickou strukturu, ve které jsou jednotlivé části spojeny jednoduchými a předem definovanými vztahy. Běžně jsou tyto vztahy jednosměrné, takže vidíme jednoznačnou hierarchii příčin a důsledků. Díky jednoduché interakci těchto součástí jsou vnitřní procesy klasických počítačů většinou vcelku srozumitelné, pokud nejsou příliš komplexní. V posledních pár desetiletích se ale ukázalo, že hierarchická struktura klasických počítačů není natolik efektivní a když je systém příliš komplikovaný, klasické návrhy omezují výkon a mohou způsobovat selhání. Problém není jen v selháních; ve složitém systému mohou také vznikat nepředvídatelné reakce vlivem vzájemného působení. Jednotlivé části mohou správně vykonávat svou funkci, a přesto mohou nastat nepředvídatelné interakce. Čím komplexnější program (např. operační systém), tím větší je problém s předvídáním následků jeho vnitřních interakcí. Pokud je systém dostatečně komplikovaný, nelze mít přehled o všech možných interakcích, jejich následcích atd. „Je důležité si uvědomit, že za naznačené problémy neodpovídá stroj nebo software sám o sobě. Slabiny jsou v naší technické metodě návrhu. Ale víme, že ne vše, co je složité, je také nespolehlivé. Mozek je mnohem složitější než počítač, přesto je mnohem méně

[96] HILLIS, W. D. *Vzor v kameni: jednoduché myšlenky, které řídí počítače*. s. 135-137
[97] Tamtéž, s. 136-137

náchylný ke katastrofálním chybám. Kontrast ve spolehlivosti mezi mozkem a počítačem dokládá rozdíl mezi výsledky evoluce a inženýrství. Jediná chyba v klasickém počítačovém programu může způsobit jeho zhroucení, zatímco mozek je většinou schopen tolerovat špatné myšlenky, nesprávné informace, a dokonce i nefungující části. Přestože neurony v mozku průběžně umírají a nejsou nikdy nahrazeny; mozek se s těmito vadami vyrovná a kompenzuje je.“[98] Jako vhodný krok k dalšímu postupu ve vývoji umělé inteligence proto může být inspirace evolucí. V systému, který by prošel evolučním procesem, by chování bylo určeno mnoha jednoduchými interakcemi, bez jakéhokoliv řízení shora dolů.[99]

Simulovaná evoluce tak představuje nový způsob návrhu softwaru. Tento způsob přitom ve vysoké míře eliminuje problémy související s klasickou metodou projektování systémů umělé inteligence. Je to způsob, při kterém je softwaru ponechána volná cesta k samovolnému vývoji. Hillis celý proces popisuje následovně: „Nejprve vygenerujeme sadu náhodných programů… dalším krokem je otestovat tuto sadu a najít programy, které jsou nejúspěšnější. Musíme tedy spustit každý program a podívat se, zda dokáže správně vykonat požadovanou operaci… protože jsou programy náhodné, žádný z nich pravděpodobně testem neprojde – ale díky pouhému štěstí budou některé blíže správnému řešení… Dalším krokem je vytvořit nové sady z úspěšných programů. Dosáhneme toho smazáním programů s menším než průměrným skóre; jen nejvhodnější programy přežijí. Novou sadu vytvoříme kopií programů, které přežily, s malými změnami, postup je podobný nepohlavnímu rozmnožování s mutací. Také bychom mohli „pěstovat" nové programy párováním přeživších z minulé generace – tento postup je analogický pohlavnímu rozmnožování. Dosáhneme toho zkombinováním posloupnosti instrukcí z obou „rodičovských" programů – vytvoříme „potomka". Rodiče pravděpodobně přežili, protože obsahovali užitečné posloupnosti instrukcí, a máme tedy slušnou šanci, že potomek zdědí od každého z rodičů ty nejužitečnější znaky.“[100]

[98] HILLIS, W. D. *Vzor v kameni: jednoduché myšlenky, které řídí počítače*. s. 140
[99] Tamtéž, s. 11-13
[100] Tamtéž, s. 141-142

„Napodobení evoluce samo o sobě neřeší problém vytvoření umělé inteligence; nicméně ukazuje nám správnou cestu. Klíčovou myšlenkou je přesunout břemeno složitosti z hierarchie návrhu na kombinatorickou sílu počítače. Napodobení evoluce je v podstatě určitou heuristickou prohledávací technikou, která prohledává prostor možných koncepcí návrhu. Heuristiky použité k prohledávání prostoru jsou *Zkus návrh podobný nejlepším, jaké jsi zatím našel* a *Zkombinuj prvky dvou úspěšných návrhů*.“[101] Umělá evoluce v této podobě volí změny slepě v reakci na chyby, to znamená bez úvahy nad tím, jaký bude výsledek.

Vzhledem k rychlosti moderních počítačů lze tento selekční proces provést statisíckrát. S každou následující generací zde probíhá postupné zdokonalování (generace od generace) pomocí „umělého“ výběru. Dle Hillisových zkušeností byly programy vybrané tímto evolučním procesem v důsledku výkonnější než ty, které byl schopen napsat sám. Dokonce dodává, že ani po přezkoumání jejich posloupnosti instrukcí z velké části nerozumí tomu, jak pracují.[102]

Rozdíl mezi lidskými bytostmi a výsledky pokusů o umělý vývoj obdobně komplexních systémů je však patrný. V Hillisově příkladu jsme viděli systém, který se vyvíjel podle předem určené strategie, která byla vybraná inženýry: vyber nejlepší a ostatní smaž. Je spíše empirickou otázkou, zda touto strategií jednou dosáhneme autonomního umělého myšlení. Ale ještě zajímavější otázkou je, zda bychom mohli považovat výsledek takového vývoje, který proběhl v umělém médiu (jako je robot nebo počítačový ekosystém), za živý? Narážíme tak na problematiku ohledně vztahu mezi životem a myslí. Zvažme virtuální ekosystém zvaný Tierra od Thomase Raye, ve kterém digitální organismy (každý z nich je v podstatě jednoduchým programem) soutěží o čas na CPU. Organismy se mohou reprodukovat (kopírovat) a podléhají změnám náhodnými mutacemi. Systém je implementován do paměti počítače a organismy soutěží, mění se a vyvíjí. Po analýze výsledné populace našel Ray řadu úspěšných a často neočekávaných strategií přežití, přičemž každá z nich využívala nějaké slabiny v dominantní

[101] Tamtéž, s. 143-144
[102] Tamtéž, s. 142

strategii. Nějaké organismy parazitovaly na jiných organismech, které si v návaznosti na to vyvíjely obrané mechanismy, popř. se učily parazitovat na svých parazitech. Tak vzniká otázka: jedná se pouze o virtuální simulované organismy, nebo je to skutečný ekosystém, osídlený reálnými organismy „žijícími" v neobvyklém místě digitální paměti počítače? Ray věří, že takové systémy mohou přinejmenším nést vlastnosti charakteristické pro život – jako např. sebereplikace, evoluce atd.[103]

2.4 Strojové učení: nabývání vědomí?

Stejně jako hierarchická struktura klasických digitálních počítačů tak i klasické pojetí učení, jakožto pravidly řízené aktualizace systému vět, se ukázalo jako chybné v mnoha ohledech. Klasický přístup není schopen obsáhnout všechny případy učení. Od přelomu století však proběhl zásadní vývoj v oblasti kognitivní neurobiologie a umělé inteligence v konekcionistickém podání. To kognitivním vědcům poskytlo solidní podklad pro postupné rozplétání problému poznávání. Předpokladem je, že proto musí existovat typ učení, který je základnější než proces manipulace s větami. Pravděpodobně je pod úrovní propozičních výroků a vět ještě jakési dynamické učení, které operuje primárně na sublingvistických faktorech.[104]

Jak bylo řečeno, klasické programy jsou dobré v tom, na co jsou naprogramovány, ale reagují neinteligentně, když čelí novým situacím. A to zejména proto, že se většina programů klasické umělé inteligence nemůže sama modifikovat a adaptovat na své prostředí. [105] Klasické systémy a programy pracovaly vždy dle pokynů programátora, a tudíž se nemohly samy vyvíjet ani modifikovat. Pravděpodobně nejčastějším argumentem proti umělé inteligenci se tak stalo přesvědčení, že stroje mohou dělat jen to na co je naprogramujeme. Ale existují i programy, které se umí samovolně přizpůsobovat, opravovat své chyby

[103] CLARK, A. *Mindware: an introduction to the philosophy of cognitive science*. s. 117
[104] CRICK, F. *Věda hledá duši: Překvapivá domněnka.* s. 252-253
[105] RUMELHART, D. E. *The Architecture of Mind: A Connectionist Approach*. s. 222-224

a hledat možná řešení úloh. Např. systémy učící se na principu zpětné vazby, které upravují své reakce na základě chyb, které je dělí od jejich cílů. Toto spojení mezi chybou a následnou reakcí může být flexibilní. Systém v takovém případě postupně přizpůsobuje své reakce a časem dosahuje lepších výsledků. Učení systému je pak podobné lidskému učení, jelikož je založeno na předchozí zkušenosti. Hlavní výhodou takových flexibilních systémů je, že se dokáží přizpůsobit neznámé situaci.[106]

2.4.1 Zpětné šíření

V rámci neuronových sítí, které jsme probrali v předchozí kapitole, je nejrozšířenější strategií učení skrze zpětné šíření. Zjednodušeně je postup následující: síť dostane tréninková data, na kterých se učí přibližovat požadovanému cíli, porovnáním vlastního výstupu s cílovým výstupem. Na základě tohoto porovnání následně vypočítává nutné změny. Typický příklad učebního systému pochází od Patricka Winstona. Jeho program se učí definice pojmů na základě kladných a záporných příkladů. Po shlédnutí příkladů toho, co pojem vyjadřuje a co naopak nevyjadřuje, se program učí jeho definici. Pokud získá chybnou definici na základě příkladů ji upravuje tak dlouho, dokud nedosáhne správné. „Každý prvek definice se program naučil uděláním nějaké chyby a příslušnou úpravou definice. Když program dojde ke správné definici, přestane dělat chyby a ponechá definici beze změny. Může pak správně identifikovat jako bránu každou bránu, kterou mu ukážeme, i když příslušné uspořádání kostek nikdy předtím neviděl. Naučil se pojem „brána".“[107]

V rámci neurálních sítí existují dva základní druhy učení: bez dohledu a s dohledem. Během učení bez dohledu síť nedostává žádné instrukce zvenčí o tom čemu a jak se učit. Síť se tudíž do jisté míry učí a organizuje sama (viz kapitola Samoorganizující se systémy). Učení s dohledem naopak probíhá tak, že síť

[106] HILLIS, W. D. *Vzor v kameni: jednoduché myšlenky, které řídí počítače*. s. 124-126
[107] Tamtéž, s. 126

dostane informace zvenčí, které usměrňují její chování tím, že označují její výstup za správný nebo nesprávný. U tohoto druhu učení musí mít síť k dispozici jistou množinu vstupů, která reprezentuje přiměřený vzorek informací, se kterými se síť po tréninku setká. Procvičování na trénovací množině je nezbytné vždy, protože síťová spojení jsou nastavena náhodně (na rozdíl od mozkových, které jsou ovlivněny genetikou).

