

Univerzita Palackého v Olomouci

Filosofická fakulta

Katedra filosofie

# **Turingův test**

**Filosofické aspekty umělé inteligence**

**Filip Tvrdý**

Disertační práce

2011

## **Abstrakt**

Disertační práce se zabývá problematikou připisování myšlení jiným entitám, a to pomocí imitační hry navržené v roce 1950 britským filosofem Alanem Turingem. Jeho kritérium, známé v dějinách filosofie jako Turingův test, je podrobena detailní analýze. Práce popisuje nejen původní námitky samotného Turinga, ale především pozdější diskuse v druhé polovině 20. století. Největší pozornost je věnována těmto kritikám: Lucasova matematická námitka využívající Gödelovu větu o neúplnosti, Searlův argument čínského pokoje konstatující nedostatečnost syntaxe pro sémantiku, Blockův návrh na použití brutální síly pro řešení imitační hry, Frenchova teorie subkognitivních informací a Michieho skepticismus ohledně možnosti umělého vědomí. Závěr práce zachycuje současný stav recepce Turingova testu a představuje pokusy o jeho praktickou realizaci, například v každoroční soutěži o Loebnerovu cenu. Autor práce zastává názor, že ani po více než šedesáti letech od uveřejnění Turingova paradigmatického eseje stále neexistují žádné vážné důvody pro zamítnutí jeho tvrzení. Tradiční komputační funkcionalismus možná není ideální teorií vysvětlující činnost myslí a jako slibnější se může jevit vývoj v neurálních vědách, ale Turingův test je přesto užitečným a snad i jediným nástrojem pro detekci inteligence u lidmi vytvořených strojů.

## **Klíčová slova**

argument čínského pokoje, Blockhead, experimentální filosofie, Gödelovy věty o neúplnosti, filosofie mysli, funkcionalismus, imitační hra, inteligence, internalismus a externalismus, kognitivní věda, komputační teorie mysli, Loebnerova cena, myšlenkové experimenty, myšlení, neurofilosofie, nutná a postačující podmínka, počítače, subkognice, syntaktická sémantika, teorie vyššího řádu, Turingův stroj, Turingův test, umělá inteligence, vědomí, Wasonův výběrový test

## **Abstract**

The doctoral thesis considers problems of ascribing thought to other entities, using the imitation game proposed in 1950 by the British philosopher Alan Turing. His criteria, known in the history of philosophy as the Turing test, are thoroughly analysed. The thesis describes not only original objections of Turing himself, but particularly later discussions during the second half of the 20th century. The greatest attention is paid to these critiques: Lucas' mathematical objection based on Gödel's incompleteness theorem, Searle's Chinese room argument claiming the insufficiency of syntax for semantics, Block's proposal to use the brute-force attack on the imitation game, French's theory of subcognitive information and Michie's scepticism about the possibility of artificial consciousness. The conclusion presents the contemporary reception of the Turing test and attempts of its practical realization, for example in the annual Loebner Prize Competition. The author of this thesis holds the view that more than sixty years after the publication of Turing's paradigmatic essay there are still no sustainable reasons for refusing his claims. Traditional computational functionalism may not be the ideal theory explaining the activities of minds - the progress in neurosciences might be more promising - but the Turing test is nonetheless a very useful tool for the detection of intelligence in computers constructed by humans.

## **Keywords**

artificial intelligence, Blockhead, Chinese room argument, cognitive science, computational theory of mind, computers, consciousness, experimental philosophy, functionalism, Gödel's incompleteness theorems, higher-order theory, imitation game, intelligence, internalism and externalism, Loebner Prize, necessary and sufficient condition, neurophilosophy, philosophy of mind, subcognition, syntactic semantics, thought, thought experiments, Turing machine, Turing test, Wason selection task

Prohlašuji, že jsem disertační práci vypracoval samostatně, pouze s pomocí zdrojů uvedených v seznamu použité literatury.

## Obsah

1. Úvod: problémy s identifikací myslí.....	6
2. Alan Turing: život a dílo.....	11
3. Imitační hra: Turing 1950.....	18
4. Raná recepce: 50. až 70. léta .....	48
5. Absence sémantiky: Searle 1980 .....	59
6. Brutální síla: Block 1981.....	77
7. Subkognice a vědomí: French 1990 a Michie 1993 .....	92
8. Pozdní recepce a praktické aplikace: 90. léta a dál .....	106
9. Závěr: priorita vědy před filosofií.....	119
10. Bibliografie .....	121
11. Příloha: chatbot.....	134

## 1. Úvod: problémy s identifikací myslí<sup>1</sup>

Není nic překvapivého na tvrzení, že jsme obklopeni mnoha myslícími, inteligentními entitami. Většina z nás se dokonce s takovými entitami, u kterých jsme s to identifikovat mysl a které považujeme za vědomé, každodenně stýká. Úmyslně přitom prozatím nerozlišuji mezi pojmy "myslet", "být inteligentní", "disponovat myslí" a "disponovat vědomím". V běžném použití jazyka jsou totiž tyto rozdíly jen málo postřehnutelné a vlastně nepodstatné. Těžko si lze představit nemyslíci, ale inteligentní bytost, nebo bytost, která sice myslí, ale není si toho vědoma.<sup>2</sup> Kdo tedy patří do té již zmiňované třídy entit, které považujeme za inteligentní?

Jsou to především ostatní lidé. Nemusíme se pouštět do žádných komplikovaných filosofických spekulací, abychom přiznali mysl všem, nebo téměř všem lidem. Problematickou podskupinou budou zřejmě někteří mentálně postižení lidé, řekněme například katatonici, dále lidé v kómatu a pravděpodobně i děti do určité fáze prenatalního či postnatalního vývoje. Ostatním ale mysl rádi přiznáváme, a dokonce si můžeme vybrat z několika filosofických důvodů, proč tomu tak je.

První řešení je z pozic zdravého rozumu, to znamená navazující na nauku o *common sense*, jak ji nalezneme v osvícenské filosofii Thomase Reida nebo analytické filosofii G. E. Moora. Podle této nauky existují tvrzení, o kterých je zbytečné diskutovat, polemizovat s nimi, snažit se je dokázat či vyvrátit. Moore je označoval jako "truismy" a řadil mezi ně širokou škálu tvrzení jako například: mám tělo, narodil jsem se a žiji na Zemi, existují dějiny, existují jiní lidé, mám zkušenosti různých druhů, jiní lidé mají zkušenosti různých druhů a podobně (Moore 1925, p.194). Každé malé dítě či nevzdělaná venkovanka vědí, že ostatní lidé myslí, a proto můžeme toto tvrzení považovat za truismus.

---

<sup>1</sup> Mé poděkování a vděk si zaslouží několik lidí, kteří mi nezištně poskytli podporu. Je to především Iva Urbanová za revizi celého textu a mnoho cenných připomínek k jeho závěrečné podobě. Dále pak Martina Juříková za konzultaci ohledně interpretace Gödelovy věty o neúplnosti, Zuzana Budínská za pomoc s právníckou terminologií, Pavel Šimáček za rady při psaní pasáží týkajících se šachu a Pavel Šuráň za opravu chyby ve Wasonově výběrovém testu. Všechny případné omyly, kterých jsem se při psaní dopustil, padají samozřejmě na mou hlavu.

<sup>2</sup> Tímto tématem se budu detailně zabývat v sedmé kapitole, kde se pokusím vyřešit otázky spojené s provázaností inteligence a vědomí. Do té doby žádám laskavého čtenáře o trpělivost.

Jiné řešení je pragmatické. To zdůrazňuje, že pochybování o tom, mají-li jiní lidé mysl, je neúčinné. Poprvé se zřejmě implicitně objevuje v Humových *Zkoumáních o lidském rozumu*, v pasáži o oprávněnosti přehnaného skepticismu. Hume tvrdí, že z přehnaného skepticismu - tj. toho, který pochybuje o všem, třeba i o existenci jiných myslí - nikdy nemůže pocházet žádné trvalejší dobro. I když je rozumově nevyvratitelný, přesto selhává v běžném životě, kde "i ten nejzatvrzelejší skeptik je pak k nerozeznání od ostatních smrtelníků" (Hume 1996, p.216). Podobně argumentuje i o dvě století později Quine, když neméně pragmaticky tvrdí, že je třeba rezignovat na snahu o přímý, nezprostředkovaný přístup k mentálním stavům jiných lidí. Za všech okolností jsme totiž odkázáni jen na vnější, hlavně jazykové projevy ostatních, ze kterých můžeme pouze na základě analogie usuzovat, jaké jsou jejich myšlenky a pohnutky. Tento postoj Quine označuje jako lingvistický behaviorismus a považuje jej za jediné možné stanovisko (Quine 1994, p.46). Tím, co a jestli vůbec něco se honí v hlavách našim spoluobčanům, si úplně jistí nebudeme nikdy. Přesto je praktické chovat se k nim tak, jako by mysleli a cítili bolest, protože jinak by se mohlo stát, že začnou oni pochybovat o *naší* schopnosti myslet a cítit bolest.

Poslední řešení je nejjednodušší a spočívá v laickém poukázání na neurální podobnosti mezi jednotlivými lidmi. Jsme si podobní, vytvoření ze stejného materiálu, máme stejně uspořádané vnitřní orgány a stejné chemické složení. Jsme vybaveni stejnými mozky se stejným nebo obdobným rozložením mozkových center. Vypadáme podobně zvenku i zevnitř, funkčně si odpovídají naše zažívací i oběhová ústrojí, bylo by tedy velmi nepravděpodobné - řekněme nemožné - že bychom se závažně lišili mentálně. Já myslím, a nutně předpokládám, že můj podobně zkonstruovaný kamarád myslí též.

Je vidět, že pokud nechceme být malichernými filosofickými kverulanty, nemáme s identifikací myslí u jiných příslušníků lidského rodu větší problém. Věci se ale začnou komplikovat, jestliže se pokusíme připisovat status myslících bytostí za hranicemi našeho biologického druhu. Okamžitě si totiž všimneme, že všechna tři řešení, která jsme vcelku upokojivě aplikovali na případu jiných lidských bytostí, selhávají. Co se týče zdravého rozumu, ukáže se jeho nevyhnutelně nejistá a relativistická povaha. Tvrzení, které jeden člověk považuje za samozřejmé, triviální či banální, může druhému člověku připadat jako neobhajitelné. Moore zřejmě úmyslně vybral sadu zcela nekontroverzních soudů o obvyklých otázkách, lidé ale mají tendenci zaštiťovat se zdravým rozumem i v případech, kdy jej

aplikovat není možné. Zamysleme se jen nad tím, jak dopadá argumentace pomocí zdravého rozumu v rasových, politických či náboženských otázkách, například ve větách "Lidé všech ras jsou si rovni.", "Bohatší musejí cítit solidaritu se sociálně slabšími." nebo "Křesťanství je nejdokonalejší zjevené náboženství." (Upozorňuji, že předkládaná tvrzení nemusí být nutně totožná s přesvědčeními autora této práce.) Stejně tak existují ti, kteří považují za zcela samozřejmé, že zvířata myslí a snad i rozumí lidské řeči, zatímco pro jiné je představa myslících zvířat, třeba hrochů, absurdní. Pěkně to vyjadřuje starý vtip, podle kterého si filosofové, kteří mají doma psa, myslí, že psi myslí, zatímco filosofové, kteří doma psa nemají, myslí, že psi nemyslí. S odkazem na zdravý rozum si tedy nevystačíme.

Stejně pochybná je i pragmatická argumentace. Viděli jsme, že je užitečné považovat ostatní nám podobné lidi za myslící. U zvířat ale nevíme, co je pro nás vlastně užitečné. Zdánlivě nejužitečnější by bylo považovat je za nemyslící, čímž bychom uchránili před zánikem zoologické zahrady a vyvarovali bychom se globálního vegetariánství, ale stejně užitečná je snaha vědců o hledání pravdy, která může skončit nalezením důkazu o zvířecí schopnosti myslet. Moc nám také nepomůže biologické srovnání nervové soustavy zvířat, protože někteří živočichové jsou nám podobní a někteří vůbec ne. To ale neznamená, že by stejné mentální stavy nemohly být vyvolány jinak konfigurovanými nervovými soustavami. Poslední staletí a desetiletí jsou ve znamení rozšiřování práv slabých a ponížených. Výrazně se zvýšila i ochrana práv zvířat, pořád ale žijeme v dosti schizofrenní situaci, kdy za ušlapání ježka může být člověk nepodmíněně odsouzen k dvěma letům vězení (ust. §203 zákona č. 140/1961 Sb., trestní zákon, ve znění pozdějších předpisů), ale na druhou stranu naše společnost široce toleruje velkochov ustájeného dobytka, klecový odchov drůbeže nebo diskutabilní způsoby porážení hospodářských zvířat. Descartes tím, že radikálně oddělil svět *res extensa* od *res cogitans*, následně zařadil zvířata výhradně do rozprostraněnosti a upřel jim participaci na mentálních stavech, ospravedlnil náš eticky obtížně obhajitelný postoj ke zbytku živočišné říše. V současné době je módní kritizovat tento aspekt karteziánství a prohlašovat, že aspoň některá zvířata si zaslouží určitý stupeň rovnoprávnosti s lidmi, nedokážeme se ale dohodnout, jaká zvířata a jaký stupeň. Těžko lze předpokládat, že žížala disponuje schopností myslet; na druhou stranu lze stejně těžko tvrdit, že šimpanz ne. Znamý aktivista ve věci práv zvířat Peter Singer začíná svou nesmírně vlivnou knihu *Osvobození zvířat* kapitolou o tom, že jsou si všichni tvorové rovni, a končí ji výzvou, aby se každý stal



vegetariánem čili jediným eticky ospravedlnitelným typem konzumenta (Singer 2001). Určitou naději pro nás, kteří si ale nechceme upřít občasný hovězí steak, představuje Dennettova teorie mnoha druhů myslí. Podle ní není "mysl" univerzální vlastnost, kterou je entita buď vybavena, anebo není. Ve skutečnosti existuje myslí značné množství. Klíčový je pro Dennetta pojem "intencionální postoj", kterým označuje naši schopnost interpretovat činnost živých organismů i neživých věcí tak, jako by byla záměrná. Intencionalita neexistuje reálně, je jen lidskou kategorií, jejíž pomocí se snažíme odhadnout chování nejrůznějších entit (Dennett 1997).

Všimněme si, že naše manipulace s pojmy, které entitám přisuzují mentální stavy, je značně nepřehledná a nespravedlivá, mohli bychom říct až šovinistická. Někdy totiž máme tendenci příliš široce aplikovat pojem "mysl" i na lidi, kteří na to zřejmě nemají právo (například už oni zmiňovaní lidé v kómatu), někdy naopak tento pojem odebíráme zvířatům, která na něj zřejmě právo mají (například vepřům, podle zvířecích měřítek údajně nejinteligentnějším obyvatelům vesnických dvorků). A situace se zkomplikuje ještě víc, pokud se pokusíme do naší teorie identifikace myslí zařadit entity, u kterých je jejich nárok relativně nedávný: stroje, počítače, roboty.

Naše předchozí pokusy o uchopení myslí jsou zcela nepoužitelné. Z pohledu zdravého rozumu nejsou počítače - především ty vypnuté - o nic víc myslící než pluhy nebo bicykly. Pragmaticky o participaci počítačů na mentálních stavech také nestojíme. Proč rozšiřovat naši už tak dosti početnou třídu vlastníků myslí, navíc v situaci, kdy jsou dnes hranice této třídy více než nejisté? A neurální srovnání není možné, protože není co s čím srovnávat. Vždyť naše nervová soustava je s architekturou počítačů naprosto nesouměřitelná, už jen kvůli materiální odlišnosti: naše je založená na komplikovaných organických sloučeninách na bázi uhlíku, zatímco ta počítačová na relativně jednoduchých anorganických sloučeninách na bázi křemíku. S etikou si také moc nepomůžeme. Na rozdíl od roztomilých zvířátek počítače ani v nejmenším nevykazují známky únavy nebo bolesti, těžko u nich hledat stopy sklíčenosti nebo naopak dobré nálady. A ani interpretace pomocí intencionálních postojů nebude příliš filosoficky produktivní, protože bude vždy zatížena obrovskou metaforičností. I ve chvílích, kdy proklínám svůj mobilní telefon, přece vím, že skutečným viníkem nočního vyzvánění není sám mobil, ale rozverný známý. A i když trpělivě přemlouvám svůj počítač, aby nedělal hlouposti a nabootoval operační systém, vím, že se mnohem víc než o jeho psychický stav

jedná o ten můj. Někteří představitelé silné umělé inteligence se nám přesto snaží tvrdit, že *svým způsobem* "myslí" i termostat. Tvůrce pojmu "silná umělá inteligence" John McCarthy ve svém slavném bonmotu říká: "Můj termostat má tři přesvědčení: Že je tady uvnitř příliš teplo, že je tady uvnitř příliš zima a že je tady uvnitř, jak má být." (Searle 1994a, p.31; překlad Marek Nekula) Tato devalvace predikátů označujících mentální činnost je ale zavádějící a má jen málo společného s otázkou, mohou-li stroje, nebo přesněji řečeno počítače, myslet.

Tato práce se zabývá nejznámějším pokusem o řešení problému, zdali můžeme počítače zařadit do elitní skupiny hrdých vlastníků mysli, společně s většinou lidí, menšinou zvířat a případně některými jinými typy entit, jako jsou třeba hypoteticky existující, biologicky nám zcela vzdálení mimozemšťané. Je to řešení, které na konci čtyřicátých let vypracoval jeden z největších duchů dvacátého století, anglický matematik Alan Turing. Předkládaný text má chronologickou strukturu: nejprve bude vysvětleno vlastní Turingovo stanovisko (kapitola 3), později popsány důležité milníky v diskusi o legitimitě imitační hry. Především se zaměřím na kritické texty Johna Searla, Neda Blocka, Roberta Frenche a Donalda Michieho (kapitoly 5, 6 a 7). Pokusím se ukázat, že ani po více než šedesáti letech od uveřejnění Turingova paradigmatického eseje se nepodařilo najít způsob, jak jeho závěry vyvrátit. Součástí práce je i popis současných snah o praktickou realizaci požadavků, které na umělou inteligenci Turing nakladl (kapitola 8). Závěr pak tvoří syntetické shrnutí (kapitola 9) a ilustrační ukázka reálného pokusu o imitační hru mezi člověkem a počítačovým programem (kapitola 11).

## 2. Alan Turing: život a dílo

Jen málo myslitelů ve dvacátém století - z filosofů snad jen Wittgenstein - mělo podobně pohnutý život jako Alan Turing. Zdá se dokonce, že běžné veřejnosti je Turing známý spíš kvůli svým personálním než kvůli podílu na rozvoji kryptografie, kybernetiky a filosofie mysli. Jeho život obsahuje všechny ingredience, které má obsahovat dobrodružný nebo bulvární příběh: oslnivou kariéru, válečné zásluhy, podíl na špionáži, ukradené tajemství, perzekuci kvůli sexuální diferencii, veřejné zostuzení, pád na dno a tragické finále, které Turingovi dnes dodává aureolu nepochopeného mučedníka. Přesto byl Turing pozapomenut až do osmdesátých let, kdy došlo k jeho znovuobjevení, zprostředkovanému ve filosofii kritikou komputacionalismu a v historiografii činností anglického žurnalisty Andrewa Hodgese.<sup>3</sup>

Turing se narodil roku 1912 v Londýně, v rodině patřící ke střední třídě. Vystudoval matematiku v Cambridgi, mezi jeho učitele patřili mimo jiné ekonom John Keynes a matematik Max Newman. Větší vliv ale na něj měly práce Gottloba Fregy a především Russellova a Whiteheadova *Principia Mathematica*. Po dokončení studií v roce 1935 zůstal na Cambridgi a už v 22 letech zde získal místo asistujícího profesora. Patřil k okruhu Wittgensteinových studentů a nejméně jeden rok byl pravidelným účastníkem jeho přednášek o základech matematiky.<sup>4</sup> V roce 1937 publikoval svůj vůbec první článek "On Computable Numbers, with an Application to the *Entscheidungsproblem*", ve kterém

---

<sup>3</sup> Autorkou prvního Turingova životopisu je jeho matka Ethel Sara Turingová, která roku 1959 publikovala memoárovou knihu *Alan M. Turing*. Nejzásadnější Turingova biografie vyšla až roku 1983 pod názvem *Alan Turing: The Enigma*, jejím autorem je právě Andrew Hodges (1992). Na ní je také založena úvodní životopisná črta v kapitole 2. Hodges později vydal ještě jednu, mnohem kratší práci *Turing: A Natural Philosopher* (1997), která má být stručným úvodem do jeho myšlení a která vyšla v edici "The Great Philosophers" nakladatelství Routledge. Hodges je také správcem Turingova archivu na internetové adrese <<http://www.turing.org.uk>>. V češtině je k dispozici pouze překlad popularizující monografie *Muž, který věděl příliš mnoho* (Leavitt 2007). Od roku 1992 do roku 2001 vyšly všechny čtyři svazky úplného vydání Turingových textů *Collected Works of A. M. Turing* a teprve nedávno byl publikován výbor Turingových nejdůležitějších prací *The Essential Turing* (Copeland 2004), který obsahuje nejen ony dvě často přetiskované eseje, ale i spousty dalších textů, z nichž mnohé nebyly dosud uveřejněny. Nejvýznamnějším sekundárním zdrojem je masivní dvousvazkový sborník *The Legacy of Alan Turing* (Millican & Clark 1996), který obsahuje 27 esejí z 90. let minulého století.

<sup>4</sup> Wittgensteinova reakce na Turingovy teorie z oblasti kybernetiky je popsána v Shankerovi (1998).

originálním způsobem vyřešil Hilbertův problém úplné formalizace matematiky a přitom mimochodem položil teoretický základ všem později zkonstruovaným počítačům. Krátce přednášel v Princetonu, kam odjel na pozvání Alonza Churcha. Roku 1938 byl naverbován britskou armádou a v průběhu druhé světové války pracoval pro kryptografické oddělení v Bletchley Parku. Zde se podílel na prolomení kódu šifrovacího stroje Enigma, jehož prostřednictvím spolu komunikovaly německé ponorky. Kód byl definitivně rozluštěn kolem roku 1942 a hlavní podíl na tomto úspěchu lze připsat Turingovi, i když navazoval na předchozí práci polských matematiků v čele s Marianem Rejewskim. Odhaduje se, že rozluštění Enigmy mělo podstatný vliv na průběh války v Evropě a Atlantickém oceánu: podle amerického prezidenta Eisenhowera by se válka bez přispění britských kryptografů prodloužila asi o dva roky (Flowers 2006, p.81), podle Turingova spolupracovníka Harryho Hinsleyho by dokonce mohlo jít až o roky čtyři (Hinsley 2001, pp.11-13)<sup>5</sup>. Turingův podíl dlouho nebyl kvůli přísnému utajení zveřejněn, a proto ani doceněn, což u něj postupem času vedlo k značné frustraci. Po válce Turing působil na Manchesterské univerzitě, kde spolupracoval na vytvoření univerzálního počítačového stroje, jak jej popsal v eseji z roku 1937. První počítač ENIAC byl postaven ve Spojených státech roku 1945 bez přímého Turingova přispění. Zakladatel počítačové vědy John von Neumann ale do značné míry vycházel z Turingových myšlenek, i když jeho vliv na svou práci nikdy explicitně nepřiznal. Jen pro zajímavost uvádím, že roku 1948 se zdatný běžec Turing téměř kvalifikoval pro maraton na olympiádě v Londýně.

Roku 1950 publikoval Turing svou jedinou esej uveřejněnou ve filosofickém časopise, "Computing Machinery and Intelligence". V ní popsal některé spekulativnější aspekty svých objevů v oblasti umělé inteligence a jako první myslitel vůbec si položil otázku, mohou-li počítače myslet. Text byl - stejně jako první článek - zpočátku jen málo chápán a stal se spíš předmětem zájmu populárně-naučných novinových příloh a nepřiliš seriózní futurologie. Turingova činnost v oblasti umělé inteligence byla nakonec přerušena osobní tragédií. V předjaří roku 1952 byl Turing zatčen, obviněn a odsouzen za homosexualitu. Turing se svou sexuální orientací nikdy netajil, a byl proto značně překvapen jednáním policie a soudů;

---

<sup>5</sup> Sir Hinsley tento názor vyjádřil v nepublikované přednášce "The Influence of Ultra in the Second World War", kterou přednesl 19. října 1993 v cambridgeském Babbage Lecture Theatre. Její transkript je k dispozici na <<http://www.cl.cam.ac.uk/research/security/historical/hinsley.html>>.

v Británii byla homosexualita trestná do roku 1957 a Světová zdravotnická organizace ji vyřadila ze seznamu nemocí a vad dokonce až v roce 1986. Uvěznění se Turing vyhnul jen díky tomu, že podstoupil estrogenovou terapii, která u něj měla potlačit sexuální pud. Turing léčbu velmi těžce nesl: ztloustl, vypadaly mu vlasy a narostly ženské sekundární pohlavní znaky; v důsledku soudního procesu byl také vyloučen z britské armády a byla mu omezena výuka na univerzitě. Zemřel roku 1954 na následky otravy vitriolem, přičemž nejpříjemnějším vysvětlením je patrně sebevražda, méně pravděpodobným vražda nebo nehoda. Pro sebevraždu svědčí extravagantní způsob otravy: Turing se inspiroval Disneyho animovaným filmem o Sněhurce a do vitriolu namočil jablko, které snědl. Je zvláštním rysem britského impéria, že se ke svým hrdinům často obrací zády a zachází s nimi jako s netvory a zločinci.<sup>6</sup>

Obsahem této práce je jen a pouze ta část Turingova díla, která se zabývá umělou inteligencí a která je v podstatě zcela vyčerpána esejí "Computing Machinery and Intelligence". Jen pro úplnost zde vysvětlím, v čem spočívají jeho rané texty týkající se matematiky. I ty byly totiž v šedesátých letech využity filosofií myslí, což vedlo k nejrůznějším konfuzím a dezinterpretacím. Je třeba si uvědomit, že se Turing v "On Computable Numbers" zabývá výsostně logickým tématem, které původně nemělo s filosofií myslí vůbec nic společného. V následující části nepoužiji žádný matematický ani logický aparát a problém se pokusím vysvětlit přirozeným jazykem; samotný *Entscheidungsproblem* pro naše potřeby není podstatný, důležitější je Turingův koncept  $\alpha$ -stroje, který dnes známe spíše pod označením Turingův stroj.

---

<sup>6</sup> Tragický rozměr Turingova života je příčinou jevu, který je mezi filosofií nevídaný - Turing se stal tématem mnoha uměleckých děl. Nesourodý a určitě neúplný seznam by mohl vypadat třeba takto: v roce 1980 napsal Ian McEwan televizní hru *The Imitation Game*, v roce 1986 pak Hugh Whitmore divadelní hru o Turingově životě *Breaking the Code*. Turingovy myšlenky se objevují i ve sci-fi románu *The Turing Option*, který napsali v roce 1993 Harry Harrison a Marvin Minsky. Sci-fi povídka Grega Egana "Oracle" z roku 2000 zase popisuje fiktivní setkání Turinga s C. S. Lewisem. Román Roberta Harrise *Enigma* (1995), který velmi volně vychází z práce kryptografického oddělení v Bletchley Parku, byl roku 2001 zfilmován Michaellem Aptedem s Kate Winsletovou v roli ženy, která je předmětem milostného zájmu - zde heterosexuálního - objevitele tajemství německé šifry. Turing vystupuje jako jedna z hlavních postav i v historickém techno-thrilleru *Cryptonomicon* (1999) od amerického spisovatele Neala Stephensona. Naprostou kuriozitou je limitované EP kalifornského elektronického dua Matmos z roku 2006, které se jmenuje *For Alan Turing* a jehož titulní píseň je téměř kompletně zahrána na zrekonstruovaný šifrovací přístroj Enigma.

Na přelomu 19. a 20. století došlo ve filosofii matematiky k významné revoluci: díky činnosti Cantora, Fregy, Meinonga, Poincarého, Husserla, Hilberta, Russella a Brouwera si filosofové a matematici začali klást otázky po samotné povaze axiomatických systémů jako je matematika nebo logika. Tyto otázky se týkaly vztahu logiky a matematiky, reference matematiky k realitě a především možnosti její úplné formalizace. Nejambicióznější program si stanovil na počátku 20. let 20. století německý matematik David Hilbert, který se v rámci svého formalismu pokusil dokázat, že matematika je pouhou hrou insignifikantních pojmů, které neodrážejí nic reálného a které jsou zcela určeny vzájemnými pravidly. Aby mohla být matematika považována za plně formalizovaný axiomatický systém, bylo třeba podle Hilberta vyřešit dvacet tři problémů, z nichž nejdůležitějším se ukázal být druhý týkající se důkazu konzistentnosti axiomů aritmetiky. Naneštěstí pro Hilberta, na počátku 30. let formuloval Kurt Gödel svou druhou větu o neúplnosti, podle které ve všech axiomatických soustavách existují věty, které jsou v jejich rámci nedokazatelné, ale i nevyvratitelné (např. "Tuto větu nelze dokázat.").

Jiným problémem Hilbertova programu, který ale nebyl součástí oněch původních 23, byl tzv. rozhodovací problém ("Entscheidungsproblem"). Ten spočíval ve snaze dokázat tvrzení, že ke každému matematickému výroku existuje ryze formální, mechanický proces, který by mu přiřadil pravdivostní hodnotu. Pokud by byl takový postup nalezen, znamenalo by to, že matematika je naprosto soběstačný a úplný axiomatický systém nezávislý na logice či jiné teorii. V letech 1935-36 vytvořili Alonzo Church a Alan Turing nezávisle na sobě dvě různá řešení rozhodovacího problému, která jsou vzájemně převeditelná a dokazují, že takovýto algoritmus nalézt nelze. V podstatě můžeme obě řešení považovat za rozvinutí Gödelovy teorie neúplnosti, ovšem technicky mnohem subtilnější. Church vymyslel  $\lambda$ -kalkul, jehož pomocí lze vyjádřit jakoukoli spočetnou funkci. Kalkul je velmi podobný funkcionálním programovacím jazykům, přímo z něj vychází třeba dodnes používaný *Lisp*. V rámci kalkulu Church převedl rozhodovací problém na hledání ekvivalence dvou výrazů pomocí univerzálního algoritmu a zjistil, že takový algoritmus neexistuje (Church 1936, p.358). Turing v "On Computable Numbers" uvažoval podobným způsobem, ale jeho řešení je pro laika pochopitelnější, protože obsahuje velmi silnou metaforu pro průběh algoritmu, který přirovnává k mechanickému stroji provádějícímu jednoduché, předprogramované úkony. Tento idealizovaný stroj má konečný počet vnitřních stavů a obsahuje potenciálně

nekonečnou pásku rozdělenou do polí. Pole jsou označena symboly z konečné abecedy, ale pro fungování stroje je možné používat jen dva (např. 0 a 1). Stroj je vybaven čtecí hlavou, která je schopna přečíst symbol na právě analyzovaném poli, a kromě toho umí symboly mazat a přepisovat. Reakce na čtené symboly je definována strojní neboli stavovou tabulkou; po přečtení symbolu se stroj podle předdefinovaných instrukcí rozhodne, má-li symbol smazat, přepsat nebo nechat být, a má-li se hlava posunout doleva, doprava nebo zůstat na tom samém poli. Chování stroje je tak zcela determinováno jeho konfigurací, tj. současným stavem a právě čteným symbolem. Popsaný stroj je schopen provádět vše, co dokáže jakýkoli reálně existující počítač. Ani pomocí takového stroje ale není možné dospět k algoritmu, který by definitivně vyřešil rozhodovací problém.

Kromě aplikací v logice a filosofii matematiky se dočkal Turingův stroj nečekané popularity ve filosofii mysli. Stal se totiž hlavní inspirací pro Hilaryho Putnama, když formuloval svou teorii funkcionalismu (Putnam 1960). Funkcionalismus vznikl jako protiklad dobového naturalismu, především teorie identity představované J. J. C. Smartem. Podle Putnama nelze mentální stavy redukovat na neurální stavy, protože můžeme předpokládat - byť jen logicky možnou - existenci bytostí, jejichž mentální stavy jsou realizovány zcela jinými způsoby než v případě lidí. Tak například není správné "druhisticky"<sup>7</sup> tvrdit, že pocit bolesti je totožný se stimulací určitých nervových vláken mozku, protože mozky jednotlivých živočichů se značně liší a navíc mohou bolest cítit i entity, které vůbec nedisponují mozky podobným tomu našemu. Je proto třeba vytvořit abstraktnější teorii, která by tento problém překonala a byla "liberální", "druhově nezávislá". Putnam v sérii esejí z 60. a 70. let vytváří teorii, podle které mentální stavy nejsou fyzicko-chemické stavy mozku, ale funkcionální stavy celého organismu (Putnam 1967; Putnam 1975a). Tento klasický funkcionalismus staví na analogii mezi myslícími bytostmi a stroji, především digitálními počítači, a proto bývá často nazýván strojovým funkcionalismem. Mysl je podobná Turingovu stroji a lidé jsou automaty vytvořené z masa a krve. Řečeno současnou počítačovou terminologií, mysl je software a tělo hardware. Zatímco vstupy a výstupy Turingova stroje jsou zajištěny čtením z pásky a zápisem na pásku, myslící organismy jsou vybaveny senzorickými a motorickými orgány.