Ačkoliv se procedura učení zpětným šířením stala pravděpodobně nejpopulárnější metodou pro trénování sítí a je využívána při trénování v problematických oblastech včetně rozpoznávání znaků, řeči, předpovědi chaotických funkcí atd., existují stále problematické oblasti. Zásadním omezením učení skrze zpětné šíření je problematika doby učení (jak omezit tréninkový čas u složitějších problémů); a problematika generalizace (jak si můžeme být jisti, že síť vyškolená na podmnožině příkladů bude správně zobecňovat své poznání na nové příklady). Dlouhá doba učení je ústřední obtíží procedury zpětného šíření, jelikož středně těžké problémy běžně vyžadují desítky až stovky tisíc tréninkových kol. Tyto nároky jsou kompenzovány rychlostí počítačů, ale problém jako takový stále přetrvává. Stejně tak problém generalizace není možné zcela eliminovat. Nejdůležitějším aspektem sítí je, že na základě tréninku získávají poznání, které jsou schopné používat v reakci na nové případy, které ještě nebyly prozkoumány. Potíží je, že u většiny problémů je velká svoboda v tom smyslu, že existuje mnoho různých řešení – a každé řešení vytváří rozdílný způsob generalizace. Ne všechny však mohou být správné.[108]

2.4.2 Samoorganizující se systémy

Nevýhodou systémů, jejichž učení je založeno na pozitivních a negativních příkladech, je nutná přítomnost trenéra neboli autority posuzující správnost příkladů. Existuje ale i typ samostatně učebních neuronových sítí. Takovou síť

[108] RUMELHART, D. E. *The Architecture of Mind: A Connectionist Approach*. s. 228-232

nazýváme samoorganizujícím se systémem. Umělé neurony jsou v těchto sítích organizovány ve dvourozměrném poli, ve kterém se sousedící neurony vzájemně ovlivňují. „Kdykoli neuron udělá „chybu" a aktivuje se jinak než jeho sousedé, neuron zvýší váhu vstupů, které odpovídají výstupům jeho sousedů, a sníží váhu ostatních vstupů. Jeho sousedi se samozřejmě rovněž snaží najít své správné propojení, takže na počátku takřka slepý vede slepého, ale postupně některé neurony najdou své správné vstupy a stanou se tak účinnými trenéry pro své sousedy. Jako obvykle se upravují jen ty neurony, které udělaly chybu.“[109] A tento proces vzájemného tréninku neuronů probíhá, dokud síť jako celek nedojde k nejvhodnějšímu výsledku.

V posledních letech dosahují tyto samoorganizující se systémy překvapivých výsledků a narůstá také jejich využitelnost v reálném světě. Jedním z důležitých kroků vpřed, v rámci této oblasti, jsou veškeré systémy schopné samostatně organizovat vlastní paměť. Nejnovějším z nich je např. systém DNC (Differentiable Neural Computer), který je zkombinován z neuronové sítě a klasické počítačové paměti. Toto propojení mu dává potenciál k vlastnímu rozvoji na základě získaných zkušeností. Veškeré zpracování informací, zapisování i čtení informací z paměti včetně struktury organizace informací v paměti, se síť naučila sama bez lidského přičinění. Taková síť je tudíž schopná aktualizovat a popřípadě replikovat své znalosti, a co víc, netrpí dlouholetým neduhem klasických neurálních sítí – zapomínáním. Dalo by se proto říci, že se jedná o skutečný samoučící se počítač, který neustálým opakováním hledá optimální způsob pro práci s vlastní pamětí.[110]

Reálným důkazem efektivity samoorganizujících se systémů, se před nedávnem stal program AlphaGo Zero z laboratoře DeepMind. Tento program byl v pořadí čtvrtou generací programu na ovládnutí hry Go. Hra Go je vhodná pro testování univerzální umělé inteligence především proto, že její kombinatorika je mnohonásobně složitější než např. šachy (o kterých si filosofové až do přelomu

[109] HILLIS, W. D. *Vzor v kameni: jednoduché myšlenky, které řídí počítače*. s. 133
[110] https://deepmind.com/blog/differentiable-neural-computers/

století mysleli, že nemohou být ovládnuty počítačem). Po desetiletí se proto hra Go považovala za nejnáročnější oříšek pro umělou inteligenci, jelikož počet možných tahů a vyhodnocování síly jednotlivých pozic je výpočetně náročné. AlphaGo Zero se oproti předchozí verzi liší tím, že jsou do ní nahrány pouze pravidla hry; zatímco do předchozích verzí byly nahrány jednotlivé záznamy her, ze kterých se systém učil lidské strategie a techniky. Svou strategii tedy vyvíjí samostatně, bez vnějšího učitele. Program začíná s náhodnou hrou a hraním proti sobě samému získává zkušenosti nutné k učení. Tak se v podstatě stává svým vlastním učitelem. Metodou pokusu a omylu se síť pod náporem tréninku aktualizuje a proměňuje. V každé iteraci se zvyšuje výkon, a tím i kvalita her, což vede k neustále silnějším verzím hráče AlphaGo Zero. Dalo by se říci, že tato verze je silnější než její předchůdci právě proto, že není omezena limity lidského poznání. Již po třech dnech tréninku se programu AplphaGo Zero podařilo porazit jeho předchozí verzi v poměru 100:0. Během několika milionů her AlphaGo vs AlphaGo se tak program postupně naučil hru Go od nuly a během několika dní vymyslel i taktiky, se kterými se lidé nikdy nesetkali. Výzkumníci DeepMind jsou proto přesvědčeni, že pokud budou podobné techniky aplikovány i na jiné strukturované problémy, může to zásadně rozšířit varianty řešení, které by buď vymysleli lidé nebo lidmi usměrněné programy.[111]

2.4.3 Podobnosti a odlišnosti s lidským učením

Při pokusech napodobit lidskou mysl může být také užitečné ptát se, co ji přivedlo do stavu, v jakém se nachází. Dle Turinga se lidská mysl skládá ze tří složek: Počáteční stav mysli při narození, vzdělání kterým prošla, a další zkušenosti mimo vzdělání. Nebylo by tedy lepší se pokusit místo simulace mysli dospělého člověka pokusit simulovat mysl dítěte, která by se následně podrobila odpovídajícímu vzdělání? Turingova vize byla taková, že dětský mozek by mohl obsahovat kromě mnoha prázdných listů jen jednoduchý mechanismus, který by

[111] https://deepmind.com/blog/alphago-zero-learning-scratch/

šlo naprogramovat snadněji. Tudíž by se problém konstrukce lidské mysli rozdělil na program dítěte a vzdělávací proces. Učební proces strojů by se samozřejmě lišil od lidského. Ale vzdělání jako takové je dle Turinga možné všude, kde probíhá obousměrná komunikace mezi žákem a učitelem. Takový stroj by se dal učit za použití trestů a odměn, které budou udělovány v závislosti na plnění příkazů a budou přenášeny pomocí neemocionálních kanálů.[112]

Taková byla vize praotce umělé inteligence. Zatímco výuku vedenou pomocí trestů a odměn využíváme už notnou dobu, idea vytvoření dětské mysli a následně komplexního procesu vzdělání se prozatím jevila jako nedosažitelná. V roce 2015 však tuto myšlenku oživil Marek Rosa ve svém programu GoodAI[113]. Snahou tohoto výzkumu je vytvoření univerzální umělé inteligence, přičemž jejich postup se vrací k původní myšlence rozdělení konstrukce umělé inteligence do dvou procesů. První fázi řídí snaha o vytvoření architektury umělého mozku, ne nutně typu neuronové sítě. Jelikož je tato snaha inspirována fungováním lidského mozku, který se umí učit graduálně – nové poznatky se kumulují a jejich získávání je s přibývajícími znalostmi stále efektivnější – musí mít také schopnost učit se z prostředí. Ve druhé fázi jde o vytvoření systému, který v sobě bude zahrnovat repertoár dovedností a schopností podobných lidským. Tyto dva milníky jsou na sobě závislé a pokud nebudou splněny oba, nebude možné vytvořit obecnou umělou inteligenci. Mohlo by se zdát, že sama první část je dostatečná, ale bez druhé složky by se jinak inteligentní systém nebyl schopen orientovat v lidském světě. Ke splnění druhého milníku je dle Rosy potřeba vytvořit něco jako školu umělé inteligence – což je virtuální prostředí, kde se umělá architektura učí poznatky postupně tak, jak prochází prostředím. Jednotlivé získané schopnosti umožňují získávání dalších kapacit ještě efektivněji. Tak postupně dochází k výchově umělé inteligence, která světu okolo nás bude rozumět velmi podobným způsobem jako my.[114]

[112] TURING, A. M. *Computing machinery and intelligence*. s. 454-457
[113] https://www.goodai.com/
[114] https://www.youtube.com/watch?v=-obfl0iFLAI

Na základě tohoto výzkumu se zdá, že je možné (a v dohledné době snad i proveditelné), aby se systémy umělé inteligence naučily čerpat zkušenosti z vnějšího světa a na základě těchto zkušeností se poté rozhodovaly a řešily problémy. A jak dodává Shanker: „Pokud je učení takovou změnou v systému, která produkuje více či méně permanentní změnu v jeho schopnosti adaptovat se na své prostředí, vyplývá z toho, že neexistuje kategorický rozdíl mezi systémy biologického učení a ideálními učebními stroji.“[115]

Ve dřívějších výzkumech se systémy umělé inteligence, bylo naopak nezbytné, aby experimentátor odstranil většinu detailů při orientaci umělé inteligence ve virtuálním prostředí, za účelem vytvoření jednoduchého popisu. Jediným vstupem dřívějších (i mnoha současných) programů byl omezený set jednoduchých tvrzení, odvozených lidmi z reálných dat. Problémy s rozpoznáváním, prostorovým porozuměním apod. byly ignorovány. Lidský popis světa ale může být zavádějící, protože svět jak jej vnímáme my, je určen našimi cíli a potřebami. Proto je nutné konstruovat a testovat programy v reálném světě, nebo v jiném dostatečně bohatém prostředí. Ne v extrémně zjednodušené realitě, protože z ní nelze systémy jednoduše přesunout do reálného světa.[116]

Částečným pokrokem na tomto poli bylo spuštění experimentu DeepMind Lab, který se snaží o posunutí hranic umělé inteligence vyvíjením systémů pro řešení komplexních problémů, bez předchozího učení postupu řešení. K tomu, aby byla umělá inteligence schopna obstát v reálném světě, je třeba vytvořit obecnou inteligenci, která bude zvládat rozličné úkoly za měnících se podmínek. V tomto případě je proto efektivnější, aby programy nebyly předprogramovány k žádným konkrétním akcím, ale vše se naučily až ze vstupů a signálů z prostředí. Zatím naše technika není natolik pokročilá, aby bylo možné vypustit umělou inteligenci do reálného světa, proto bylo vytvořeno bohaté simulované prostředí DeepMind Lab. Prostředí je naprogramováno ve formě trojrozměrného světa plného objektů, ve kterém se agent může pohybovat a plnit úkoly. V každém okamžiku agenti

[115] SHANKER, S. G. *Wittgenstein's Remarks on the Foundations of AI*. s. 632
[116] BROOKS, R. A. *Intelligence without Representation*. s. 399-407

pozorují svět, jako obraz vykreslený z vlastní perspektivy. Výzkum obecné umělé inteligence zdůrazňuje nutnost plánování strategií a plně autonomních agentů, kteří musí sami zjišťovat, jaké úkoly je třeba provést k prozkoumání svého prostředí. Tento experiment má již nyní dopad na naše vnímání umělé i přirozené inteligence. Rozvíjet inteligenci v trojrozměrném světě a z pohledu první osoby je, jak se ukázalo, snazší a přirozenější v porovnání s ostatními strategiemi. Už jen proto, že jediný příklad univerzální inteligence který známe, vznikl kombinací evoluce a podobného učení. Je možné, že velká část inteligence je tudíž přímým důsledkem bohatství našeho prostředí a je nepravděpodobné, že by vznikla bez něj. Zvažme alternativu: kdybychom vyrostli ve světě, který by vypadal jako Space Invaders nebo Pac-Man, jak pravděpodobné by bylo, že bychom dosáhli obecné inteligence?[117]

[117] https://deepmind.com/blog/open-sourcing-deepmind-lab/

3 Lidské myšlení a strojové „myšlení“: podobnost čistě náhodná?

Řekli jsme již mnoho o historii ideje umělé inteligence, vývoji technologií i moderních programech schopných samostatně vykonávat mnohé úkoly, a dokonce i samostatně vybírat, které z nich je třeba provést pro splnění svých cílů. Tento pokrok byl do značné míry inspirován lidskou formou myšlení. Přesto jsme ale zatím nevyřešili jednu z nejpalčivějších otázek umělé inteligence: jaký je vztah mezi počítačovým zpracováním informací a lidským myšlením? Chápání sebe samých a našich duševních výkonů je jistě ovlivněno stavem moderní vědy. Myšlení je v každé době připodobňováno k výkonu aktuálních technologií. Dnešní dilema je ale o to větší, že nehledáme pouze ekvivalent našeho myšlení, nýbrž se snažíme moderní technologii předat část sebe samých. Pokoušíme se strojům předat naši schopnost myšlení. Částečným úspěchem je, že jsou nyní počítače běžně schopné zpracovávat některé z našich myšlenek. Stejně tak jako jiné lidské nástroje, i počítače pomáhají zefektivnit naše schopnosti. Na rozdíl od ostatních nástrojů jsou ale počítače schopné uchopit lidské myšlenky a rozvíjet je samostatně, často i za hranice našeho vlastního poznání.[118]

Existují ale meze, které stroje skutečně nemohou překročit? Nebo jsou tyto omezení uměle nastaveny jen ze strachu z narušení jedinečného postavení lidské bytosti? Jak bylo ukázáno v předchozí kapitole, dnes jsme již schopni konstruovat stroje a psát programy, jež jsou schopné řešit širokou škálu typicky lidských problémů.[119] A nejen to. Jsou schopné se takovým řešením naučit sami bez lidského přičinění. V zásadě platí, že učení, rozhodování, plánování apod. už není výsadou biologických organismů. A ačkoliv je rozum od dob Aristotela považován za podstatu člověka, umělá inteligence se snaží obejít lidské tělo i lidskou mysl, s cílem dosáhnout čisté racionality. Výzkum umělé inteligence je v podstatě sázkou na to, že člověk dokáže formalizovat své myšlení i chování a předat ho

[118] HILLIS, W. D. *Vzor v kameni: jednoduché myšlenky, které řídí počítače*. s. 14
[119] Tamtéž, s. 68

stroji. Anebo pochopit a okopírovat veškeré mozkové procesy a následně jejich činnost simulovat v umělém médiu. Pokud v jedné z těchto snah uspějeme a vytvoříme umělou mysl, zásadně to změní naše chápání sebe sama.[120]

V rámci pokroku ve vývoji neuronových sítí můžeme už nyní bez přehánění říci, že umíme simulovat nejdůležitější vlastnosti neuronů. Samozřejmě ne do detailu a rozhodně nám prozatím uniká celistvé porozumění jejich chování, ale základní funkcionalitu simulovat umíme. Teoreticky tedy můžeme simulovat síť mnoha miliard umělých neuronů, což by ve výsledku znamenalo, že jsme potenciálně schopni simulovat základní funkcionalitu mozku. A pakliže bychom mozek chápali jako biologický stroj, který je zodpovědný za lidské myšlení, nebude v principu existovat žádná nepřekonatelná propast mezi lidským a strojovým myšlením. Takové tvrzení, že myšlení je pouze složitým výpočetním procesem založeným na neurální aktivitě, je ale problematické. Ačkoliv je moderní vědou podporované, intuitivně se nám může jevit jako neúplné.[121]

Jak se zdá, vývoj komplexní umělé inteligence na lidské úrovni vyžaduje hlubší poznání nás samých a našeho myšlení. Proto si před vytvářením jakýchkoliv modelů myšlení musíme položit pár otázek: je lidské myšlení plně formalizovatelné a je tedy možné ho popsat konečným souborem pravidel, nebo závisí i na aspektech které nelze formalizovat? Je myšlení redukovatelné na činnost neuronů? A pokud ano, znamená to, že je determinované? Nebo snad myšlení vyžaduje celé tělo v interakci s vnějším prostředím? Mohou mít jiné formy existence stejné myšlení jako my? A v neposlední řadě, je lidský typ myšlení jediným a nutným kritériem inteligence obecně?

Výzkum umělé inteligence v sobě zahrnuje mnohé z těchto otázek, a odpovědi hledá především skrze snahu o porozumění návrhu lidské mysli. Jedná se o snahu poznat strukturu a mechanismy mysli; ne z důvodu její analýzy, ale za účelem jejího opětovného sestavení v umělém médiu.[122] O inteligenci v plném

[120] DREYFUS, H. L. *What computers still can't do: a critique of artificial reason*. Introduction
[121] HILLIS, W. D. *Vzor v kameni: jednoduché myšlenky, které řídí počítače*. s. 80-81
[122] HAUGELAND, J. *What Is Mind Design?* s. 1

lidském rozsahu se ale prozatím nedá hovořit. Nicméně pomalu vznikají první zárodky fungujících univerzálních programů. A jelikož se jediný funkční předobraz univerzální inteligence nachází v lidské mysli, není překvapivé, že se mnohé proudy umělé inteligence inspirují nejen aktuálním stavem lidské mysli ale i jejím vývojem, ve snaze tento proces reprodukovat.[123]

3.1 Užitečnost a pozitiva této snahy (lidské myšlení jako jediný funkční model)

Už od dob Turinga je cílem většiny výzkumníků umělé inteligence vytvoření strojů, které se budou chovat a myslet jako lidé. Přirozeně, jiný vzor nemáme. Tato snaha sice nutně nezahrnuje celistvé pochopení lidské mysli, ale ptát se *jak by to a to dělali lidé* je nejjednodušším způsobem modelování umělého myšlení. V podstatě tak téměř vždy začínáme s více či méně povedenou simulací lidského myšlení nebo jeho vnějších projevů.[124]

Na základě výzkumů a pokusů z posledních pár desetiletí je patrné, že pokud vytvoříme obecnou umělou inteligenci, bude se s největší pravděpodobností zakládat na neuronových sítích, jejichž cílem je co možná nejvěrohodnější modelování funkcionality skutečných neuronů. Je přirozené, že se vývoj umělé inteligence bude stáčet směrem k napodobení jediného vzoru inteligence, který máme. To by ale nemělo vyvolávat přesvědčení, že nemohou existovat jiné struktury generující myšlení, ať už biologické či mechanické. Pouze to značí, že je o mnoho jednoduší tvořit umělou inteligenci podle předlohy než bez ní.

Současný pokrok v rámci neuronových sítí je tak pozoruhodný právě na základě použití idejí z neurovědy. Čerpání z neurovědy je přitom důležité ze dvou důvodů. Neurověda může ověřit techniky umělé inteligence, které už existují. Jinými slovy, pokud zjistíme, že algoritmy v umělých neuronových sítích

[123] BROOKS, R. A. *Intelligence without Representation*. s. 395
[124] WEIZENBAUM, J. *Mýtus počítače: počítačový pohled na svět*. s. 34

napodobují nějakou funkci mozku, je to jistým potvrzením, že tento algoritmus bude fungovat i v komplexních systémech. A navíc může neurověda poskytnout zdroj inspirace pro nové algoritmy i architektury umělých sítí.[125]

Vedlejším efektem vytváření umělé inteligence pomocí napodobování té lidské se tak stává, že nám neuronové sítě, na základě podobnosti s mozkovou strukturou, mohou posloužit jako cesta k poznání lidského myšlení. Dean Zipser předpokládá, že pokud je struktura sítí alespoň vzdáleně podobná reálnému uspořádání neuronů v mozku, pak je pravděpodobné že síť pomocí algoritmu minimalizujícího omyly (typicky zpětného šíření) dojde k řešení, které bude užitečné, protože bude podobné biologickému řešení.[126] Jednotky sítí jsou totiž daleko přístupnější než pravé neurony, a přesto je jejich základní funkcionalita obdobná. Díky tomu můžeme do jisté míry predikovat a ověřovat základní teorie o činnosti mozku. Porovnávání algoritmů s činnostmi lidského mozku je tudíž využíváno jak za účelem rozvoje umělé inteligence, tak i neurovědy.[127]

Ať už jsou ale algoritmy minimalizující omyly více či méně biologicky reálné, jsou užitečné zejména při samovolném vývoj neuronových sítí. Jak jsme viděli, existují sítě, které jsou schopné bez zásahu vnější autority dosáhnout požadovaných výsledků. Nejedná se zde o snahu přesně simulovat procesy v lidském mozku, ale spíše o snahu simulovat vývojový proces, který vedl až k dosažení autonomního myšlení, na základě schopností jako je učení, rozpoznávání apod. Ačkoliv tedy neuronové sítě vychází z modelu lidského myšlení, mohou nacházet i vlastní cesty k řešení problémů, které nejsou (kromě systémových omezení) zásadně vymezeny lidským poznáním nebo představivostí. S nadsázkou můžeme říci, že před našimi zraky probíhá jakási umělá forma evoluce. Sítě se vyvíjí metodou pokusu a omylu, adaptují se na aktuální podmínky a jejich hlavním cílem je úspěšné plnění úkolů. V průběhu tohoto procesu se postupně vyvíjejí a zdokonalují, podobně jako biologické organismy. I taková strategie může být prospěšná, jelikož samovolný vývoj systémů může vytvořit

[125] https://deepmind.com/blog/ai-and-neuroscience-virtuous-circle/
[126] CRICK, F. *Věda hledá duši: Překvapivá domněnka.* s. 200
[127] CHURCHLAND, P. M. *On the Nature of Theories: A Neurocomputational Perspective*. s. 272-276

zcela odlišné postupy, než by přineslo lidské programování. Můžeme proto očekávat, že takový vývoj dosáhne překvapivých výsledků, jenž z naší pozice nebudeme schopni předem předvídat.

3.2 Negativa (formalizace myšlení)

Ačkoliv jsme svědky prudkého rozvoje umělé inteligence, stále není zcela jasné, zda lze počítačům předat veškeré obsahy lidského myšlení. V předchozí kapitole jsme řekli, že se výzkumníci při vývoji umělé inteligence často tážou *jak by to a to dělali lidé*. Tato otázka je jistě opodstatněná, ale zároveň může do zkoumání umělé inteligence vnést řadu problémů. Pokud se chceme inspirovat lidským návrhem, musíme zároveň předpokládat možnost formalizace lidského myšlení. To ale vyvolává otázku: lze celé lidské myšlení skutečně popsat pomocí konečného souboru pravidel? Pokud při sázce na formalizaci uděláme chybu, bude to mít za následek slepou uličku ve vývoji umělé inteligence. Dost možná totiž nemusí být problém ve vytvoření umělé mysli jako takové, ale ve snaze zahrnout do ní veškeré lidské obsahy. Jistě není nepředstavitelné, že by mohl existovat i jiný druh myšlení. Zdá se ale nepravděpodobné, že by odlišná mysl byla schopna stejným způsobem zpracovávat veškeré obsahy, které jsou v lidské mysli.