---

<sup>7</sup> Termín "druhismus" [speciesism] poprvé použil britský psycholog Richard D. Ryder roku 1973 pro označení neoprávněné nadřazenosti, kterou člověk cítí nad jinými druhy, a s oblibou jej používá Peter Singer.

Stejně jako Turingovy stroje jsou ale organismy řízeny programem, který jednotlivým vstupům přiřazuje příslušné výstupy. Není přitom podstatné, z čeho je hardware vyroben. Mysl může být fyzicky konstruována mnoha různými způsoby: organickými neurony založenými na sloučeninách uhlíku, neorganickými mikročipy založenými na křemíku, mohla by být z kovu, ze skla, dokonce ze dřeva. Často se proto mluví o "vícenásobné realizovatelnosti" mysli (Putnam 1967). Se značnou mírou nadsázky Putnam tvrdí, že "bychom mohli být vytvořeni z ementálu a nevadilo by to" (Putnam 1975a, p.291). Stojí za povšimnutí, že přinejmenším na několika místech argumentuje Turing zcela analogicky Putnamovi. V rozhlasové diskusi na stanici BBC Turing tvrdil následující:

"Nezajímá nás, že má mozek konzistenci studené ovesné kaše. Nechceme říkat 'Tento stroj je docela pevný, takže to není mozek, takže nemůže myslet.'" (Turing 2004a, pp.494-495)<sup>8</sup>

Jinde zase Turing upozorňuje na odlišnost nervového systému, založeného většinou na chemicko-elektrickém principu, a dobových počítačů, které byly povětšinou elektrické. Skutečná podobnost totiž podle něj spočívá v "matematických analogiích funkcí" (Turing 1950, p.439).

Funkcionalismus se stal jedním z nejpopulárnějších způsobů řešení problému mysl-tělo, John Searle jej dokonce označil za ortodoxii současné analytické filosofie mysli (Searle 1992, p.7). Přesto nebo právě proto je třeba všimnout si určitých odlišností mezi původní Turingovou formulací a Putnamovou reinterpretací. Hlavním rozdílem je, že Turing usiluje o teoretické a později i praktické sestavení počítače, který by fungoval stejně jako lidský mozek. Putnam se o umělou inteligenci příliš nezajímá, a dokonce ji označuje za "mnoho povyku pro docela málo"<sup>9</sup>. Tvrdí, že Turingem popsaný počítač není třeba vůbec sestavovat, protože jej už máme k dispozici - je jím právě lidský mozek. Zatímco Turing pracuje na vytvoření umělé mysli, Putnamovi stačí popsat mysl přirozenou, tu naši. Důležité si je také uvědomit, že koncepty Turingova stroje a Turingova testu<sup>10</sup> spolu mají jen velmi málo

---

<sup>8</sup> Překlad této i všech dalších citací FT, pokud není uvedeno jinak.

<sup>9</sup> "Much Ado about Not Very Much" je název Putnamovy eseje týkající se umělé inteligence (Putnam 1988a).

<sup>10</sup> Od této chvíle budu pro zkrácení používat pojem "T-stroj" pro označení Turingova stroje a "T-test" pro označení Turingova testu.



společného. Jedinou souvislostí je, že pokud někdy vůbec nějaký stroj projde T-testem, bude s velkou pravděpodobností založen na principech T-stroje.

Na závěr je ještě důležité dodat, že Putnam svůj někdejší funkcionalistický program opustil. Již od počátku jej považoval za obtížně verifikovatelnou hypotézu a postupem času k ní nacházel mnoho protipříkladů, které jej v 80. letech přinutily teorii zcela opustit (Putnam 1988b). Hlavním důvodem bylo uvědomění si utopického charakteru funkcionalismu, který je příliš vágní a není s to dokonale popsat program, který způsobuje mentální stavy bytí jen jediné myslící bytosti. Navíc začal Putnam zastávat sémantický externalismus, který významy jazykových výrazů posouvá do vnějšího světa - "významy nejsou v hlavě" (Putnam 1975b, p.227). Funkcionální analýza vstupů a výstupů myslícího organismu tak není dostatečným nástrojem vysvětlení psychologických jevů, protože přesvědčení myslících bytostí nelze vysvětlit bez reference k sociálnímu a fyzikálnímu prostředí, ve kterém žijí. I přes Putnamovu rezignaci je funkcionalismus stále velmi populární a objevuje se v mnoha variantách, především homunkulární a teleologické, a také ve velmi vlivném komputacionalismu.

### 3. Imitační hra: Turing 1950

Turingovi bylo třicet osm let, když vyšla jeho esej "Computing Machinery and Intelligence". Těžko v té době mohl tušit, že napsal jeden z nevlivnějších filosofických textů druhé poloviny 20. století. Ani z dnešního pohledu není zcela jasné, proč se esej stala klasickou: Turingova předpověď ohledně budoucích úspěchů počítačů v imitační hře se ukázala mylnou, některé pasáže jsou téměř nesrozumitelné, v textu se objevují jen málo zábavné pokusy o humor, autor si neláme hlavu s formálními požadavky vědecké práce a v podstatě vůbec necituje, a podobně. Hlavní síla eseje ale spočívá v geniální jednoduchosti odpovědi na řešenou otázku, která lidstvo trápila po staletí. Již v raně novověké filosofii se objevuje názor, že myšlení má něco společného s počítáním, že je svého druhu komputačním procesem. Jako první to zřejmě tvrdil Hobbes: "Pod pojmem rozumová úvaha rozumím počítání." (1988, p.36; překlad Vojtěch Balík), později a důrazněji La Mettrie, který považoval všechny mentální procesy za strojové operace vykonávané "pružinami" (La Mettrie 1958, p.76). V 17. století se objevuje dokonce první primitivní varianta T-testu, a to v často citované pasáži Descartovy *Rozpravy o metodě*:

[K]dyby existovaly stroje, podobající se našim tělům a napodobující naše úkony potud, pokud by to mravně bylo možné, měli bychom vždy dva velice vážné důvody, abychom poznali, že proto ještě nejsou skutečnými lidmi. První důvod je, že by nikdy nemohly užívat slov ani jiných znaků, skládající je, jako činíme my, abychom své myšlenky vyložili jiným. Neboť lze dobře chápat, že stroj může být udělán tak, aby pronášel slova, ba dokonce aby pronášel některá ve spojení s tělesnými úkony, souvisejícími s nějakými změnami jeho orgánů: jako například když se ho dotkneme na určitém místě, aby se zeptal, co mu chceme říci, když na jiném místě, aby křičel, že ho to bolí, a podobně; nemůže však být udělán tak, aby slova různě sestavoval a takto odpovídal na vše, co se řekne v jeho přítomnosti, jak to i nejtupější lidé mohou činit. (Descartes 1992, p.41; překlad Věra Szathmáryová-Vlčková)

Descartes pokračuje úvahou o všestranném použití rozumu, který se tak liší od orgánů stroje se specializovaným určením. Vzhledem k jeho nauce o dvojí substanci nemůže ani překvapit, že je vůči myslícím strojům skeptický. Podobné stanovisko zaujal v roce 1668 i karteziánec Géraud de Cordemoy, který předchází T-test v neisofistikovanější podobě. Pokud by byl někdy sestaven stroj, který by používal jazyk stejně jako my, museli bychom mu přiznat duši, tedy myšlení. Cordemoy ale jedním dechem dodává, že reálná existence takového stroje je

nemožná. Snad by bylo možné vytvořit stroj, který by mluvil, ale nikdy by nebyl schopen rozumět slyšenému (de Cordemoy 2003, pp.6-14; Copeland 2000, pp.527-528).

I když bývá Turingova esej většinou považována za příspěvek k hledání odpovědi na otázku, zdali mohou stroje myslet, ve skutečnosti se v ní zabývá ještě dvěma dalšími problémy. Původní otázka "Mohou stroje myslet?" je totiž formulovaná velmi vágním způsobem. Nejprve je třeba zjistit, co je přesně "stroj" a co to znamená "myslet". Turingova definice stroje byla přitom pro budoucnost výpočetní techniky mnohem důležitější než samotný koncept imitační hry; zabývejme se tedy proto jí. Turing nezačíná právě nejslibněji, protože za "stroj" nejprve považuje jakýkoli lidmi stvořený stroj včetně lidí samotných. Těžko říct, jedná-li se o špatný vtip, ale tým vědců, který stroj staví, nesmí používat pohlavní reprodukci nebo klonování lidských bytostí. Ze své definice tak vyřazuje "lidi zrozené běžným způsobem" a navrhuje rozlišení mezi "digitálními" a "lidskými" počítači (Turing 1950, p.436). Následující argumentace už je rozumnější, protože Turing určí tři esenciální součásti, ze kterých se stroj musí skládat. Je to paměť, řídicí jednotka a kontrola.<sup>11</sup> Paměť může být tvořena třeba papírovou páskou nebo děrnými štítky, její součástí je i "kniha pravidel" nebo "tabulka instrukcí", což je vestavěný program počítače, který určuje jeho chování ve všech myslitelných situacích. Turing takto předchází pozdější dělení paměti na ROM - zde fixní tabulka instrukcí - a RAM čili přepisovatelný zbytek paměti s nestálým obsahem. Řídicí jednotka je obdobou dnešních procesorů a vykonává činnost stroje podle instrukcí, tedy například matematické operace nebo čtení z a zapisování do paměti. Kontrola nakonec dohlíží na správnost činnosti řídicí jednotky podle instrukcí paměti. Od roku 1937 se vlastně nic nezměnilo a jedná se o náčrt fyzické realizace nám už známého T-stroje. Snad může dnešnímu čtenáři připadat Turingův popis počítače naivní, musíme si ale uvědomit, že pro adresáta eseje z roku 1950 byly počítače buď něčím zcela neznámým, nebo velmi exotickým. Autor proto také vyzývá čtenáře, aby uvěřili v možnost konstrukce digitálních počítačů a jejich schopnost imitovat činnosti člověka. Na závěr této pasáže Turing vyrukuje s nejoriginálnější částí své definice a tvrdí, že jeho stroje jsou *počítače s neomezenou kapacitou*, což jsou *stroje nespojitých stavů* a zároveň *univerzální stroje*.<sup>12</sup> Počítače s neomezenou kapacitou mají paměť, která je teoreticky nekonečná v tom smyslu, že ji lze

---

<sup>11</sup> V originále "store", "executive unit" a "control".

<sup>12</sup> Tedy "infinite capacity computers", "discrete state machines" a "universal machines".

kdykoli rozšířit o další konečnou část. Stroje nespojitých stavů jsou ty, které se pohybují z jednoho konečného stavu do jiného. V případě digitálních počítačů založených na binaritě se tedy jedná o přechod ze stavu 0 do stavu 1 a zpět, přičemž mezi stavy neexistují žádné přechody. Turing podotýká, že stroje nespojitých stavů jsou jen hypotetické, protože ve skutečnosti se vše pohybuje kontinuálně; nespojitost čili diskrétnost je jen užitečným zjednodušením. Podle poslední charakteristiky jsou digitální počítače také univerzálními stroji, protože jsou schopny imitovat jakýkoli stroj nespojitých stavů a mohou být tak naprogramovány pro jakýkoli účel.<sup>13</sup>

Zatímco definice stroje je dnes v počítačové vědě téměř bezesbytku přijímána, Turingův pokus o definici myšlení je mnohem kontroverznější a setkal se s celou řadou oprávněných námitek, které je třeba brát vážně. Turing se domnívá, že vágní pojem "myslet" je možné nahradit specifičtějším "uspět v imitační hře", který v sobě zahrnuje přesná pravidla pro určení splnění podmínek. Imitační hra je nejprve popsána ve variantě, která nemá co dočinění s myšlením ani počítači. (Není přitom zcela jasné, zdali byla hra někdy skutečně hrána nebo zdali se jedná o autorův výmysl.) Výchozí verze hry se účastní tři hráči: muž (*A*), žena (*B*) a tazatel (*C*)<sup>14</sup>, jehož pohlaví není pro hru podstatné. Tazatel je v jiné místnosti než účastníci hry a jeho úkolem je pomocí série otázek zjistit, kdo je žena a kdo muž. Muž předstírá, že je žena, zatímco žena se snaží tazateli pomáhat při rozhodování. Komunikace je prováděna písemně nebo zprostředkovaně, například pomocí dálnopisu nebo terminálu, aby se tazatel nemohl řídit barvou hlasu nebo jinými fyzickými znaky. Turing navrhuje příklad otázek, které bychom dnes zřejmě považovali za nedostačující nebo genderově předpojaté, jako třeba "Jak dlouhé máte vlasy?" Zároveň správně předpokládá, že pro ženu je neúspěšnější strategií pravdomluvnost (např. "Mé vlasy jsou zastřižené do mikáda a dlouhé asi 23 centimetrů."), zatímco muž musí nutně lhát. Takto popsaná nevinná hříčka dostane dramaticky odlišný význam, pokud hru nehrají muž a žena, ale člověk proti počítači. V druhé variantě testu tazatel už nemá za úkol určit pohlaví účastníků hry, ale rozhodnout, kdo je skutečným příslušníkem lidského druhu a kdo se za něj jen vydává.

---

<sup>13</sup> Těmto strojům se dnes často říká "univerzální Turingovy stroje" a jsou samotným Turingem definovány jako "a single machine which can be used to compute any computable sequence" (Turing 1937, p.241).

<sup>14</sup> V angličtině jsou použity pojmy "competitors" a "interrogator". Z nedostatku české terminologie budu *A* a *B* označovat "účastníci", *C* pak jako "tazatele".

V úvodu této kapitoly jsem zmínil, že se Turing občas nevyjadřuje zcela jasně a na jedno takové pochybení bude vhodné poukázat hned. Problém vyplývá z ambivalence anglického termínu "a man", který může podle kontextu označovat buď muže, nebo obecněji člověka. Turing chce v pozměněné verzi hry nahradit roli A počítačem, takže by ji přísně vzato měli hrát počítač A, žena B a tazatel C. Když jsou později zmiňovány například "the intellectual capacities of a man", které by měl být počítač schopem emulovat, není úplně zřejmé, jedná-li se o "intelektuální schopnosti člověka" nebo "intelektuální schopnosti muže". Jakkoli mi tato nepřesnost přijde zanedbatelná a zřejmě založená na nepozornosti pisatele, kterému otázky genderu asi nepřipadaly dosti důležité, stala se právě ona základem pro "nestandardní" interpretaci (Hodges 1997, p.37). Přestože zbytek eseje tomuto čtení odporuje, existují autoři a autorky, kteří se snaží chápat test jako hru o identifikaci ženského myšlení v počítači. Genova (1994b) se domnívá, že imitační hra nemá určit biologický druh ("species"), ale pohlaví ("gender") jedince A. To, že má stroj hrát roli ženy, není podle ní projevem Turingova sexismu, ale údajně má jít o úmyslné napadení tradičního rozlišení pohlavních rolí; Turing tak prý ukazuje, že gender je sociálním konstruktem a nikoli biologickou daností. Turingovu snahu o vytvoření umělé, dokonalé a bezpohlavní bytosti nazývá utopickým aktem, srovnatelným s tím Pygmalionovým. Je nasnadě, že tato dekonstruktivistická analýza původní text dezinterpretuje k nepoznání.<sup>15</sup> Přesto nebyla opuštěna a už o rok později Hayes a Ford (1995) považují T-test za škodlivý. Podle autorů má T-test negativní dopady na vývoj umělé inteligence, a to z důvodů patriarchálního, šovinistického přístupu k ženám. Aniž by se příliš obtěžovali argumentací, v první části své eseje kategoricky tvrdí, že se "test netýká vytvoření umělého člověka, ale mechanického transvestity" (Hayes & K. Ford 1995, p.973). Cílem umělé inteligence totiž není stroj, který by byl umělou ženou, ale pouze něco, co by mluvilo jako žena, aby to podvedlo ostatní. Turing si prý vybral hru o identifikaci pohlaví úmyslně, protože homosexualita jej postavila před

---

<sup>15</sup> V současné feministické teorii snad nenajdeme ostudnější nepochopení myšlenek analytických filosofů, možná společně s manipulativním zneužitím Austinovy teorie mluvních aktů Judith Butlerovou; obojí je pěkným dokladem banalizace témat ženského hnutí. Pro další diskusi o Genovině genderové interpretaci viz totéž číslo časopisu *Social Epistemology* (Anderson 1994; Keith 1994; Genova 1994a).

celoživotní a bolestný problém: jak se chovat jako žena, i když jste ve skutečnosti muž.<sup>16</sup> Podobně uvažuje i Lassègue (1996), když srovnává myšlenky z "Computing Machinery" s událostmi Turingova dospívání; v jeho psychoanalytické, zřejmě vážně míněné komparaci se tak dozvíme třeba o okolnostech Turingovy první lásky nebo vlivu obřízky na jeho sexuální orientaci.<sup>17</sup> Nejrozumnějším zamyšlením nad pohlavními aspekty T-testu je článek Susan Sterrettové (2000), která sice také dává přednost pohlavní interpretaci před druhovou, ale činí tak bez ohledu na feminismus pro dobro samotné umělé inteligence: úspěšná imitace ženy vypovídá o inteligenci stroje mnohem víc než genderově neutrální konverzace s tazatelem.

Jak upozornili už mnozí jiní (např. Piccinini 2000; Shieber 2004, pp.103-104), Turing používá substantivum "a man" v generickém významu a jinde je při popisu T-testu mnohem přesnější. V již citované rozhlasové přednášce z roku 1952 zcela opomíjí genderovou variantu imitační hry a mluví jen a pouze o identifikaci člověka a stroje (Turing 2004a, p.495). Dalo by se říct, že přísné čtení příznivců genderové interpretace je vážným porušením Davidsonova principu vstřícnosti, který by měl být maximou pro jakoukoli komunikaci. Autorem zamýšlený význam musí mít přednost před významem doslovným; pravděpodobnější interpretace má prioritu před těmi méně pravděpodobnými. Vraťme se proto k původnímu Turingovu návrhu, jak by měla imitační hra probíhat.

Tazatel tentokrát nemá určit pohlaví účastníků, ale jejich příslušnost k lidskému druhu; tomuto účelu bude také přizpůsobovat své otázky. Bude se snažit konverzovat o tématech, která podle tradičního myšlení může skutečně chápat jen člověk. T-test je tematicky neomezený, a proto je možné klást dotazy z naprosto všech oblastí. "Lidskost" účastníka je posuzována podle kvality odpovědí, jejich přirozenosti, smysluplnosti, vtipnosti apod., tedy úplně stejně jako posuzujeme intelekt našeho partnera v běžné mezilidské konverzaci. Turing uvádí příklad malého výseku rozhovoru, který je záměrně nahodilý:

---

<sup>16</sup> Bylo by nespravedlivé, kdybych Hayesovi a Fordovi nepřiznal kredit za některá jejich smysluplnější tvrzení: například analogie současné umělé inteligence s někdejšími pokusy létat je podle mého názoru velmi výstižná (1995, pp.975-976).

<sup>17</sup> Podobně jako Genovin příspěvek vyvolal i Lassègue vlnu kritiky, přičemž diskuse trvá v podstatě dodnes (Leiber 2000; Lassègue 2000; Leiber 2006).

Otázka: Napište, prosím, sonet na téma železničního mostu přes řeku Forth.  
Odpověď: Tak v tom se mnou nepočítejte. Neumím psát poezii.  
Otázka: Sečtete 34957 a 70764.  
Odpověď: (Pauza asi 30 sekund) 105621.  
Otázka: Hrajete šachy?  
Odpověď: Ano.  
Otázka: Mám krále na E1 a žádné jiné figury. Vy máte jen krále na E3 a věž na A8. Jak budete hrát?  
Odpověď: (Pauza asi 15 sekund) Věž na A1, šach-mat. (Turing 1950, pp.434-435)

Všimněme si, že v této krátké ukázce jsou naznačena tři témata, která se liší mírou "lidskosti". Psaní poezie je považováno za kreativní činnost *par excellence*, která je lidskou výsadou. Počítání bylo už v Turingově době u počítačů samozřejmostí, otázka má proto spíš zjistit rychlost reakce: příliš rychlá odpověď by napovídala, že partnerem v komunikaci je stroj (nebo geniální počtář). V roce 1950 se spíš spekulovalo o tom, že by počítače byly schopné hrát šachy na úrovni člověka. První šachové programy vznikly až na přelomu 50. a 60. let a převaha počítačů nad šachovými velmistry začala až roku 1997, vítězstvím Deep Blue nad tehdejším úřadujícím mistrem světa Garry Kasparovem. Dnes je proto jasné, že "chess is too easy"<sup>18</sup> a schopnost hrát šachy má jen pramálo společného s pravou umělou inteligencí. O několik stránek dál je popsána ještě jedna ukázka imitační hry, která se opět zaměřuje na schopnost počítače chápat poezii. Cílem je prokázat, že účastník konverzace nemůže jen bezmyšlenkovitě "papouškovat" to, co předtím slyšel:

Tazatel: První verš vašeho sonetu zní "Mám srovnávat tě s krásným letním dnem?" Nemyslíte, že by bylo stejně vhodné nebo dokonce lepší použít "podzimní den"?<sup>19</sup>  
Účastník: Báseň by pak neměla správný rytmus.  
Tazatel: A co "zimní den"? To by se rytmicky hodilo.  
Účastník: Ano, ale nikdo nechce být přirovnán k zimnímu dni.  
Tazatel: Řekl byste, že vám pan Pickwick připomíná Štědrý den?  
Účastník: Svým způsobem ano.  
Tazatel: Přitom Štědrý den je zimní a nemyslím si, že by panu Pickwickovi toto přirovnání vadilo.  
Účastník: To nemůžete myslet vážně. Zimním dnem přece většina lidí myslí typický zimní den, a ne sváteční jako je ten Štědrý. (Turing 1950, p.446)

---

<sup>18</sup> Tak pojmenoval svůj komentář ke hře Deep Blue a Kasparova filosof Selmer Bringsjord (1998).

<sup>19</sup> V originále "spring day"; bohužel české "jarní" má stejný počet slabik jako "letní" nebo "zimní", tedy dvě. Podobně nepřesně překládám i "Christmas" jako "Štědrý den", i když se v Anglii slaví o den později než u nás. Překlad Shakespearova XVIII. sonetu Martin Hilský.

Lze říct, že Turing redukuje úspěch v imitační hře na verbální kritérium, na schopnost komunikovat prostřednictvím jazyka. Otázkou zůstává, zdali Turing skutečně považoval toto ryze behaviorální kritérium za definici inteligence. Začátek eseje by tomu snad i napovídal: nejprve je třeba nalézt definice slov "myslet" a "stroj" (Turing 1950, p.433). Později je vágní věta "Mohou stroje myslet?" nahrazena exaktnější "Mohou hypoteticky existovat digitální počítače, které by si dobře vedly v imitační hře?" (Turing 1950, p.442). Zdá se proto, že pojmy "myslet" a "uspět v imitační hře" jsou synonymní, že delší formulace může sloužit jako definice kratší formulace. Turing bývá běžně označován za čelního představitele analytického behaviorismu, společně s Rylem a Quinem snad za toho vůbec nejdůležitějšího. Jeho imitační hra je pak považována za typicky behavioristickou, operacionální definici myšlení (např. Millar 1973, p.595; Hodges 1992, p.415; French 2000b, p.115). Sám autor to ale v textu nikdy explicitně neuvádí, zřejmě úmyslně, protože situace je mnohem složitější, než se může na první pohled zdát. První, kdo se domníval, že T-test není možné považovat za definici myšlení, byl James H. Moor (1976). Podle Moora se nejedná ani tak o definici, jako spíš o induktivní metodu kumulující evidenci pro vyslovení hypotézy, že stroj myslí. Úspěch v T-testu neznamena definitivní odpověď na otázku "Mohou stroje myslet?", ale spíš něco na způsob "Je průměrná úspěšnost tazatele při odhalení počítačového účastníka po  $n$  minutách nebo  $m$  otázkách podstatně vyšší než 50%?" Pravděpodobnost správnosti takto formulované hypotézy se přitom zvyšuje s narůstající délkou testu a počtem otázek, T-test tudíž vyhovuje běžným standardům vědeckého testování hypotéz. Moor navíc předkládá svou vlastní definici, která je dle jeho názoru naprosto neproblematická: "Myšlení je zpracování informací pomocí metod, které zahrnují rozpoznávání, představivost, vyhodnocování a rozhodování." (Moor 1976, p.250)<sup>20</sup> Podle tohoto pojetí "myslí" v omezené míře i dnešní počítače, otázkou tedy je, budou-li stroje někdy schopny myslet na takové úrovni složitosti jako normální dospělí lidé - a k tomu má právě přispívat induktivně chápaný T-test.

Jakkoli je toto řešení svou jednoduchostí sympatické, neodpovídá zřejmě Turingovým původním ambicím, a je proto naneštěstí nepřijatelné. Pokud je každé následující kolo imitační hry jen dalším kouskem do induktivní skládačky naší teorie, pak je posledním dílkem odhalení totožnosti účastníka imitační hry. Zjistíme-li, že tímto účastníkem je stroj, je to

---

<sup>20</sup> V originále: "To think is to process information in ways which involve recognition, imagination, evaluation and decision."



důvod pro falzifikaci, odmítnutí hypotézy.<sup>21</sup> Indukce není dostatečným vysvětlením T-testu, protože kromě ní by bylo třeba ještě nějaké další kritérium, které by nám umožnilo označit stroj za nepochybně myslící. Turing věřil, že tímto definitivním kritériem je právě T-test, nic víc není třeba. Domnívám se, že existuje jen jedna uspokojivá interpretace, a to ta, která chápe T-test jako postačující podmínku myšlení. Jen pro pořádek připomínám, že propozice  $p$  je logicky *nutnou* podmínkou propozice  $q$  v případě, kdy  $q$  nemůže být pravdivá, pokud je  $p$  nepravdivá ( $p \leftarrow q$ ). Naopak propozice  $p$  je logicky *postačující* podmínkou propozice  $q$  v případě, kdy  $p$  nemůže být pravdivá, pokud je  $q$  nepravdivá ( $p \rightarrow q$ ). A nakonec podmínka je logicky *nutná a postačující* tehdy, když platí obojí zároveň ( $p \leftrightarrow q$ ). Mějme proto věty:

- (1) Entita X uspěla v T-testu.
- (2) Entita X myslí.

Věta (1) není logicky nutnou a postačující podmínkou věty (2), výroky tedy nejsou ve vztahu ekvivalence a nemohou být proto považovány za základ pro definici. Věta (1) je pouze postačující podmínkou věty (2), z čehož vyplývá:

- (3) Každá entita, která uspěje v T-testu, myslí.
- (4) Neplatí, že každá myslící entita uspěje v T-testu.

Jak později uvidíme, je tento způsob čtení obranou proti většině námitek, se kterými kritici T-testu v průběhu následných desetiletí vyrukovali (pro detailnější diskusi viz Oppy & Dowe 2011, oddíl 4; případně Block 1981, pp.14-16).

Turing věnuje podstatnou část eseje vyvrácení devíti potenciálních námitek. (Obdobně bude o třicet let později postupovat i Searle v článku "Minds, Brains and Programs".) Lze říct, že vlastně předešel všechny relevantní způsoby kritiky více či méně úspěšně. Podívejme se na ně detailněji.

---

<sup>21</sup> Podobně argumentuje Douglas F. Stalker v diskusi, která proběhla na stránkách *Philosophical Studies* (Stalker 1978; srov. i Moor 1978; Moor 2001; Copeland 2000, pp.522-524).

### 3.1 Teologická námitka

První Turingova námitka není považována za příliš důležitou a v sekundární literatuře je jen velmi málo zmiňována. Jedná se totiž o argumentaci prostřednictvím křesťanské zvěsti:

Myšlení je funkcí lidské nesmrtelné duše. Bůh dal nesmrtelnou duši každému muži a každé ženě, ale nikoli jiným zvířatům či strojům. Proto zvířata ani stroje nemohou myslet. (Turing 1950, p.443)

Ani sám Turing této úvaze zřejmě nepřikládá velký význam a tvrdí: "Nejsem z toho schopen přijmout ani jednu část." Náboženství odmítá jako celek, protože v minulosti často bránilo vědeckému a společenskému pokroku. Kromě toho je přesvědčen, že živočichové patří do stejné skupiny jako lidé; rozdíl mezi živými bytostmi a jejich neživými protějšky je mnohem markantnější než rozdíl mezi lidmi a zbytkem přírody. Křesťanská ortodoxie je při klasifikaci entit obdařených duší stejně arbitrární jako ta muslimská, která upírá duši ženám. Turing se čtenáři svěřuje: "Teologické argumenty na mě nedělají valný dojem, ať už mají podporovat cokoli." Přesto se snaží nabídnout i teologické řešení a předcházet tak případnou kritiku, která by z toho směru mohla přijít. Předpoklad, že stroje nemohou mít duši, je v rozporu s učením o všemohoucnosti boha. Podle Turinga je monoteistický bůh křesťanů omezen jen logickou nemožností, což ale není případ vlastnictví duše zvířaty nebo stroji. Pokud by se bohu zachtělo, mohl by - třeba naším prostřednictvím - stvořit cokoli, tedy i myslící stroj.

Předložený důkaz není příliš silný, protože operuje se špatně nadefinovanou a zřejmě i špatně chápanou všemohoucností boha. Není to Turingova chyba, v celých dějinách filosofie a přirozené teologie se snad nenašlo řešení tohoto problému, které by bylo po logické a etické stránce uspokojivé. Mnohem podivnější ale je, proč se Turing vůbec pouštěl na tenký led teologických úvah tohoto typu. Jedná se o jedno ze tří zvláště temných míst Turingovy eseje, jejichž opodstatnění je jen málo pochopitelné - kromě této námitky se jedná ještě o nekomentovanou pasáž zmiňující stěhování duší (Turing 1950, p.444) a rozsáhlou část o mimosmyslovém vnímání (námitka 9, Turing 1950, pp.453-454). Turingův životopisec Hodges, který je na jeho skryté pohnutky asi největší autoritou, se domnívá, že se v tomto případě jedná o ironii, o pokus zesměšnit tradičně scholastický způsob dokazování v otázkách moderní vědy (Hodges 1997, pp.39-40). Navrhuji ale ještě jeden možný způsob čtení, který se sice vymyká původnímu vyznění textu, ale je filosoficky produktivnější a dodržuje elementární ctnost vědeckého diskurzu: za žádných okolností nezatahovat do diskuse

pseudoargumenty vycházející z náboženství. Pokud si odmyslíme z této námitky vše teologické, tj. především boha a nesmrtelnou duši, je možné redukovat ji na Descartovo stanovisko. Turing se možná negativně vymezuje vůči dualismu, který je v příkrém rozporu s jeho vlastními cíly. Tím, že stírá hranici mezi hmotným a myslícím, se hlásí k nějaké formě naturalismu, která je pro jeho filosofii nezbytná. Turingův behaviorismus navíc zcela popírá možnost existence netělesných, a tudíž nepozorovatelných duší. Teologickou námitku bychom tak měli spíš chápat jako odmítnutí karteziánství, nikoli travestii křesťanství.