Přesto se výzkumníci umělé inteligence často vydávají touto cestou a více či méně úspěšně řeší problém formalizace. Z každodenní zkušenosti vyplývá, že zatímco jsme už dnes schopni jisté aspekty lidské mysli formalizovat a vložit do počítače, jiné prozatím formalizovat neumíme. Dalo by se tedy říci, že z hlediska vývoje umělé inteligence můžeme nyní rozlišovat mezi dvěma složkami duševních procesů. První složka zahrnuje to, co je objektivně popsatelné. Typicky formalizovatelnou složkou mysli je logika, která může být vcelku jednoduše

implementována do strojů. Druhou složkou je potom vnitřní prožitek subjektu, který nejsme schopni plně vyjádřit slovy, natož ho vložit do počítače.[128]

V moderní vědě převládá důvěra ve vypočitatelnost skutečnosti, z každodenní zkušenosti ale vyplývá, že existují aspekty lidské mysli, které prozatím formalizaci unikají. Výsledek vědecké snahy o plnou formalizaci je nyní takový, že pokud chce věda něco vypovídat o lidských prožitcích, dělá to na základě jejich ztotožnění s poznatky získanými při pozorování vnějších projevů člověka. Ale takové ztotožnění je problematické, protože se intuitivně jeví jako neúplné. Přesto je společnost přesvědčena, že každý efektivní postup může být přenesen do počítače. A člověk i zbytek přírody fungují v zásadě efektivně, tudíž je možné jejich činnost simulovat pomocí počítače.[129]

Pokud budeme veškeré aspekty života chápat jako vypočitatelné, potom budou za jediné legitimní teorie o lidském chování považovány ty, se kterými lze počítačově manipulovat. To je ale podle kritiků umělé inteligence poněkud zúžené pojetí. Tvrdí přitom, že lidé mají kromě vědeckého popisu ještě jiný druh pojímání věcí, který je mimo standardy dokazatelnosti. A to je považováno za zásadní rozdíl mezi člověkem a strojem. Stroj podle nich není schopen dosáhnout této celistvosti zahrnující obavy, rizika, důvěru atd.[130]

Jedním z nejvýznamnějších kritiků v tomto směru byl Searle, který tvrdil že operace digitálního počítače lze formálně popsat. Jednotlivé kroky např. specifikujeme abstraktními symboly, které však nemají význam ani sémantický obsah. K ničemu se nevztahují. Ale mít mysl, podle něj, znamená víc než jen vykonávat formální procesy. Protože mentální procesy mají určité obsahy. A to znamená, že každý mentální stav má kromě formálních vlastností i nějaký psychický obsah. Dalo by se tedy říci, že mysl má kromě syntaxe i sémantiku. A počítače, dle této kritiky, pouze simulují formální vlastnosti duševních procesů. Jinými slovy Searle věřil, že ačkoliv mohou mít počítače shodný postup, přesto

[128] HAVEL, I. *Přirozené a umělé myšlení jako filosofický problém.* s. 17
[129] WEIZENBAUM, J. *Mýtus počítače: počítačový pohled na svět*. s. 20-26
[130] Tamtéž, s. 72-121

nemusí vykonávat tu samou činnost, protože jim schází duševní obsahy. Porozumění těmto obsahům je přitom propojeno s poznáním vnějšího světa. Řešení tohoto problému by tak mohlo spočívat ve vývoji univerzální inteligence, která bude zapuštěna do skutečného světa, z nějž bude nabývat své vlastní *duševní* obsahy.[131]

Pokud však chceme mluvit o univerzální inteligenci, kterou by bylo možné vypustit do světa, je nutné, aby se byla schopna vypořádat s obecnými problémy každodenní praxe. Bohužel hlavní proud umělé inteligence v minulosti vedl spíše k vytváření programů, které zvládaly pouze úzký segment problémů. Problémem většiny počítačů je, že v otevřeně strukturovaných problémech nedokážou rozhodnout, jaké skutečnosti jsou relevantní pro jejich řešení. Skutečnosti totiž nejsou relevantní na pevno, ale z hlediska účelů. Stroj, který by byl schopen zvládat obecně strukturované problémy otevřeného světa, musí nejprve sám rozpoznat relevanci dat pro řešení úlohy. Trénování obecné inteligence v otevřeném prostředí, ať už virtuálním či skutečném, se proto jeví jako slibná cesta ve vývoji umělé inteligence. Pokud totiž program není zasazen do bohatého prostředí, nebude schopen naučit se rozlišovat, jaké fakty jsou relevantní. A soubor všech faktů v jakékoliv situaci je natolik rozsáhlý, že pokud by měl počítač zkoumat relevantnost každého z nich podle předepsaných pravidel, nikdy by nedosáhl výsledku v reálném čase.[132]

Člověk umí vybírat mezi fakty možná proto, že je na rozdíl od stroje zasazen do světa a je vždy v určité situaci. V situaci typické pro něj jakožto člověka. I kdyby nějaký program reprezentoval veškeré lidské poznání a stereotypy, reprezentuje je pouze zvenčí, protože se nenachází ve stejné situaci jako člověk. Situace, ve kterých lidé jednají, jsou patrně od základu organizovány z hlediska lidských potřeb, které dávají skutečnostem jejich význam a dělají je tím, čím jsou. Přesto umělá inteligence často začíná až na úrovni již vytvořených faktů. Nicméně fakty vyňaté z kontextu se stávají jen masou neutrálních údajů. Oblast lidského

[131] SEARLE, J. R. *Mysl, mozek a věda*. s. 32-52
[132] DREYFUS, H. L. *What computers still can't do: a critique of artificial reason*. s. 168-171

poznání ale není neutrální, je strukturována našimi zájmy. Jinými slovy, pokud hrají nereprezentovatelné aspekty lidského života tak zásadní roli v situovanosti a situovanost je předpokladem lidského inteligentního chování, není možné simulovat celek lidského myšlení bez lidského těla a světa ve kterém by se pohybovalo.[133]

Na začátku kapitoly jsme si položili otázku, *jaké oblasti lidské činnosti mohou být formalizovány?* Vývoj umělé inteligence naznačuje, že je bez větších obtíží možné, digitálně simulovat aktivity, které mají jasná pravila, např. některé hry. Přirozený jazyk je zde nahrazen formálním a rozpoznávání vzorů je zpravidla podloženo předem definovaným seznamem znaků. Neformální oblast chování je potom ta, do které spadá většina aktivit každodenní lidské činnosti ve světě, které jsou pravidelné ale nelze jednoduše určit soubor pravidel, kterými se řídí. Problémy v této oblasti jsou obecně strukturované a vyžadují rozpoznání relevantního od irelevantního, ještě před samotným řešením problému. Výzkum umělé inteligence běžně nerozlišuje mezi těmito oblastmi a předpokládá, že veškeré lidské inteligentní chování je stejného typu. Čímž dochází ke zobecnění úspěchu z první oblasti na oblast druhou.[134]

Z výše uvedených poznatků vychází, že jelikož je lidská inteligence patrně závislá na situaci, ve které se nachází, není možné ji oddělit od zbytku lidského života. [135] Tato skutečnost přesto nemá moc společného s možností existence umělé inteligence jako takové. Cílem této úvahy není vyvrácení samotné možnosti formalizace lidského myšlení. Ale spíše poukázání na fakt, že vzhledem k (prozatím nepřekonaným) problémům s formalizací, nejjednodušší řešení zřejmě nespočívá v digitální simulaci aspektů lidské mysli, ale spíše v zasazení umělé inteligence do její vlastní situace s vlastními cíli. Se vší pravděpodobností tímto způsobem nedosáhneme inteligence, která by byla podobná nebo dokonce totožná s lidskou. Jiný kontext „života“, potřeby, cíle a zkušenosti bezpochyby přinesou zásadně odlišnou strukturu umělého poznání a myšlení. Přesto se snaha o vyvíjení

[133] DREYFUS, H. L. *What computers still can't do: a critique of artificial reason*. s. 172-174
[134] Tamtéž, s. 203-206
[135] DREYFUS, H. L. *From Micro-Worlds to Knowledge Representation: AI at an Impasse*. s. 181-182

umělé inteligence v kontextu jejích vlastních potřeb a cílů jeví jako slibnější varianta, než pokusy o plnou formalizaci lidského života.

3.3 Redukce myšlení na funkci neuronů

Přestože věda jednou možná odhalí veškeré procesy v lidském mozku a bude je schopna simulovat na stroji, otázkou zůstává, jak to celé vlastně souvisí s myslí. I kdybychom byli schopni do nejmenšího detailu okopírovat mozkové součástky a vykonávat s jejich pomocí veškeré známé procesy, stačilo by to k vytvoření plnohodnotné mysli?

Mysl byla odpradávna považována za podstatu člověka, která byla zastřená tajemstvím. Většina náboženských i filosofických směrů napříč historií se shodnou na existenci mysli. Snaha objevit podstatu člověka a tím pádem povahu mysli byla univerzální. Rozdíly se týkaly *pouze* její charakteristiky. Až dnes se ujímá názor, že mysl je pouze shlukem neuronů. Tato změna názoru je z velké části zapříčiněna rozvojem moderní vědy, která nestaví Zemi do středu vesmíru a člověka na vrchol tvorstva. Moderní neurověda zdánlivě nepotřebuje mysl, aby byla s to vysvětlit lidskou činnost. Základní myšlenkou západního smýšlení o člověku se tak stává, že lidské jednání, cítění atd. je výsledkem interakce nervových buněk.[136]

Narůstající obliba redukcionismu stojí jistě za povšimnutí. Jeho zastánci tvrdí, že jakýkoliv psychický stav koreluje s určitou činností neuronů, a tudíž je k porozumění mysli dostačující porozumět interakci neuronů. V tomto pojetí si samozřejmě pod myslí nelze představit jakousi podstatu člověka, ale spíše sled myšlenek a pocitů, které utváří lidskou psychiku. Myšlení, cítění a ostatní mentální fenomény jsou přitom kauzálním důsledkem mozkových procesů, které se odehrávají na úrovni neuronů. A jelikož jsou stavy vědomí v podstatě stavy mozku, není podle redukcionismu žádný důvod k tomu, abychom je nemohli považovat za objektivní fakty. To znamená, že chtějí vztah mozku a mysli vysvětlit

[136] CRICK, F. *Věda hledá duši: Překvapivá domněnka.* s. 13-17

pomocí termínů, které používáme k osvětlování ostatních fyzikálních jevů. Dle redukcionismu zbývá odhalit, jak přesně funguje spojení mezi jednotlivými biologickými procesy a psychickými fenomény. A poté zmizí veškerá tajemství obklopující fenomén mysli.[137]

3.4 Svoboda lidské mysli

Jsme svobodní? Tato otázka sužuje filosofy už od nepaměti. S příchodem moderní vědy, redukcionismu, materialismu a fyzikalismu se ale stává ještě více naléhavou. Pokud věda prokáže, že je naše myšlení a jednání způsobeno chováním vzájemně interagujících neuronů, představa svobodné mysli se rozplyne. Přitom právě rozdíl mezi svobodou a determinovaností je od počátku vývoje umělé inteligence pokládán za jeden z hlavních rozdílů mezi lidskými bytostmi a uměle vytvořenými stroji. Svoboda spojená s autonomním myšlením a rozhodováním založeným na intelektu – to je to, co člověku zdánlivě propůjčuje jedinečné místo v kosmu. Jak se ale na otázku svobodné vůle dívá současná věda? Mohou duševní stavy skutečně změnit cokoliv v reálném světě? Naše vnitřní hlas napovídá, že ano. Přesto je tato myšlenka v rozporu se současným vědeckým předpokladem, že vše, co se děje ve fyzickém světě musí mít fyzické příčiny. Tedy že náš svět je takový jaký je, protože mikrosvět který ho podkládá je takový jaký je. Z pohledu moderní vědy je chování celků vysvětlitelné z hlediska základních částí, jejichž chování zkoumá fyzika. Ale funguje to stejně i v případě lidského myšlení a jednání? Naše intuice radí, že se můžeme svévolně rozhodovat o svém chování a tím pádem se příčiny událostí nemohou nacházet jen v oblasti determinovaného mikrosvěta.[138]

Nicméně atraktivita fyzikalismu je silná. Zdá se totiž, že pokud odmítneme ideu dualismu, nezbyde nám nic jiného než přijmout nějakou z forem fyzikalismus, materialismu nebo naturalismu. Jinými slovy, pokud odmítneme, že

[137] SEARLE, J. R. *Mysl, mozek a věda*. s. 11-23
[138] COCKBURN, D. *An introduction to the philosophy of mind*. s. 76-77

mysl je nehmotná věc mimo fyzikální zákony, která člověku udílí svobodnou vůli, poté nám nezbývá mnoho variant. Pravděpodobně potom na základě vědeckých poznatků usoudíme, že mysl je mozkem, jehož chování je určeno chováním jeho částí. Která z těchto variant je přijatelnější?[139]

Dualismus je myšlence svobody nakloněn, jelikož mysl je tím, co člověku uděluje svobodnou vůli. Descartes věřil, že mysl je imateriální substance, která existuje zcela v odlišném světě od toho, který známe z každodenní zkušenosti. Tím pádem se na ni nevztahují ani fyzikální zákony materiálního světa. Představa byla následující: v duši se odehrávají duševní události, které způsobují fyzické události těla. Komunikace mezi tělem a duší přitom probíhá skrze šišinku mozkovou, která v závislosti na duševních stavech ovlivňuje pohyby v mozku, jenž mají za následek požadované pohyby těla. Duševní události nespadají pod fyzikální zákony a tím pádem nejsou determinovány stejně tak jako fyzické události. Kauzální řetězec, který vede ke konečnému chování lidské bytosti tedy začíná u šišinky, jejíž pohyb nemá kauzální příčinu. To by znamenalo že v podvěsku začínají fyzické události, které už nejsou dál fyzikálně vysvětlitelné. Descartes to však nevnímal jako porušování fyzikálních zákonů. Naopak věřil, že existují události, které nemohou být vysvětleny z hlediska fyzikálních zákonů právě proto, že jejich příčina nespadá do materiálního světa. Ačkoliv je představa imateriální mysli jednou z mála možností, jak člověka vyčlenit z deterministického světa a vysvětlit tak svobodnou vůli, zásadně koliduje s moderními vědeckými poznatky.[140]

Pokud věříme v existenci imateriální mysli, a přesto odmítneme Descartovu tezi, že by nehmotná mysl mohla vyvolávat fyzické stavy těla, získáme jistou formu epifenomenalismu. V rámci epifenomenalismu je duševní povaha věcí kauzálně určena fyzikální povahou. To znamená, že jakékoliv dvě události, které jsou stejné ve fyzických ohledech, musí být dle epifenomenalistů totožné i v těch duševních ohledech. Kauzální účinnost je v této doktríně jednosměrná: duševní

[139] COCKBURN, D. *An introduction to the philosophy of mind*. s. 86
[140] Tamtéž, s. 78-86

události jsou pouze epifenoménem, tedy stínem, který se odvíjí od dějů v mozku. Příčiny se vždy nacházejí v ne-duševní sféře a duševní i fyzické události jsou tak důsledky stejné příčiny – aktivity v mém mozku. Jinými slovy, veškeré příčiny lidské činnosti se nacházejí pouze v oblasti neuronů. A proto nemají duševní události žádný vliv na dění ve fyzikálním světě; mysl tak nemůže být základem svobodného jednání. Takové pojetí ale zásadně naráží na naše běžné chápání lidských bytostí. Je problematické říct, že duševní stavy jsou určeny fyzickými; tedy že jakékoliv dvě události, které jsou stejné ve fyzikálních ohledech jsou nutně stejné i v duševních ohledech. Znamenalo by to totiž, že kopie člověka by měla nutně ty stejné myšlenky jako originál. Ale naše myšlenky a prožitky jsou zajisté určeny i kontextem našeho života. Pokud tomu tak je, potom není možné odvodit duševní stav čistě z fyzického stavu těla, aniž bychom přitom nic nevynechali.[141]

Ačkoliv je v epifenomenalismu duševní kauzálně určeno fyzickým, neustále rozlišuje mezi těmito dvěma druhy událostí a chápeme je jako odlišné. Proti tomu se vymezuje fyzikalismus, který věří, že duševní je realizováno fyzickým. Mysl se tak stává ekvivalentem mozku. To co skutečně existuje, jsou základní částice a jejich vlastnosti, které popisuje fyzika. Protože vše, co se děje ve fyzickém světě, musí mít příčiny ve fyzickém světě. Jinými slovy svět je takový jaký je, protože mikrosvět je takový jaký je. Jak ale podotýká Cockburn, ze skutečnosti, že každý pohyb lze vysvětlit z hlediska fyzických příčin nevyplývá to, že celistvé chování je způsobeno nějakou sekvencí v říši mikro-entit. Nemusí zkrátka existovat vysvětlení komplexních pohybů z hlediska chování mikro-entit o nichž mluví fyzika. To vytváří jistý prostor pro příčiny jednání, které nemusí být chápány fyzikalisticky.[142]

V rámci fyzikalismu bychom např. mohli říci, že naše svoboda je omezena tím, že vše co děláme závisí na chemických látkách v mozku. Tudíž nikdo z nás není nikdy skutečně svobodný. Naše intuice jde však proti tomuto přesvědčení. Běžně přijímáme, že vše co děláme je závislé na fyzických či chemických

[141] Tamtéž, s. 80-84
[142] Tamtéž, s. 82-91

procesech v našich tělech. Přesto se ale dle Cockburna zdá, že tělesné stavy jsou spíše nutnou podmínkou nikoliv příčinou jednání. A mezi těmi je třeba rozlišovat. Máme totiž tendenci myslet si, že vědecké zkoumání může dokázat, že skutečné příčiny lidského chování nejsou kompatibilní s ideou svobody. Věda ukázala, že velká část našeho chování závisí na řadě fyziologických podmínek, které nemůžeme ovlivnit. To nás může přivést až ke smýšlení o lidském chování, jako o přírodním fenoménu, který lze předvídat a řídit. Ale při detailním studiu toho, jak závisí konkrétní formy chování na konkrétním fyziologickém stavu, věda neukazuje, že fyziologický stav je přímou příčinou chování.[143]

3.4.1 Determinovanost vs. nepředvídatelnost (nejen) strojového myšlení

Ačkoliv stále existují rozličné názory na svobodu jednání, lidská inteligence je obecně považována za jediný funkční model autonomního uvažování. Proto není překvapivé, že jsou při snaze o dosažení umělé inteligence napodobovány konkrétní funkce, struktura nebo alespoň vnější projevy mozku. Přes veškeré podobnosti moderního hardwaru či softwaru s lidským mozkem a myšlením však přetrvává otázka, zda mohou stroje vytvořené člověkem někdy dosáhnout vlastních kvalit. Co se týká vytvoření kopie biologického originálu, jistě to není nemožné. Neskrývá se ale v našich myslích něco víc, co není možné redukovat na procesy v mozku? Jak bylo řečeno, člověk odedávna intuitivně věří v existenci svobodné vůle, která podkládá veškerá naše rozhodnutí a vymaňuje nás tak z pout determinismu fyzikálního světa. Nemůže ale být naše svobodná vůle jen klamem, způsobeným nedostatečnou znalostí veškerých proměnných? Jak poznamenal Alana Turing, i když je prokázáno, že schopnosti jakéhokoliv stroje jsou omezené, je otázkou, zda schopnosti lidského intelektu jsou ve skutečnosti neomezené.

[143] Tamtéž, s. 129-135

Přesto naše přetrvávající víra v lidskou svobodu dává vzniknout námitkám proti umělé inteligenci typu Lady Lovelace – tedy že stroj nemůže dosáhnout lidských kvalit, jelikož nemůže nikdy nic sám začít. Nemá vlastní iniciativu k tvořivosti, tak jako člověk. Stroj se nerozhoduje o svých činech, může vykonat pouze to, co mu přikážeme.[144] V době Turinga to jistě byla uvážená námitka, z dnešní perspektivy se však její význam vytrácí. U komplexních počítačových systémů se již nedá říct, že programy vykonávají pouze to, co je jim explicitně sděleno. Jakmile program dosáhne jisté míry složitosti, mohou jeho části vzájemně interagovat a vytvářet tak nepředvídatelné výstupy. V tomto smyslu nás chování samotného systému může jistě překvapit.

Přesto tento typ námitek zaznívá i v dnešní době. Např. Joseph Weizenbaum připouští možnost, že lidské chování je zapříčiněno řetězcem výpočtů uskutečňovaných v rozhodovacích uzlech. Zároveň však zdůrazňuje, že i přesto existuje distinkce mezi mechanickým a biologickým jednáním, která je založena právě na tom, že zdrojem mechanického jednání je vždy příkaz, zatímco lidské jednání stojí na rozhodnutí. A právě tato odlišnost podkládá možnost svobodné vůle.[145]

Ale je lidské rozhodování skutečně odlišné od jednání počítače, který se řídí pravidly vymezující jeho činnost? I když připustíme Cockburnovu myšlenku, že je rozdíl mezi nutnými podmínkami a příčinami k jednání, vše v námi známém vesmíru bude stále vymezeno (i když ne zapříčiněno) fyzikálními zákony. V tomto ohledu jsme fyzikálními zákony omezeni stejně tak, jako jsou počítače omezeny jejich technickým návrhem.[146]

[144] TURING, A. *Computing Machinery and Inteligence*. s. 450
[145] WEIZENBAUM, J. *Mýtus počítače: počítačový pohled na svět*. s. 100-101
[146] HILLIS, W. D. *Vzor v kameni: jednoduché myšlenky, které řídí počítače*. s. 72-73

3.4.2 Reálná nepředvídatelnost systémů

Přestože jsou počítačové systémy vystavěny tak, aby vykonávaly předem určené funkce, je tu jistá nepředvídatelnost reálného výsledku. Tato nepředvídatelnost může mít jak pozitivní, tak negativní důsledky.