### 3.2 Námitka "hlavy v písku"

Skutečným důvodem, proč je mezi lidmi tolik populární teologická námitka, je podle Turinga antropocentrické elitářství. Lidé mají pocit, že jsou ontologicky nadřazení zbytku světa, a neradi vidí, když je jejich výsadní postavení někým nebo něčím ohrožováno. Dokonce by se dalo říct, že čím menší je skupina "my", tím větší pocit výjimečnosti zažíváme. Podle známé poučky lze historii novověku úspěšně chápat jako dějiny vyčleňování člověka z centra univerza: Koperník odstranil Zemi ze středu Sluneční soustavy a Galilei jeho teorii radikalizoval tvrzením, že Slunce je jen jednou hvězdou z myriád jiných. Darwin zpochybnil nadřazenost člověka nad ostatními živočišnými druhy a naznačil hypotetickou existenci společného předka, který je předchůdcem všeho živého na Zemi. Rovnostářské dvacáté století nivelizovalo nerovnosti uvnitř našeho druhu, protože došlo k zrovnoprávnění lidí různých pohlaví, ras a etnicit. Turingovo teoretizování o myslících strojích je tak útokem na poslední palisádu výjimečnosti, která nám zůstala: pokud jeho názory přijmeme za své a přiznáme myšlení i mechanickým strojům, pak bude do našeho kdysi výlučného "my" patřit kde co. Turing nazývá tento pocit námitkou "hlavy v písku" ("the 'Heads in the Sand' Objection"); lidé se totiž chovají podobně jako přísloveční pštrosi, ukrywající své hlavy před nebezpečím do písku:

Důsledky plynoucí z existence myslících strojů by byly příliš hrozný. Doufejme a věřme, že toho nejsou schopny. (Turing 1950, p.444)

Na první pohled je vidět, že se nejedná ani tak o argument proti samotné možnosti existence myslících strojů, ale o vyjádření praktických obav nad jejími možnými důsledky. Není obvyklé, aby byla tato námitka podrobněji zkoumána sekundární literaturou, o to běžněji se objevuje

- mluvím jen z osobní zkušenosti - při neformální konverzaci s lidmi. Nedůvěra k vytvoření umělé inteligence má podle mě několik podob: (a) Umělá inteligence, představovaná třeba humanoidními roboty, je nepotřebná, protože už dnes žije na světě šest a tři čtvrtě miliardy lidí ochotných udělat pro peníze cokoli. Proč vyvíjet finančně nesmírně náročné elektronické sluhy, když ve třetím světě najdeme spoustu zájemců o tutéž práci, kteří ji navíc vykonají za minimální mzdu? (b) S tím je spojená obava ze sociálních dopadů hromadného zavádění inteligentních strojů, která je obdobou ludditského hnutí 19. století. Stejně jako luddité průmyslové revoluce rozbíjeli stroje v továrnách, protože jim údajně brali práci, mají dnes lidé obavu z technologického outsourcingu. Představme si jen ten nejjednodušší příklad za všechny: počítač, který úspěšně projde T-testem, by mohl být ihned zaměstnán třeba v telemarketingu nebo v telefonických poradenských centrech. Ze dne na den by tak přišly o práci miliony lidí, kteří by se mohli považovat právě za oběti nastupující umělé inteligence. (c) Někteří lidé se obávají, že by se systémy umělé inteligence mohly člověku nejen vyrovnat, ale časem jej i překonat. Bylo by tak zřejmě jen otázkou času, kdy by se stroje přestaly spokojovat s podřízeným postavením a převzaly by vedoucí postavení ve společnosti. Jakkoli nám tyto úvahy mohou připadat jako science-fiction, někteří autoři je neberou na lehkou váhu. Tak třeba Hayes a Ford místo vývoje lidem podobných myslících strojů navrhují zabývat se vytvářením vysoce specializovaných expertních systémů, které by v T-testu rozhodně neuspěly, ale zato by byly schopny bezproblémově vykonávat to, k čemu jsou určeny (Hayes & K. Ford 1995, pp.974-975). Britský specialista na kybernetiku Kevin Warwick si zjevně libuje v popisování neradostné budoucnosti, která jako by vypadla z filmu *Terminátor*. Podle něj je docela dobře možné, že stroje převezmou vládu nad světem někdy kolem roku 2050. Jeho argumentace je přitom velmi jednoduchá:

1. My lidé jsme nyní dominantní formou života na zemi, protože jsme nejinteligentnější.
2. V dohledné budoucnosti se mohou stroje stát inteligentnějšími, než jsme my.
3. Pak se stroje stanou dominantní formou života na zemi. (Warwick 1999, p.198; překlad Alena Čechová)

Všechny tři varianty této námitky, jak jsem je zde popsal, nelze brát dle mého názoru příliš vážně. Strach z neznámého je antropologickou konstantou stejně jako lidská posedlost očekáváním apokalypsy. Novodobí luddité zmatečností připomínají své někdejší předchůdce: stroje ve skutečnosti neberou lidem práci, ale umožňují jim zabývat se činnostmi, na které

předtím nebyl čas. Po rozšíření strojové výroby následovala expanze terciární sféry služeb a přepokládám, že po nástupu umělé inteligence by mohlo dojít k masovému nárůstu kvartální sféry vědy a výzkumu. Veškeré prognózy o budoucím vývoji umělé inteligence a jeho dopadu na morálku, společnost a politiku jsou ovšem jen intelektuálním cvičením. Nakonec je zcela lhostejné, zdali jste technooptimistou nebo technopesimistou, protože vědecký pokrok je neustávající proces a nezastaví se jen kvůli tomu, že má nějaká skupina humanitně orientovaných myslitelů obavy z budoucího zneužití. Věda není dobrá ani špatná, je hodnotově neutrální; etické posouzení se bude vždy týkat jen praktických aplikací jejích teoretických poznatků. Řešení tohoto problému nám ve své eseji nenabízí ani Turing, jen doporučuje nalezení útěchy v myšlence stěhování duší. Zřejmě jen on sám by nám mohl vysvětlit, co tím myslel.

### 3.3 Matematická námitka

Moderní matematická logika ve 20. století ukázala určitá omezení strojů nespojitých stavů, která vyplývají například z Gödelova teorému o neúplnosti<sup>22</sup>:

[V] každé dostatečně silné logické soustavě lze formulovat tvrzení, která nemohou být v rámci soustavy dokázána ani vyvrácena, pokud sama soustava není nekonzistentní. (Turing 1950, p.444)

Kromě Gödela k podobnému závěru dospěli i Church, Kleene nebo Rosser, Turing ale dává v tomto případě přednost svému vlastnímu postupu, který zahrnuje stroj, tedy digitální počítač s neomezenou kapacitou (viz kapitola 2 této práce). Turing z věty o neúplnosti vyvozuje, že "existují určité věci, které stroj nemůže udělat" (Turing 1950, p.444). V průběhu imitační hry by počítač na některé otázky odpovídal nesprávně nebo neodpověděl vůbec, i kdyby měl k dispozici neomezené množství času.<sup>23</sup> Tyto otázky by se týkaly většinou autoreference: zeptáme-li se například počítače X "Odpoví někdy X na nějakou otázku 'ano'?", je podle Turinga matematickým faktem, že stroj nebude schopen vytvořit odpověď. Turing tuto námitku přijímá, vždyť ji koneckonců vymyslel sám, ale odmítá souhlasit

---

<sup>22</sup> Jedná se ve skutečnosti o dva teorémy, ale Turing mezi nimi nerozlišuje. Pokud by používal dnes přijímanou terminologii, mluvil by zřejmě o první větě o neúplnosti.

<sup>23</sup> Proto se této vlastnosti T-stroje říká problém zastavení ("the halting problem")

s nedokázaným tvrzením, že se toto omezení netýká lidského rozumu. Pokud totiž prozkoumáme náš intelekt pozorněji, nalezneme podobná omezení jako u počítačů - Turing ale přesněji nespecifikuje jaká.

Třetí námitka se stala, i když nechápu přesně proč, z původních devíti tou vůbec nejpůvodnější. Od šedesátých let do dneška vyšlo kolem stovky vědeckých článků, ve kterých jsou prezentovány argumenty pro a proti; naprostá většina se přitom zabývá podružnými, technickými detaily a používá spíše větší než menší množství nástrojů formální logiky.<sup>24</sup> Nenaplnil se přitom Turingův odhad, že příznivci matematické námitky budou uznávat imitační hru jako východisko pro diskusi o umělé inteligenci. Právě naopak: tato slabina mechanických myslí bývá používána k úplné destrukci počítačového přístupu k myšlení. Úplně první vyjádření skepticismu tohoto typu nalezneme v dnes už klasické knize Ernesta Nagela a Jamese R. Newmana z roku 1958, která zpopularizovala Gödelův matematický objev:

Gödelovy závěry souvisí s otázkou, zda lze zkonstruovat počítač, který by se co do matematické inteligence vyrovnal lidskému mozku. Dnešní počítače mají v sobě zabudovanou fixovanou množinu příkazů; tyto příkazy odpovídají fixovaným pravidlům odvození formalizované axiomatické procedury. Tyto stroje tak odpovídají na problémy tím, že krok za krokem provádějí operace, přičemž každý krok je řízen zabudovanými příkazy. Jak však prokázal svým teorémem neúplnosti Gödel, v elementární teorii čísel je nesčetně mnoho problémů, které jsou mimo dosah fixované axiomatické metody, problémů, na které tyto stroje odpověď dát nemohou, ať jsou jejich zabudované mechanismy jakkoli složité a důmyslné, a jakkoli rychlé jsou jejich operace. Je-li dán určitý problém, lze postavit takový stroj na jeho vyřešení; není však možné postavit žádný takový stroj, který by vyřešil každý problém. Jistě, lidský mozek má zabudovaná svá omezení a mohou existovat takové matematické problémy, které není schopen vyřešit. I tak to ale vypadá, že mozek ztělesňuje strukturu operačních pravidel daleko silnější, než je táž struktura u běžně chápaných umělých strojů. Nemáme proč se v této chvíli domnívat, že lidská mysl může být nahrazena roboty. (E. Nagel & Newman 2006, pp.87-88; překlad Rostislav Niederle)

Skutečným průkopníkem tohoto způsobu kritiky se stal britský filosof J. R. Lucas, který v roce 1961 formuloval nejběžnější variantu námitky a od té doby ji hájil pomocí krátkých odpovědí, tzv. "rejoinders". Lucas nejprve stručně demonstroval Gödelův teorém na větě "Tato formule

---

<sup>24</sup> Podrobná, ale ani zdaleka ne úplná bibliografie je k nalezení na stránce *MindPapers* na adrese <<http://consc.net/mindpapers/6.1b>>.

je v této soustavě nedokazatelná." Pak vysvětlil, proč je správné aplikovat teorém v oblasti umělé inteligence: "Gödelův teorém se musí týkat kybernetických strojů, protože k esenciálním vlastnostem stroje patří, že je konkrétní realizací formálního systému." (Lucas 1961, p.113) Z toho vyplývá, že myslí počítačů nejsou úplné; vždy v nich existuje aspoň jedna věta  $p$ , ale spíš třída vět  $P$ , které jsou pravdivé, ale v rámci soustavy  $S$  nedokazatelné. Lidská mysl není tímto nedostatkem ovlivněna, čímž se fundamentálně liší od myslí počítačových. Znamená to tedy, že počítačové modely nejsou zcela adekvátní svým lidským protějškům, Gödelův teorém nám ukazuje jejich Achillovu patu. Lidé vždy dokážou aspoň něco, co počítače nesvedou.

Lucasův důkaz je elegantní a vypadá přesvědčivě, přesto je podle mého názoru zatížen fatálním nedostatkem. Podle Lucase nejsou počítače schopny dokázat platnost  $p$ , zatímco lidé mohou důkaz pohodlně provést, protože mají možnost vystoupit "ven" z formálního systému. Neustále opakuje, že racionální bytosti stojí *mimo* tento systém.<sup>25</sup> Je zarážející, že téměř třicet let po Tarského rozlišení mezi objektovým jazykem a metajazykem Lucas nechápe, jak tento zdánlivý paradox vyřešit. Ono vystoupení ven ze systému totiž není ničím jiným než nahlédnutím na objektový jazyk pomocí metajazyka. Je pravda, že věta  $p$  je nedokazatelná v soustavě  $S_1$ , ale my můžeme tuto nedokazatelnost konstatovat v nadřazené soustavě  $S_2$ . A i když se v soustavě  $S_2$  setkáme opět s neúplností, můžeme postup opakovat povýšením na úroveň  $S_3$  a tak dál *ad infinitum*. Proč neumožnit strojům totéž, co děláme běžně my lidé? Není složité uložit do programu počítače teorii hierarchie jazyků. Počítače dokonce už dnes něco podobného obsahují: je to ona dobře známá hierarchie strojového kódu, vyššího programovacího jazyka a vlastního programu. Vysvětlovat odlišnost lidských a počítačových myslí odkazem na Gödelův teorém je proto iluzorní. Není mi známo, že by nějaký autor argumentoval proti matematické námitce stejným způsobem jako já. O něco podobného se pokusil jen C. H. Whiteley ve své kratičké, ironické reakci na Lucasův původní článek. Whiteley hned na začátku tvrdí, že nám Lucas nepředkládá důkaz, ale logickou obdobu salónního triku. Podle Lucase stroj nemůže provést důkaz  $p$ , protože je strojem. Whiteley se oproti tomu domnívá, že chyba není ve stroji, ale ve vlastním zadání příkladu. Mějme například větu "Tato formule je Lucasem nedokazatelná." Sám Lucas větu nebude

---

<sup>25</sup> V originále "rational being standing outside the system" (např. Lucas 1961, pp.113, 117, 120, 124).

schopen dokázat, a to ze stejných důvodů jako stroj v jeho článku. Naopak my všichni ostatní můžeme s klidným svědomím tu samou větu tvrdit a považovat za pravdivou. Znamená to snad, že Lucas nemyslí na stejné úrovni jako my, protože existuje aspoň jedna věta, které my rozumíme a on ne? (Whiteley 1962, p.61)

I když mi tento způsob argumentace připadá dostatečný, uvádím ještě dva další důvody pro odmítnutí matematické námitky. První: abychom mohli identifikovat větu  $p$  nebo třídu vět  $P$ , které nelze dokázat ani vyvrátit, musíme mít velmi detailní, snad úplné informace o soustavě  $S$ , tedy o obsahu mysli. V případě počítačů je to možná proveditelné, protože jsou zatím docela jednoduchými systémy (aspoň ve srovnání s lidskými mozky) a máme možnost se setkat s jejich tvůrci čili programátory. U lidí to ale rozhodně možné není; nejen že jsou jejich mysli mnohonásobně složitější, ale navíc jejich vývoj probíhal po dobu milionů let metodou pokusu a omylu, prostřednictvím mechanismu přirozeného výběru. To znamená, že i lidská mysl může být počítačem, jen neznáme její program.<sup>26</sup> Druhý: tvrzení, že počítačové mysli jsou konzistentními systémy, zatímco lidské mysli jsou nekonzistentní, je intuitivní a v empirii pro ně nenalzáme příliš mnoho opory. Idealizované počítače a ideální lidé jsou dokonale bezrozpornými mysliteli, reálný svět takovou dokonalostí nedisponuje. Počítače se sice dopouštějí menšího množství chyb než lidé, ale podstatné je, že občas chybujeme všichni.<sup>27</sup>

Je zajímavé položit si tradiční právnickou otázku *cui bono*? Proč vlastně Lucas brojí proti komputacionalismu (v jeho terminologii "mechanicismu") pomocí natolik obskurního prostředku, jako je metamatematická poučka o hranicích axiomatické metody? Odpověď na tuto otázku nám odhalí poslední stránky Lucasovy eseje:

Jestliže je důkaz nesprávnosti mechanicismu platný, má to obrovský význam pro celou filosofii. Od časů Newtona byli filosofové posedlí příznakem mechanického determinismu. Pokud jsme chtěli být vědečtí, zdálo se, že

---

<sup>26</sup> To má zřejmě na mysli Paul Benacerraf, když píše, že pokud by byla mysl T-strojem, selhala by při plnění sókratovského programu "poznej sám sebe" (Benacerraf 1967, p.30).

<sup>27</sup> Lucas ve svém rekapitulačním článku (Lucas 1996, pp.106-107) identifikuje pět druhů kritiky, se kterými se jeho argument během třiceti pěti let musel vypořádat. Jsou to připomínky týkající se: (a) idealizace mysli a stroje (Hanson 1971), (b) modality a konečnosti (Boyer 1983), (c) transfinitní aritmetiky (Good 1967), (d) možnosti formalizace racionálního odvozování (Webb 1968), (e) konzistence (Chihara 1972); uvedená bibliografie je jen výběrová a nepřekrývá se přesně s texty, které uvádí Lucas.



musíme na lidské bytosti pohlížet jako na determinované automaty, nikoli jako na autonomní morální agenty [...]. Ale nyní, i když stále přetrvává mnoho argumentů proti lidské svobodě, ztratil svou sílu argument mechanicismu, pravděpodobně ten nejpřesvědčivější ze všech. (Lucas 1961, pp.126-127)

Nemohu si pomoci, ale zdá se mi, že celá matematická námitka je vlastně jen variantou hlavy v písku. Nejedná se tentokrát o obavy z potenciální nadvlády strojů nad lidmi, ale o udržení představy člověka jako ontologicky samostatné a eticky svobodné entity. Není divu, že další Lucasovou knihou byla *The Freedom of Will*, ve které se snaží pomocí Gödelova teorému argumentovat proti fyzikálnímu determinismu a ve prospěch svobodné vůle. Díky větě o neúplnosti se lidé vymykají zbytku přírody, jsou racionální a mohou jednat mravně (Lucas 1970, p.171). Toto stanovisko je navíc spojeno s Lucasovým nenápadným antiscientismem: "vědecké bádání nemůže vyčerpat nekonečnou rozmanitost lidské mysli" (Lucas 1961, p.127). Lucas nenavrhuje žádnou konstruktivní alternativu vůči vědecké teorii mysli, jen se nedokáže smířit s tím, že by mohla věda vysvětlit něco tak intimního a tajemného, jako je mentální život člověka. Je to defétismus téměř fenomenologický: ať si přírodní věda dělá, co chce, přirozený svět naší zkušenosti nepochopí nikdy.

Přes neutuchající odpor nebyl Lucasův argument opuštěn, ale dočkal se oživení v pracích Rogera Penrose (1989; 1994). Penrose píše velmi rozsáhlé knihy, ve kterých sleduje mnoho různých cílů a dotýká se mnoha témat, proto si budu všimát jen částí, ve kterých se zabývá vztahem Gödelova teorému k teorii mysli. I Penrose nejprve v *The Emperor's New Mind* vysvětlí, co to vlastně Gödelův teorém je (Penrose 1989, pp.138-141), a později - shodou okolností přesně po 400 stranách - po Lucasově vzoru nastíní, jak teorém vyvrací komputační teorii (Penrose 1989, pp.538-541). Rekonstruujme Lucasův příklad pomocí Penrosovy terminologie: v každém formálním systému se nacházejí matematické propozice jako Gödelova věta - Penrose je označuje jako propoziční funkci  $P_k(k)$ . Tyto propozice nelze dokázat žádným, jakkoli rozsáhlým algoritmem. Matematikové jsou si ale přesto vědomi jejich pravdivosti, kterou jsou schopni nahlédnout ("seeing") blíže nespecifikovaným způsobem. Matematické pravdy jsou tudíž něčím víc než pouhými produkty algoritmů. Z toho vyplývá, že vědomí má nealgoritmickou povahu.

Penrose sice pozměnil slovník, kterým důkaz vyjadřuje, princip je ale tentýž. Změny jsou jen kosmetické: "vědomí" nahradilo "myšlení", "algoritmický" zastupuje "mechanistický". Na Penrosovu verzi se proto vztahují identické námitky jako na Lucasův

originál. Totéž lze v podstatě říci i o inovované variantě z *Shadows of the Mind* (Penrose 1994, kapitola 2 a 3).<sup>28</sup> Penrosova motivace pro odmítnutí komputacionalismu je snad ještě obskurnější než ta Lucasova. Tento vynikající matematik a spolupracovník Stephena Hawkinga usiluje o vyvrácení teorie, podle které je mysl programem mozku, aby ji mohl nahradit jinou, podle které je mysl projevem subatomárních částic podílejících se na chodu mozku. Kromě postřehu, že správná teorie mysli musí brát v potaz jevy popisované kvantovou mechanikou, není Penrose schopen říct o ní cokoli detailnějšího.

Na závěr této části si dovoluji ještě jednu úvahu, a to o ošidnosti používání Gödelova teorému mimo kontext logiky a matematiky. Sokal a Bricmont věnovali celou jednu kapitolu ve své knize *Intellectual Impostures* příkladům nepochopení Gödelovy sentence. Lucas nebo Penrose určitě nemohou být obviňováni z neznalosti logiky a nelze je házet do jednoho pytle s intelektuálně nespolehlivými šarlatány jako je Kristeva nebo Debray. Jejich chybou je ale příliš doslovná aplikace logiky na realitu, snaha naroubovat dokonale zmapovaný deduktivní systém na z větší části neznámý svět. Řečeno společně se Sokalem a jeho partnerem:

Metateorémy matematické logiky, jako je Gödelův teorém [...], mají poněkud odlišný status oproti konvenčním matematickým teorémům. Mělo by být proto zdůrazněno, že tato vysoce specializovaná odvětví základů matematiky mají jen velmi malý dopad na zbytek matematického výzkumu a téměř žádný dopad na přírodní vědy. (Sokal & Bricmont 1998, p.128n)

Existují-li důvody pro odmítnutí komputacionalismu a T-testu, není třeba je hledat v logicky konzistentním světě platónských idejí.<sup>29</sup>

### 3.4 Argument z vědomí

Následující námitka bývá, stejně jako ta předchozí, velmi často přejímána odpůrci mechanistického, funkcionálního přístupu k mysli. Turing ji tentokrát neformuluje sám, ale

---

<sup>28</sup> Pro kritiku důkazu z roku 1989 viz Dodd (1991) a Sloman (1992); pro verzi z roku 1994 viz Feferman (1995), Bringsjord and Xiao (2000) a Lindström (2001). Zatím posledním kolem sporu jsou McCall (1999) a jeho kritici George and Velleman (2000).

<sup>29</sup> K podobnému závěru dospěli i Hofstadter (1979) a Franzén (2005): aplikace Gödelova teorému mimo oblast axiomatických disciplín, jako třeba v teorii mysli, je možná jen metaforicky, nikoli doslovně.

cituje její tradiční variantu, kterou našel v přednášce britského neurologa Geoffreyho Jeffersona:

Dokud nebude stroj schopen napsat sonet nebo složit koncert na základě svých myšlenek a emocí, nikoli náhodným výběrem symbolů, nemůžeme přistoupit na tvrzení, že stroj rovná se mozek. To znamená, že stroji nestačí sonet jen napsat, ale musí i vědět, že jej napsal. Žádný mechanismus nemůže cítit (nejen uměle dávat najevo, což je snadné) uspokojení z úspěchu nebo smutek ze spálené elektronky, nemůže být potěšen lichotkami, zarmoucen svými chybami, okouzlen opačným pohlavím, rozzloben nebo deprimován, když není schopen dosáhnout toho, co chce. (Turing 1950, pp.445-446; Jefferson 1949, p.1110)

Jefferson argumentuje dosti intuitivně a směšuje v jedné námitce více oddělených témat. Jedná se vlastně o tři problémy: (a) neschopnost stroje mít vědomí, respektive uvědomovat si sám sebe; (b) neschopnost stroje vytvářet umělecká díla, jako například poezii nebo hudbu; (c) neschopnost stroje pociťovat emoce, jako třeba smutek nebo zlost. Část týkající se umění Turing odbývá ukázkou rozhovoru, ve které tazatel s účastníkem analyzuje sonet a kterou jsem citoval už dříve. Pokud je výkon počítače v imitační hře dostatečně přesvědčivý, nezbyvá nám než uznat jeho schopnost vyznat se v poezii. Není přitom podstatné, jestli byl sonet napsán na základě pisatelových emocí nebo pomocí chladného, racionálního kalkulu - to přece nedokážeme ani v případě Poeova "Havrana"! Snaha určovat "lidskost" účastníka hry pomocí umění je vůbec dosti pochybná, zvláště dnes, kdy existují programy generující literaturu, hudbu i obrazy. Morgensternovo "Veliké lalulá", Cageovo "4'33" nebo Malevičův "Černý čtverec na bílém pozadí" mohl klidně vytvořit počítač, abych uvedl jen ty nejkřiklavější případy. Není složité představit si naopak díla, která působí lidsky, i když jsou dílem stroje. Příkladem může být náhodně generované haiku, algoritmicky komponovaná hudba nebo počítačem vizualizované fraktály.

Nejpodstatnější částí této námitky je ale samozřejmě otázka vědomí a sebevědomí. Jefferson rozlišuje mezi náhodným a úmyslným napsáním sonetu, Turing ale upozorňuje na obtížnost takové dichotomie. Podle něj totiž jediný způsob, jak si být skutečně jistý, že má stroj vědomí, je být tímto strojem. Problematiku vědomí do diskuse o T-testu uvedl především Donald Michie v návaznosti na Thomase Nagela a já se jí budu podrobně zabývat později (v kapitole 7 této práce). Proto stručně: Turing správně poukazuje na fakt, že připisování vědomí jen na základě autority první osoby je extrémním stanoviskem, které není

schopno identifikovat vědomí ani u jiných lidí; jedná se tudíž o formu solipsismu. Je prakticky nemožné žít ve společnosti, ve které osoba X upírá osobě Y schopnost myslet a osoba Y upírá totéž osobě X. Správnější je přijmout zdvořilou konvenci, podle které myslí bez rozdílu všechny osoby. Jak jsem už ale upozornil v úvodu, tento přístup lze aplikovat v případě nám podobných bytostí, ale nikoli v případě počítačů. Jediným způsobem, jak s těmito nám zcela nepodobnými entitami operovat, je vymyslet kritérium - třeba imitační hru - které nám umožní považovat stroje za určitých podmínek za myslící. Co se týče emocí, pak je situace naprosto obdobná. Protože nejsme schopni odlišit skutečný smutek od toho předstíraného, reálnou zlost od té kaširované, musíme se spoléhat na vnější, behaviorální projevy sledovaných subjektů. Můžeme si s nimi třeba o jejich emocích promluvit a pokusit se odhalit, které jsou ryzí a které falešné. A jsme opět zpátky u imitační hry jako posledního kritéria, za které nelze jít.

### **3.5 Argument z různých neschopností**

Počítače sice dokážou spoustu věcí, ale některých údajně nebudou schopny nikdy. Turing předkládá značně nesystematický seznam šestnácti omezení:

Být milý, duchapřítomný, krásný, přátelský, iniciativní, mít smysl pro humor, odlišit dobro od zla, dělat chyby, zamilovat se, pochutnávat si na jahodách se smetanou, být milován, učit se ze zkušenosti, používat správně slova, přemýšlet sám o sobě, mít mnoho různých druhů chování jako člověk, dělat něco skutečně nového. (Turing 1950, p.447)

Některá omezení jsou podle Turinga natolik závažná, že jim věnuje samostatnou pozornost, jako třeba problém originality ("dělat něco skutečně nového"), kterým se detailně zabývá hned v následující námitce lady Lovelaceové. Jiná zřejmě považuje za nedůležitá, a proto se jimi nezabývá vůbec (např. "mít smysl pro humor"). Turing předpokládá, že všechna nebo téměř všechna tato tvrzení jsou založena na nesprávném zobecnění, které vychází z indukce. To, že kolem roku 1950 byly všechny počítače ošklivé a "hloupé", ještě neznamená, že tomu tak bude napořád. Stav do značné míry závisel na relativně malé paměťové kapacitě tehdejších strojů, což se ale podle Turinga v budoucnosti zajisté změní. Tak například prý jen nedostatečná paměť znemožňuje počítačům "mít mnoho různých druhů chování jako člověk"; ať už tím Turing myslí cokoli, vždyť chování všech účastníků imitační hry je striktně

omezeno na slovní výstup na monitoru nebo dálnopisu. Podívejme se ale, jak jsou na tom s uvedenými omezeními dnešní počítače.

Hned na úvod je třeba uvést na pravou míru některé zjevně mylné položky na seznamu. Velká většina uživatelů počítačů se shodne na tom, že Apple iMac je krásný. Víru v neomylnost počítačů zase nemůže brát vážně nikdo z nás, kdo aspoň jednou viděl "modrou obrazovku smrti".<sup>30</sup> Turing navíc správně poznamenává, že v imitační hře je neomylnost spíše na škodu; počítače by měly být pro ni naprogramovány tak, aby naopak chyby simulovaly, třeba v komplikovaných početních příkladech. Dále: sofistikovanější počítačové systémy jsou už dnes schopny učit se ze zkušenosti, ať už se jedná o řešení složitých probabilistických úloh nebo snahu obehřát nás v pokeru. Triviální příklad čerpání informací ze zkušenosti můžete sledovat při používání internetového vyhledávače. Google si pamatuje, jaké odkazy jste preferovali při předešlých vyhledáváních, a přizpůsobuje jim budoucí výsledky. Pokud jsem v minulosti dával najevo větší zájem o morální psychologii než o britskou beletrii dvacátého století, zobrazí se mi po zadání dotazu "Greene" spíš Joshua než Graham.

Některá jiná tvrzení jsou zase založena na velmi úzkém, šovinistickém pojetí myšlení. Tak například kdekdo s alergií na laktózu není schopen pochutnávat si na jahodách se smetanou, to mu nebo jí neubírá na lidskosti či inteligenci. Pokud budeme brát poznámku vážněji, pak stejně nemá valného smyslu: počítače si principiálně nemohou na ničem pochutnávat, protože v naprosté většině případů nejsou vybaveny chuťovými senzory (i když by je teoreticky mít mohly, jak zmiňuje Turing). Nemají k tomu žádný důvod, protože mezi jejich funkce nepatří metabolismus založený na zpracování tuhé a tekuté potravy. Stejně tak je absurdní požadovat po počítačích schopnost milovat a být milován; ke svému štěstí - anebo smůle - nevznikají díky sexuální reprodukci, a proto je evoluce nemusela vybavit patřičnými emocionálními obezličkami, které mají za úkol zastírat pohlavní pud.

---

<sup>30</sup> "Modrá obrazovka smrti (BSoD, Blue screen of death) je slangové označení chybového hlášení, které operační systém Microsoft Windows zobrazí v situaci, kdy došlo k závažné systémové chybě, ze které není schopen se zotavit. Toto chybové hlášení se zobrazí přes celou obrazovku, bílým písmem na modrém pozadí (odtud název)." ("Modrá obrazovka smrti." *Wikipedie: Otevřená encyklopedie*. 2. července 2009)

Považuji celou řádku vlastností jako "být milý, duchapřítomný, přátelský, mít smysl pro humor, odlišit dobro od zla, přemýšlet sám o sobě" za pouhé symptomy širšího "používat správně slova". Abychom mohli počítač nazvat některým z těchto přívlastků, musíme se s ním nejprve utkat v imitační hře, která dokáže nejlépe - a možná jako jediná - vyhodnotit jeho schopnost používání jazyka. Někteří autoři operují s tezí, že určité konverzační mody, mezi něž patří zřejmě i humor, jsou natolik komplikované, že je počítače nebudou zvládat nikdy. Podle Frenche (1990) jsou totiž právě tato témata založena na subkognitivních znalostech, které jsou determinovány sociokulturním prostředím i naší přirozeností. Tímto směrem argumentace se ale budu zabývat později (v kapitole 7 této práce). Nakonec nám tak ze všech šestnácti Turingem uvedených omezení zůstávají dvě poslední: "být iniciativní" a "dělat něco skutečně nového". A právě originality se týká další námitka.

### **3.6 Námitka lady Lovelaceové**

Britský matematik Charles Babbage popsal v roce 1822 diferenční stroj ("Difference Engine") a později v roce 1834 analytický stroj ("Analytical Engine"), které jsou všeobecně považovány za první počítače v dnešním slova smyslu. Babbage bohužel nedokázal ani jeden ze svých návrhů realizovat a po jeho smrti byl projekt opuštěn. Není zcela jasné, zdali mohl být stroj v tehdejší době a s tehdejší technologií vůbec zkonstruován. Plán byl totiž velmi ambiciózní: mělo se jednat o kalkulátor schopný počítat logaritmické a goniometrické funkce, k čemuž sloužila paměť o kapacitě tisíce padesátimístných čísel. Vstup byl prováděn pomocí děrných štítků, které se osvědčily už u dobových tkalcovských stavů a které ještě v 50. letech 20. století používaly sálové počítače, na nichž pracoval Turing. Výstup měl být zajištěn nějakým typem ukazatele nebo tiskárny. Celý stroj včetně "procesoru" měl ryze mechanický design: funkci stroje zajišťovala nesmírně složitá soustava 96 ozubených kol, 24 hřídelí a nespočtu táhel, čepů, válců. Podle střízlivých odhadů mohl tento parou poháněný kolos vážit až dvě tuny. Roku 1842 analytický stroj vylíčil matematik Luigi Menabrea, jehož francouzský text do angličtiny neprodleně přeložila a poznámkami opatřila Ada Lovelaceová. Celým jménem se jmenovala Augusta Ada Kingová, hraběnka z Lovelace, byla jedinou legitimní dcerou básníka George Byrona a zároveň jednou z nejvýznamnějších matematiček své doby. Její dodatky k Menabreaovi byly rozsáhlejší než původní text a obsahovaly mimo

jiné popis použití analytického stroje při výpočtu Bernoulliho čísel, díky čemuž bývá Lovelaceová považována za první programátorku vůbec. V závěrečném diagramu jsou totiž zřejmě poprvé využity dnes běžné postupy programování, jako jsou podmíněné a nepodmíněné příkazy, podprogramy nebo programovací cykly s opakováním. Proto byl na její počest pojmenován silný programovací jazyk Ada, vytvořený v sedmdesátých letech dvacátého století ministerstvem obrany Spojených států a dodnes používaný například v letectví.

Právě od Ady Lovelaceové převzal Turing další hypotetickou námitku proti svému testu. Lovelaceová svůj postřeh formuluje v poslední poznámce G:

Analytický stroj nemá schopnost vytvořit nic *původního*. Udělá jen to, co mu *přikážeme*. Může se *zabývat* analýzou, ale nedokáže předjímat analytické vztahy nebo pravdy. Jeho schopnosti spočívají v tom, že nám asistuje při *získávání* znalostí, které už máme. (Menabrea & Lovelace 1961, p.284)<sup>31</sup>

Ve zjednodušené variantě lze toto tvrzení shrnout do stručnějšího "stroj nedokáže dělat něco skutečně nového", které se objevilo už v předchozím argumentu z různých neschopností. Počítače nejsou s to vyprodukovat nic, co by předtím nebylo součástí jejich programu. Turing na jejich obranu vyrukuje s nepřesvědčivým argumentem, který je navíc cirkulární. Pokusím se ho stručně a jednoduše parafrázovat:

Představme si hypotetický stroj spojitých stavů, který by byl schopen vytvořit něco skutečně nového. I tento stroj by mohl být emulován univerzálním digitálním počítačem, například analytickým strojem, který - jak už víme - dokáže přesně reprodukovat chování jakéhokoli stroje. Proto tento hypotetický stroj schopný kreativity může existovat.

Problém spočívá v tom, že onen hypotetický stroj obdařený kreativitou není vůbec popsán, natož dokázán. Zkusme nahradit vlastnost "vytvořit něco skutečně nového" méně zavádějící dovedností, například "vykouzlit králíka z klobouku", a nesprávnost úvahy se vyjeví naprosto nepochybně.