S pozitivními důsledky nepředvídatelnosti programů se inženýři setkávají běžně při použití heuristik. V počítačových programech se nevyskytují pouze algoritmy, s přímými příkazy, ale i tzv. heuristické algoritmy, které zastupují algoritmy v úlohách, na které není možné získat absolutně správnou odpověď (popřípadě by její nalezení trvalo příliš dlouho). V takovém případě provádí počítač, namísto přesných výpočtů, promyšlené odhady. „Používá-li počítač heuristiky, je schopen jak překvapení (příjemných), tak chyb – čímž se z něj stává trochu více člověk a méně stroj.“[147]

Kromě zajímavých výsledků mohou z nepředvídatelných typů chování vzejít i nepředvídatelné chyby. Programátor je schopen eliminovat jakoukoliv chybu, vyskytující se v programu, musí ale tuto chybu dopředu očekávat. Na předpověď následků veškerých interakcí je však lidská fantazie nedostatečná. Nedokonalost a chybovost počítačů je tedy zapříčiněna nedostatečností softwarového (a méně často i hardwarového) návrhu. Počítače jsou tím nejkomplexnějším lidským výtvorem. A naše technologické postupy prozatím nejsou na dostatečné úrovni, potřebné k bezchybnému návrhu takto komplexních objektů. Nejsme reálně schopni předpovědět následky milionů současně probíhajících, logických operací a jejich následných kombinací. „Pokud se objeví nepředvídané interakce (a ony se objeví), pak předpoklady, na nichž abstrakce stojí, selžou a následky mohou být katastrofické. Chování velkého počítačového systému, i když nedojde k chybám, je někdy nepředvídatelné – a to je hlavním důvodem, proč není možné navrhnout dokonale spolehlivý počítač.“[148]

[147] HILLIS, W. D. *Vzor v kameni: jednoduché myšlenky, které řídí počítače*. s. 94
[148] Tamtéž, s. 108

Důsledkem složitosti systémů a vzájemné interakce částí se tedy stává, že i v případě kdy se neobjeví žádná chyba a jednotlivé části správně vykonávají svou funkci, se mohou vyskytnout předem nepředpokládané interakce. Čím komplexnější program je, tím rychleji přibývají nepředvídané následky interakcí. Pokud je systém dostatečně komplexní, nelze mít o všech jeho potencionálních interakcích a následcích přehled. Jelikož však víme, že existuje i o mnoho složitější systém, který pracuje takřka bezchybně, je jisté, že chyba není v neproveditelnosti tohoto úkolu ale v nedostatečnosti technických návrhů softwaru. Systém, o kterém mluvíme, je samozřejmě lidský mozek - ten je i přes značnou složitost spolehlivější a jeho dispozice ke kolapsům menší. Klasické umělé konstrukce jsou oproti biologickým náchylné ke katastrofálním selháním, jelikož mají striktní hierarchickou strukturu, kterou biologické uspořádání postrádá. Evoluce, navrhla dokonalý model, kterému se snažíme přiblížit. Systémy jsou sice stavěny tak, aby se chybám zamezilo, nemůžeme však zamezit chybám nepředvídatelným. Hlavním problémem tak zůstává, že fungování systému nelze v jeho celistvosti porozumět dostatečně na to, abychom byli schopni předvídat všechny potencionální důsledky.[149]

Jak již bylo řečeno, nepředvídatelné interakce ale mohou podnítit i velmi zajímavé, neočekávané výsledky a produkovat tak zdání nedeterminovaného chování. Daniel Hillis předpokládá, že takové chování se bude rozvíjet nejrychleji v rámci růstu internetové sítě: „Věřím, že se Internet nakonec rozroste o počítače zabudované v telefonních systémech, automobilech a jednoduchých domácích spotřebičích. Stroje budou své vstupy číst přímo ze skutečného světa a nebudou závislé na lidech jako prostřednících. Tak jak se informace dostupné na Internetu obohacují a interakce mezi propojenými počítači se stávají složitějšími, předpokládám, že Internet začne vykazovat vlastní chování, které půjde mnohem dále, než bylo explicitně v systému naprogramováno. (…) Jakmile si počítače na síti začnou vyměňovat vzájemně působící programy místo pouhé elektronické

[149] Tamtéž, s. 139–140

pošty, domnívám se, že se Internet bude chovat méně jako síť a více jako paralelní počítač. Tuším, že vlastní chování Internetu bude mnohem zajímavější.“[150]

3.4.3 Determinovanost programů: je lidské myšlení jiné?

V předchozích podkapitolách jsme se zabývali možností nepředvídatelného jednání u determinovaných systémů a zjistili jsme, že tato problematika se týká spíše rozdílu mezi reálnou nepředvídatelností a naší neznalostí zákonů určujících výsledek. Vezměme toto téma z opačného konce a zamysleme se nyní nad tím, zda se tento fakt vztahuje pouze na námi vytvořené dílo, nebo obsahuje i zbytek hmotného světa, včetně člověka samotného. Nebyla by pak svobodná vůle jen klamnou představou? A má tato představa ještě vůbec místo v našem mechanickém chápání kosmu? Jak píše Polanyi „vědecký světový názor stvořil mechanické pojetí člověka, v němž už není místa pro vědu samu (…) toto pojetí popřelo jakoukoliv vnitřní sílu myšlení, a tak popřelo jakékoli důvody pro svobodu myšlení“.[151]

Takové pojetí člověka se šířilo ruku v ruce s vývojem umělé inteligence. Člověk pravděpodobně začíná pociťovat ztrátu svého výsadního postavení nejvíce v momentě, kdy se zdá, že myšlení již nadále nebude pouze lidskou výsadou. To vše přišlo s prvními oboustranně komunikačními stroji, počínaje Weizenbaumovou ELIZOU. Lidé ji viděli jako přelom v umělé inteligenci, přestože od pravé inteligence měla ještě velmi daleko. Přesto bylo v ovzduší tehdejší doby cítit očekávání, že stroje by v dohledné době mohly zastat některé lidské funkce. Překvapivě mezi ně byly zařazeny i funkce zahrnující běžné lidské chování a myšlení. Colby napsal článek, ve kterém přirovnává lidského psychoterapeuta k „zařízení pro zpracování informací a rozhodování s danou sadou rozhodovacích pravidel (…) Při tomto rozhodování je veden hrubými

[150] Tamtéž, s. 120
[151] POLÁNYI, M. *The Tacit Dimension*. s. 3-4

empirickými pravidly, která mu říkají, co je vhodné říkat v jistých kontextech, a co ne. Zabudovat tyto procesy na úrovni odpovídající lidskému terapeutovi do programu bude značně složité, my se však pokoušíme tímto směrem postupovat.“[152]

Z tohoto názorového prostředí vznikla debata se dvěma stranami: na jedné straně stojí názor, že počítače by měly a časem budou vykonávat vše, co může vykonat člověk. A na druhé straně jsou ti, kteří věří, že funkcionalita počítačů bude vždy značně omezena. Když se dostaneme do jádra této diskuze, jde v podstatě o otázku redukovatelnosti lidského myšlení na rozsáhlý ale omezený soubor pravidel. Před prudkým vývojem moderní vědy a technologií tu bylo místo pro lidskou autonomii a svobodu rozhodování. Ale kosmologie vytvořená moderní vědou je založena na kauzální nutnosti. Vše co se děje, musí mít příčiny a následky, děje se to nutně.[153]

Pokud přijmeme názor, že lidské myšlení a z toho vzešlé chování se odvíjí od sady pravidel, můžeme člověka definovat jako stroj. Pak by veškeré hranice mezi mechanickými a biologickými stroji padly. Tento pohled zastává Francis Crick, který věří, že člověk je v podstatě biologický stroj. V jeho pojetí je lidské jednání a myšlení určeno chováním jednotlivých neuronů a jejich vzájemných interakcí. Crick zastává názor, že chování mozku (nebo přinejmenším jeho obecné principy) bychom tudíž měli být schopni pochopit z chování velkého počtu neuronů a ze znalosti jejich interakcí.[154] Vazby a chování jednotlivých neuronů jsou přitom alespoň z počátku určeny geny, které jsou pro změnu určeny zkušenostmi našich předků. Jiné získáváme vlastní zkušeností. Oba druhy jsou stabilní. Detaily zapojení jednotlivých neuronů jsou přitom patrně ovlivněny učením.[155]

Jakkoliv je Crickův popis mozku technický, je tu prostor i pro poněkud nezvyklý návrh „svobodné vůle“. Crickova představa „svobodné vůle“ stojí na

[152] COLBY, K. M., J. B. WATT a J. P. GILBERT. *A Computer Method of Psychotherapy*. s. 148–152
[153] WEIZENBAUM, J. *Mýtus počítače: počítačový pohled na svět*. s. 18-19
[154] CRICK, F. *Věda hledá duši: Překvapivá domněnka.* s. 21
[155] Tamtéž, s. 211

ideji, že část našeho mozku se zabývá plánováním potencionálních akcí, čehož si můžeme a nemusíme být vědomi. Nicméně si nejsme vědomi výpočetní aktivity oné oblasti mozku, pouze konkrétních rozhodnutí, která vyprodukuje. Výpočetní činnost závisí na struktuře této oblasti, která je tvořena z části geneticky, z části na základě zkušeností, jakož i interakci s ostatními oblastmi mozku. Jsme si tudíž vědomi pouze rozhodnutí, nikoliv procesu, který k nim vedl. To v nás vyvolává představu o svobodné vůli, přestože technicky vzato jde o pouhou iluzi, která je způsobena nevědomými procesy v mozku.[156]

Zajímavé na této domněnce je, že není nijak omezena pouze na lidské myšlení. Crick je přesvědčen, že „Bude-li mít stroj (…) tyto vlastnosti, bude mu připadat, že má svobodnou vůli, za předpokladu, že svou osobnost může personifikovat – tj. že bude mít představu „jáství“. Bezprostřední příčina rozhodnutí může být jasně vyjádřena (…) nebo může být deterministická, nicméně chaotická – velmi malé proměny, mohou způsobit obrovské rozdíly v konečném výsledku. To dá vůli „svobodné“ vzezření, neboť právě toto bude v podstatě příčina nepředpověditelnosti výsledku. Vědomé činnosti mohou samozřejmě rovněž mechanismus rozhodování ovlivnit (…) Stroj toho druhu se může pokusit vysvětlit si, proč učinil určitý druh volby (za pomoci introspekce).“[157] Je možné, že v takovém případě odhalí pravdu. Pokud ne, bude si kvůli nevědomosti těchto procesů pravděpodobně vymýšlet zástupnou příčinu volby.

3.4.3.1 Determinovanost jako základní rys fyzikálního světa

Běžnou námitkou proti myšlení u umělé inteligence je fakt, že zatímco lidské jednání je neformální, stroje se řídí pouze pomocí pravidel. Jak jsme ale viděli, je pravděpodobné, že lidské jednání je jako vše ve fyzikálním světě usměrněno sadou přírodních zákonů. Činnost člověka je s největší

[156] Tamtéž, s. 263–264
[157] Tamtéž, s. 264

pravděpodobností řízena mozkem a nervovou soustavou, což jsou ve své podstatě běžné fyzikální objekty, spadající pod fyzikální zákony. Jak říká Daniel Hillis „Možná že jsme zvířata, ale náš mozek je v určitém smyslu stroj (…) Myslím, že vědomí je důsledkem působení normálních fyzikálních zákonů a důkazem složitého výpočetního postupu, ale to pro mě neznamená, že je vědomí méně záhadné nebo nádherné – spíše naopak. Mezi signály našich neuronů a pocity v našich myšlenkách je tak velká mezera, že se ji možná nikdy nepodaří lidským porozuměním překlenout. Říkám-li tedy, že mozek je stroj, nemá to být urážka mysli, ale ocenění potenciálu stroje. Nemyslím si, že lidská mysl je méně, než za co ji pokládáme, ale že stroj může být mnohem, mnohem více.“[158]

Také bychom se mohli vzdálit od kauzálního determinismu a říci, že stroj není o nic více determinován technickým návrhem, než je člověk determinován genetickým návrhem. S myšlenkou genetické determinace přišla sociobiologie, v čele s Edwardem Wilsonem, který tvrdil, že biologie v určitém smyslu určuje náš osud. Máme ale potom vůbec možnost svobodných rozhodnutí? „Jestliže naše geny jsou zděděné a jestliže naše prostředí je řetěz fyzikálních událostí, uvedených do pohybu před naším narozením, jak může v mozku existovat skutečně nezávislý činitel? Činitel sám je tvořen interakcí genů a prostředí. Ukázalo by se, že naše svoboda je sebeklam.“[159]

Pokud jsou obecné rysy našeho druhu determinovány neboli zakódovány v našich genech, zároveň to znamená existenci pevných a popsatelných faktů o lidské přirozenosti. Pokud bychom byli schopni vypočítat veškeré počáteční podmínky, bylo by možné předpovědět hrubé obrysy našeho „osudu“. I když je lidské chování velmi složité a komplexní, je podle sociobiologie determinováno a limitováno genetikou a oblastí fyzikálních jevů. To omezuje počet možných výsledků. Lidská psychika je ale v rámci našich aktuálních možností příliš složitá na to, aby se dala předvídat. Představa svobody je tedy v tomto ohledu fixována na neznalost psychických procesů, nicméně naše chování je přesto determinováno

[158] HILLIS, W. D. *Vzor v kameni: jednoduché myšlenky, které řídí počítače*. s. 148
[159] WILSON, E. O. *O lidské přirozenosti: máme svobodnou vůli, nebo je naše chování řízeno genetickým kódem?* s. 75

charakteristikou lidského druhu. A toto předurčení, k jistým typům chování než jiným, je určeno evolucí a přirozeným výběrem. Nejde ale o determinaci jednotlivých rozhodnutí, spíše o rastr chování typický pro příslušný druh.[160]

Základním předpokladem tak zůstává, že vše ve vesmíru je vymezeno fyzikálními zákony. A lidské myšlení a jednání je také do jisté míry určeno genetickými předpoklady. Moderní věda nám předkládá obraz světa, kde každá příčina má svůj následek a veškeré jevy jsou fyzikálně, či biologicky popsatelné. Pokud bychom tedy znali veškeré proměnné, mohli bychom na jejich základě předvídat následky. Zrovna tak „digitální počítače jsou předvídatelné i nepředvídatelné – stejně jako zbytek hmotného světa. Řídí se deterministickými zákony, ale tyto zákony mají složité důsledky, které je velmi těžké předpovědět.“[161] Ze souhrnu výše zmíněných tezí vyplývá, že není důvod zavrhovat vyhlídky na autonomní umělou inteligenci v důsledku nepřítomnosti „svobodné vůle“.

3.5 Člověk jako kritérium myšlení. Opravdu?

Pokud chceme zjistit, zda má mysl kdokoli nebo cokoliv kromě člověka, je přirozené se nejprve ptát, zda má mysl podobnou té naší. Protože to je ta jediná, o které něco víme. Pokud by byla jiná mysl odlišná, dost možná bychom ani nebyli schopni určit, zda tento jiný druh mysli někdo má, pokud bychom ho sami neznali. Naše mysl se tak stává standardem, ze kterého vycházíme. Určení toho, zda někdo či něco má mysl, je přitom zásadní pro naše vnímání takových bytostí. Vlastnost *mít mysl* totiž poskytuje jakousi záruku určitého morálního statusu. Hluboko v našem myšlení je zakořeněno, že pouze ti co mají mysl, mohou mít nějaký zájem

[160] WILSON, E. O. *O lidské přirozenosti: máme svobodnou vůli, nebo je naše chování řízeno genetickým kódem?* 76–82
[161] HILLIS, W. D. *Vzor v kameni: jednoduché myšlenky, které řídí počítače*. s. 73

na vlastním bytí. A jako takoví jsou v protikladu ke zbytku nemyslícího vesmíru. To je jeden z důvodů, proč se nás otázka jiných myslí tak zásadně dotýká.[162]

Lidská mysl se, stejně jako všechny ostatní aspekty lidské bytosti, patrně vyvinula pod určitými selekčními tlaky a její aktuální stav je tudíž závislý na prostředí ve kterém se vyvíjela. Pokud tedy hledáme nebo se aktivně pokoušíme vytvořit mysl u strojů, měli bychom vycházet z jejich přirozeného prostředí. Na základě zkoumání vývoje vlastní mysli předpokládáme, že se složitost myšlení vyvíjí současně se složitostí prostředí. Možná bychom, místo snahy formalizovat a převádět lidské myšlení do strojů, měli pracovat na zasazení autonomní umělé inteligence do bohatého prostředí. Zdá se totiž daleko pravděpodobnější, že nalezneme zárodky autonomního myšlení u programu, který prošel vlastním vývojovým procesem než u programu, do kterého se programátoři snažili vtisknout typicky lidské charakteristiky. Otázkou zůstává, jestli takové procesy, které nebudou odpovídat lidskému myšlení, budeme vůbec považovat za myšlení. Tu a tam máme totiž tendenci zapomínat, že náš způsob smýšlení o světě není jediný možný a neměl by tedy být považován za nutný předpoklad myšlení jako takového.[163]

Přesto se v tradiční umělé inteligenci běžně předpokládá, že počítače, roboti atd. vnímají svět stejně jako lidé a měli by s ním i stejně zacházet. Jinak řečeno, jsme to my, kdo hodnotí co je správné a co není. To samozřejmě není problém, pokud stroje konstruujeme za účelem rozšíření vlastních schopností. Pokud ale chceme vytvořit plně autonomní systém, který bude schopen myslet a „žít vlastním životem“, není to adekvátní. Konec konců i v přírodě existuje mnoho jiných druhů, které vnímají svět zcela odlišně od nás. Protože náš zkušenostní svět je nepochybně spjat s naší tělesnou konstrukcí, která tak určuje a limituje naše poznání a myšlení. Poznání, které získáváme ze světa skrze vlastní tělo, není objektivním poznáním nějakého předem daného světa. Naopak, svět, který

[162] DENNETT, D. C. *Druhy myslí: k pochopení vědomí*. s. 13-15
[163] Tamtéž, s. 121-122

poznáváme a o kterém smýšlíme, je tvarován naším specifickým způsobem vnímání.[164]

Odlišná inteligence tudíž nemusí a patrně ani nebude poznávat a vnímat svět jako my. Tím pádem nemůže mít ani stejně strukturované myšlení. Ačkoliv se tedy hlavní proud umělé inteligence pokouší o napodobení lidské mysli, protože jiný funkční vzor nemá, odlišná mysl bude s největší pravděpodobností formována zcela odlišným způsobem. Aby vznikla komplexní mysl, nikoliv pouze nedokonalá napodobenina imitující vnější znaky, musí projít vlastním vývojem, který bude určen jejími cíli a potřebami v daném prostředí. Ani jedno z toho však nemůže být naprogramováno zvnějšku člověkem. Ten může pouze poskytnout vhodné prostředí k jejímu vlastnímu rozvoji. Jak tvrdil už Hillis, náš největší přínos bude pravděpodobně v nastolení správných podmínek pro vznik umělé inteligence.

Zda se takový systém vyvine až v komplexní mysl poznáme především podle toho, zda bude schopen samostatně fungovat ve svém prostředí. Přesto je výzkum klasické umělé inteligence ovlivněn přesvědčením, že systém bude považován za inteligentní, pouze pokud bude schopen vyhovět Turingovu testu, nebo jemu podobným; tedy pokud bude jednat a myslet jako člověk. Zda je tento cíl splnitelný, je empirickou otázkou. Zda je prospěšný, to je ale zcela jiná věc. Jak jsme již řekli, taková forma inteligence by nebyla ničím jiným než neúplnou imitací. Stroj nemůže získávat stejné zkušenosti a poznání jako člověk. Můžeme mu je sice zprostředkovat, ale podle všeho nebudou kompatibilní s tělem a historií počítače. Ve výsledku tak bude počítač schopen napodobit pouze některé vnější projevy lidské inteligence. Kdybychom však zkonstruovali systém, který by prošel vlastním vývojem, patrně by byl vybaven odlišnými prostředky vnímání a jednání, formující pro něj specifický druh mysli.[165]

Tradiční umělá inteligence možná trvá na svém návrhu napodobení lidského myšlení jen kvůli zjednodušenému pojetí lidské inteligence. Už samotné

[164] HAVEL, I. *Přirozené a umělé myšlení jako filosofický problém.* s. 37-38
[165] Tamtéž, s. 39-40

měření inteligence na jakési pomyslné lineární škále, je velmi zavádějící představa. Takové pojetí inteligence je značně omezující a především neúplné. Měří se pouze ty rysy, které jsou důležité z pohledu západní společnosti. Inteligence ale není trvalá ani kulturně neměnná. Naopak, inteligence je kulturně orientovaná a nelze ji proto nijak obecně porovnávat. Běžná představa je ale opačná a do jisté míry je základem zavádějící otázky, zda mohou být počítače inteligentní. Pokud vezmeme v potaz, že nelze celistvě porovnat inteligenci jedinců ze dvou naprosto odlišných kultur, pro odlišné druhy to bude platit dvojnásob. Každý druh řeší odlišné problémy, které jsou nutné k přežití v prostředí, do nějž je zasazen. Odlišné druhy tudíž musí mít odlišné poznání, závislé na jejich prostředí a vnímání světa. Čili, i kdyby počítače prošli vývojem, který by vyústil v inteligenci, byla by jejich inteligence zcela odlišná od té naší.[166]

3.5.1 Zamyšlení se nad možností odlišného druhu inteligence

Jak by tedy měla vypadat umělá inteligence, která se nebude pokoušet striktně napodobit lidské myšlení? Obecně přijímanou moderní tezí je funkcionalismus, který říká, že to co dělá mysl myslí, je závislé na tom co může vykonávat. Jakákoliv mysl, alespoň z toho co víme, by měla být schopna zpracovávat informace a na základě toho řídit jednání organismu. K tomu je potřeba vnímat okolní prostředí a pohybovat se v něm za účelem přežití. A zdá se, že právě schopnost plnit úkoly spojené s přežitím v dynamickém prostředí, poskytuje základ pro vývoj inteligence. Za účelem vytvoření autonomní inteligence je proto patrně nezbytně nutné, zasadit ji do dostatečně bohatého a dynamického prostředí. V takovém prostředí bude umělá inteligence nucena vyrovnávat se se změnami prostředí pomocí rozšiřování svého poznání a

[166] WEIZENBAUM, J. *Mýtus počítače: počítačový pohled na svět*. s. 74-84

postupnou adaptací. Také by měla být schopna zachovávat více cílů a v závislosti na okolnostech aktivně usilovat o různé z nich.[167]

K tomu je nutné, aby měla vlastní intencionalitu. Zaměřenost je to, co systému umožňuje jednat s ohledem na vlastní cíle, tedy jednat racionálně. Pokud se umělý systém chová samostatně, racionálně a konzistentně za rozličných okolností, pak můžeme říci, že má zaměřenost.[168] A to mimo jiné znamená, že je schopen sám určit činnosti nutné k udržení vlastní existence. A tak se v zásadě stává autonomním.

Pro dosažení takové umělé inteligence je třeba, aby byla umístěna v konkrétním prostředí. V ideálním případě by umělá inteligence měla být zasazena do reálného světa. Cílem je vytvoření robota, který by byl schopen vnímat a pohybovat se v reálném světě, na nějž by reagoval. Protože lidmi zprostředkované poznání světa nebude nikdy dostatečné pro vývoj umělé mysli. Searle měl jistě pravdu v tom, že agent může porozumět významu věcí a událostí jen když je zasazen do prostředí, kde se tyto věci vyskytují. Zasazení umělé inteligence do reálného světa ale brání jak obtíže při vyvíjení plně multifunkčních robotů, tak etické otázky spojené s vypuštěním nepředvídatelných strojů mezi lidi. Proto se prozatím musíme spokojit s vývojem umělé inteligence ve virtuálním prostředí. Čím bohatší prostředí, tím spíše se umělá inteligence stane univerzální. Tak jako tak je vždy nutné, aby se program pohyboval v prostředí, ve kterém vyvstávají nové situace, na něž musí reagovat, a kde má jeho jednání kauzální účinky.