---

<sup>31</sup> Můj překlad je dosti volný: "The Analytical Engine has no pretensions whatever to *originate* anything. It can do whatever we *know how to order* it to perform. It can *follow* analysis; but it has no power of *anticipating* any analytical relations or truths. Its province is to assist us in making *available* what we are already acquainted with." Turing navíc ve své eseji cituje jen první dvě věty, které zřejmě považuje za nejpodstatnější (Turing 1950, p.450).

Turing vzápětí napadne i obvyklou představu o fungování představitivosti u příslušníků druhu *Homo sapiens sapiens*. Je dost dobře možné, že i u lidí je takzvaná tvořivost zapříčiněna jen učením a následováním obecných pravidel. Bohužel Turing tomuto tématu přestane ihned věnovat pozornost a námitku přetransformuje do mnohem slabší podoby, podle které nás stroje "nedokážou překvapit". Na základě svých čistě soukromých zážitků pak tvrdí, že jej počítače překvapují téměř neustále, a dodává, že pocit překvapení může být vyvolán živými bytostmi i neživými artefakty. Pokud použiji vlastní ilustrativní příklad, může mě překvapit nevěrná manželka, které jsem doposud věřil, nebo dobře napsaná kniha od autora, nad kterým jsem už dávno zlomil hůl. Podle Turinga proto není důvod, proč by nás nemohl překvapovat i počítač. Jedná se o jedno z mála míst v eseji, kde autor neargumentuje úplně poctivě. Nerozlišuje třeba vůbec mezi dvěma úrovněmi překvapení: kniha sama o sobě nás překvapit nedokáže, je jen prostředkem pro sdělení myšlenek překvapujícího člověka.

Bringsjord et al. (2001) považují Turingovu odpověď za "chudokrevnou", "tajemnou" a "nekompetentní". Domnívají se, že právě námitka Lovelaceové je nejsilnějším a dostatečným důvodem pro odmítnutí celého T-testu. Žádné stroje, ani dnešní pokročilé počítače, nevytvářejí nic originálního, i když nás dokážou překvapit. Pokud například robot v automobilce umístí rezervní pneumatiku na nárazník a nikoli do úložného prostoru vozidla, jedná se o výsledek syntaktické chyby v programu a ne o projev tvořivého rozhodování. Nikdo by nemohl tvrdit, že tovární robot vyrobil nový model auta. Autoři proto navrhuji mnohem přísnější kritérium pro posuzování inteligence, takzvaný Lovelaceové test. Umělý agent *A*, který byl navržen lidským architektem *H*, projde úspěšně L-testem tehdy a jen tehdy, když:

1. *A* vyprodukuje výstup *o*;
2. *o* není výsledkem náhodné chyby hardwaru, ale procesu, který *A* umí zopakovat;
3. *H* nedokáže vysvětlit, jak *A* vyprodukoval *o*. (Bringsjord et al. 2001, p.9)<sup>32</sup>

Závěr tohoto technicky zaměřeného článku, který se zabývá mimo jiné konkrétními případy programů generujících příběhy nebo programováním v jazyce *Lisp*, je neobvykle metafyzický.

---

<sup>32</sup> Zkratky jsou odvozeny z anglických pojmů "artificial agent" (*A*), "human architect" (*H*), "output" (*o*).



Aby stroj prošel Lovelaceově testem, musí disponovat svobodnou vůlí, aspoň ve významu schopnosti autonomního agenta způsobovat kauzální události, které nejsou vyvolány jinou událostí (tzv. "agent causation" vs. "event causation").

Nechci se pouštět do zatím nevyřešeného sporu o *ontologický* status svobodné vůle, i když se se vší důvěrou nechávám přesvědčit tvrdým determinismem ("hard determinism"): díky deterministické povaze fyzikálních těles a částic jsme zprostředkovaně determinováni i my lidé. Takový názor je pochopitelně monistický a naturalistický, neuznává jakékoli ideální, emergentní nebo supervenientní entity. Ty mimochodem nikdy nikdo nepozoroval, a proto je jejich existence značně pochybná. Zároveň odmítá některé filosofické způsoby výkladu kvantové mechaniky, které interpretují nepředpověditelnost událostí mikrosvěta indeterministicky. Jedná se o nepochopení: to, že nedokážeme predikovat chování subatomárních částic, ještě neznamena, že se tyto částice pohybují nahodile, chaoticky nebo dokonce podle principu svobodné vůle! Pro tyto účely se spokojím s mnohem skromnější formou determinismu, který pracovně označím jako *antropologický*. Stejně jako jsou počítače determinovány svým programem a vstupními informacemi, jsou i lidé determinováni genetickou výbavou a sociokulturním prostředím, ve kterém vyrůstají a žijí. Pokud bychom měli tento poznatek vyjádřit žargonem biologie: je determinován genotyp i fenotyp. Náš program je zatím oproti počítačům nesmírně komplikovaný, a proto mají ostatní lidé někdy pocit, že vytváříme něco nového nebo že je překvapujeme; všechny kousky informace tu už ale byly přítomny i před naším originálním kouskem a předpřipravena byla dokonce i pravidla, pomocí nichž jsme tyto kousky poskládali dohromady.<sup>33</sup> Jakékoli jiné pojetí originality musí do své teorie zahrnout něco zázračného nebo nepřírozeného. K úhybnému manévru tohoto typu ale vědec nemůže a ani nechce poklesnout.

---

<sup>33</sup> Podobně argumentuje i Daniel Dennett ve svém prezidentském projevu k východní divizi *American Philosophical Association* z roku 2000 (přetištěno jako Dennett 2001). Dennett přisuzuje vznik kreativity evolučnímu mechanismu přirozeného výběru, nezastává ale tak striktně deterministické stanovisko jako já; je totiž zastáncem kompatibilismu.

### 3.7 Argument ze spojitosti nervové soustavy

Podle Turingovy definice jsou počítače digitální, čímž se údajně liší od lidského nervového uspořádání:

Nervová soustava jistě není strojem nespojitých stavů. Malá chyba v informaci o velikosti příchozího nervového impulsu působícího na neuron může způsobit velký rozdíl ve velikosti odchozího impulsu. Lze proto namítnout, že chování nervové soustavy nemůže být imitováno pomocí stroje nespojitých stavů. (Turing 1950, p.451)

Turing tento rozdíl konstatuje, ale námitku nepovažuje za podstatnou pro průběh imitační hry. Tazatel nemůže v T-testu zjistit, zdali je účastník založen na spojitě nebo nespojitě technologii, je-li živým analogovým organismem nebo neživým digitálním strojem. Na verbálních odpovědích se tento rozdíl vůbec neprojeví.

Turingovo řešení problému je dostatečně uspokojivé, zkusme se ale ponořit hlouběji, do samotných předpokladů této námítky. Jsem přesvědčen, že existují přinejmenším dva důvody, proč je rozdíl mezi spojitou nervovou soustavou a strojem nespojitých stavů jednoduše překonatelný; je možné, že žádný rozdíl vůbec neexistuje. Prvním důvodem je snadná převoditelnost mezi digitálním a analogovým způsobem uchování informací. Vše, co může být přenášeno analogově, je možné přenést i digitálně. Médium není sdělením, chtělo by se říct natruc Marshallu McLuhanovi. Ilustrujme si tento fakt na příkladu z běžného života: posluchači je zcela lhostejné, poslouchá-li Davisovo *Kind of Blue* z gramofonové desky nebo kompaktního disku, obsažená informace - tj. 45 minut a 44 sekund toho nejlepšího jazzu - je v obou případech identická.<sup>34</sup> Je dokonce možné digitálně emulovat analogová zařízení, čehož si všímá už Turing, když naznačuje potenciální digitální nápodobu diferenčního analyzátoru. Reálným příkladem ze současnosti jsou četné počítačové programy, které ryze digitálně rekonstruují zvuk starších analogových syntezátorů.<sup>35</sup> Digitální záznam má ještě jednu skvělou výhodu: opakovaným kopírováním se přesnost původní informace nezhoršuje, jak teprve nedávno a ke své hrůze zjistili majitelé autorských práv na cokoli tváří v tvář

---

<sup>34</sup> Na tom, zdali je kvalitnější LP nebo CD se nedokážou shodnout ani kovaní audiofilové. Osobně se domnívám, že digitální přenos čehokoli je za všech okolností přesnější. Údajnou "plastičnost" a "teplý zvuk" vinylové nahrávky považuji za placebo efekt milovníků starých pořádků.

<sup>35</sup> Například program ReBirth RB-338 společnosti Propellerhead Software, který dokáže věrně emulovat klasické syntezátory Roland TB-303, TR-808 a TR-909.

digitálnímu pirátství. Této výhody si koneckonců všimla i evoluce, když se snažila vyřešit problém ztráty dat při jejich přenosu napříč miliony let.<sup>36</sup> A to je také druhý důvod pro falzifikaci sedmé Turingovy námitky: i živé organismy jsou svým způsobem digitální. Genetická informace je v nukleových kyselinách DNA a RNA kódována ve čtyřkové soustavě, pomocí bází cytosinu, guaninu, adeninu a thyminu, respektive uracilu. Velmi lapidárně tento fakt vyjádřil Richard Dawkins: "Život jsou pouze tisíce a tisíce bajtů digitální informace." (Dawkins 1996, p.32; překlad Lucie Šoltýsová) Svůj přírůstek doplňuje tvrzením, že od roku 1953, kdy James D. Watson a Francis Crick objevili a popsali fungování DNA, je biologie digitální disciplínou.

Do důsledků promyšleno, je i naše myšlení částečně digitální, přesněji řečeno je kódováno binárně. Přenos nervových impulsů je zajištěn dvojím způsobem: uvnitř neuronů se šíří jako elektrický náboj, v synapsích mezi neurony osmoticky uvolněním neurotransmiteru, jako je například acetylcholin. Aspoň v některých úsecích, tedy uvnitř neuritů, se nervový vzruch šíří pomocí dvojkové soustavy: signál je v neuronu buď přítomen, či nikoli, žádný jiný stav není možný. Je proto snad oprávněné považovat lidský mozek také za stroj nespojitých stavů, ve kterém možná dochází jen k větší produkci nepřesností a chyb než v případě běžných digitálních počítačů.

### **3.8 Argument z neformálnosti chování**

Zatímco počítače jsou pevně svázané svým programem, lidé mají podle některých myslitelů možnost samostatně se rozhodovat v případech, pro které pravidla neexistují:

Není možné vytvořit množinu pravidel, která by popisovala, co by měl člověk dělat v každé myslitelné situaci. (Turing 1950, p.452)

Jsme schopni improvizovat, jako když se na světelné křižovatce na semaforu zároveň rozsvítí červená i zelená. Za těchto okolností je správné zůstat stát, ale stroj by se takto prý rozhodnout nedokázal.

---

<sup>36</sup> Pro přesnost dodávám, že evoluce si samozřejmě ničeho "nevšímá" a nic "neřeší". Tyto obraty používají personifikaci jen ze stylistických důvodů. Lze říct, že tak k evoluci zaujímáme intencionální postoj.

Už nejméně dvakrát (v kapitolách 3.3 a 3.6 této práce) jsem poukazoval na skutečnost, že i lidské myšlení může být určováno programem, který je pouze mnohem složitější než ten počítačový. Zaměřím proto tentokrát svou pozornost na jiný způsob, jak se vypořádat se záhadou neformálnosti. Turing věnuje poslední část eseje problematice strojů, které se dokážou učit ze zkušenosti. Doporučuje, abychom se pro účely imitační hry nepokoušeli naprogramovat počítače, jež by napodobovaly mysl dospělého člověka, ale abychom místo toho usilovali o imitaci dětské mysli. Do počítače by byly uloženy jen informace, které odpovídají výchozímu stavu mysli při narození; představujme si je třeba jako kantovskou apriorní poznávací výbavu nebo darwinistické, evolučně získané instinkty. Kromě toho musí být zajištěno "vzdělávání" a jiné zkušenosti, kterým by byla počítačová mysl vystavena (Turing 1950, p.455). Zkušenost by přitom nemusela mít podobu stimulace neurálních receptorů v senzorních orgánech, ale spíše by spočívala v možnosti přístupu do databází, komunikačních kanálů a jiných zdrojů informací. Turing připomíná případ Helen Kellerové, která v raném věku ztratila zrak a sluch, ale i jako hluchoslepá pokračovala ve vzdělání a nakonec získala vysokoškolský titul.<sup>37</sup>

Problém neformálnosti je tak v podstatě vyřešen. Počítač nemusí být připraven na všechna potenciálně možná uspořádání světa, se kterými by mohl být konfrontován. Stačí, aby mu byla naprogramována schopnost učit se ze zkušenosti a získávat ponaučení z předchozích chyb. Vraťme se k výchozímu příkladu s křižovatkou a představme si, že počítačový program personifikovaný avatarem má za úkol pohybovat se po virtuálním městě, pokud možno bez újmy na "zdraví" a "životě". Počítač ví, že pokud svítí na semaforu červená, je žádoucí nehýbat se z místa, zatímco zelená velí přejít vozovku. V případě rozsvícení obou světel by byl počítač nucen volit, třeba pomocí generace náhodného čísla: na stupnici jedna až sto znamená výsledek do čísla 50 "zůstaň", vyšší číslo pak "přejdi". Dojde-li při přechodu silnice na červenou a zelenou ke střetu s prostředím, např. se simulovaným automobilem, a k penalizaci v podobě ztráty bodů, "zdraví" nebo "života", má počítač možnost se z toho pro příště poučit. Stejně tak se chovají lidé, ať už při hraní počítačových her nebo v interakcích s reálným světem. Naše budoucí jednání je ovlivněno minulou zkušeností, a to i díky systému

---

<sup>37</sup> Význam Helen Kellerové pro oblast umělé inteligence detailně rozebírá William Rapaport (2006a); jeho argumentací se budu podrobně zabírat až v souvislosti s vyvrácením Searlova myšlenkového experimentu v kapitole 5.

sankcí a odměn, umělému nebo přirozenému: lízátko vs. poznámka do žákovské knížky, vidina reprodukčního úspěchu vs. riziko zranění či smrti. Turing důvtipně podotýká, že v případě počítačů přejímá úlohu přirozeného výběru úsudek lidského examinátora. Přímo se nabízí srovnání s progresivním odvětvím teorie her, a to se simulací evolučních strategií, které se snaží přežít v hostilním prostředí omezených zdrojů a v konkurenci jiných strategií. Základy tohoto přístupu položil Robert Axelrod, když se pokoušel vysvětlit vznik altruistické kooperace pomocí evolučních mechanismů (Axelrod 1980; Axelrod 1984). Pro tento účel použil počítačovou simulaci, ve které umožnil interagovat mezi sebou patnácti strategiím, jež se lišily chováním při hraní vězňova dilematu s opakováním kol. I takto primitivní programy tvořené jen několika řádky jazyků Fortran nebo BASIC byly schopné učit se ze zkušenosti: těm strategiím, které je jednou obelstily, už příště nedůvěřovaly a oplácely jim stejnou mincí. V tomto experimentu nakonec uspívaly a přežívaly jen ty strategie, které byly nejlépe přizpůsobeny jeho podmínkám. (Shodou okolností to byly ty nejslušnější, které nikdy nezrazovaly jako první a zradu oplácely nanejvýš jednou.)

Lze si představit, že nějak podobně by mohly být environmentálně "vychovávány" i mnohem složitější programy účastníci se imitační hry. Fascinující jsou v tomto ohledu závěry švýcarského počítačového odborníka Daria Floreana a jeho spolupracovníků. Podle autorů výzkumu už dnes dostatečně chápeme neurofyziologické procesy, které umožňují komunikaci, ale příliš nerozumíme okolnostem jejího vzniku. Navrhli proto elegantní experiment, ve kterém uzavřeli několik robotů do prostředí s limitovaným množstvím "potravy" a vybavili je schopností vizuální signalizace. Pomocí těchto signálů mohli roboti upozorňovat ostatní subjekty experimentu na přítomnost potravy. Floreano s překvapením zjistil, že v průběhu relativně krátkého času se v kolonii samovolně vytvořila schopnost komunikovat (Floreano et al. 2007). V sérii pozdějších pokusů se u nich dokonce vytvořila schopnost lhát a zamlčovat informace (Mitri et al. 2009), nebo naopak kooperovat v souladu s Hamiltonovými pravidly příbuzenského výběru (Waibel et al. 2011). Autoři sice netvrdí, že jejich roboti myslí, ale podle mého názoru je možné mluvit o inteligenci na úrovni společensky žijícího hmyzu, s nímž je spojuje právě eusocialita.

Počítače si tedy mohou spoluvytvářet pravidla pro rozhodování, obrazně řečeno mohou přepisovat svůj program a vyvíjet se směrem, který původní autor programu nepředpokládal. Turing to popisuje pomocí příměru s černou skříňkou: v případě učících se

programů nemůže programátor vědět, co přesně se v počítači děje. Čím déle získávání zkušeností trvá, tím se zvyšuje nepředvídatelnost počítačového jednání. Ještě jednou se tak vracíme k předchozí námitce týkající se kreativity: není snad právě tato nepředvídatelnost definicí tvořivosti, kterou běžně používáme v každodenním životě?

### 3.9 Argument z mimosmyslového vnímání

Poslední námitka je, stejně jako ta první, komentátory považována jen za málo přínosnou. Turing se v ní docela nepochopitelně zabývá problematikou mimosmyslového vnímání (ESP) a uvažuje o jeho potenciálním vlivu na průběh imitační hry. Připouští minimálně čtyři formy, kterými jsou telepatie, jasnozřivost, předtucha a psychokineze; nejrozsáhlejší evidence podle něj existuje pro existenci telepatie. Tyto jevy nedokáže současná věda vysvětlit, což se v budoucnosti může změnit a paranormální jevy se stanou součástí "normální" vědy. Turing popisuje, jak by například použití telepatie mohlo ovlivnit úspěšnost lidského účastníka hry - v dosti nejasné pasáži naznačuje otázky, které v sobě asi nějakým způsobem zahrnují použití Zenerových karet. Telepatický účastník by měl pak větší šanci správně hádat jejich hodnoty, což by ho značně zvýhodnilo oproti jeho telepatii neobdařenému počítačovému protivníkovi. Mimosmyslové vnímání je tak jediným důvodem pro zostření T-testu: pokud telepatie existuje, měla by hra probíhat v místnosti, která je vůči ní odolná.

Serióznímu mysliteli musí připadat takové úvahy naprosto absurdní. Nikdy nebyla prokázána existence žádného paranormálního jevu a samotné mimosmyslové vnímání bylo v minulém století - přinejmenším od Harryho Houdiniho - podrobena zdánlivě nekončícímu proudu kritiky a skepticismu. Už v roce 1922 vypsala časopis *Scientific American* odměnu 2500\$ za nevyvratitelný důkaz paranormality, od roku 1996 zvýšila nadace iluzionisty Jamese Randiho vyplácenou částku na rovný milion dolarů. Zdá se téměř nemožné, že by si toho metodologicky střízlivý a racionálně orientovaný Turing nebyl vědom. Nabízí se možnost považovat celou devátou námitku za sarkasmus, podobně jako to bylo v případě námitky první. Autor nám ale nikdy nenaznačí, že svůj postřeh nemyslí vážně, nikdy na čtenáře spiklenecky nezamrká a neuvede věc na pravou míru. Autoři encyklopedického hesla "The Turing Test" na *Stanford Encyclopedia of Philosophy* sice připouštějí možnost skryté ironie

("perhaps it is intended to be tongue-in-cheek"), ale spíš se přiklání k názoru, že tuto námitku myslí Turing vážně (Oppy & Dowe 2011, oddíl 2.9). Nejzajímavější interpretaci nabízí jako obvykle Hodges: podle něj Turing sice nebyl křesťan, ale zajímal se o spiritismus. Velký vliv na něj mělo úmrtí jeho mladého přítele Christophera Morcoma (+1930), jež v něm vyvolalo zájem o posmrtný život, komunikaci s mrtvými a mimosmyslové vnímání (Hodges 1992, pp.63-64). V té době byly značně populární teorie, které propagoval zakladatel parapsychologie Joseph Banks Rhine. Ten působil na Duke University v Severní Karolíně, vydával *Journal of Parapsychology* a spoluzakládal *Parapsychologickou asociaci*; to vše dodávalo jeho aktivitám punc vědeckosti. Zdá se, že v ostatních ohledech velmi skeptický Turing v možnost mimosmyslového vnímání opravdu věřil a považoval je za jev, který bude jednou integrován do naší teorie o fungování fyzikálního světa. Přiznejme našemu hrdinovi aspoň jednou právo fatálně se mýlit a ponechme gentlemansky stranou dosti nepatřičnou, ale přesto vtíravou myšlenku: pokud jsou telepatie údajně schopni lidé, proč ne i počítače?

## 4. Raná recepce: 50. až 70. léta

Esej "Computing Machinery and Intelligence" se během následujících desetiletí setkala s ohromující odezvou, počet reakcí jde do stovek, ne-li tisíců časopiseckých a knižních příspěvků. Domnívám se, že se jedná o vůbec nejdůležitější filosofickou esej druhé poloviny dvacátého století<sup>38</sup>, srovnatelnou nanejvýš s Quinovými "Two Dogmas of Empiricism" (která byla mimochodem napsána a odpřednášena také v roce 1950, i když publikována až o rok později). Přijetí hlavních myšlenek článku se bohužel neobešlo bez určitých dezinterpretací a kontroverzí, většina komentářů vydaná v letech 1950-1980 je spíše příkladem nepochopení Turingova vizionářského textu. O zdařilejších reakcích jsem se zmínil v předchozí části (především Lucas 1961; Moor 1976), teď se pokusím shrnout nejpodstatnější myšlenky z těch méně šťastných. V téže době se objevily i první pokusy o naprogramování komunikačního programu, který by odolal T-testu; těmito praktickými aplikacemi (Colby et al. 1971; Colby et al. 1972; Winograd 1972; Weizenbaum 1966) se ale budu zabývat až v kapitole 8 této práce. Diskuse se nejprve odehrávala na stránkách oxfordského časopisu *Mind* a částečně v cambridgeském *Philosophy*. Ještě pořád se používal termín "imitační hra", o "Turingovu testu" se začalo psát až od 70. let: zřejmě nejstarším výskytem fráze je název Moorova článku z roku 1976.

### 4.1 Schopnost chybovat

Úplně první reakcí na Turingův článek byla kratičká, dvoustránková poznámka, kterou napsal dnes už zcela zapomenutý myslitel, genetik Leonard Pinsky (1951). Pinsky T-test zcela okatě parodoval a zároveň si vyřizoval účty s takzvanými *terapeutickými pozitivisty*, nelze ale

---

<sup>38</sup> Taková kategorická tvrzení jsou v humanitních disciplínách jen obtížně dokazatelná, protože společenské vědy nedisponují podobně silným nástrojem jako je *Science Citation Index* na *Web of Knowledge*; tamější *Social Sciences Citation Index* a *Arts & Humanities Citation Index* nedokážou zpracovat díla publikovaná před rokem 1980. Podle mnohem méně přesného vyhledávače *Google Scholar* byla Turingova esej citována nejméně 3350.



říct, že úspěšně.<sup>39</sup> Terapeutičtí pozitivisté zastávali názor, podle kterého byla až přibližně do roku 1933 veškerá filosofická činnost založena na nesprávném použití rozumu. Je proto třeba revidovat Aristotelovu definici, že člověka od zbytku světa odlišuje schopnost používat rozum - ve skutečnosti je člověk charakterizován schopností používat rozum nesprávně! Pinsky dále navrhl provést následující experiment: počítači je předložena Turingova esej, po jejímž přečtení stroj začne přemýšlet o možnosti stroje, který by přemýšlel. Stejně jako neuspěly při řešení metafyzických otázek tohoto typu celé generace filosofů, musí selhat i počítač. Pinsky dokonce předpokládal, že by se stroj - bůhví vlastně proč - mohl nervově zhroutit a následně by byl podroben psychoanalytické terapii. Chabým závěrem je, že aspoň v tomto případě počítač použil rozum nesprávně, a proto je podle terapeutické interpretace Aristotelovy definice myslící bytostí rovnou člověku. Ani v době svého vydání nemohl být tento posměšný špílec brán příliš vážně; Stuart Shieber jej neváhal označit za pohádku, se kterou dnes už nevíme, co si počít (Shieber 2004, p.141). Není se proto čemu divit, že se k Pinského článku později už nikdo nevrátil. Záhadou zůstává jen fakt, proč vlastně editoři časopisu *Mind* uznali za vhodné text otisknout. I to ale možná svědčí o celkové míře nepochopení Turingovy argumentace u filosofické veřejnosti, která měla jen velmi povrchní informace o fungování digitálních počítačů a žádné osobní zkušenosti s jejich používáním.

#### 4.2 Jazyková hádanka myšlení

V časopise *Philosophy* se zanedlouho rozpoutala o poznání serióznější diskuse v návaznosti na úvahu s přitažlivým názvem "Can Machines Think?" (Mays 1952) Její autor Wolfe Mays v sobě nezapřel Wittgensteinova žáka a existenci myslících strojů řešil po vzoru svého učitele jako jazykovou hádanku. Celý problém tkví ve slovesu "myslet", které Turing používá v rozporu s běžně rozšířeným územ. Termín "myslet" neznamena "uspět v imitační hře", ale jedná se o mentální sloveso vyhrazené pouze lidem. Mays předpokládal, že se na tom nic nezmění ani v roce 2000, kdy podle Turinga už budou existovat stroje vítězí v T-testu. Vypomohl si přitom oxfordským *New English Dictionary*, jenž definuje stroj takto:

---

<sup>39</sup> Termín "terapeutický pozitivismus" zavedl B. A. Farrell pro označení směru, který je dnes známý jako oxfordská škola přirozeného jazyka. Jedná se o myslitele navazující na Wittgensteina, Farrell mezi ně řadí třeba Johna Wisdoma nebo Normana Malcolma (Farrell 1946a; Farrell 1946b).

[S]pojení součástí pohybujících se mechanicky, na rozdíl od bytostí, které mají život, vědomí a vůli. Používá se také pro označení osob, které jednají jen podle zvyku, v souladu s pravidlem, bez inteligence; nebo pro ty, jejichž jednání vykazuje neproměnlivou preciznost a strojovou jednotvárnost. (Mays 1952, p.149)<sup>40</sup>

Stroje jsou tedy z definice entity, které nemohou mít mentální stavy. Pokud něco myslí, pak to není stroj. Vnější projevy jednání nesmíme zaměňovat za vnitřní myšlenkové pochody. Mays navrhoval reformu přirozeného jazyka a místo termínu "stroj" chtěl v tomto kontextu používat čapkovské "robot" pro označení ještě nesestrojených umělých entit, které by snad byly schopny myslet, a proto by stály na půl cesty mezi člověkem a strojem. Ani tento způsob vyjadřování není naneštěstí zcela správný, protože i sloveso "myslet" je třeba nahradit nějakým novotvarem, který by se vztahoval jen na roboty. Samotným jádrem problému je skutečnost, že při popisu světa používáme dvojí jazyk: fyzikalistický a psychologický. Mays přitom odmítl Carnapovo redukcionistické přesvědčení o převoditelnosti psychologického jazyka do toho fyzikalistického. Některé predikáty - například právě "myslet" - jsou totiž svou povahou neredukovatelné. Klíčovou chybou je směšování obou jazyků, kterého se neúmyslně dopouští i Turing, když při popisu neživých strojů používá psychologické výrazy jako "rozhodovat se", "chybovat" nebo "poslouchat příkazy". Dokonce tak nevinné slovo jako "paměť" může v případě označování datové jednotky počítače způsobovat konfúze, které jsou podle Maysy projevem animismu.

Kromě tohoto hlavního argumentu, vycházejícího z kritického přezkoumání významu termínů "stroj" a "myslet", uvedl autor ještě celou řadu dodatečných. Ty už jsou ale mnohem méně podstatné: překlad z přirozeného jazyka lidí do formálního jazyka počítačů bude zatížen nepřesnostmi; rozum nelze oddělit od emocí, na druhou stranu v imitační hře dochází ke směšování logiky s psychologií; introspektivní prožívání mentálních událostí nelze popsat objektivně; naše poznání není atomární, ale skládá se z celků složených podle zásad gestaltismu; a podobně. Kromě obligátní varianty argumentu z vědomí nestojí tyto připomínky za zvláštní pozornost. Pozoruhodně mylné jsou ale autorovy prognózy ohledně budoucího vývoje počítačů (Mays z nejasných důvodů důsledně používá tvar "computer").

---

<sup>40</sup> V originále: "[Machine is] a combination of parts moving mechanically as contrasted with a being having life, consciousness and will. Hence applied to a person who acts merely from habit or obedience to a rule, without intelligence, or to one whose actions have the undeviating precision and uniformity of a machine."

Autoritativně tvrdil, že vývoj technologie není neomezený, a vyjadřoval se skepticky k možnosti počítače "velkého jako Empire State Building, poháněného energií Niagarských vodopádů" (Mays 1952, p.154). Této chyby v úsudku se dopouští mnoho společenských myslitelů, kteří nemají přílišné ponětí o vědeckém pokroku. V tomto případě autor díky nedostatku představivosti nechápal možnost *miniaturizace* počítačových komponentů a snažil se čtenáře ohromit představou monstrózního, prakticky neuskutečnitelného stroje. Citoval dokonce amerického filosofa techniky Lewise Mumforda, podle kterého některé typy strojů už dosáhly ve vývoji svého limitu a nebudou se dále zdokonalovat: jedná se například o knihtisk nebo telefon. Tento absurdní postřeh z roku 1934 mě jen utvrzuje v přesvědčení, že by se filosofové - a filosofové techniky zvláště - měli naučit zdržovat se úsudku mnohem častěji, než jsou zvyklí. Na počátku padesátých let byl Turing jedním z mála obyvatel Británie, kteří se vyznali v počítačích; o to překvapivější je skromnost jeho předpovědí o jejich budoucí podobě. Zatímco se filosofové předhánějí v extravagantních návrzích, inženýrsky uvažující Turing si vystačí s tvrzením o nutnosti dostatečné paměťové kapacity těchto strojů. Mays se touto neokázalostí nedokázal inspirovat, i když byl Turingův kolega z Manchesterské univerzity. V samotném závěru si neodpustil ještě jeden melodramatický apel na čtenářovu fantazii. Pokud se lidé fundamentálně neliší od strojů, jsou vlastně i oni stroje řízené programem. Naplňuje se tak temná předpověď George Orwella z románu *1984*, kde "lidé mohou být chápáni jen jako nevýkonné digitální počítače a Bůh je Vrchním programátorem" (Mays 1952, p.162).<sup>41</sup>

Ponechme tato emotivní vyjádření stranou a zaměříme se ještě jednou na Maysovu původní sémantickou analýzu. Domnívám se, že se autor dopouští nejméně dvou logických chyb. Především není úplně korektní považovat používání psychologických predikátů o neživých věcech za animistický omyl. Turing volí mentalistický slovník, protože je úspornější a pro čtenáře srozumitelnější. I v případě, že jsem se ontologicky zavázal třeba k naturalistické teorii mysli a považuji všechny mentální stavy za pouhé projevy nervové soustavy, budu v běžné konverzaci používat možná zavádějící, ale snadno pochopitelné pojmy lidové psychologie. Sdělení "jsem zamilovaný" zní přeci jen o dost lépe než "v mozku se mi

---

<sup>41</sup> Ať čtu Orwella jakkoli, chápu ho jako výsostného kritika politické totality, nikoli bojácného odpůrce vědy a techniky. Mays má zjevně slabost pro antiutopické romány, což se projeví, když srovnává Turingův návrh počítačově simulovat dětskou mysl s Huxleyho *Báječným novým světem*.

krátkodobě zvýšily hladiny dopaminu, oxytocinu a vazopresinu".<sup>42</sup> Analogická fráze "počítač poslouchá příkazy" metaforicky označuje situaci, kdy uživatel vkládá do paměti data pomocí klávesnice. Mays svou argumentaci zakládá na ještě hlubším nedorozumění: běžné významy slov jsou dány konvencí aktuálního jazykového společenství, tudíž z nich nemůžeme vyvozovat normativní závěry. Jazyk není neměnný, a proto dokazovat něco pomocí slovníku, jehož první svazek vyšel v roce 1888, je čirý anachronismus. Zkusme vytvořit jiný příklad tohoto přístupu a ukažme si jako *reductio ad absurdum* jeho neudržitelnost. Od Kanta se ve filosofii řeší dramatický spor, zdali je možné vystavět etiku bez náboženského základu. Naše otázka tedy zní: "Mohou být ateisté morální?" Nahlédněme opět do slovníku, tak například mé čtvrté vydání *American Heritage Dictionary* definuje ateismus takto:

Nevíra nebo popření existence boha či bohů. Doktrína, podle které neexistuje bůh či bohové. Bezbožnost; nemorálnost.<sup>43</sup>

Podle Maysova způsobu čtení jsou ateisté z definice lidé, kteří morální nemohou být. Pokud se někdo chová morálně, pak to není ateista. A jestliže ateista dodržuje nějaký systém pravidel, pak to není morálka. Problém se zdá být vyřešen. Ve skutečnosti se nejedná o legitimní filosofickou argumentaci, ale o prachobyčejnou sofistiku, navíc ne moc zdařilou. Maysova esej vyvolala ve své době mírnou vlnu zájmu, a dokonce nesouhlasu, stejně jako v případě Pinského nemají tyto texty vliv na dnešní podobu diskuse a jsou zapomenuty či ignorovány. Je to škoda, přinejmenším v případě jasnozřivé obhajoby T-testu, kterou napsal Frank Honeywill George.<sup>44</sup>

### 4.3 Šlapaná

Další příklad nepochopení Turingova myšlenkového postupu předvedl Keith Gunderson ve svém zábavném článku z roku 1964. Gunderson, podobně jako před ním

---

<sup>42</sup> Takto aspoň popisuje pocity zamilovanosti antropoložka Helen Fisherová, která se problematikou romantické lásky zabývá systematicky (Fisher et al. 2006).

<sup>43</sup> V originále: "Disbelief in or denial of the existence of God or gods. The doctrine that there is no God or gods. Godlessness; immorality." Nutno říct, že ještě ve třetím vydání bylo "immorality" nahrazeno "wickedness" čili špatností.

<sup>44</sup> První kolo: F. H. George (1956), Ritchie (1957); druhé kolo: F. H. George (1957), Ritchie and Mays (1957), H. R. T. Roberts (1958).

Pinsky, satirizuje T-test a pokládá si otázku: "Dokáže kámen napodobit člověka?" Pro tento účel navrhuje své vlastní originální kritérium a nazývá je *šlapaná* ("toe-stepping game"). Nejprve popíše nám povědomou genderovou variantu testu. Hry se účastní tři hráči: muž (A), žena (B) a tazatel (C). Tazatel se nachází v místnosti, která je s místností hráčů spojena malým otvorem těsně nad podlahou, do kterého může tazatel vložit téměř celé chodidlo. Hráči A a B bez udání pořadí šlapou tazateli C na nohu, jeho cílem je poznat, zdali na nohu šlápla žena či muž. Žena tazateli pomáhá, a proto šlape "upřímně". Muž naopak podvádí a snaží se imitovat typicky ženský způsob došlapu. Představme si teď pozměněnou variantu hry, kdy jednoho z hráčů nahradí zvláštní zařízení zvané šutr-box ("rock-box"). Jedná se o krabici naplněnou kameny různé váhy, velikosti a tvaru, jež je navíc vybavena elektrickým okem. Zařízení je schopno položit kámen na nohu tazatele vždy, když prostrčí chodidlo otvorem ve zdi, a po chvíli kámen opět zvednout. Tentokrát má tazatel pochopitelně zjistit, zdali mu na nohu šlápl člověk či šutr-box. Původní víceznačná otázka "Dokáže kámen napodobit člověka?" je tak nahrazena mnohem srozumitelnější "Dokáže kámen uspět ve šlapané?" Gunderson předpovídá, že do padesáti let budou existovat sofistikované šutr-boxy s dostatečnou kapacitou kamenů, které tazatele přesvědčí o své lidskosti.<sup>45</sup>

Autor se samozřejmě pomocí této okatě nesmyslné hříčky snaží ukázat nedostatečnost Turingova kritéria. Stroje sice mohou uspět v imitační hře, ale to ještě nic nevyovídá o jejich schopnosti myslet. Podle Gundersona nestačí jen srovnat výsledky činnosti strojů a lidí, ale je třeba vzít v potaz i způsob, jakým k výsledkům dospívají. Dveře nám otevře buď vrátný, nebo fotobuňka - to ale ještě neznamena, že se jedná o rovnocenné entity. Stejnou skladbu nám může přehrát gramofon nebo Budapeštský smyčcový kvartet - přístroj si ovšem nezaslouží obdiv za precizní přednes, natož potlesk. Turing se dopouští zásadní chyby, když přeceňuje význam pouhých hrubých výsledků a opomíjí, co přesně k těmto výsledkům vedlo. Gunderson jde ve své kritice ještě dál a tvrdí, že úspěch v imitační hře je jen jedním z projevů inteligentního chování. I kdyby stroj tuto podmínku splnil,

---

<sup>45</sup> Gundersonova parodie je dokonalá a používá zcela totožné obraty jako Turing: "I believe that in less than fifty years' time it will be possible to set up elaborately constructed rock-boxes, with large rock-storage capacities, so that they will play the toe-stepping game so well that the average person who would get his toe stepped on would not have more than 70 per cent, chance of making the right identification after about five minutes of toe-stepping." (Gunderson 1964, p.237) Samotnou Turingovou předpovědí se budu zabývat na začátku kapitoly 9.

nedokázal by obstát v nespočtu jiných situací, které vyžadují inteligenci. Autor si opět vypomáhá anekdotickým příběhem ze života a přirovnává Turinga k obchodnímu cestujícímu, který nabízí vysavač *Swish 600*. Ten ve skutečnosti umí pouze vysávat prach, ale přesto je inzerován jako víceúčelový. Schopnost vysávat prach pak není příkladem činnosti, kterou vysavač dokáže vykonávat, ale výčtem všech činností, které vysavač dokáže vykonávat. Stejně tak schopnost počítače uspět v imitační hře je tím jediným, co stroj dokáže; tato dovednost rozhodně nemůže nahradit všechny aspekty myšlení, které přísluší živým lidským bytostem. V závěru vzdává Gunderson strojům podivný hold: to, že počítače nemyslí, nemusí nutně znamenat, že jsou horší než lidé. Naopak to může být projevem vyššího stupně dokonalosti, když dovedou tolik věcí bez přemýšlení.

Může se zdát, že Gunderson sám sebe nebere příliš vážně, ale později ze svých stanovisek neustoupil ani o píď, a proto je třeba věnovat jeho námitkám patřičnou pozornost.<sup>46</sup> Začnu od toho nejjednoduššího a prohlásím šlapanou, i v její genderové verzi, za naprostý nesmysl. Přijměme předpoklad, že mezi muži a ženami existují vrozené rozdíly ve fyziologické konstituci i mentální výbavě. Imitační hru je možné vyhrát, tazatel má k dispozici nástroje pro úspěšnou identifikaci pohlaví účastníků. Sice by si asi nevystačil s Turingem navrhovanými dotazy týkajícími se délky vlasů, ale zřejmě by mohl vyrukovat s tématy intimní povahy, která může zcela uspokojivě znát jen příslušník toho kterého genderu.<sup>47</sup> Oproti tomu ve šlapané nemá tazatel sebemenší šanci: nelze poznat, zdali vám na nohu šlápl muž nebo žena. Naše sensorické orgány tomu nejsou uzpůsobeny a neexistuje ani zvláštní způsob, kterým se šlápnutí liší pohlavně. Jediným faktorem může být vyšší průměrná váha mužů, kterou ale nejsme s to správně vyhodnotit a lze ji navíc snadno předstírat. Gunderson totiž opomíjí zcela fundamentální rozdíl mezi imitační hrou a jakýmkoli jiným testem. Turing si vybral lingvistický projev, protože je naprosto jedinečný a pro člověka charakteristický. Šlápnout na nohu nebo otevřít dveře vám může ledacos, ale promluvit si o počasí nebo postěžovat si na vládu nedokáže kromě lidí nic - zatím. Schopnost komplexní komunikace

---

<sup>46</sup> Jedinou souvisejší kritikou Gundersonova stanoviska je Stevenson (1976) a několik zmínek v Moorovi (1976).

<sup>47</sup> Po vzoru esencialistických feministek druhé vlny můžeme zvolit otázky týkající se *embodimentu* čili zkušenosti tělesnosti, jako je například menstruace, porod, laktace, ejakulace apod. Purtil navrhuje mnohem důvtipnější a ne tolik osobní dotazy: "Kolik stojí pár nylonových punčoch?" nebo "Kde je v této budově dámská toaleta?" (Purtill 1971, p.290)

není srovnatelná s žádnou jinou lidskou vlastností. Lze ji tedy oproti Gundersonovu přesvědčení považovat za víceúčelovou: úspěch v imitační hře zajišťuje zvládnutí všech ostatních činností, které vyžadují inteligenci. Počítač, který prošel T-testem, sice neumí uvařit čaj nebo otevřít láhev piva, ale může poradit, jak na to. Imitační hra je tak dostatečným důvodem pro přisouzení ostatních projevů inteligence.

Přes všechny nedostatky, které v Gundersonově argumentaci shledávám, mu nedokážu upřít uznání za jeho závěrečnou poznámku o možném přeceňování inteligence. Interpretováno darwinisticky, schopnost vykonávat činnosti bez přemýšlení může být evoluční výhodou. Lidský druh je s přehledem nejinteligentnější v dějinách této planety, mnoho úspěchů ani štěstí mu to ale nepřineslo. Máme antropocentrismem zapříčiněnou představu, že jsme pány světa, vrcholem hierarchického uspořádání všeho živého, ale tento pocit převahy může být do značné míry neoprávněný a falešný. Naši nejstarší předci se objevili před asi dvěma miliony let a ve své moderní podobě existujeme přibližně teprve dvě stě tisíc roků. Co je to proti hrůzně bezmyšlenkovité efektivitě žraloků a krokodýlů, kteří přetrvávají 100, respektive 200 milionů let bez podstatnějších změn? Dovolím si zaspekulovat bez sebemenší opory v empirických datech: pokud na nějaké exoplanetě existuje život, pak je pravděpodobně na hony vzdálen inteligenci lidského typu. A pokud by někdy v budoucnosti lidský druh zanikl a světové dominance by se ujal nějaký jeho nástupce, asi by přirozený výběr znovu neprošlapával slepou cestičku na inteligenci založené civilizace a spolehl by se na jednodušší metody přežití.<sup>48</sup> Vsázím na mravence nebo něco jim podobného.

#### **4.4 Kauzalita a inteligence**

V roce 1971 se do debaty o oprávněnosti T-testu přidal i americký filosof (a spisovatel fantasy) Richard Purtill, když obrátil pozornost k problematice kauzality. Podle něj v imitační hře nelze určit, které odpovědi na tazatelovy otázky pochází od člověka a které od stroje. Ve skutečnosti pochází obě sady odpovědí od člověka, protože i do počítače byly člověkem

---

<sup>48</sup> Historik vědy Jacob Bronowski k tomu napsal: "Příroda s námi provádí jedinečný pokus, aby dokázala, je-li racionální inteligence lepší než reflexní jednání." (Bronowski 1985, p.426; překlad Jaroslava Davidová-Moserová)

uloženy. Tazatel se může nanejvýš pokusit o nalezení řešení otázky: "Které člověkem vytvořené odpovědi byly přeneseny pomocí počítače a které ne?" Z imitační hry se tak stává intelektuální zápas mezi tazatelem a programátorem, počítač není pro průběh vůbec podstatný (Purtill 1971, p.291). Na Turingovu obranu musím podotknout, že lidé jsou v úplně stejné situaci jako stroje. Zatímco počítači byly informace zprostředkovány programátorem, do lidí je po celé roky ukládali rodiče, sousedé, členové dětské herní skupiny, učitelé, spolužáci, přátelé a přítelkyně, prostě každý, s kým přišli během života do kontaktu. Kromě toho se samozřejmě na zvyšování informovanosti podíleli i spisovatelé a vědci, živí či mrtví, a to díky knihám, televizním dokumentům, rozhlasovým přednáškám, webovým stránkám, a tak podobně. Když reaguji na otázku "Co Jaspers míní pojmem *die Achsenzeit*?", odpovídám skutečně já nebo můj učitel Karel Floss, který do mě správnou odpověď v říjnu 1994 vložil?<sup>49</sup>

Purtill se pokouší svou teorii podpořit pomocí dalšího science-fiction myšlenkového experimentu, které jsou mezi teoretiky T-testu mimořádně populární.<sup>50</sup> V imitační hře by čistě hypoteticky mohlo uspět prazvláštní zařízení, představované řadou číslovaných poštovních schránek, ze kterých by nějaký jednoduchý mechanismus vybíral předem napsané odpovědi. Není podán žádný detailní popis tohoto stroje a Purtill má navíc velmi zkreslené představy o významu adjektiva "jednoduchý". Na první pohled se jedná o laciný trik, který budou filosofičtí kritici T-testu používat docela často. Počítače jsou malé a pro laika tajuplné, proto je třeba převést problém do mnohem většího měřítka a záhadné součástky nahradit něčím pochopitelnějším, třeba přátelskými, červeně natřenými poštovními schránkami. Podle Purtillovy argumentace i hlupák musí vědět, že poštovní schránky *přece* nemyslí, a s ulehčeným pousmáním může hodit celou představu myslících strojů za hlavu. Ve skutečnosti jsou všechny stroje - poštovní schránky, počítače a dokonce mozky - fyzikálními entitami o různém stupni složitosti. Dnešní počítače jsou mnohonásobně komplikovanější než Purtillem popsany stroj, a lidské mozky ještě víc, mnohem víc. Jejich princip je ale v podstatě analogický: vstupní data jsou zpracována pomocí proměnlivých algoritmů a výsledné informace jsou různými způsoby dávány najevo. Připomínám, že i Babbegův analytický stroj byl univerzálním digitálním počítačem. Hodí se možná vzpomenout Clarkův

---

<sup>49</sup> Téměř identicky argumentuje ve své kritické poznámce Geoffrey Sampson (1973, p.593).

<sup>50</sup> Pokusím se o předložení jejich úplného seznamu: "rock-box" (Gunderson 1964), "mailbox" (Purtill 1971), "Chinese box" (Searle 1980b), "Blockhead" (Block 1981).



třetí zákon: "Každá dostatečně pokročilá technologie je nerozeznatelná od magie." (Clarke 1973, p.21) Filozofové často zapominají, že i k těm nejneepochopitelnějším jevům lze vždy najít - dříve nebo později - naturalistické vysvětlení.

V závěru eseje autor poněkud otupuje ostří své kritiky a vysloví domněnku, že někdy ve vzdálené budoucnosti snad lidé sestrojí počítače, které budou myslet; v dohledné době to ale nebude. Jednou možná budou lidé cestovat rychlostí blížíící se světelné, stejně jako v případě myslících počítačů nám ale současný stav vědeckého poznání nebude k ničemu. Purtil se proto staví velice skepticky k praktické realizaci T-testu:

Představte si, že bych oslovil počítačového odborníka a nabídl mu úžasnou odměnu, pokud se mu podaří sestrojít počítač, který by odpověděl na otázku, jež bych mu položil. Odborníkovi bych o příslušné otázce neposkytl vůbec žádné informace, a to až do okamžiku dokončení stroje. Ukažte mi někoho, kdo by na mou nabídku přistoupil, a já vám ukážu člověka, který ztratil kontakt s realitou. [...] Pokud má dnes někdo představu o principech, na kterých by měl být takový stroj postaven, sním celou svou knihovnu počítačové literatury. (Purtill 1971, p.293)

S tímto dramatickým tvrzením lze do určité míry souhlasit. Téměř šedesát let od publikování "Computing Machinery and Intelligence" se zatím žádnému počítači nepodařilo projít jakkoli omezenou formou T-testu, natož jeho úplnou verzí. Zdá se, že dosažení Turingem vytyčené mety bude mnohem náročnější, než se zpočátku mohlo zdát.

#### 4.5 Antropocentrismus

Posledním relevantním komentářem ze sedmdesátých let je další velmi stručný článek, který v časopise *Mind* uveřejnil P. H. Millar (1973). Autor poukazuje na dva aspekty imitační hry, které zatím zůstaly bez povšimnutí. Jako první se zabývá lidským účastníkem T-testu, o kterém se Turing vůbec nezmiňoval. Millar upozorňuje, že na osobě toho druhého záleží, protože funguje jako kontrolní skupina, byť jednočlenná.<sup>51</sup> V případě, že bude lidský účastník imitační hry nějakým způsobem handicapován, může tazateli připadat méně přesvědčivý než jeho počítačový protihráč. Kdyby se T-testu účastnili třeba Marťané nebo

---

<sup>51</sup> Až mnohem později začali teoretici umělé inteligence věnovat stejnou pozornost i tazateli v imitační hře, jehož úloha je ještě důležitější. Srov. např. Floridi et al. (2009).

Pygmejové, neměli by šanci v něm uspět. Celý test je podle Millara zatížen značnou mírou antropocentrismu, protože nehledá u stroje jakoukoli inteligenci, ale nutí ho imitovat právě tu lidskou. Stejně jako nemůže existovat standardizovaný test inteligence napříč lidskými kulturami, není možné vytvořit jedno uniformní kritérium pro všechny potenciálně myslící entity. I tato námitka je podle mého názoru snadno překonatelná: podle interpretace, kterou jsem nabídl v kapitole 3 této práce, je T-test postačující, nikoli nutnou podmínkou myšlení. Marťanovi, který v imitační hře neuspěje, se definitivně neuzavírá přístup do skupiny majitelů myslí; můžeme na něj aplikovat nějaká jiná kritéria, jako je fyziologická podobnost našich nervových soustav a podobně. Nebo počkáme, až si osvojí jazyk a internalizuje sociokulturní znalosti, které jsou v naší společnosti běžně rozšířeny, a může se pokusit o reparát. Abych se vyvaroval těchto abstraktních příkladů, použiji konkrétnější: roční dítě se nemůže ani účastnit imitační hry, natož v ní uspět, ale přesto je považujeme za obdařené myslí. Druhou Millarovou výtkou vůči T-testu je Turingova mlčenlivost o krocích, které vedou ke konstrukci myslícího počítače. Turing stanoví cíl svého programu, tj. vytvoření umělé inteligence, ale nepopíše způsob jeho dosažení. Chci připomenout, že původní esej vyšla ve filosofickém časopise a čtenáři by proto asi příliš neocenili, kdyby její autor používal technický žargon a uváděl dlouhé ukázky v programovacím jazyce. Turing jako filosof nastínil budoucí vývoj své disciplíny, pak jej jako programátor tvrdou a mravenčí prací realizoval v laboratoři - o tom se ale v jeho eseji nepíše. Škodolibě podotýkám, že podle všech myslitelných ukazatelů byli počítačovní odborníci za posledních šedesát let nesrovnatelně úspěšnější než jejich filosofické protějšky; nemáme si tedy nač stěžovat.

Žádná z reakcí na Turingův esej, které se objevily v padesátých až sedmdesátých letech, nepředkládá dostatečné důvody pro odmítnutí imitační hry jako kritéria inteligence. Většinou se jedná o amatérské výpady založené na nedokonalém chápání výpočetní techniky, a proto je můžeme bez výčitek svědomí považovat za bezpředmětné. Krušné časy pro T-test nastaly až roku 1980, a právě tímto zlomovým okamžikem se budu zabývat v příští kapitole.

## 5. Absence sémantiky: Searle 1980

John Searle si určitě zaslouží čestné místo v panteonu nejvýznamnějších filosofů naší doby. Přestože je Američan a většinu života přednáší na své domovské univerzitě v kalifornském Berkeley, bývá jako přímý žák a spolupracovník Johna Austina řazen do okruhu filosofů oxfordské školy, která se zabývala filosofickou analýzou přirozeného jazyka. Jeho prvním úspěchem byl podíl na formulování teorie mluvních aktů, která je sice obvykle připisována Austinovi, Searle ale zřejmě měl na její finální podobě zásadní podíl. Stal se také nejrazantnějším propagátorem závěrů svého učitele a bránil jeho odkaz před nejrůznějšími nepochopeními a dezinterpretacemi. Derridův diletantský pokus o využití teorie mluvních aktů ve prospěch relativizace pravdy Searla přivedl ke kritice postmoderní filosofie a literární teorie. Od roku 1977, kdy proběhlo první kolo dnes už legendární disputace mezi oběma mysliteli na stránkách časopisu *Glyph*, byl Searle pasován do role hlavního odpůrce postmoderního programu z řad analytických filosofů. Obviňuje Derridu a jeho sympatizanty z intelektuální nepoctivosti, banálnosti, relativismu, iracionalismu a dokonce - slovy Michela Foucaulta - z obskurantního terorismu. Tento sám o sobě docela obskurní termín označuje samoúčelnou nesrozumitelnost Derridových textů a jeho neochotu hájit své názory v otevřené diskusi, kdy jediným obranným manévrem je obvinění odpůrců z ignorance. Searle, podobně jako Habermas v kontinentální filosofii, doporučuje jako protiváhu postmoderny racionalistickou tradici osvícenství, za jejíhož představitele se považuje.

Zdaleka nejdůležitějším Searlovým přínosem pro filosofii je jeho kritika kognitivní vědy založené na komputacionismu a zásluha o znovuobjevení intencionality v rámci analytické filosofie. Roku 1980 publikoval paradigmatický článek "Minds, Brains and Programs"<sup>52</sup>, který může být chápán jako první odmítnutí T-testu provedené na seriózních základech. Searle řadí imitační hru do takzvané *silné umělé inteligence*, kterou na rozdíl od *slabé umělé inteligence* kategoricky odmítá. Slabá umělá inteligence považuje počítač za užitečný nástroj při studiu mysli, jenž nám umožňuje lépe formulovat a testovat hypotézy.

---

<sup>52</sup> Původní verze z časopisu *Behavioral and Brain Sciences* obsahuje i reakce sedmadvaceti filosofů a kognitivních vědců a autorovu odpověď. Další reprinty tuto část už neobsahují, jako například Searle (1990b).

Oproti tomu silná umělá inteligence tvrdí, že správně naprogramovaný počítač disponuje myslí, rozumí a má kognitivní stavy. Za příklad tohoto radikální přístupu si autor vybral program počítačového odborníka Rogera C. Schanka a psychologa Roberta P. Abelsona, který simuluje lidskou schopnost rozumět příběhům. Searle se tedy nepouští přímo do kritiky T-testu, ale domnívá se, že stejné výtky lze vztáhnout i na Winogradův program *SHRDLU* nebo Weizenbaumův program *ELIZA*, což jsou běžnější konverzační chatboty, které o úspěch v T-testu usilují.<sup>53</sup> Ve skutečnosti se podle Searla jeho připomínky týkají vůbec všech pokusů o nápodobu mysli, které jsou založeny na principu T-stroje, to znamená všech dnešních digitálních počítačů. Schankův a Abelsonův program podstupuje test funkční gramotnosti, protože dokáže odvozovat z vložených příběhů to, co v nich není explicitně sděleno. Vysvětleme si jeho činnost na tomto minimalistickém příkladu: "Muž přišel do restaurace a objednal si jídlo; obsluha mu přinesla hamburger, který byl spálený na uhel; muž se rozzuřeně vyřítit z restaurace, bez zaplacení."<sup>54</sup> Správně fungující program by měl být schopen odpovídat na otázky týkající se děje, jako například "Dojedl muž hamburger?" Zastánci silné umělé inteligence se proto domnívají, že takový program rozumí příběhu a zároveň lze jeho pomocí vysvětlit lidskou schopnost rozumění. Searle ovšem s tímto závěrem přesvědčivě polemizuje.

Hlavním Searlovým argumentem je fenomenálně známý *myšlenkový experiment čínského pokoje*. V něm je Searle zamčen v pokoji s velkým množstvím textu v čínštině. Neumí mluvit ani psát čínsky, dokonce ani nerozpozná čínštinu od japonštiny nebo nesmyslných klikyháků. Po chvíli dostane další text v čínštině společně se souborem pravidel pro přiřazování druhé skupiny textu k první skupině. Pravidla jsou v angličtině a Searle jim jako rodilý mluvčí rozumí. Přiřazování na základě pravidel je formální, protože symboly neidentifikuje podle významu, ale jen podle tvaru. Po chvíli dostane třetí dávku čínského textu, společně s anglickými instrukcemi, podle kterých je schopen tuto dávku textu přiřazovat k první a druhé a výsledek vracet zadavateli. Zadavatelé pak pojmenují první dávku textu "skript", druhou "příběh", třetí "otázky", symboly, které jim vrátí, "odpovědi" a

---

<sup>53</sup> Termíny "chatbot" a "chatterbot" v angličtině označují program, který je určen k vedení komunikace s lidským protějškem a může se tak účastnit T-testu (viz kapitola 8 této práce).

<sup>54</sup> Parafrazuji Searlův vlastní příklad (Searle 1980b, p.417), který je ale založen na autentických Schankových "restauračních skriptech" (Schank & Abelson 1977).

pravidla v angličtině "program". Postupem času se pravidla natolik zdokonalí a Searle získá takový cvik, že jeho odpovědi budou nerozlišitelné od odpovědí rodilého mluvčího čínštiny; budou stejně správné, jako kdyby odpovídal na anglické otázky anglicky. Přesto nerozumí čínštině a chová se přesně jako počítač: provádí komputační operace podle formálních pravidel (Searle 1980b, pp.417-418). Searle později učinil svůj příklad transparentnějším a srozumitelnějším, začal ještě více klást důraz na rozdíl mezi *syntaxí* a *sémantikou*:

A teď si představte, že by vás někdo zavřel do místnosti, v níž je několik košů naplněných čínskými znaky. A představte si, že (stejně jako já) nerozumíte ani slovo čínsky, že však máte k dispozici manuál pro manipulaci s čínskými znaky, napsaný ve vašem jazyce. Pravidla přitom určují způsob manipulace se znaky čistě formálně - pouze s ohledem na jejich syntax, a nikoli na jejich sémantiku. [...] Předpokládejme, že vám nějací lidé do místnosti podají další čínské znaky a že vám poskytnou též pravidla pro vydávání znaků z místnosti ven. Předpokládejme, že znaky podávané dovnitř se bez vašeho vědomí chápou jako 'otázky' a symboly podávané ven jako 'odpovědi'. Předpokládejme dále, že programátoři byli při psaní programů tak úspěšní a váš výkon při manipulaci se symboly tak dokonalý, že vaše odpovědi brzy nebude možno odlišit od odpovědí rodilého mluvčího čínštiny. A tak jste zavřeni v místnosti, skládáte znaky a podáváte je ven jako odpověď na znaky zvenčí. Avšak na základě situace, kterou jsem právě popsal, pouze manipulací formálními znaky byste se nikdy nemohli naučit čínsky. (Searle 1994a, pp.33-34; překlad Marek Nekula)

Searle dokonce provedl formalizovaný zápis, který je depersonifikovanou a zestručnělou verzí příběhu (Searle 1994b, pp.546-547):<sup>55</sup>

Programy jsou formální (syntaktické).

Mysli mají obsahy (sémantiku).

Syntax není dostatečná pro sémantiku.

Programy nejsou mysli.

Ještě se vraťme k Schankovu programu a povšimněme si, že Searle v čínském pokoji nerozumí příběhům v čínštině, stejně jako program nerozumí příběhům v angličtině. A úplně stejně na tom jsou chatboty snažící se uspět v T-testu. Ať je jejich projev jakkoli humanoidní, přesto nemohou jejich syntakticky vytvořené programy pochopit sémantický obsah vět, které produkují. Zdá se tedy, že je případ uzavřen: protože dokáže T-testem projít i

<sup>55</sup> Původně, v poněkud pozměněné formě, Searle (1990a, p.27) a Searle (1991, p.526); mnohem podrobněji Searle (1994a, pp.40-43).

jednoznačně nemyslící stroj, nemůžeme jej dále považovat za postačující podmínku pro přiznání myšlení. Anebo můžeme?

Podívejme se nejprve na zdroje tohoto myšlenkového experimentu. Searlův postup bývá sice považován za originální, ale ve skutečnosti můžeme vystopovat jeho předchůdce.<sup>56</sup> Předchozí vlivy lze rozdělit do dvou skupin: za prvé to je myšlenka, že počítačové úkony uvnitř stroje může vykonávat člověk, a za druhé tvrzení o absenci sémantiky u strojům podobných entit. Nejstarší zmínkou o tom, co se děje uvnitř údajně myslícího stroje, je zřejmě sedmnáctý oddíl Leibnizovy *Monadologie*, ve kterém se autor zamýšlí nad povahou lidského vnímání:

Dále se musí nutně přiznat, že percepci a to, co na ní závisí, nelze vysvětlit z *mechanických důvodů*, tj. z tvarů a pohybů. Kdybychom si mysleli nějaký stroj, který by byl zařízen tak, že by dokázal myslet, vnímat a percipovat, mohli bychom si jej při zachování těchto vztahů myslet zvětšený tak, že bychom do něho mohli vstoupit jako do nějakého mlýna. Předpokládáme-li to, pak, kdybychom prohlíželi jeho vnitřek, nenašli bychom v něm nic než kusy, které na sebe narážejí, nikdy něco, z čeho by bylo možné vysvětlit percepci. Je jí tedy nutno hledat v jednoduché substanci, nikoli ve složeném či ve stroji. (Leibniz 1982, pp.158-159; překlad Jindřich Husák)

Podobnost mezi Leibnizovou a Searlovou argumentací je až neuvěřitelná, a to včetně skeptického stanoviska vůči možnosti myslícího stroje. Neutrální postoj naopak zastává Turing, který ve svých textech operuje s takzvaným papírovým počítačem. Jedná se o člověka, který s pomocí mechanických prostředků - například poznámkového bloku, tužky a gumy - provádí počítačové operace definované programem. Jedinec ovládaný papírovým počítačem se při řešení problému nespolehá na vlastní úsudek, ale následuje jednoduché instrukce, které rozhodují za něj. Člověk využívající tyto mechanické prostředky se tak překvapivě stává univerzálním počítačem, který je schopen zastat činnost jakéhokoli jiného počítače, i když s vynaložením mnohem delšího času. Turing například předpokládá, že je možné sestavit papírový počítač, který by dokázal hrát šachy (Turing 2004b, p.412). Ještě v osmdesátých letech byly papírové počítače docela běžným nástrojem pro výuku programování a mnozí dnešní uživatelé počítačů nostalgicky zavzpomínají na papírový počítač CGS, na kterém si během pozdní normalizace mohli trpěliví čtenáři *VTM* zahrát

---

<sup>56</sup> V tomto ohledu se nechávám inspirovat encyklopedickým heslem "The Chinese Room Argument" ze *Stanford Encyclopedia of Philosophy* (Cole 2009; oddíl 2).

strhující hry jako *Přistání na Měsíci*.<sup>57</sup> Nejdokonalejší analogií k čínskému pokoji je metaforický popis T-stroje, který ve své přednášce použil matematik Hartley Rogers více než dvacet let před Searlem. Rogersův přírůbek je téměř identický se Searlovým a zahrnuje muže, který je uzavřen v krabici se stolem, tužkami a papírem. Krabice obsahuje dvě štěrbinou označené jako *vstup* a *výstup*. Na základě instrukcí je muž schopen přiřadit k číslům, získaným na vstupu, jiná čísla, která odevzdává štěrbinou výstupu (Rogers 1959, pp.115-117). Fakt, že Searle neuvádí ani jeden z těchto textů jako pramen inspirace, je zarážející.

Druhým zdrojem jsou úvahy o tom, že činnost počítače může být sice na pohled neodlišitelná od lidské, ale počítači chybí schopnost chápat její význam. Už Ritchie v dříve citovaném textu intuitivně tvrdil, že mrknutí člověka do publika během šachové partie má význam, v případě robota nikoli (Ritchie 1957, p.66). Ritchie ale otázku významu nepřilíší pochopitelně spojil se svobodou rozhodování, a proto se Searlově preciznější argumentaci ani nepřiblížil. Mnohem zásadnější, ale opět nepřiznaný, je vliv eseje Neda Blocka "Troubles with Functionalism" z roku 1978.<sup>58</sup> Block ještě před svou fenomenálně známou kritikou T-testu v "Psychologism and Behaviorism", kterou se budu zabývat v příští kapitole, navrhl možnost hypotetické simulace lidského mozku pomocí celého obyvatelstva Číny. Tomuto příkladu se někdy říká *argument čínského národa* a Searle je mu zavázán určitě víc, než jen co do jména. Block uvažuje takto: lidský mozek je tvořen asi stovkou miliard neuronů, ale pro naše potřeby si vystačíme se zjednodušeným modelem o jedné miliardě. Představme si, že se nám podaří přesvědčit vládu Čínské republiky, aby rezignovala na komunismus a za svou oficiální doktrínu - aspoň dočasně - přijala funkcionalismus. Každý obyvatel Číny představuje jednu nervovou buňku a s ostatními komunikuje pomocí obousměrné vysílačky, kterou má k dispozici. V případě kontaktu mají Číňané za úkol předat signál svým předem stanoveným kolegům, a tak se "nervový vzruch" šíří napříč celým národem. Block má za to, že podle funkcionalistů by Čína byla funkcionálně ekvivalentní s lidským mozkem, což ovšem považuje za absurdní (Block 1978, pp.277-280; Block 2007, pp.70-72). Všimněme si, že i když se detaily obou experimentů liší, Blockův a Searlův cíl je totožný: ukázat, že komputace není

---

<sup>57</sup> Pod vznešeným názvem "Computer Games Station" (CGS) se skrývala programovatelná papírová kalkulačka, kterou si mohl uživatel vystřihnout z obálky časopisu, vybavit papírovou páskou a používat při programování primitivních počítačových úloh (Škoda 1982, pp.68-80).