Musíme přitom předpokládat, že vývoj inteligence ve virtuálním prostředí (ve kterém budou odlišné podmínky od našeho světa) bude mít za následek vznik daleko vzdálenější formy inteligence než té, která by se vyvinula ve stejném prostředí jako lidská inteligence. Vývoj systému probíhající ve zcela rozdílném prostředí bude jistě usměrněn odlišnými potřebami a tím bude formovat i odlišný typ myšlení a vnímání. Přestože by se u takové umělé inteligence nevyvinula stejná

[167] BROOKS, R. A. *Intelligence without Representation*. s. 396-402
[168] HAUGELAND, J. *What Is Mind Design?* s. 5-6

vnitřní struktura nebo vnější projevy myšlení, jako máme my, nepopírá to její inteligenci jako takovou. Inteligence každého druhu se nutně vyvíjí v kontextu pro něj specifického vnímání a poznávání světa. A naše vnímání světa rozhodně není objektivně správnější nebo lepší než potencionální vnímání, poznávání a z toho vyplývající myšlení umělých bytostí.

4 Závěr

Ve své diplomové práci jsem přezkoumala vznik a historii ideje umělé inteligence. Následně jsem shrnula nejpoužívanější strategie a technologie využívané při vytváření umělé inteligence. Poté jsem se obrátila k současným snahám o rozvoj obecné umělé inteligence, která by mohla být základem autonomní mysli. Na základě tohoto bádání jsem porovnala lidské a strojové aspekty myšlení a ve výsledku se zaměřila na užitečnost vývoje umělé inteligence podle vzoru, který nalézá v lidské mysli.

V úvodu práce jsem si položila několik otázek, ne všechny však mohu na základě svého zkoumání zodpovědět. Ačkoliv se pokrok ve vývoji umělé inteligence neustále zrychluje a je už možné simulovat některé z aktivit a schopností, které jsou považovány za typicky lidské, není stále zcela jasné, zda některé z nich stačí pro myšlení jako takové. Počítače se mohou samovolně učit, a co víc, vykazují při učení podobné tendence jako člověk. Umí plánovat své jednání, dokonce získávají představivost. Ale nezdá se, že by některý z těchto výkonů byl sám o sobě dostatečný k přisouzení myšlení. Základním rysem mysli je patrně její schopnost adekvátně reagovat na veškeré problémy vycházející z prostředí, ve kterém se vyskytuje, a to bez omezeného pole působnosti. Vytvoření obdobně obecné umělé inteligence je nepochybně mnohem složitější než simulace jednotlivých schopností.

S ohledem na historické pokusy a aktuální stav výzkumu musím říci, že simulaci obsahů lidské mysli nebo procesů lidského mozku nepovažuji za užitečnou cestu při snaze o vývoj komplexní umělé mysli. Vzhledem k tomu, že nejsme schopni (přinejmenším nyní) formalizovat veškeré obsahy lidské mysli, úsilí o vložení jakékoliv formální struktury lidského myšlení do stroje, by pravděpodobně mělo za následek pouze neúplnou imitaci vnějších projevů. Podobná situace nastává i při pokusech o vymodelování neuronové sítě, která by zrcadlila veškeré procesy lidského mozku. Naše myšlení, zdá se, není

redukovatelné pouze na interakce mezi neurony. Svou roli hrají i další faktory jako např. tělo, nebo prostředí ve kterém se nachází.

Aktuální stav lidské mysli je zcela jistě závislý na vývoji, kterým prošla. Strukturu našeho myšlení formovaly potřeby a cíle, jež se utvářely v souladu s prostředím, ve kterém jsme se po miliony let vyvíjeli. Jinými slovy, lidská mysl je spojena s lidskou historií, tělem i prostředím. Myšlení proto nelze oddělit od zbytku lidské bytosti a vložit jej v nepozměněné formě do odlišného média. Budoucnost výzkumu umělé inteligence proto nevidím ve splnění nerealistického požadavku na vytvoření umělé inteligence, která by vnímala svět tak jako člověk a měla stejné vnější projevy. Naopak, pokud skutečně jednoho dne vznikne autonomní umělá mysl, patrně bude muset projít svým vlastním vývojem. Náš příspěvek k tomto vývoji může spočívat v poskytnutí příhodného prostředí a zahájení umělé evoluce. Výsledkem takového procesu ale pravděpodobně bude zcela odlišný druh myšlení, než je ten náš. Otázkou pro další diskuzi zůstává, zda je takový výsledek naším cílem.

Vedlejším efektem mého výzkumu se tak stává, že hlavní přínos pokusů o přesné modelování struktury a procesů lidského mozku nevidím v potencionálním vytvoření umělé mysli. Ale spíše v prohloubení znalostí o lidské mysli a mozku, jelikož se tento druh výzkumu ukázal být nápomocný při potvrzování neurovědeckých teorií. Pokud však chceme vytvořit skutečně autonomní mysl, nikoliv nástroj sloužící k sebepoznání nebo usnadnění vlastních životů, nezdá se příliš prospěšné, formovat ji podle předobrazu lidské mysli.

5 Prameny a literatura

Prameny

CRICK, Francis. *Věda hledá duši: překvapivá domněnka*. Praha: Mladá fronta, 1997. Kolumbus, sv. 136. ISBN 80-204-0633-6.

Dreyfus, H. L. *What computers still can't do: a critique of artificial reason*. New York: MIT Press, 1992. ISBN 06-011082-1.

HAUGELAND, J. What Is Mind Design? In: *Mind design II: philosophy, psychology, artificial intelligence*. Cambridge: MIT Press, 1997, s. 1-28. ISBN 0262581531.

HILLIS, W. D. *Vzor v kameni: jednoduché myšlenky, které řídí počítače*. Praha: Academia, 2003. Mistři vědy. 158 stran. ISBN 80-200-1067-x.

SAYGIN, A. P., CICEKLI, I. a AKMAN, V. (2000): Turing Test: 50 Years Later. *Minds and Machines* 10, 463-518.

SEARLE, J. R. *Mysl, mozek a věda*. Přeložil Marek NEKULA. Praha: Mladá fronta, 1994. Váhy (Mladá fronta). ISBN 80-204-0509-7.

TURING, A. M. Computing machinery and intelligence. *Mind*. Oxford University Press, 1950, 59(236), 443-460.

WEIZENBAUM, J. *Mýtus počítače: počítačový pohled na svět*. Přeložil Jiří Fiala. Břeclav: Moraviapress, 2002. Knihovna Ceny Nadace Dagmar a Václava Havlových VIZE 97. ISBN 80-86181-55-3.

Literatura a internetové zdroje

BARRESI, J. Prospects for the Cyberiad: Certain Limits on Human Self-Knowledge in the Cybernetic Age. *Journal for the Theory of Social Behavior*. 1987, (17), s. 19–46.

BLOCK, N. Psychologism and Behaviorism. *Philosophical Review.* 1981, (90), s. 5–43.

BROOKS, R. A. Intelligence without Representation. In: *Mind design II: philosophy, psychology, artificial intelligence*. Cambridge: MIT Press, 1997, s. 395-420. ISBN 0262581531.

CLARK, A. *Mindware: an introduction to the philosophy of cognitive science*. New York: Oxford University Press, 2001. ISBN 978-0-19-513857-3.

COCKBURN, D. *An introduction to the philosophy of mind*. New York: Palgrave, 2001. ISBN 0–333–78637–8.

COLBY, K. M., J. B. WATT a J. P. GILBERT. A Computer Method of Psychotherapy. *The Journal of Nervous and Mental Disease*. 1966, 142(2), 148-152. ISSN 0022-3018.

DENNETT, D. C. *Druhy myslí: k pochopení vědomí*. Praha: Academia, 2004. Mistři vědy. ISBN 80-200-1177-3.

DREYFUS, H. L. From Micro-Worlds to Knowledge Representation: AI at an Impasse. In: *Mind design II: philosophy, psychology, artificial intelligence*. Cambridge: MIT Press, 1997, s. 143-182. ISBN 0262581531.

FRENCH, R. Subcognition and the Limits of the Turing Test, *Mind.* 1990, 99(393), s. 53–65.

GUNDERSON, K. The Imitation Game. *Mind*. 1964, (73), 234–245.

HARNAD, S. The Symbol Grounding Problem, *Physica D.* 1990, (42), s. 335–346.

HAVEL, I. M. *Přirozené a umělé myšlení jako filosofický problém* [online]. Glosy.info. 3.12.2004. [cit. 2018-01-18]. Dostupné na WWW: <http://glosy.info/texty/prirozene-a-umele-mysleni-jako-filosoficky-problem/>. ISSN 1214-8857.

HAYES, P. a FORD, K. Turing Test Considered Harmful. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*. 1995, (1), s. 972-973.

CHURCHLAND, P. M. On the Nature of Theories: A Neurocomputational Perspective. In: *Mind design II: philosophy, psychology, artificial intelligence*. Cambridge: MIT Press, 1997, s. 251-292. ISBN 0262581531.

MILLAR, P. H. On the Point of the Imitation Game, *Mind*. 1973, (82), s. 597.

MILLER, G. A. *Language, Learning and Models of the Mind*, nepublikovaný rukopis, červen 1972

NEWELL, A. a SIMON, H. A. Computer Science as Empirical Inquiry: Symbols and Search. In: *Mind design II: philosophy, psychology, artificial intelligence*. Cambridge: MIT Press, 1997, s. 81-110. ISBN 0262581531.

PEREGRIN, J. Filosofie mysli. (přednáška) Hradec Králové: UHK, 02. 05. 2013.

POLÁNYI, M. *The Tacit Dimension*. New York: Anchor Books, 1967.

PURTILL, R.L. Beating the Imitation Game. *Mind.* 1971, (80), s. 291-293.

RODNEY, A. B. Intelligence without Representation. In: *Mind design II: philosophy, psychology, artificial intelligence*. Cambridge: MIT Press, 1997, s. 395-420. ISBN 0262581531.

ROSA, M. *Umělá inteligence a budoucnost lidstva* [online]. Youtube .2016 [cit. 2018-03-28]. Dostupné z: Dostupné na WWW: <https://www.youtube.com/watch?v=-obfl0iFLAI>.

ROSENBERG, J. F. Connectionism and Cognition. In: *Mind design II: philosophy, psychology, artificial intelligence*. Cambridge: MIT Press, 1997, s. 293-308. ISBN 0262581531.

RUMELHART, D. E. The Architecture of Mind: A Connectionist Approach. In: *Mind design II: philosophy, psychology, artificial intelligence*. Cambridge: MIT Press, 1997, s. 205-223. ISBN 0262581531.

RUMELHART, D. E. a J. L. MCCLELLAND. *Parallel distributed processing: explorations in the microstructure of cognition*. Cambridge: MIT Press, 1986. ISBN 9780262181204.

SEARLE, J. R. Minds, Brains, and Programs. In: *Mind design II: philosophy, psychology, artificial intelligence*. Cambridge: MIT Press, 1997, s. 183-204. ISBN 0262581531.

SHANKER, S. G. *Wittgenstein's Remarks on the Foundations of AI*. Routledge, 2002.

WILSON, E. O. *O lidské přirozenosti: máme svobodnou vůli, nebo je naše chování řízeno genetickým kódem?*. Praha: Lidové noviny, 1993. Edice 21. ISBN 80-7106-076-3.

WITTGENSTEIN, L. *Remarks on the Foundations of Mathematics*, eds., G. H. von Wright, R. Rhees, a G. E. M. Anscombe, Oxford: Basil Blackwell, 1978.

WITTGENSTEIN, L. The Blue and Brown Books, Oxford: Basil Blackwell, 1960.

Computer simulating 13-year-old boy becomes first to pass Turing test. The Guardian[online]. 2014 [cit. 2018-02-28]. Dostupné na WWW: <https://www.theguardian.com/technology/2014/jun/08/super-computer-simulates-13-year-old-boy-passes-turing-test>.