<sup>58</sup> Pozdější reprinty prošly výraznými revizemi, nejaktuálnější verzí je Block (2007).

dostatečná pro produkci mysli. Searle v čínském pokoji je stejně bezradný a nevědomý jako Číňan v celonárodní simulaci mozku. V roce 1990 publikoval Searle revidovanou verzi svého důkazu, která je zaměřena na kritiku konekcionismu. Tentokrát v pokoji - říkejme mu *čínská tělocvična* - není jen jeden člověk, ale celá skupina rodilých mluvčích angličtiny, kteří jiný jazyk neovládají. Každý člen skupiny internalizoval jednu malou část instrukcí programu pro chápání příběhů v čínštině, ale ani v tomto případě nemůžeme říct, že by členové skupiny čínštině rozuměli, nerozumí jí ani skupina jako celek (Searle 1990a, p.28). Searlova neochota uznat zásluhy svých předchůdců se projevila i tentokrát, v textu totiž opět nenajdeme ani zmínku o Blockovi a jeho původním experimentu.<sup>59</sup>

Dříve, než Searle svou esej publikoval, diskutoval její obsah s nejrůznějšími filosofi. Výběr těch nejčastějších a nejpřesvědčivějších námitek tvoří - podobně jako v "Computing Machinery and Intelligence" - podstatnou část Searlova článku. Jejich názvy vymyslel Searle a pro úplnost do závorky uvedl jméno univerzity, odkud námitka pochází (Berkeley, Yale, MIT nebo Stanford). Není snad třeba zdůrazňovat, že - opět jako Turing - žádnou námitku nepovažoval za natolik závažnou, aby kvůli ní rezignoval na snahu demaskovat neudržitelnost komputačního přístupu k mysli. Stručně je všechny představím, ale podrobně se budu zabývat jen těmi, které považuji za podstatné. Zatímco v Turingově případě bylo třeba ubránit T-test proti všem způsobům kritiky, u Searla bude stačit nalézt pouze jeden doklad neudržitelnosti argumentu čínského pokoje.<sup>60</sup>

## 5.1 Systémová námitka

*Je pravda, že osoba zavřená v čínském pokoji nerozumí čínštině, ale je pouze součástí rozsáhlejšího systému, který už čínštině rozumí. Tento systém tvoří například osoba, příručka pravidel, papíry a tužky pro provádění výpočtů atd.*

I kdyby individuum internalizovalo všechny elementy systému, naučilo se pravidla nazpaměť a všechny výpočty by provádělo v hlavě, stejně by se na jeho nerozumění

---

<sup>59</sup> Na toto Searlovo opomenutí upozorňuje Rapaport (2006a, p.382), zatímco Rey (1986, p.181) explicitně mluví o nepoctivosti.

<sup>60</sup> V následující pasáži jsem si kvůli přehlednosti dovolil zvolit zkratkovitý způsob zápisu: vlastní námitka bude shrnuta v kurzívě, Searlova reakce pak v běžném písmu.



čínštině nic nezměnilo. Námitka je podle Searla zcela absurdní už od počátku, protože pokud nerozumí čínštině samotná osoba, pak jí nemůže rozumět ani tato osoba plus listy papíru. Žádná součást systému ani on sám jako celek totiž nezná referenci slov, která používá. Systémový přístup k problematice navíc může chápat každý nekognitivní systém jako kognitivní, pokud zahrnuje přijímání podnětů a vydávání výstupů (např. žaludek nebo už zmiňovaný termostat). Odlišit mentální a nementální systémy je přitom podle Searla jedním z hlavních úkolů teorie mysli a silná umělá inteligence v něm fatálně selhává.

## 5.2 Námitka robota

*Argument čínského pokoje je oprávněný v případě stávajících počítačů, ale lze ho překonat tím, že program umístíme do počítače v robotovi, který je vybaven senzomotorickými orgány. Robot by mohl pomocí čidel vnímat vnější svět a končetinami by na něj mohl kauzálně působit, byl by tak schopen vykonávat činnosti podobné lidským.*

Jedná se o silnější variantu systémové námitky v duchu homunkulárního funkcionalismu, která nahrazuje kancelářské potřeby mechanickým tělem robota. Searle ji velmi stručně zamítá jako neopodstatněnou, protože přidání percepčních a motorických schopností nemění nic na problému rozumění. V experimentu nezáleží na tom, zavřeme-li Searla do místnosti, která je vybavena i externí televizní kamerou a robotickou paží. I v tomto případě bude osoba zavřená uvnitř čínského pokoje manipulovat symboly podle striktně formálních pravidel.

## 5.3 Námitka simulátoru mozku

*Vytvoříme program, který nereprezentuje informace o světě, ale simuluje skutečné nervové vzruchy na synapsích mozku rodilého mluvčího čínštiny, když poslouchá příběhy a odpovídá na otázky. Potom stroji poskytneme jako vstup příběhy v čínštině a díky počítačové simulaci mozkových procesů získáme odpovědi na příslušné otázky.*

Autoři této námitky se podle Searla na hony vzdalují původním ambicím funkcionalismu, který chce vysvětlit mentální stavy bez odkazování na neurální činnost mozku. Pokud je třeba neurofyziologického popisu, pak nepotřebujeme funkcionalismus ani umělou inteligenci. Kromě toho přidává další argument a navrhuje nahradit čínský pokoj systémem vodovodních trubek a ventilů, které ovládá člověk podle instrukcí v angličtině.<sup>61</sup> Jestliže by skutečně nastala situace, že by tento systém produkoval správné odpovědi na otázky v čínštině, přesto by jim nerozuměl. Problém takového simulátoru spočívá v tom, že simuluje jen formální funkce mozku a nikoli jeho kauzální vlastnosti, které vytvářejí intencionální stavy.

#### 5.4 Kombinovaná námitka

*Předchozí tři námitky nejsou zcela přesvědčivé, ale pokud je propojíme, získají na významnosti. Robot vybavený senzomotorickými orgány a počítačem simulujícím mozek je jako celek intencionálním systémem.*

Searle tvrdí, že při povrchním setkání s takovým robotem mu budeme opravdu přisuzovat racionalitu a intencionalitu, aspoň dokud se o něm postupem času nedozvíme víc. Jedinou signifikantní vlastností robota je nerozlišitelnost chování, která ale není dostatečným důvodem pro přiznání schopnosti myšlení. Pokud bude chování robota řízeno pouze manipulací s neinterpretovanými formálními symboly, pak je to jen velmi promyšlená mechanická loutka bez mysli. To je mimochodem velmi důležitá připomínka vůči T-testu: Turing předpokládá, že nerozlišitelnost chování člověka a stroje je zárukou myšlení, zatímco Searle vyžaduje při přisuzování myšlení a intencionality ještě znalost vnitřního uspořádání entity. Jiné živočišné druhy (například opice nebo psy) považujeme za intencionální bez větších problémů, protože nejsme schopni vysvětlit jejich chování jinak než na základě intencionality a

---

<sup>61</sup> Na začátku této pasáže jsem sliboval jen pouhé shrnutí námitek a autorových reakcí, nemohu si zde ale odpustit poznámku k intelektuální nepoctivosti, které se Searle nedokáže vyhnout - ostatně jako mnozí další kritici komputacionismu. Podobně jako Purtillo využívá Searle úhybný manévr a odvádí pozornost čtenáře od elegantního digitálního počítače k neohrabanému zařízení, které přece *nemůže* myslet (viz kapitola 4).

navíc víme, že jsou nám zástupci těchto druhů podobní, co se týče materiálu, z něhož jsou složeni. U robotů naopak můžeme s jistotou říct, že jsou fyziologicky odlišní, a jsme i schopni odhadnout, čím je způsobeno jejich zdánlivě intencionální jednání - syntaktickým programem.

## 5.5 Námitka jiných myslí

*Informace o myšlení jiných lidí získáváme jen na základě jejich chování, musíme proto behaviorálně přisoudit myšlení i strojům, jejichž chování působí rozumným dojmem.*

Tuto námitku, která je blízká Turingovu čtvrtému argumentu z vědomí, považuje Searle za nepodstatnou a nevěnuje jí zvláštní pozornost. V kognitivních vědách předpokládáme existenci myslících entit samozřejmě, ale zajímá nás, co přesně znamená mít schopnost myslet. Myšlení však není pouhým komputačním procesem, proto jím nemohou disponovat počítače.

## 5.6 Námitka mnoha příbytků<sup>62</sup>

*Naše současné analogové a digitální počítače nejsou pro umělou inteligenci dostatečné, ale v budoucnosti sestrojíme technologie se schopností kauzálního působení a intencionality, které budou myslet.*

Ani v tomto případě Searle nepřikládá námitce důležitost, protože je opět v rozporu se samotnou definicí komputacionismu. Zastánci silné umělé inteligence tvrdí, že myšlení je zcela redukovatelné na komputační procesy definované čistě formálně. Budoucí počítače vybavené intencionalitou ale musí mít něco víc než syntaktický program, který není dostatečnou podmínkou myšlení.

---

<sup>62</sup> Z důvodů, které dost dobře nechápu, pojmenoval Searle tuto námitku s odkazem na evangelium (Jan 14:2). Pro převedení názvu námitky jsem použil ekumenický překlad *Nového zákona* (Praha: Česká liturgická komise, 1989).

Na závěrečných stranách své eseje Searle připomíná, že stroje *mohou* myslet, protože my lidé jsme přesně takovými myslícími stroji. Naše schopnost intencionality je způsobena tím, že jsme organismy s biologickou, chemickou a fyzikální strukturou. Formální vlastnosti nemohou produkovat intencionalitu, protože nemají neformální kauzální sílu. Searle je ovšem při pozitivním popisu mysli až podezřele tajnůstkářský. Není schopen říct nic konkrétnějšího k tomu, jaké mají být jeho počítače schopné intencionality, kauzálních interakcí se světem a myšlení.<sup>63</sup> Pořádně ani nevysvětlí, co vlastně myslí těmi "kauzálními vlastnostmi mozku". Pokud máme Searlův text považovat za definitivní odmítnutí T-testu, komputační teorie a funkcionalismu, musíme jej podrobit důkladné analýze.

Myšlenkový experiment čínského pokoje vyvolal vlnu bouřlivých diskusí a byl mnohými kritizován. Searle považuje svůj argument za nepochybně platný, jak prohlásil v nedávném rozhovoru pro *Philosophers' Magazine*:

Doufám, že jsme se už zotavili z behavioristické fáze ve filosofii mysli. Myslím, že jsme ji překonali stejně jako podobný omyl, komputacionalistickou fází. Je nesprávné považovat mysl za pouhé chování a dispozice k chování, ale stejně tak nesprávné je domnívat se, že je mysl pouhý počítačový program nebo soubor počítačových programů. Jednalo se o katastroficky mylné předpoklady, ale jak už jsem řekl, dostáváme se přes ně. (Baggini 2008, pp.57-58)

Tvrzení o nedostatečnosti syntaxe pro sémantiku se stalo jakousi mantrou odpůrců komputačního přístupu, aniž by bylo někdy přesněji řečeno, co se tím vágním zaklínadlem vlastně myslí. Nehledě na autorovu sebejistotu nepovažuji jeho argument za udržitelný. V posledních letech se objevila celá řada myslitelů, kteří Searlův myšlenkový experiment z nejrůznějších důvodů napadali. Robert Damper dokonce současnou situaci popsal tak, že "podle obecného konsensu mezi komentátory je argument čínského pokoje chybný".<sup>64</sup> Neexistuje sice žádná důvěryhodná statistika, která by Damperova slova potvrdila, ale vzhledem k paradigmatickému postavení komputacionismu v kognitivních vědách považuji autorův odhad za dosti pravděpodobný. Rozsah této práce mi neumožní zabývat se všemi

---

<sup>63</sup> Donald Michie tento nedostatek komentoval slovy: "we must hope that a suitable Searle Test will be forthcoming to complement Turing's" (Michie 1993, p.8). Není snad třeba dodávat, že toto čekání bylo dosud neúspěšné.

<sup>64</sup> V originále: "the general consensus among commentators is that the CRA [Chinese room argument] is flawed" (Damper 2006, p.163).

typy kritických reakcí, i kdybych toho byl vůbec schopen.<sup>65</sup> Zaměřím se proto jen na tři, které jsou podle mého názoru dostatečně silné, aby důkaz čínského pokoje vyvrátily. Uvedu je v pořadí podle důležitosti od té nejméně podstatné po nejzávažnější: (a) Domnívám se, že myšlenkové experimenty jako ten Searlův jsou jen obtížně ověřitelné, a nemohou proto sloužit jako základ pro seriózní odmítnutí teorie. (b) Tvrdím, že přinejmenším jedna z původních námitek je oprávněná, a to ta systémová. (c) Tvrdím také, že aspoň jedna premisa Searlova důkazu je chybná, konkrétně předpoklad týkající se údajné nedostatečnosti syntaxe pro sémantiku. Podívejme se na jednotlivé způsoby kritiky podrobněji.

Zlé jazyky říkají, že myšlenkové experimenty jsou myšlenkové právě proto, že ve skutečnosti nefungují. Tento bonmot je s velkou pravděpodobností přehnaný, myšlenkové experimenty jsou hojně užívány i v rámci přírodních věd a třeba teoretická fyzika nebo kosmologie jsou na ně do značné míry odkázány.<sup>66</sup> Přesto bychom měli být při jejich formulaci a hodnocení nanejvýš opatrní, abychom se nezapletli do sofismat a intelektuálních hříček. I současní filosofové mnohdy připomínají scholastické myslitele při pokusech apriorně vyvrátit existenci jevů, které by měly být spíše posuzovány pomocí empirických metod. F. H. George už v roce 1956 v souvislosti s umělou inteligencí varoval před "filosofy, kteří věří, že je *a priori* nemožné sestrojít létající stroj" a dával přednost "tvrdohlavým empirikům, kteří pokračují ve snaze vynalézt jej" (F. H. George 1956, p.247). Při analýze fenoménu čínského pokoje je třeba mít na mysli, že možnosti ruční simulace počítačích procesů jsou značně omezené a Searlův experiment nemůže být ve skutečnosti nikdy realizován. Matematický fyzik Frank J. Tipler si ve své jinak dosti obskurní knize *The Physics of Immortality* dal práci s výpočtem, jak by ruční simulace lidského mozku musela vypadat. Tipler odhaduje kapacitu mozku na  $10^{15}$  bitů; při průměrné délce knihy asi  $10^6$  by záznam všech těchto informací vyžadoval asi 100 milionů svazků, které by se vešly do třiceti pěti patrové knihovny. Přístupová rychlost celé naší paměti je asi 100 sekund (*sic!*) a během této doby by člověk v papírovém počítači musel být schopen každou knihu najít, vytáhnout z police, prolistovat a

---

<sup>65</sup> Částečný přehled je možné nalézt na serveru *MindPapers* na adrese <<http://consc.net/mindpapers/6.1c>>. Obecné uvedení do problematiky nabízí Cole (2009), reprint původního textu a klasické reakce otiskli Boden (1990) nebo Rosenthal (1991), výběr novějších komentářů se nachází v antologii Preston & Bishop (2002).

<sup>66</sup> V příští kapitole se budu obecnou kritikou myšlenkových experimentů zabývat detailněji, a proto si na tomto místě vystačím jen s dílčími poznámkami k samotnému čínskému pokoju.

vrátit zpět. Při průměrné váze knihy půl kilogramu by energetický výdaj na pouhé přenášení knih činil  $3 \times 10^{19}$  joulů za 100 sekund, to znamená výkon  $3 \times 10^{11}$  megawattů. Běžný lidský výkon je asi 100 wattů, a proto by pro pohon ruční simulace mozku bylo třeba energie  $3 \times 10^{15}$  lidských bytostí - dnes jich na planetě žije přibližně  $7 \times 10^9$  - nebo 300 milionů jaderných elektráren. Tipler svůj výpočet ukončí tvrzením, že pro člověka by bylo snazší přeskočit Měsíc, protože je to energeticky méně náročné (Tipler 1995, p.40). Výpočet nemůžeme brát příliš vážně, protože jen v mé verzi se aspoň desetkrát vyskytne termín "asi" - to není ani důvěryhodná aproximace, natož přesný výsledek. Přesto si můžeme udělat přibližný obrázek o nemožnosti papírového počítače simulujícího mozek:

Searlův experiment čínského pokoje nás nutí představit si logickou kontradikci: normální lidskou bytost, která dělá něco, co normální lidská bytost nedokáže. (Tipler 1995, p.42)

Při mnohem skromnějším a zřejmě jediném uskutečnění experimentu, který zorganizovali na konci minulého století Rom Harré a Han-Ting Wang, musel být čínský slovník omezen na pouhých třináct čínských ideogramů a dvě odvozovací pravidla. Teprve za těchto podmínek byli participanti schopni provádět ruční komputace, i když žádný z nich neuměl čínsky a ani se to v průběhu pokusu nenaučil (Harré & Wang 1999). Nerad bych se pouštěl do diskuse týkající se logické a ontologické nemožnosti určitých entit a snad to ani není třeba. Důležité je, že existence počítače schopného uspět v T-testu je spíše empirickým než logickým problémem. Ptám se po jeho reálné existenci v aktuálním světě, hypotetická existence ve světech modální logiky mě jako naturalizovaného filosofa - s prominutím - nezajímá.

Jak už jsem předeslal, má první připomínka vůči čínskému pokoji je ta nejméně podstatná. Mnohem důležitější je podle mého názoru námitka, kterou sám Searle označuje za systémovou. Ta je také mezi kritiky myšlenkového experimentu zřejmě vůbec nejrozšířenější, a proto si pro své potřeby vybírám jen její nejklassičtější a přitom velmi srozumitelnou variantu (Rey 1986).<sup>67</sup> Podle Georgese Reye má Searle pravdu, když člověku uzavřenému v čínském pokoji upírá schopnost rozumění příběhům v čínštině. Připomíná ale, že ze systémového pohledu je člověk pouze jednou, i když klíčovou součástí na úrovni

---

<sup>67</sup> Autor svůj přístup později rozvedl v Rey (2002). Podobně argumentují i mnozí další, poprvé zřejmě Wilensky (1980), dále například Dennett (1991, pp.435-440), Crane (2003, pp.123-128) nebo Gallagher (2009).

procesoru (CPU). Pokud vezmeme v úvahu Turingovu definici univerzálního T-stroje, pak je člověk pouhou řídicí jednotkou, zatímco celý počítač (tj. čínský pokoj) se skládá ještě z paměti s tabulkou instrukcí - to jsou právě všechny ty příručky obsahující pravidla pro manipulaci s čínskými "klikyháky". Searle intuitivně tvrdí, že i kdyby se člověk v čínském pokoji naučil všechna pravidla nazpaměť, internalizoval je, ani pak by čínštině nerozuměl. Rey podobně intuitivně tvrdí opak, ale svůj postoj dokládá následující úvahou: co je podle Searla vlastně třeba pro rozumění jakémukoli jazyku, když to není znalost gramatických pravidel? Představme si, že by účastník experimentu mohl vyhlížet z pokoje ven na situace, o kterých činí výpovědi. V takovém případě by se učil jazyk ostentativně, přesně jak to dělají malé děti při zvládnání mateřštiny nebo terénní lingvisté při vytváření překladových příruček pro exotické jazyky (Quine 1960). Dřív nebo později by začal čínštině rozumět. Není zas tak složité představit si, že by byl skutečný počítač či robot vybaven smyslovými čidly, jež by mu umožňovala "vyhlížet" ven metaforickými okny čínského pokoje. Zdá se, že Searle úmyslně kalí vodu a komplikuje naše porozumění jazykové komunikaci. Toto jeho matení při popisu mysli je podle Reye projevem podezřelého druhismu a nepřináší nic pozitivního či nového.<sup>68</sup>

Rey upozorňuje ještě na jeden paradox týkající se Searlova myšlenkového postupu, spočívající v autorově údajném přijetí Fodorova metodologického solipsismu, který bývá často označován jako epistemologický či sémantický internalismus, případně individualismus. Spor mezi externalismem a internalismem byl iniciován sérií článků Hilaryho Putnama ze 70. let minulého století, především fenomenálně známým esejem "The Meaning of 'Meaning'" (1975b). Do čela opozice se postavil Jerry Fodor v "Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology" (1980), který odmítal považovat externí faktory za podstatné pro určování obsahů našich propozičních postojů. Nabídl oproti tomu karteziánskou teorii přísně rozlišující mezi vnitřními mentálními stavy a vnějším prostředím. Searle se k tomuto programu na začátku 80. let oficiálně i přes dílčí výhrady přihlásil a po Fodorově vzoru internalismus označoval za hlavní výzkumnou strategii kognitivní vědy (Searle 1980a). Jeho kritika počítačného pojetí mysli ovšem protirečí externalismu i internalismu. Nesouhlasí s externalistickým stanoviskem, jež považuje mluvčího a vnější sadu

---

<sup>68</sup> Rey píše: "Searle's denial that such a machine would really have mental states would seem as dogmatic as the pre-Galilean insistence that extra-terrestrial bodies couldn't have mass, or the anti-Darwinian insistence that humans couldn't be a kind of ape." (Rey 1986, p.180)

pravidel za systém ovládající čínštinu; odmítá i internalistické stanovisko, které si chce vystačit při analýze významu jen s mentálními obsahy našich myslí. Rey se domnívá, že Searlovy vágní požadavky na mysl s kauzálními vztahy k vnějšímu světu se ukazují být bezpředmětné a vlastně zcela postrádají jakoukoli výpovědní hodnotu. Navíc je možné protiargumentovat, že i dnešní počítače jsou kauzálními agenty. Reagují totiž na příčiny z vnějšího světa, i kdybychom je měli minimalisticky redukovat na vstupy z klávesnice, a jsou schopny produkovat kauzální účinky, jako například výstupy na monitorech, tiskárnách či jiných perifériích. Searle tak nedokáže vysvětlit, za jakých okolností je schopen myslet digitální počítač, ale ani za jakých okolností lze myšlení připsat lidem. Sice tvrdí, že lidské mozky jsou počítači svého druhu, není ale vůbec jasné, čím se od běžných počítačů liší.

Předchozí úvahy o externalismu doplním stručnou poznámkou o teorii rozšířené mysli, jejímiž autory jsou Andy Clark a David Chalmers a jež navazuje na Putnamovu, Burgovu či Davidsonovu epistemologii (Clark & Chalmers 1998; Clark 2008). Přijetí této intenzivně diskutované hypotézy by totiž učinilo rázný konec nejen Searlově argumentaci prostřednictvím čínského pokoje, ale jakýmkoli internalistickým pokusům o explikaci mysli. Clark s Chalmersem představují několik ilustrativních příkladů, z nichž některé ani není třeba považovat za myšlenkové experimenty. Nejdůležitější je příběh Otty, jenž onemocněl Alzheimerovou chorobou. Stejně jako mnozí jiní takto postižení pacienti Otto trpí poruchami krátkodobé i dlouhodobé paměti. Spoléhá se proto na informace z vnějšího prostředí, především na poznámkový blok, do kterého si píše vše podstatné a který neustále nosí s sebou. Autoři tvrdí, že Otto se v důsledku nijak neliší od Inge, která je zdravá a ukládá si informace nutné k životu do běžné paměti. Z těchto důvodů je možné považovat Ottův poznámkový blok za externí součást jeho mysli, kterou Clark a Chalmers označují jako "rozšířenou". Hranice našeho těla určená lebkou nebo pokožkou není epistemologicky závazná, naše intelektuální schopnosti mohou externalisticky spočívat i ve vnějších objektech. V posledních desetiletích tak činíme čím dál častěji, protože používáme mnoho kognitivních pomůcek, jako například logaritmická pravítka, kalkulátory, chytré telefony nebo přenosné počítače, ale i staré dobré tužky, poznámkové sešity a nalepovací lístky s upomínkami. Searle v čínském pokoji je závislý na právě takových kancelářských potřebách, a můžeme jej proto zahrnout do systému rozšířené mysli, jež exotický jazyk s přehledem ovládá. Podle autorů teorie je dokonce i tento pohled poněkud zkreslující, protože jazyk není



"zrcadlem našich vnitřních stavů, ale jejich doplňkem" (Clark & Chalmers 1998, p.18). Ze všech prostředků, které naše mysl používá při potýkání se se zbytkem světa, je jazyk zdaleka nejefektivnějším nástrojem. Poctivě připravený seznam s nákupem nám může mimořádně ulehčit bloudění po chodbách supermarketu, ale sestavení seznamu je podmíněno znalostí mnohem robustnějšího nástroje, tedy jazyka. Searle se tak ptá naprosto nesprávným způsobem, protože nejde o to *rozumět* jazyku pomocí externích prostředků, ale *používat* externí prostředek jazyka žádoucím způsobem.<sup>69</sup>

Se systémovou a robotickou námitkou úzce souvisí poslední způsob falzifikace Searlova experimentu, kterým se budu zabývat. Spočívá v odmítnutí poněkud vágní premisy, že syntax není dostatečná pro sémantiku. Na první pohled se zdá, že tyto dvě lingvistické disciplíny spolu mají jen pramálo společného: syntax analyzuje formální vztahy slov ve větě, zatímco sémantika operuje s významem jazykových výrazů. Přesto se v nedávné době objevily názory, jež tuto dichotomii relativizují či dokonce zcela popírají. Největším průkopníkem tohoto přístupu je americký informatik a filosof William J. Rapaport, který v posledních patnácti letech publikoval celou řadu textů popisujících vzestup sémantiky z ryze syntaktických prostředků (např. Rapaport 1986; Rapaport 1995; Rapaport 1998; Rapaport 2000; Rapaport 2002; Rapaport 2003; Rapaport 2006a). Velmi podobě argumentuje i David Chalmers, když v knize *The Conscious Mind* (1996, pp.326-327) posměšně parafrázuje Searlův argument:

Recepty jsou syntaktické.

Syntax není dostatečná pro drobivost.

Koláče jsou drobivé.

Implementace receptu není dostatečná pro drobivost.<sup>70</sup>

---

<sup>69</sup> Nutno podotknout, že teorie rozšířená myslí zůstává po více než deseti letech od svého prvního zveřejnění dosti kontroverzní. Vlivným a čtenářsky přístupným zamítnutím je recenze Clarkovy knihy z pera Jerryho Fodora (2009).

<sup>70</sup> V originále se jedná o dvojici analogicky formulovaných důkazů: "1. A computer program is syntactic. 2. Syntax is not sufficient for semantics. 3. Minds have semantics. 4. Therefore, implementing a program is insufficient for a mind." - "1. Recipes are syntactic. 2. Syntax is not sufficient for crumbliness. 3. Cakes are crumbly. 4. Therefore, implementing a recipe is insufficient for a cake."

Chalmers v podstatě opakuje systémovou námitku, protože rozlišuje mezi *programem* a *implementací programu*.<sup>71</sup> Rapaport jde víc do hloubky, neboť program považuje za statický objekt, jenž musí být realizován aktivním procesorem. Teprve součinnost těchto dvou prvků dává vzniknout celému systému, v jehož případě už můžeme mluvit o vlastnictví kognitivních a sémantických schopností. Navíc ale ukazuje, že rozumění přirozenému jazyku je jen zdánlivě sémantickou záležitostí. Přesněji řečeno - rozumění je způsobeno syntaktickou sémantikou, která spočívá v počítačové manipulaci se symboly, ničím jiným. Jako doklad pro svá překvapivá tvrzení Rapaport použil tři teze, z nichž první je zdaleka nejdůležitější (Rapaport 2006a, pp.385-387):

1. Podle klasické definice Charlese Morrisa je sémantika naukou o vztazích mezi dvěma množinami entit (fyzikální objekty a jejich mentální reprezentace), zatímco syntax je disciplína o vztazích uvnitř jedné množiny (mentální reprezentace). Myslící bytosti ovšem významy internalizují, převedou totiž vztahy mezi vnitřními pojmy a vnějšími předměty do nové skupiny interních entit. Rapaport zjednodušeně tvrdí, že sémantické i syntaktické vlastnosti jazykových výrazů jsou realizovány týmiž konfiguracemi neuronů centrální nervové soustavy, a proto mezi nimi nemůže být principiální rozdíl.

2. V novověké filosofii je takřka univerzálně rozšířená představa, že lidé disponují epistemickým přístupem pouze k interním reprezentacím předmětů a nikoli k samotným objektům vnějšího světa. Tato teorie je přijímána všemi variantami insulárního empirismu, protože všichni významní britští filosofové té doby považovali za elementární stavební jednotky našeho poznání nesložená fantazmata (Hobbes), percepce (Locke) nebo impresie (Hume). Teorie nalezla zastání i v kantovské tradici operující s fenomenálními obsahy nebo v rané analytické filosofii novorealistů, kteří tyto entity označovali jako smyslová data. Quine se zase domnívá, že vstupní branou k poznání světa jsou stimuly na nervových zakončeních našich sensorických orgánů. Filosofové mohou a snad by i měli být fyzikalisty v oblasti ontologie, ale musí být senzualisty v rámci epistemologie, protože neexistuje jiný zdroj poznání o vnějším světě než empirická zkušenost.

---

<sup>71</sup> "The argument does not distinguish between recipes, which are syntactic objects, and implementations of recipes, which are full-bodied physical systems in the real world." (Chalmers 1996, p.327)

3. Rozumění je rekurzivní, a proto je možné interpretovat oblast syntaxe v pojmech sémantiky. Syntaktické porozumění je ale elementárnější a tvoří fundament pro celý jazyk včetně sémantiky.

Je zřejmé, že Rapaport i přes svou spřízněnost s Chalmersovými závěry zastává epistemologický a sémantický internalismus. Domnívám se, že z této diskrepance vyplývají důsledky především pro samotný spor internalismu a externalismu, který si dovoluji považovat jen za málo podstatný. Zdá se mi, že jádro diskusí spočívá ve slovíčkaření o nastavení hranic mezi tím, co je pro člověka interní a externí. Tato téměř scholastická disputace ovšem nijak neposouvá míru pochopení našich kognitivních schopností. Internalismus příliš akcentuje triviální tvrzení, že veškeré mentální události jsou produkty fyzikálních stavů lidských mozků. Externalismus naopak příliš akcentuje neméně triviální tvrzení, že my lidé - včetně našich mentálních stavů - jsme součástí téhož fyzikálního světa jako vše ostatní. Otázka, zdali jsou chemické složení vnějších předmětů a sociální dělba práce jazykového společenství nutnými determinanty sémantického významu, spadá zřejmě do oblasti hermeneutické interpretace, a proto není z vědeckého hlediska příliš důležitá.

Rapaport má oproti Searlovi značnou výhodu, protože své názory nedokládá abstraktními myšlenkovými experimenty, ale konkrétními příklady z historie. Tvrdí, že některé lidské bytosti se ve skutečnosti nacházejí v čínských pokojích, což jim přesto nezabraňuje naučit se používat jazyk a rozumět jeho sémantickému obsahu. Modelovým příkladem je podle něj Helen Kellerová, která ve věku několika měsíců onemocněla blíže nespecifikovanou chorobou - pravděpodobně spálou či zánětem mozkových blan - a prožila téměř celý svůj život bez sluchu a zraku. Péčí své vychovatelky Anne Sullivanové se přesto vzmohla k mimořádnému intelektuálnímu výkonu a naučila se komunikovat se svým okolím pomocí prstové abecedy, což jí později umožnilo vystudovat vysokou školu a napsat několik autobiografických i jiných knih.<sup>72</sup> Kellerová ve svých pamětech detailně popisuje, jak začala postupně spojovat taktilní a olfaktorické stimuly s taktilními znaky jazyka hluchoslepých. Klíčovým okamžikem bylo pochopení vztahu mezi hmatovým počítkem panenky a prstově vyhláskovaným slovem "panenka". Ještě více přelomovým byl zřejmě moment, kdy Kellerová

---

<sup>72</sup> Podrobnosti z života Helen Kellerové je možné načerpat především z knihy *The Story of My Life* (1905). Rapaport není zcela originální, když vyzdvihuje přínos Kellerové ke kognitivním vědám; podobně již před ním argumentovali třeba Swan (1992) a Leiber (1996).

nahlédla souvislost mezi látkovými termíny a konkrétním prožíváním příslušných stimulací, jak se to poprvé projevilo při porozumění pojmu "voda". Pro Rapaporta je příklad Kellerové mimořádně důležitý, protože díky svému postižení už od počátku hledala vztahy uvnitř jedné množiny taktilních zkušeností. Rapaport správně připomíná, že v jejím případě nebylo možné zřetelně odlišit lingvistickou a nelingvistickou část světa, a proto je její osvojení si jazyka ryze syntaktickým procesem. Syntax se tak může stát dostatečnou podmínkou sémantiky, výchozí Searlův argument se ukazuje být zcela neudržitelný.<sup>73</sup>

Uzavřu tuto kapitolu poněkud radikálně, obviněním Searla z použití psychologického triku. Podle mého názoru autor zneužívá fakt, že nám laikům čínština připadá nesmírně komplikovaná, a proto máme tendenci pochybovat o možnosti jejímu porozumění uvnitř čínského pokoje. Mnohé by se ale vyjasnilo, kdyby bylo v příkladu manipulováno s obrázkovými piktogramy, ze kterých se čínské znakové písmo původně vyvinulo. Po Helen Kellerové prošlo podobným vývojem jako ona mnoho dalších, jako například ukrajinská hluchoslepá vědkyně Olga Ivanovna Skorochodová. Pokud bychom nechali rozhodnutí o jejich schopnosti rozumět jazyku pouze na Searlovi, pak by se žádný z příslovečného čínského pokoje nikdy nedostal.

---

<sup>73</sup> Pro další diskusi o oprávněnosti analogie mezi Helen Kellerovou a čínským pokojem viz Ford (2011) a Rapaport (2011).

## 6. Brutální síla: Block 1981

Osmdesátá léta minulého století se nesla ve znamení kritiky behaviorálního a funkcionálního přístupu k inteligenci, za jehož představitele byl Turing považován. Kromě Searlových prací je nejvýznamnějším příspěvkem v tomto tažení mnohokrát přetiskovaný a hojně citovaný článek "Psychologism and Behaviorism" od amerického filosofa Neda Blocka. V tomto eseji je předložen slavný myšlenkový experiment vyvracející T-test, později označovaný jako "Blockhead argument".<sup>74</sup> Autorova motivace je ovšem mnohem ambicióznější než pouhé zamítnutí T-testu jako kritéria pro přiznání myšlení digitálnímu počítači. Skutečným cílem je dekonstrukce behaviorismu ve prospěch přístupu, jenž Block označuje jako psychologismus. Je totiž údajně možné, aby byly dva systémy totožné v chování, i když je jeden z nich inteligentní a druhý nikoli.<sup>75</sup> Block proto zdůrazňuje, že připsání inteligence nesmí být založeno na pouhých vnějších projevech příslušné entity, ale i na pochopení vnitřních procesů zpracování informací, jimiž jsou vnější projevy způsobeny. Tímto žádoucím přístupem ke studiu myšlení je právě psychologismus, koncipovaný jako protiváha behaviorismu, funkcionalismu a komputacionismu. Hned na úvod si dovoluji poznamenat, že Blockova kritika je snad legitimní v případě klasického behaviorismu J. B. Watsona a B. F. Skinnera, který opravdu pracoval s minimalistickou rovnicí  $S \rightarrow R$ , podle níž stimul vyvolává reakci. Už v nebehaviorismu je ale tato koncepce doplněna na  $S \rightarrow (O) \rightarrow R$ , přičemž střední "O" prostředkuje mezi stimulem a reakcí a bývá interpretováno buď jako "osobnost", nebo obecněji jako "organismus". Funkcionalismus je poučenou verzí starších behaviorálních teorií a zajímá se o povahu komputačních procesů, k nimž uvnitř myslících entit dochází. Block přitom explicitně tvrdí, že námitky proti behaviorismu lze aplikovat i proti funkcionalismu (Block 1981, p.7). Do jisté míry tedy dezinterpretuje metodologická východiska funkcionalismu a mohl by snad být obviněn z útoku na slaměného panáka ("straw

---

<sup>74</sup> Termín poprvé použil Jackson (1993, p.240). Neuškodí zamyslet se nad etymologií tohoto novotvaru. "Blockhead" je pochopitelně narážkou na autorovo příjmení, ale zároveň je i anglickým hovorovým pojmem, který v překladu znamená "hlupák" nebo "pitomec".

<sup>75</sup> Analogicky argumentuje i David Chalmers, i když v otázce přisuzování vědomí. Na notoricky známém příkladu s filosofickými zombiemi ukazuje, že živý člověk se od nemrtvého nemusí lišit behaviorálními projevy, přestože zombie nedisponuje vůbec žádnými mentálními stavy (např. Chalmers 1996, pp.94-99 a jinde).

man"), tedy logické chyby založené na konstrukci stanoviska, jež není oponentem vůbec zastáváno.

Block nepovažuje standardní námitky proti behaviorismu za příliš přesvědčivé a tvrdí, že T-test je vůči nim imunní. Jako příklad uvádí Chisholmův a Geachův argument, podle kterého není možné přiřadit ani ty nejjednodušší mentální stavy ke konkrétním behaviorálním dispozicím, nebo Putnamův myšlenkový experiment o spartánských hercích, kteří dokážou svým hereckým uměním zabránit vnějškovým projevům svých vnitřních mentálních stavů. Většina námitek je založena na *operacionálním* pojetí T-testu, které považuje jeho úspěšné složení za nutnou podmínku inteligence, zatímco správnější je chápat T-test *behaviorálně*, jako postačující podmínku. Entita, která neprošla T-testem, může být přesto za určitých okolností považována za inteligentní. Může být třeba paralyzována nebo se jí prostě nebude chtít. Stačí tedy, pokud je příslušný systém vybaven behaviorální dispozicí ke splnění T-testu, není již třeba jej reálně splnit.

Zároveň ale Block poukazuje na problematičnost některých aspektů T-testu. Nejvíce potíží spatřuje v Turingově neuspokojivé definici inteligence<sup>76</sup>, která až příliš spoléhá na osobnostní rysy lidských tazatelů. Jednou z nevýhod je, že tazatel může sice počítačového účastníka zpočátku považovat za inteligentního, ale díky určitým zvláštnostem v odpovědích později poznat, že se jedná o stroj. Počítači se prostě nemusí dařit lhát o své povaze dostatečně efektivně. Podle Blocka jsou lidé v otázce připisování inteligence neživým entitám dosti šovinističtí, což může tazatele vést k podceňování testovaného počítače a následnému odmítnutí považovat jej za myslící. Ještě důležitější nevýhodou je, že tazatel může být v celku snadno obelstěn i nemyslicím strojem, u něhož inteligence naprosto absentuje. Lidé podléhají celé řadě kognitivních předpojatostí<sup>77</sup>, mezi něž patří i tendence k neoprávněné antropomorfizaci nemyslicích agentů. Block uvádí anekdotické příběhy o důvěřivých lidech,

---

<sup>76</sup> Je třeba pro pořádek připomenout, že Turing svůj test za definici inteligence vůbec nepovažoval, ale interpretoval jej pouze jako postačující podmínku pro její připsání (viz kapitola 3 této práce). Block tak dokonale naplňuje temnou poznámku Daniela Dennetta: "It is a sad irony that Turing's proposal has had exactly the opposite effect on the discussion of that which he intended." (Dennett 1995, p.122) Podobně se vyjadřuje i Stuart Shieber: "But philosophers just won't listen. They insist on investigating the issue of whether the Turing Test is a good definition of intelligence, despite Turing's best efforts to avoid definitions entirely." (Shieber 2007, p.690)

<sup>77</sup> Jako "kognitivní předpojatost" překládám anglický termín "cognitive bias", abych je odlišil od kognitivní chyby ("cognitive error") a kognitivního zkreslení ("cognitive distortion").

kteří považovali za myslící i primitivní Weizenbaumův program *ELIZA*, který byl tvořen pouhými dvěma sty řádky v programovacím jazyce BASIC (viz kapitola 8 této práce). *ELIZA* byla přesvědčivá především v krátkých rozhovorech, ve kterých se hrávala roli naslouchajícího účastníka používajícího jednoduché doplňující otázky. Podle Blocka lze předpokládat, že komplikovanější program s pestřejší škálou reakcí by obelstil každého nebo téměř každého tazatele. Soudci v T-testu se potácejí mezi dvěma nebezpečnostmi, mezi *Skyllou* přílišné přesnosti a *Charybdou* neoprávněné benevolentnosti:

Stručně shrnuto, lidští soudci mohou být nespravedlivě šovinističtí při zamítání skutečně myslících strojů, anebo přespříliš liberální při přijetí chytře zkonstruovaných, nemyslících strojů. (Block 1981, p.10)

Pro Blocka je podstatnější druhé nebezpečí, protože první je možné překonat důslednější definicí inteligence, která by se nespolehala pouze na posuzování lidskými interprety. Například lze (konverzační) inteligenci definovat jako schopnost produkovat smysluplné sekvence verbálních reakcí na sekvence verbálních stimulací. Ani tato reformulace nám ale nezabrání, aby nás nedokázala podvést zařízení s jednoduchou vnitřní architekturou, jež imitují inteligentní odpovědi v mezilidském rozhovoru. Nejabsurdnějším příkladem takového stroje je právě *Blockhead*, který podle svého autora dokáže korektně reagovat na jakýkoli dotaz, i když po prozkoumání jeho vnitřního uspořádání musíme uznat, že neoplývá inteligencí v žádném běžném významu toho slova.

*Blockhead* má naprogramovány všechny řetězce smysluplných vět, které lze použít v přibližně hodinové konverzaci. Hodinový limit je přitom arbitrární a má sloužit jen k snazšímu pochopení Blockova myšlenkového experimentu. Hypotetický počet všech řetězců smysluplných vět je sice obrovský, ale přesto konečný. Stroj funguje tak, že nejprve vyslechne tazatelovu první větu, najde ji ve svém seznamu řetězců a odpoví libovolnou jinou větou z příslušného řetězce. Počká si na tazatelovu následnou větu, opět ji vyhledá v příslušném seznamu a náhodně vybere další vhodnou odpověď. Konverzaci je možné představit si jako strom s mnoha větvemi, které se postupem času stále dělí. Počítač ve větvích jen mechanicky nalézá vhodné odpovědi, které do něj byly dříve uloženy lidskými programátory. Aby nebyla práce tvůrců programu příliš komplikovaná, mohou pro zjednodušení simulovat reakce pouze jedné osoby s vlastní minulostí, přičemž Block žertem

navrhuje svou tetu Bertu nebo třeba Robinsona Crusoe.<sup>78</sup> Podle Blocka je zřejmé, že schopnost produkovat smysluplné odpovědi není postačující podmínkou inteligence, a proto je její pojetí vyplývající z T-testu nevyhovující. Obrazně řečeno, Blockhead je stejně inteligentní jako obousměrná vysílačka, tudíž vůbec ne.<sup>79</sup> Rozdíl spočívá jen v tom, že Blockův stroj má všechny možné odpovědi archivovány člověkem ještě před začátkem testu, zatímco v případě vysílačky je tentýž člověk vytváří průběžně, v reálném čase. Ani v jednom případě tudíž není možné mluvit o inteligenci ve vlastním slova smyslu. Veškerá inteligence, která je mu připisována, je ve skutečnosti inteligencí jeho programátorů. V jasné narážce na Searla Block svou argumentaci rozšiřuje a tvrdí, že identický myšlenkový postup nemusí být použit jen v případě strojů, ale dokonce i u lidí, kteří se naučili nazpaměť něco, čemu vůbec nerozumí. Jsou to například ti, kteří podle zvuku namemorovali odpovědi v čínštině, aniž by jazyk doopravdy ovládali. Tak by například bylo možné, aby idiot s fotografickou pamětí vedl brilantní filosofickou disputaci podle konverzačních řetězců, které pro něj předem připravil tým vynikajících odborníků v oboru. Není navíc třeba omezovat experiment jen na oblast konverzační inteligence, protože v souladu se Searlovou robotickou námitkou by bylo možné sestrojít nemyslíčího robota, který by se choval rozumně ve všech představitelných situacích, pokud by měl dopředu naprogramovány všechny myslitelné sekvence činností.

Podobně jako v případě čínského pokoje není Blockův myšlenkový experiment zcela originální. Původní zdroje jsou ale tentokrát mnohem obskurnější a není vůbec jisté, zdali byl s nimi autor seznámen. První formulace pochází už z poloviny padesátých let, kdy Claude E. Shannon a John McCarthy napsali:

Nevýhoda Turingovy definice myšlení spočívá v tom, že bylo principiálně možné navrhnout stroj s úplnou množinou arbitrárně zvolených odpovědí na všechny možné vstupní stimuly. [...] Takový stroj by pro každou vstupní situaci (včetně těch minulých) byl schopen ve "slovníku" vyhledat správnou reakci. S vhodným slovníkem by takový stroj naplnil Turingovu definici, ale nevyhovoval by našemu běžnému a intuitivnímu pojetí myšlení. Z toho vyplývá, že fundamentálnější definice musí nějak zahrnovat způsob, jakým stroj dospívá ke svým odpovědím. Analogicky řečeno, musí zachytit rozdíl mezi

---

<sup>78</sup> O pár let později Block přejmenoval svou tetu na domáctější Bubbles (Block 1995, pp.382-384), a právě pod názvem "Aunt Bubbles" bývá argument někdy používán (např. Łupkowski 2006).

<sup>79</sup> Block používá ještě dva jiné slavné příklady: Blockhead je inteligentní jako toaster (Block 1981, p.21) nebo jukebox (Block 1995, p.383).



osobou, která řeší problém přemýšlením, a jinou, která se odpověď naučila nazpaměť. (Shannon & McCarthy 1956, p.vi)<sup>80</sup>

Polský badatel Pawel Łupkowski připomíná jiný pozoruhodný fakt, a to že několik desetiletí před Blockem něco podobného tvrdil i science-fiction spisovatel Stanisław Lem (Łupkowski 2006). Už v roce 1964 v *Summa technologiae* představil "vesmírný gramofon", který se liší od "obyčejných" počítačů a pracuje na analogickém principu jako Blockhead:

Druhý stroj je naprosto jiný. Je to do rozměrů planety (nebo vesmíru) zvětšený Gramofon. Má velmi mnoho, např. sto trilionů nahraných odpovědí na všechny možné otázky. Když položíme otázku, stroj jí "nerozumí" a jen její forma, tj. posloupnost zvuku našeho hlasu uvede do pohybu relé, které roztočí desku nebo pásek s nahranou odpovědí. Nejde o technickou stránku věci. Je jasné, že takový stroj je neekonomický, že ho nikdo nebude dělat, protože je to vlastně nemožné a hlavně se neví proč. Nás však zajímá teoretická stránka. Jestliže o tom, zda má stroj vědomí, rozhoduje chování a ne vnitřní stavba, nedojdeme brzy k závěru, že "vesmírný gramofon" je má a tím vyslovíme nesmysl? (Lem 1995, p.113)

Podobnost je opět až zarážející, ale Łupkowski správně poukazuje na absenci překladu Lemovy knihy do angličtiny, a tudíž i nízkou pravděpodobnost, že by Block příslušný text mohl číst. Pozornosti historiků Blockova argumentu zatím unikal i Richard L. Purtil, který v dříve analyzovaném článku "Beating the Imitation Game" navrhl zjednodušenou variantu imitační hry a zároveň důvody pro její zamítnutí.<sup>81</sup> Purtillova verze spočívá v tazatelově okamžitém hodnocení odpovědi počítače na jedinou otázku. Samočinný stroj by nejprve ve velkém počtu těchto her generoval náhodné odpovědi. Později by se mechanicky vyhýbal těm chybným, které byly tazatelem odmítnuty, a v budoucnu používal jen ty úspěšné. Stroj by se takto "učil" sám ze svých omylů a jeho odpovědi by vůbec nebyly determinovány programátorem, ale přesto by podle autora bylo absurdní snažit se mu připsat myšlení a inteligenci (Purtill 1971, p.292).

Není příliš překvapivé, že Block v závěru své eseje uvedl sadu osmi námitek, u nichž očekával, že zaznějí v následné diskusi. V literatuře týkající se T-testu je to standardní postup,

---

<sup>80</sup> Na tento pramen upozorňuje Copeland (2000, p.532) a Shieber (2007, p.711).

<sup>81</sup> Nejsem si jistý, zdali se jedná o zcela legitimní poznámku, ale Purtillova verze je do značné míry podobná slavné povídce "Babylónská knihovna" od Jorgeho Luise Borgese, ve které jsou náhodně vygenerovány všechny myslitelné knihy světa. Nabízí se i analogie s pověstnou opicí, jež po dostatečně dlouhém tlučení do psacího stroje napíše Hamleta (tzv. "infinite monkey theorem").

který už dříve použil Turing nebo Searle. Připadá mi, že většina Blockových připomínek není příliš silná. Opět je všechny stručně představím, a to včetně příslušných odpovědí autora. Detailněji se budu později zabývat jen šestou, která je podle mého názoru dostatečným důvodem pro zamítnutí Blockova argumentu. Hlavní směr mého útoku bude totiž veden proti samotným myšlenkovým experimentům, jež byly v posledním desetiletí podrobeny intenzivní kritice, a to především díky hnutí experimentální filosofie.

### **6.1 První námitka**

*Argument Blockheada je příliš silný, protože každý inteligentní stroj vykazuje pouze inteligenci svých programátorů.*

Není pravda, že inteligence každé entity závisí na inteligenci jejích tvůrců. Například my lidé disponujeme vlastní inteligencí, i když jsme byli vytvořeni přičiněním našich rodičů. Block se vyjadřuje stylem, který je velice podobný někdejšímu Turingovu tvrzení o nutnosti vyloučení přirozeného, tedy pohlavního způsobu produkce inteligentních entit. Domnívá se navíc, že problém Turingovy definice inteligence spočívá právě v neschopnosti odlišit mezi inteligencí stroje a inteligencí jeho programátorů.

### **6.2 Druhá námitka**

*Stroj by nebyl schopen odpovídat na otázky týkající se aktuálních událostí, které jeho tvůrci během programování neznali.*

Systém může být inteligentní, i když nemá znalosti o aktuálních událostech. Jedná se o málo podstatnou výtku, protože programátoři mohou snadno simulovat osobu, která z určitého důvodu žila v izolaci nebo nejevila o současné dění zájem. Navíc je možné program periodicky aktualizovat.

### 6.3 Třetí námitka

*Blockova internalistická definice dává termínu "intelligence" zcela nový význam, protože normálně považujeme za inteligentní externí projevy.*

Externí projevy jsou opravdu při přisuzování inteligence primární, ale za určitých podmínek se řídíme i znalostmi o interní struktuře inteligence. Příkladem může být situace, kdy je zdánlivě inteligentní chování založeno na podvodu. Block uvádí příklad s šachistou, který hraje dvě simultánní partie s předními velmistry. Jeho brilantní hra je ale založena na podvodu, protože ve skutečnosti v první hře používá tahy druhého oponenta a obráceně. Stejným způsobem podvádí i Blockhead, jehož všechny reakce jsou pouhou ozvěnou dřívějšího jednání programátorů.

### 6.4 Čtvrtá námitka

*Pokud by se prokázalo, že lidé zpracovávají informace stejným způsobem jako Blockhead, znamenalo by to, že lidé také nejsou inteligentní.*

Opět se podle Blocka jedná o slabou námitku, protože lidské zpracování informací se od popsaného postupu diametrálně liší. Dalo by se dokonce říct, že lidé jsou inteligentní z definice a ostatní entity jsou s jejich myšlením poměřovány.

### 6.5 Pátá námitka

*Block nemá žádný důvod domnívat se, že lidé zpracovávají informace jinak než Blockhead.*

Už samotný fakt, že rozumíme předloženému článku, nás přesvědčuje o opaku. Lidé nejsou pouhými hledači konverzačních řetězců ("string-searchers"), ale bytostmi s plně rozvinutými kognitivními procesy. Jejich fungování zatím zcela nerozumíme, ale jsou zajisté víc než jen výběrem z předpřipravených možností.

## 6.6 Šestá námitka

*Blockhead je nemožný, protože díky kombinatorické explozi by hodinová konverzace odhadem představovala  $10^{1500}$  možných řetězců, což je více než počet částic ve vesmíru.*

Block operuje pouze s logickou, nikoli nomologickou možností. Není pro něj podstatná *reálná existence* popisovaného zařízení, ale jen jeho *metafyzická možnost*. Kromě toho Block tvrdí, že ani samotný T-test není metodou ověření ryze empirické hypotézy. V nejlepším případě jej považuje za hypotézu kvazi-empirickou, která pracuje s inteligencí jako neempirickým principem empirické psychologie. Spekuluje dokonce o skutečné realizaci Blockheada, a to v nějaké hypotetické oblasti vesmíru, kde jsou částice dělitelné donekonečna. Později se pokusím ukázat, že právě tento přístup operující s vysoce spekulativními premisami je Achillovou patou všech myšlenkových experimentů, které podle mého názoru mohou mít nanejvýše estetickou hodnou - podobně jako třeba logické a sémantické paradoxy.

## 6.7 Sedmá námitka

*Blockova verze T-testu je časově omezená, a proto neodpovídá původnímu Turingovu zadání.*

Námitka vyžadující neomezenost T-testu je nekorektní, protože ani lidé nemají možnost uspět v časově neomezené verzi. Jsou limitováni mnoha různými faktory, z nichž zcela definitivní je omezení vlastní smrtelností.

## 6.8 Osmá námitka

*Blockhead není argumentem proti umělé inteligenci, protože podle klasické definice Alana Newella a Herberta Simona je "úkolem inteligence předejít neustálé hrozbě exponenciální exploze vyhledávání" (Newell & Simon 1976, p.123).*

Block se škodolibým uspokojením tvrdí, že definice inteligence upravená o vyhýbání se exponenciální explozi je rezignací na behaviorismus. Zastánci tohoto přístupu totiž projevují zájem o vnitřní etiologii zkoumané entity, a tím se nevědomě hlásí k Blockem proklamovanému psychologismu. Navíc se nemusí jednat o poslední úpravu definice, kterou mohou příznivci silné umělé inteligence ohýbat podle toho, jak se jim to právě hodí.

První reakcí na Blockovu esej byl stručný článek, který v roce 1982 uveřejnil tehdy ještě velice mladý filosof Robert C. Richardson. Je s podivem, že autor původního textu na tuto dobře promyšlenou kritiku nikdy nezareagoval, i když ona sama o sobě je dostatečným důvodem pro vyvrácení Blockových předpokladů. Lze říct, že Blockova představa o fungování jazyka je značně naivní, protože jej považuje za jakýsi omezený rezervoár vět, které mohou být proneseny. Ve skutečnosti je ale jazyk mechanismem, kterým lze z omezeného množství slov vytvořit potenciálně neomezené množství vět. Tato vlastnost jazyka je umožněna použitím rekurze, tedy konceptu Chomského transformační gramatiky, díky němuž jsme schopni vytvářet nové jazykové výrazy opakovaným použitím nějakého pravidla. Například počet možných adjektiv před substantivem je teoreticky neomezen, stejně jako počet příslovcí měnících význam slovesa. Rekurze je důvodem, proč může člověk pronést zcela novou výpověď, která ještě nikdy předtím nebyla vyslovena, a adresát jí bude přesto rozumět (Chomsky 2009, p.62; Richardson 1982, p.423). Není vůbec jasné, jak by asi Blockhead reagoval na Werichovu kuriózní, ale přesto pochopitelnou větu z pohádky o žlutém mužátku: "Na ostrově Hokaidó kuká v lese kokaidó." Jediný způsob, jak vytvořit stroj, jenž by dokázal dešifrovat i takto kryptické věty, je vybavit jej elementární lingvistickou schopností rekurze, na které je jakýkoli jazyk založen. Blockhead takovým strojem rozhodně není.

Pokud bych Richardsonovu kritiku rozšířil, pak celá konstrukce komunikujícího stroje jen na principu brutální síly je značně pochybným projektem. Z praktického hlediska lze říct, že mnohem snazší by bylo sestavit stroj, který by v T-testu uspěl a přitom "nepodváděl" použitím permutačních metod. Brutální síla se nepoužívá ani u problémů s konečným, i když obrovským počtem situací. Představa, že by její pomocí mohla být hrána imitační hra

s potenciálně nekonečným množstvím situací, je absurdní. Například v šachu je dnes možné pomocí brutální síly odehrát pár tahů zahájení a některé koncovky, které používají nanejvýš šest až sedm kamenů. Celá střední hra je záležitostí krátkodobého taktického manévrování či dlouhodobého strategického plánování a zatím zcela uniká možnostem mechanického vybírání ze seznamu tahů. Na Blockovu obranu lze samozřejmě uvést, že mu jde pouze o logickou možnost (viz kapitola 6.6 této práce), která odpovídá chápání T-testu jako logicky postačující podmínky pro připsání inteligence. Problém vidím ovšem především v tom, že Block považuje Blockheada za nemyslíciho jen na základě intuice, na níž stojí vlastně všechny myšlenkové experimenty. Blockheadův duchovní otec nikdy přesně nevysvětlí, jak by mělo popsané zařízení ve skutečnosti fungovat a spoléhá se na zcela minimalistický popis jeho činnosti. Pokud by byl podobný stroj někdy zkonstruován, pak by zajisté nebyl jednoduchým "vesmírným jukeboxem", jak se nám snaží Block sugerovat. Komplexnost použitých technologií by mohla nestranného pozorovatele vést k neméně intuitivní domněnce, že Blockhead myslí. Podle Stuarta Shiebera je Block typickým představitelem popíračů T-testu, kteří se rekrutují z řad myslitelů považujících - obrazně řečeno - symptomy za nedostatečné doklady onemocnění. Zastánci tohoto přístupu jsou ochotni hodinu poslouchat konverzaci v italštině, ale zároveň upřít jejímu aktivnímu účastníkovi schopnost mluvit italsky; podobně může údajně jedinec, jenž získal ve standardizovaném testu maximální počet bodů, zcela postrádat inteligenci. Shieber s tímto názorem hluboce nesouhlasí a tvrdí, že komunikaci v cizím jazyce nebo inteligentní chování nelze předstírat, protože symptomy jsou v takovém případě totožné s příčinami (Shieber 2007, pp.688-689).

Arthur C. Clarke se v povídce "... A ozve se Frankenstein" zabýval myšlenkou, že jakýkoli dostatečně složitý systém propojených prepínačů by při dosažení kritického stavu mohl nabýt vědomí. Jedná se samozřejmě o pouhou vědecko-fantastickou fikci, ale Clarkova intuice, že by celosvětová telefonní síť mohla myslet, je zhruba stejně přesvědčivá jako Blockova intuice, že by Blockhead myslet nemohl. Block připomíná nešťastné představitele rané aviatiky, kteří ještě na přelomu 19. a 20. století autoritativně prohlašovali let strojů těžších než vzduch za apriorně nemožný a svá tvrzení dokládali propracovanými matematickými výpočty. Intuice jako empiricky nepodložená tvrzení mohou snad sloužit pro prvotní orientaci ve světě, ale vědecké bádání by od nich mělo být oproštěno. Nezdůvodněné intuice jsou legitimní součástí "naivních" teorií, které používáme při

předvědeckém poznání. Je proto možné mluvit o naivní matematice, fyzice, morálce nebo teorii mysli, přibližně ve smyslu zdravého rozumu osvícenských empiriků. Když ovšem dojde ke konfliktu mezi intuicí a evidencí, má vědecká metoda samozřejmě navrch. Tragickým aspektem myšlenkových experimentů ve filosofii je principální nemožnost jejich empirické falzifikace. Filosofie se v tomto ohledu podstatně liší od teoretické fyziky, která myšlenkové experimenty také s oblibou používá. Na rozdíl od ní ale nemá filosofie experimentální větev, která by teoretické poznatky prověřovala laboratorními metodami.<sup>82</sup> Tak například Einsteinova speciální teorie relativity byla zpočátku založena jen na několika dosti divokých příkladech o dvojčatech stárnoucích jinou rychlostí. I tyto velmi proti-intuitivní myšlenky byly ale nakonec experimentálně potvrzeny, třeba pomocí přesných atomových hodin na palubě dvou dopravních letadel letících po a proti směru zemské rotace. Pokud by fyzikové uvažovali stejně jako filosofové, pak by si museli položit otázku, proč vůbec stavět urychlovače částic, když si je mohou pouze představit a ušetřit tak miliardy eur?

Na Blockův stroj se vztahují všechny námitky týkající se realizovatelnosti myšlenkových experimentů, podobně jako u Putnamovy Země-dvojčete, Searlova čínského pokoje nebo Chalmersových metafyzických zombií. Někteří komentátoři přesto vyzdvihují, že se Block snažil držet představivost na uzdě a podařilo se mu vyhnout tvrzením, která jsou zcela neodůvodněnými články víry.<sup>83</sup> Značnou nevoli vůči science-fiction myšlenkovým experimentům projevoval už Quine a někteří jeho pokračovatelé. Tak například Donald Davidson v reakci na Putnamovy příklady týkající se externalismu píše:

"[M]ám obecnou nedůvěru k myšlenkovým experimentům, jejichž autoři předstírají, že vědí, co bychom říkali v situacích, které ve skutečnosti nikdy nenastanou." (Davidson 2004, p.226; překlad Jan Kolář)

---

<sup>82</sup> Tento názor zastává přední experimentální filosof Jonathan Weinberg a jeho spolupracovníci: "The theories of the physicist, or the predictive models of the meteorologist, or even just the clear win/loss/draw signal at the end of the game for a chess player - these all provide the requisite sort of external check for which philosophers have drastically less that is comparable for their intuitive judgments about thought experiments." (Weinberg et al. 2010, p.349) Stejně odmítavý je i Edouard Machery, který pečlivě analyzuje použití myšlenkových experimentů jako důkazního břemene ve filosofii (Machery 2011).

<sup>83</sup> Na rozdíl od Searla: "This argument is based on an article of faith that is too woolly to argue against, namely, that no formal system that merely manipulates symbols could bear intelligence." (Shieber 2007, pp.689-690)

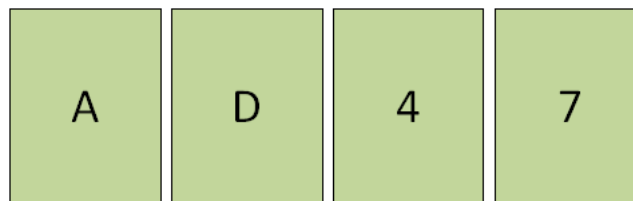
Skutečně systematickou kritiku provedli až mladší badatelé hlásící se k hnutí experimentální filosofie, nazývané v anglofonním prostředí zkráceně x-phi. Představitelé tohoto progresivního a módního směru vycházejí z prací psychologů Jonathana Haidta a Joshuy Greena, z nichž především ten druhý se proslavil snahou řešit tradiční filosofické problémy experimentálními metodami. Jako modelový příklad zvolil problém drezíny ("trolley problem"), který v 60. letech vytvořila morální filosofka Philippa Footová jako argument proti etickému utilitarismu. Greene na rozdíl od jiných nespoléhal při posuzování příkladu jen na deduktivní metody, ale vzal si na pomoc funkční magnetickou rezonanci (fMRI), již použil pro snímkování mozku u probandů řešících Footové dilema (Greene et al. 2001). Dospěl k překvapivému závěru, podle něhož jsou lidské bytosti vybaveny hned dvěma neurálními centry, která spolu kooperují a zároveň si konkurují během morálního rozhodování. Jedno centrum je racionální a používá konsekvenční etiku, zatímco druhé je emocionální a řídí se deontologií. Greene tak přesvědčivě ukázal, že staletí trvající spor utilitarismu a jeho alternativ si vlastně kladl nesprávnou otázku, na niž nebylo možné odpovědět bez metodologie přírodní vědy.

Experimentální filosofové následující Greenův průkopnický příklad začali zpochybňovat tradiční metody filosofického bádání, mezi něž je možné zařadit apriorní argumenty logiky, dialektiku, konceptuální analýzu, intuitivní řešení a v neposlední řadě právě myšlenkové experimenty. Myslitelé spojovaní s x-phi začali pro tyto postupy používat souhrnné a poněkud posměšné označení "armchair philosophy", a dokonce si jako logo hnutí vybrali obrázek ušáku v plamenech. Jejich programem je vyjít z útulně vytopených kabinetů do terénu či laboratoře a usilovat o experimentální podporu testovaných hypotéz. Poprvé na sebe výrazně upozornili sborníkem *Experimental Philosophy*, který obsahoval i manifest hnutí sepsaný Joshuou Knobem a Shaunem Nicholsem (2008). Autoři v něm vznášejí radikální požadavek na návrat před vznik analytické filosofie, kdy byla disciplína více praktická a empirická. Zřejmě hlavní metodou jsou dotazníková šetření mezi filosofickou i nefilosofickou veřejností, která mají poukázat na variabilitu intuitivních řešení, jež jsou filozofy považována za univerzální. Cílem této práce není posuzovat úspěšnost x-phi. Já sám mám podezření, že tento přístup může sloužit ke znejišťování statusu intuicí, ale jinak nevidím mnoho možností, jak pomocí sociologických šetření prohlubovat filosofické poznání. Negativní aspekt x-phi může přijít vhod právě při kritice myšlenkových experimentů, které



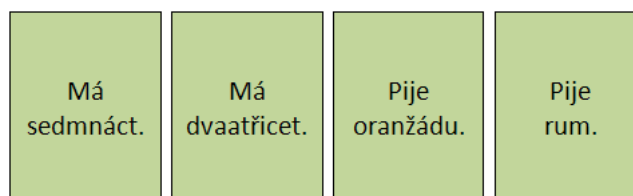
jsou většinou příliš komplikované a abstraktní na to, abychom je byli schopní analyzovat jen pomocí apriorní intuice. Jako ilustraci uvedu velmi jednoduchý příklad, na který upozornil další představitel x-phi Eric Schwitzgebel. Na konci 60. let uveřejnil americký psycholog Peter Cathcart Wason výsledky experimentu, kterému se dnes na autorovu počest běžně říká Wasonův výběrový test (Wason 1968). Jedná se o logickou úlohu, ve které má respondent prostřednictvím otáčení karet zjistit platnost pravidla. Existuje velké množství formulací testu, ale zadání může znít například takto:

Na stole leží čtyři karty, z nichž každá má na jedné straně písmeno (A nebo D) a na druhé číslo (4 nebo 7). Karty jsou položeny tak, že na dvou je vidět písmeno a na jiných dvou číslo. Minimálně kolik karet a které z nich je třeba obrátit, aby bylo potvrzeno pravidlo, že pokud je na líci samohláska, pak je na rubu sudé číslo?



Ukázalo se, že úkol je pro naprostou většinu lidí takřka neřešitelný, protože správně dokázalo odpovědět jen necelých 10% účastníků testu. Velká část z nich pochopila, že je nutné obrátit dvě karty, ale nedokázala přesně určit které. Wason z toho odvodil, že lidé jsou překvapivě neúspěšnější při ověřování hypotéz založených na implikaci, především pokud je zahrnuto použití *modus tollens*. Nejedná se o žádný převratný poznatek, protože naše kognitivní schopnosti jsou prostě omezené, a díky tomu často selháváme při řešení abstraktních problémů. Zajímavé ovšem je, že pokud příklad přesadíme do kontextu sociálních vztahů, pak mají respondenti podstatně vyšší šanci na úspěch. Zamysleme se nad následující variantou téhož příkladu:

V hospodě sedí čtyři zákazníci, některým je sedmnáct a zbývajícím dvaatřicet. Někteří popíjí nealkoholickou oranžádu, zatímco ostatní vysokoprocentní rum. Úkolem je pomocí co nejmenšího množství dotazů zjistit, zdali hostinský nenalévá mladistvým.



Odlišnost kontextu fatálním způsobem ovlivnila úspěšnost respondentů, i když řešení je v obou případech identické. Je třeba obrátit kartu s hledaným znakem a pak protipříklad: v prvním znění se jedná o karty A a 7, ve druhém jsou to karty "Má sedmnáct" a "Pije rum". Všimněme si, že druhá varianta může být velmi snadno přeformulována do abstraktní podoby (17 - 32 - O - R), přičemž úspěšnost odpovídajících by se opět markantně snížila.

Evoluční psycholožka Leda Cosmidesová na základě Wasonova výběrového testu argumentovala ve prospěch modulární teorie mysli, podle níž je lidský mozek složen z velkého množství specializovaných modulů, které fungují do značné míry nezávisle a jejichž činnost není dokonale propojena (Cosmides 1989 a další). Modul sociálního myšlení je podle Cosmidesové efektivnější než modul pro řešení abstraktních problémů, protože byl evolučně důležitější. Používáme jej totiž vždy, když se snažíme pochopit složité předivo mezilidských vztahů uvnitř společnosti a je zvláště nezbytný při odhalování podvodníků. Příklad s hostinským porušujícím sociální normu je nám proto mnohem bližší než úloha o formálních vztazích mezi čísly a písmeny. Schwitzgebel využil jeden z nejvíce zkoumaných psychologických experimentů k vlastním, filosoficky motivovaným cílům. Poukázal na skutečnost, že lidská neschopnost zabývat se odtažitými příklady bez vazby na každodenní zkušenost nás velmi handicapuje ve filosofickém bádání. Právě díky tomu nejsme schopni shodnout se na řešení komplikovaných otázek, u nichž se nemůžeme dovolávat empirické evidence. Například Kantova transcendentální filosofie je založena na obrovském množství zcela abstraktních konceptů, při jejichž manipulaci se může naše defektní myšlení dopouštět nepřeborného množství obtížně zjistitelných chyb. Ještě problematičtější je snaha dokazovat cokoli pomocí myšlenkových experimentů, jejichž zpracování je zřejmě do značné míry ovlivněno dispozicemi konkrétního řešitele. Schwitzgebel ukončuje svou úvahu dosti pesimisticky:

Většina našich filosofických ambicí je *daleko* za hranicí kognitivních schopností běžného člověka. Naše filosofické názory jsou proto do značné míry ovlivněny sociologickými a psychologickými faktory, které mají jen málo co do činění se skutečným obsahem argumentů pro a proti. To částečně vysvětluje, proč jsme v průběhu staletí byli svědky jen nepatrného filosofického pokroku. Pokud by po nás nastoupil (biologický nebo robotický) druh s většími kognitivními schopnostmi, druh, jenž by považoval Wasonův výběrový test s abstraktním kombinováním kondicionálů a negací za intuitivně jednoduchý, pak by se vysmíval našemu potýkání se s paradoxem Šípkové Růženky, Kantovou transcendentální dedukcí, paradoxy autoreference atd. Vysmíval by se nám

stejným způsobem, jako my informovaní a připravení máme (neoprávněně) nutkání posmívat se studentům nižších ročníků, kteří ve Wasonově výběrovém testu selhali. (Schwitzgebel 2011b, bez paginace)

Experimentální filosofové se hlouběji zabývají i kulturní podmíněností filosofických intuicí, přičemž vzorovým příkladem mohou být studie "Normativity and Epistemic Intuitions" (Weinberg et al. 2008) a "Gender and Philosophical Intuition" (Buckwalter & Stich 2010). Autoři prvního článku zkoumali interpretaci různých variant Gettierova problému mezi obyvateli Severní Ameriky a Indického subkontinentu a dospěli k závěru, že intuice obou skupin o řešení problému se dramaticky lišily. Ještě zajímavější jsou snad poznatky druhé studie, ve které Buckwalter se Stichem mapovali rozdílné chápání významných myšlenkových experimentů mezi příslušníky mužského a ženského pohlaví. Konkrétně se jednalo o Gettierův problém, kompatibilismus v otázce determinismu a morální odpovědnosti, Thomsonové příklad s houslistou, problém drezíny, posuzování intencionality v morálním rozhodování, Knobův efekt, Putnamovy mozky v kádi a příklad se Zemí-dvojčetem, Searlův čínský pokoj, Karneadův argument prkna a hru na diktátora. Genderové rozdíly byly ještě hlubší než při mezikulturním srovnání, i když autoři nedokázali formulovat uspokojivou hypotézu, čím jsou vlastně způsobeny. Přesto si dovolili z šetření učinit závěry týkající se disproporčního postavení žen v akademické filosofii.<sup>84</sup> Podle jejich názoru jsou ženy při studiu filosofie znevýhodněny, protože převážně mužské prostředí nespravedlivě upřednostňuje intuice, které jsou charakteristické pro muže. To pak u žen vyvolává ztrátu zájmu o disciplínu, jejíž ústřední problémy nahlížejí diametrálně odlišně. Tématu této práce nejvíce odpovídá Buckwalterova a Stichova analýza Searlova myšlenkového experimentu, v níž položili skupině studentů otázku, zdali osoba zavřená v čínském pokoji rozumí čínštině. Na sedmistupňové stupnici, kde jednička znamenala naprostý nesouhlas a sedmička naprostý souhlas, odpovídaly ženy na schopnost rozumění víc skepticky: průměrná známka byla 4.13 u mužů a 3.25 u žen.<sup>85</sup> Zdá se proto jasné, že Searlův - a zřejmě i Blockův - zdánlivě logický argument je založen na mimologických faktorech, které nemají s obsahem důkazu nic společného.

---

<sup>84</sup> Podle výpočtu Kathryn Norlockové působí na amerických katedrách filosofie 2158 žen na plný úvazek, což činí podíl pouhých 16.6% ze všech pracovních pozic; jiné údaje jsou poněkud vyšší, ale nepřesahují 20%.

<sup>85</sup> Šetření se zúčastnilo 127 respondentů, z toho 54 mužů a 73 žen (Buckwalter & Stich 2010, pp.21-23; draft 3.0).

## 7. Subkognice a vědomí: French 1990 a Michie 1993

V devadesátých letech minulého století se diskuse týkající se T-testu nejen že nezklidnila, ale naopak prohloubila a rozšířila i mimo zavedená filosofická periodika. V roce 1989 vznikl specializovaný *Journal of Experimental & Theoretical Artificial Intelligence*, který interdisciplinárně zkombinoval filosofický a inženýrský přístup k problematice, a o dva roky později spatřilo světa první číslo časopisu *Minds and Machines* specializovaného výhradně na filosofii umělé inteligence. Právě na stránkách těchto dvou časopisů se odehrála velká část sporů ohledně T-testu, i když se některé důležité texty objevily i v jiných publikacích. Z nepřeberného množství teoretických prací, o nichž zde ještě nepadla zmínka, jsem se rozhodl věnovat pozornost dvěma článkům, které k T-testu přistupují originálním, dosud nevídaným způsobem. Jedná se o "Subcognition and the Limits of the Turing Test" Roberta M. Frenche (1990) a "Turing's Test and Conscious Thought" Donalda Michieho (1993). Díky Thomasi Nagelovi a dalším myslitelům se do centra pozornosti analytických filosofů myslí dostal fenomén vědomí, intencionality a jiných mentálních aktivit. Ty údajně nejsou vysvětlitelné pomocí redukcionistických přístupů, ať už se jedná o teorii identity, eliminativní materialismus nebo nejrůznější odrůdy funkcionalismu. Tento způsob pohlížení na mentální stavy, intencionální obsahy a propoziční postoje se prosadil i v kritice T-testu. V kontextu dřívějších polemik se nejedná o zas tak velké novum, protože názor předjímal Turing už ve čtvrté námitce ke své eseji z roku 1950 (viz kapitola 3.4 této práce).

### 7.1 Subkognice

French se explicitně nezabývá přímo vědomím, ale jeho subkognitivními složkami, které nám umožňují myslet jako lidské bytosti. Na úvod vysvětluje svůj vztah k Turingovu paradigmatickému článku. Podle něj Turing činí dvě odlišná, i když vzájemně provázaná tvrzení. Tím prvním je *filosofické* přesvědčení, podle něhož je stroj, který uspěje v T-testu, nutně inteligentní. Druhé je *pragmatické* a spočívá v predikci, že v blízké budoucnosti takový stroj postavíme. Většina dosavadní kritiky T-testu se týkala filosofického tvrzení, které ale French považuje za naprosto správné: T-test je postačující podmínkou inteligence a navíc

může být základem pro její operacionální definici. Domnívá se, že "cokoli *jedná* postačujícím způsobem inteligentně, *je* inteligentní" (French 1990, p.53).<sup>86</sup> Naopak druhé, pragmatické tvrzení je podle Frenche neudržitelné, neboť T-test je nastaven tak, aby v něm uspěli nám podobní tvorové s přibližně stejnou anatomii a identickými kognitivními schopnostmi. Centrální myšlenkou eseje je, že T-test garantuje pouze lidskou, kulturně podmíněnou inteligenci, nikoli inteligenci společnou všem myslícím entitám. Pokud by testování podstoupil nehumanoidní mimozemšťan s odlišnými smyslovými orgány a jiným civilizačním zázemím, pak by mu jeho případná inteligence nebyla nic platná a zcela by propadl. French je tak ve velmi unikátní pozici, protože ostatní myslitelé se domnívají, že T-test je příliš *lehký* - dokáže jím totiž projít i čínský pokoj bez znalosti sémantiky nebo jukebox mechanicky vybírající ze škály naprogramovaných odpovědí. Na rozdíl od svých předchůdců je French přesvědčen, že T-test je naopak příliš *těžký* a nedokáže jej úspěšně složit žádný stroj, jaký bychom mohli v dohledné době zkonstruovat.<sup>87</sup> Kámen úrazu spočívá v subkognitivních informacích, které jsou pevně fixovány na zkušenost tělesnosti a vytvářejí substrát podvědomých, obtížně sdělitelných znalostí o fungování jazyka a okolního světa.<sup>88</sup> French píše:

Turingův test ve skutečnosti zkoumá síť pojmových (a sub-pojmových) asociací dvou kandidátů. Tyto síť jsou produktem celoživotních interakcí se světem, které nutně zahrnují lidské smyslové orgány, jejich umístění na těle, citlivost na různé stimuly atd. (French 1990, p.62)

Například živočich takřka identický s člověkem, který má ale oči umístěny na kolenou, by měl ve srovnání s lidmi naprosto odlišné zkušenosti. Lokace očí přitom nemá pranic společného s inteligencí, a přesto tvora s odlišnou "sítí pojmových asociací" handicapuje v T-testu. Zdá se

---

<sup>86</sup> V originále: "whatever *acts* sufficiently intelligent *is* intelligent". Je opět nanejvýš žádoucí upozornit čtenáře, že Turingovi o definici inteligence vůbec nešlo; viz úvod kapitoly 3 a poznámka pod čarou 76. Ve Frenchově podání *definiens* neodpovídá *definiendu*, protože ani inteligentní entita se nemusí chovat vždy inteligentně - třeba pokud právě spí, je intoxikována nebo neovládá jazyk, ve kterém je konverzace vedena.

<sup>87</sup> French je mírně optimistický, zdali to dokážeme ve velmi daleké budoucnosti. Jednou snad vyrobíme zařízení, jehož chování bude neodlišitelné od toho našeho. Takový stroj bude mít schopnost "padat z bicyklů, píchat se o růžové trny, cítit zápach splašků a chutnat jahody", ale rozhodně nebude ani vzdáleně připomínat dnešní počítače.

<sup>88</sup> Pro označení této nejzákladnější kognitivní složky French používá termíny "subcognition", "low-level cognitive structure", "human cognitive substrate" nebo "unconscious cognitive substrate". Podle definice je to "the subconscious associative network in human minds that consists of highly overlapping activatable representations of experience" (French 1990, pp.56-57).

navíc, že neexistuje žádné kritérium, jak odlišit subkognitivní a běžné kognitivní otázky. Každá dostatečně rozsáhlá sada dotazů jich proto bude několik obsahovat, a to spíš více než méně. Dokonce i některé otázky navrhované samotným Turingem jsou subkognitivní, jako třeba všechny týkající se poetických kvalit jazyka. Z těchto důvodů je T-test prakticky nepoužitelný, a French proto navrhuje nové metody připisování obecné inteligence, mezi něž patří analýza schopností kategorizace, učení se novým pojmům, přizpůsobení starých pojmů novým kontextům atd. O všech těchto metodách je ale dosti skoupý na slovo, protože hlavní cíl jeho práce spočívá v negativním zamítnutí T-testu, nikoli pozitivním hledání náhrady za něj.

Pro lepší pochopení svých východisek používá French ilustrativní příklad, který není ani třeba považovat za myšlenkový experiment, protože neposkytuje evidenci pro následující argumentaci. Sám autor jej označil za *podobenství o severských raccích* ("Nordic seagulls parable") a zestručnělá podoba by mohla znít třeba takto:

Na jednom severním ostrově žijí jako jediní létaví ptáci racci. Dva filosofové se snaží zjistit, co vlastně znamená pojem "létat". Vytvoří racčí test ("Seagull Test"), který má být postačující podmínkou pro připsání schopnosti letu jakémukoli stroji. Test spočívá v tom, že lidský soudce sleduje dvě obrazovky třírozměrného radaru, z nichž jedna zobrazuje racka a druhá létající stroj. Úkolem soudce je určit, na které obrazovce je racek. Stroj uspěje v racčím testu právě tehdy, když soudce není schopen odlišit stroj a racka.

French vcelku konzistentně předpokládá, že mnoho létajících entit racčím testem nikdy neprojde, například helikoptéry, trysková letadla, netopýři, brouci, albatrosi nebo kolibříci. S velkou pravděpodobností v testu uspějí jen a pouze racci, takže je správnější považovat jej nikoli za test generalizované schopnosti létat, ale spíš specifické schopnosti létat jako severský racek.<sup>89</sup> Analogie s T-testem je zcela zřejmá:

Je možné, aby určitá entita byla extrémně inteligentní, ale - pokud neodpovídá na tazatelovy otázky veskrze lidsky - neprošla by testem. Domnívám se, že jediným způsobem, jak zodpovídat otázky zcela jako člověk, je disponovat lidskou zkušeností světa. Nemáme tedy test, který zjišťuje obecnou inteligenci,

---

<sup>89</sup> V originále: "What we have is thus not a test for flight at all, but rather a test for flight as practised by a Nordic seagull." (French 1990, p.56)

ale spíše test pro inteligenci používanou lidskými bytostmi. (French 1990, p.56)<sup>90</sup>

Není správné považovat inteligenci za fenomén, který je buď přítomný, nebo nepřítomný. French upozorňuje, že inteligence se vyskytuje v různých stupních; pavouci jí mají méně než vrabci, psi méně než opice, děti méně než dospělí. K tomu lze navíc dodat, že současní psychologové se nedokážou shodnout ani na tom, zdali vůbec existuje nějaká obecná inteligence (*g*), nebo se jedná o jev způsobený konglomerátem kooperujících a specializovaných modulů.

V hlavní části eseje French uvádí příklady dotazů, jejichž zodpovězení by bylo nad možností jakéhokoli počítače. Nakonec identifikuje jejich dva hlavní druhy, i když uznává existenci celé řady dalších. První sada subkognitivních otázek spočívá ve využití psychologického mechanismu, který je znám jako asociativní priming nebo sémantická falicitace.<sup>91</sup> Díky němu si lidé mezi pojmy vytvářejí asociace různé síly. Tak například "chléb" má silnou asociaci se slovem "máslo", zatímco "pes" má naopak s "máslem" asociaci velmi slabou. Právě na tomto principu jsou založeny nejrůznější testy slovního sdružování, které jsou součástí projektivních technik v rámci kvalitativního psychologického výzkumu. Tazatelé v T- testu by stačilo znát průměrné lidské hodnoty asociací mezi termíny a potom je porovnat s odpověďmi obou účastníků. Stroj by v takovém testu nemohl uspět, protože jeho asociace by byly odlišné od běžných lidských reakcí. Jiným příbuzným případem z této sady je určování následnosti jistých úkonů, jako třeba při pečení koláče: otevřít sáček s moukou, rozbít vajíčka, smíchat mouku s vejci, dát těsto do trouby, vytáhnout hotový koláč. Podle Frenche počítač nedokáže nalézt správné pořadí aktivit, které nikdy nezažil.

---

<sup>90</sup> Poslední věta v originále: "What we have is thus not a test for intelligence at all, but rather a test for intelligence as practised by a human being." (French 1990, p.56) Podobně se autor vyjadřuje v samotném závěru eseje: "The Turing Test provides a sufficient condition for human intelligence but does not address the more important issue of intelligence in general." (French 1990, p.64)

<sup>91</sup> Ani jeden z pojmů není v české literatuře příliš rozšířen, navíc se jedná o dosti otročský převod původního "associative priming" a "semantic facilitation". Význam termínů lze pochopit z kontextu, a tak jen předesílám, že v psychologii "priming" (nebo "vybuzení") označuje "proces, ve kterém expozice podnětu ovlivňuje interpretaci nové informace" (Výrost & Slaměník 2008, p.193), případně "proces, jehož prostřednictvím specifický počáteční podnět aktivuje mentální dráhu, čímž dochází ke zvýšení schopnosti zpracovávat následné podněty vztahující se k podnětu primingu určitým způsobem" (Sternberg 2002, p.606).

Do druhé sady subkognitivních otázek autor řadí hned několik her, ve kterých účastníci hodnotí pojmy podle sémantických a pragmatických požadavků. Tak například ve hře hodnocení novotvarů ("Neologism Rating Game") jsou otázky založeny na lidské dovednosti dávat význam slovům, která nikdy předtím neslyšeli.<sup>92</sup> Účastník testu hodnotí vhodnost slova pro označení nějakého předmětu na škále od 0 do 10, přičemž 0 znamená "naprosto nepřijatelné" a 10 "naprosto přijatelné". Jako vzorek mohou sloužit následující neologismy, jež pro potřeby hry vymyslel sám French, i když jejich úplné pochopení je podmíněno kvalitní znalostí anglického jazyka. Slovo "Flugblogs" by mohlo být (a) názvem snídaňových obilnin, (b) jménem počítačové společnosti, (c) označením pro vzduchem plněné balóny určené k chůzi po vodě; slovo "Flugly" by mohlo být (a) jménem dětského medvídky, (b) příjmením účetního v komediálním filmu, (c) příjmením půvabné hollywoodské herečky. V průběhu T-testu by tazatel opět srovnával běžné lidské odpovědi s těmi, které získal od účastníků, a tak snadno určil, kdo je stroj a kdo člověk. Podle Frenche slovo "Flugblogs" foneticky asociuje nepříjemná slova, jako například "flub", "thug", "ugly", "ugh!", "blob" nebo "bog", a proto se nehodí pro komerční označení cereálií. Naopak "Flugly" zní roztomile, a je tak vhodným jménem dětské hračky, protože pozitivní asociace slov "snugly" a "cuddly" převažují nad negativní asociací "ugly".

Douglas Hofstadter navrhl jinou variantu hry ("Category Rating Game"), ve které účastník nehodnotí novotvary, ale vyjadřuje vhodnost předmětu k nějakému účelu, opět na škále od 0 do 10. French uvádí tyto příklady: "spadané listí jako úkryt", "rádio jako hudební nástroj", "zmrzlinový pohár jako lék", "koncertní křídlo jako dopravní prostředek", "dámská kabelka jako zbraň", "plnicí pero jako zbraň", "pánské sako jako příkrývka" a "borovicové větve jako matrace". V porovnání s průměrnými lidskými odpověďmi by byl počítač opět naprosto neúspěšný, stejně jako v dalších alternativách - při posuzování estetičnosti verše ("Poetic Beauty Rating Game"), směšnosti žertu ("Joke Rating Game") nebo účinnosti reklamního sloganu ("Advertising Rating Game"). French konstatuje značný skepticismus k subkognitivním schopnostem počítačů:

---

<sup>92</sup> I tato vlastnost lidské jazykové výbavy je důvodem, proč by Blockhead v imitační hře nikdy neuspěl (viz kapitola 6 této práce).



Dobré reklamní slogany, stejně jako dobré vtipy a dobré verše, jsou považovány za dobré díky myriádám podvědomých vlivů a asociací, které člověk získal během svého dosavadního prožívání světa. (French 1990, pp.61-62)

Po stručném shrnutí obsahu Frenchovy eseje je třeba vypořádat se s kvalitou jeho argumentace. Na Frenchovi je podle mého zřetelně vidět, že filosofové až příliš často trpí nedostatkem představivosti a nedokážou úspěšně predikovat budoucí vědecko-technický pokrok. Mnoho z jeho intuicí bylo totiž postupem času vyvráceno a současné počítače pouhých dvacet let od původní předpovědi svedou věci, o kterých se autorovi ani nesnilo. O to podivnější je, že French své názory nepodrobil revizi, i když k tomu měl mnoho příležitostí (např. French 2000a; French 2000b). Dalo by se říct, že filosofové nejsou schopni vytvářet expertní závěry *a priori* (viz poslední část předchozí kapitoly) a nejsou příliš dobří ani ve formulaci zkušenostně ověřitelných tvrzení - škodolibě se chce dodat, že bez použití empirických metod to není ani dost dobře možné. Frenchovi se nepodařilo odhadnout nástup internetu, který slouží připojeným počítačům jako online databáze s obrovským množstvím dat, jež lze třídit pomocí správně nastavených algoritmů a používat v komputačních procesech. Z dnešního pohledu je zvláště patetický příklad s recepty, protože na webu jsou umístěny miliony návodů na upečení čehokoli a existují elegantní analytické nástroje, jak se v nich orientovat (např. Google Recipes). Tento pokrok byl umožněn rozvojem takzvaného vytěžování dat ("data mining"), jenž začal v 90. letech minulého století. Ve zkratce řečeno se jedná o analytickou metodu pro získávání netriviálních a potenciálně užitečných informací z velkých souborů dat.

Právě pomocí vytěžování dat argumentuje proti Frenchovým závěrům odborník na informační technologie Peter D. Turney, když poukazuje na slabiny jeho myšlenkového postupu.<sup>93</sup> Turney tvrdí, že i odtělesněný ("disembodied") počítač může na subkognitivní otázky odpovídat korektním způsobem, pokud k tomu použije statistické informace extrahované z velmi rozsáhlých kolekcí textových souborů. Podrobně ukazuje, že počítač dokáže pomocí jednoduchého učícího algoritmu posuzovat sémantickou podobnost dvojic slov a frází. Hlavním zdrojem dat byl v Turneyho případě vyhledávač AltaVista, který v té

---

<sup>93</sup> Turney (2001) je reakcí na French (2000a) a v důsledku i na French (1990). Za povšimnutí stojí, že Turney stejně jako French považuje T-test za příliš silné kritérium inteligence, ale přesto se kriticky vymezuje proti navrhovaným důvodům odmítnutí T-testu.

době indexoval asi 350 milionů webových stránek, dnes by se přirozeně nabízelo využití vyhledávače Google, jenž sbírá informace o mnohonásobně vyšší počtu dokumentů. Turneyho program PMI-IR<sup>94</sup> byl sice původně určen pro rozeznávání synonym, ale po nenáročném úpravě dokázal správně řešit Frenchovy otázky týkající se hodnocení novotvarů. Na konkrétním příkladu "Flugly" dosáhl Turney výsledků, které sám French považoval za lidské, protože počítač ohodnotil přijatelnost slova jako jména hollywoodské herečky známkou 1, jména účetního v komediálním filmu 2 a jména dětského medvídky 10.

Na stejná metodická pochybení narazíme ve hře hodnocení kategorií. Nejen, že Turneyho program dokázal úspěšně řešit i tyto problémy a správně například seřadil potenciální hudební nástroje od zcela nevhodných banánových slupek až po naprosto přijatelné skořápky kokosových ořechů.<sup>95</sup> Navíc je třeba vypořádat se s řadou dalších omylů a nepřesností. French si zřejmě vůbec není vědom, že účastníci T-testu nemusejí při hodnocení novotvarů a kategorií odpovídat správně nebo běžně. Mnohem relevantnější je, jak své rozhodnutí dokáže soutěžící vysvětlit. Může například argumentovat, že propisovací pero je velmi účinná zbraň, protože v prvním díle trilogie o agentu Bourneovi viděl, jak hlavní hrdina zabil protivníka právě tímto nástrojem. Nebo může pozitivně hodnotit anekdotu se skatologickým obsahem, neboť mu takový typ připadá nesmírně zábavný. Z těchto důvodů by při zodpovídání subkognitivních otázek neuspěli lidé s neobvyklými chutěmi, bez estetického cítění či smyslu pro humor. Posuzování kategorií je velice subjektivní záležitostí a vychází z dosti problematických projektivních technik kvalitativní psychologie, které byly od svého vzniku na konci 19. století vystaveny kritice kvůli metodologickým nedostatkům a nemožnosti ověření.

Nejaktuálnějším důkazem nepřesnosti Frenchových tvrzení a neoprávněnosti jeho závěrů je počítač Watson, kterého nedávno na veřejnosti představili programátoři

---

<sup>94</sup> Název programu je odvozen ze dvou hlavních metod, které pro svou činnost používá: "pointwise mutual information" (PMI, česky "bodová vzájemná informace") a "information retrieval" (IR; česky "vyhledávání informací").

<sup>95</sup> Příklad pochází z French (2000a) a program vhodnost ohodnotil takto: banánová slupka 1, rádio 4, kokosová skořápka 10. Podobně úspěšně se PMI-IR vypořádal s dalším úkolem, ve kterém měl účastník seřadit různé vůně a zápachy od nejméně příjemných po nejpřitažlivější: nemocniční chodba 1, mokrá ručník 5, oceán 7, čerstvě upečený chleba 10, právě posekaná tráva 10 (Turney 2001, pp.415-417).

z laboratoří IBM (Thompson 2010; Markoff 2011). Jejich ambicí bylo zkonstruovat stroj, který by uspěl v televizním pořadu *Jeopardy!*, což je vědomostní soutěž podobná našemu někdejšímu *Riskuj!*. Watson pracuje přesně tak, jak Turney před deseti lety doporučoval. Údaje do něj nejsou manuálně vkládány programátory, ale počítač si je sám algoritmicky vyhledává v databázích encyklopedií, slovníků, tezurů, románů, novin a jiných dokumentů. Komputační analýza přirozeného jazyka je zatím v plenkách, ale Watsonovy výkony musí člověka přesto naplňovat úžasem. Vedoucí projektu David Ferucci se snažil, aby byl program schopen reagovat i na velmi sofistikované dotazy, které obsahují složité narážky a slovní hříčky. Nejedná se tedy jen o triviální problémy ("hlavní město Ruska"), ale i o komplexní hádanky ("název této pokrývky hlavy je úplně elementární, milý soutěžící").<sup>96</sup> Někdy jsou Watsonovy odpovědi zcela zmatečné, protože nedostatečným způsobem interpretoval zadání, ale přesto dokázal při několika příležitostech porazit lidské soutěžící. První ostré kolo proběhlo na začátku roku 2011 a kromě Watsona se ho zúčastnili dva lidští soupeři z řad bývalých rekordmanů. Soutěž trvala tři dny a Watson v ní velmi výrazně vedl již od začátku. Úkoly byly počítači podávány v podobě textových inputů, protože není vybaven prostředky pro rozlišování lidského hlasu, ale i tak byla jeho schopnost pochopení přirozeného jazyka téměř bezchybná. V závěru Watson suverénně zvítězil se ziskem 77 147 dolarů oproti 24 000 a 21 600 svých lidských oponentů. Takový vhled do jemných nuancí západní popkultury vyžaduje obrovské množství subkognitivních znalostí a je snad oprávněné domnívat se, že Frenchovy připomínky k T-testu byly s definitivní platností experimentálně vyvráceny.

## 7.2 Vědomí

Subkognice je součástí mnohem širší skupiny mentálních jevů, kterým se souhrnně říká vědomí. Jak už bylo předesláno na začátku kapitoly, hlavním iniciátorem zájmu o tento fenomén byl mezi analytickými filosofi Thomas Nagel, jenž v první polovině 70. let publikoval slavnou esej se sugestivním názvem "What Is It Like to Be a Bat?" (T. Nagel 1974). V ní nekompromisně odsuzoval dobový redukcionismus ve filosofii mysli, do něhož řadil všechny druhy naturalismu a funkcionalismu. Neurální ani komputační vysvětlení totiž vůbec neumí

---

<sup>96</sup> Správnou reakcí je v tomto případě otázka: "Co je to lovecká čapka?"

postihnout fenomén vědomí, které je neredukovatelné na nižší explikační schéma a nelze je zachytit objektivně. Subjektivní zážitky unikají možnostem vědecké metodologie, což Nagel ilustruje příkladem s nám jen málo příbuznými savci z řádu letounů. Netopýři vnímají vnější svět především pomocí sonaru, který není ani trochu podobný žádnému z našich smyslů.<sup>97</sup> Tento fakt znamená, že nikdy zcela nedokážeme pochopit vnitřní život netopýra. Můžeme si nanejvýš představit, jaké by bylo *chovat se* jako netopýr, ale už ne, jaké je to *být* netopýrem. Naše představa bude kvůli absenci echolokace vždy velmi nepřesná a vzdálená privátním prožitkům jedinců tohoto unikátního druhu. Různé inteligentní formy se nemohou vzájemně pochopit díky odlišné struktuře myslí, svedou si jen připisovat určité základní typy mentálních stavů. Neurofyziologické studium netopýřích mozků nebo funkcionální analýza jejich vnějškových projevů nás podle Nagela nepřiblíží ani o píď k pochopení, co zažívá netopýr letící stmívající se krajinou za hmyzí potravou. Redukcionistické teorie proto nebudou pro zkoumání vědomí nikdy dostatečné a bude třeba je nahradit novou "objektivní fenomenologií", která by nezávisela na empatii nebo představivosti. Jak už to ale v projektech tohoto typu chodívá, Nagel není ochoten ani přibližně narýsovat základy této vysněné disciplíny. Je prokletím filosofů, že badatelské programy raději navrhují, než aby je mravenčí prací naplňovali.

Proti možnosti strojů vybavených vědomím po Turingovi poprvé argumentoval Michael Scriven v pozapomenuté eseji "The Mechanical Concept of Mind". Jeho myšlenkový postup je pro příznivce antiredukcionistických teorií mysli charakteristický, a to i díky ledabylé práci s pojmy a intuitivnímu zastávání obtížně obhajitelných pozic. Scriven představuje antropocentrismus ve své nejčistší podobě, protože nastavuje kritéria pro připsání vědomí tak, aby na ně dosáhli jen příslušníci lidského druhu. Počítače a roboti nemohou disponovat vědomím, i kdyby tomu nasvědčovaly všechny jejich vnější projevy.

---

<sup>97</sup> S tímto tvrzením vážně polemizuje Richard Dawkins, jenž se navigačním systémem netopýrů podrobně zabývá v druhé kapitole své knihy *Slepý hodinář*: "Netopýři i my potřebujeme typ vnitřního modelu, který umí zobrazovat polohu předmětů v trojrozměrném prostoru. Pak je ovšem naprosto irelevantní, že netopýři si vytvářejí svůj vnitřní model s pomocí ozvěn, kdežto my vsadili na světlo. V každém případě je cestou do mozku informace o vnějším světě převáděna na stejný typ nervových vzruchů. Proto se domnívám, že netopýři 'vidí' úplně stejně jako my, byť svět 'tam venku' převádějí do nervových vzruchů docela jiným fyzikálním médiem - ultrazvukem, nikoli světlem." (Dawkins 2002, p.46; překlad Tomáš Grim) Přiznám se, že mám v tomto ohledu tendenci více důvěřovat zoologu Dawkinsovi než filosofu Nagelovi, jenž v roce 2009 nechvalně proslul svými protidarwinistickými postoji a chválou inteligentního designu.

Nutnou podmínkou pro připsání vědomí je údajně život, přičemž stroje nepatří do kategorie živých ani mrtvých entit: "Roboti jsou stroje složené pouze z mechanických a elektrických součástí, nemohou být naživu." (Scriven 1953, p.233) Scriven by byl ochoten své názory změnit jen v případě, že bychom ve vesmíru objevili planetu, na níž mechanické bytosti rodí a vychovávají potomky, učí je myslet a mluvit, stárnou a umírají. Ani v tom případě by nebylo přípustné používat pro jejich označení pojem "robot", musel by být aplikován jiný termín jako třeba "android". Roboti totiž z definice vědomí nemají a ani mít nebudou, dokud je budeme sériově vyrábět v automatizovaných továrnách. Autora takto silných prohlášení přitom vůbec netrápí fakt, že biologové dosud nedospěli k uspokojivé definici života. Neexistuje žádný konsens, jaké vlastnosti živých bytostí by měly být považovány za esenciální: přítomnost proteinů nebo nukleových kyselin, schopnost metabolizovat nebo se reprodukovat, podléhání evolučnímu mechanismu přirozeného výběru nebo vzdorování druhému termodynamickému zákonu a podobně. Podmiňovat nedefinované vědomí nedefinovaným životem je *ignotum per æque ignotum* a Scriven nás takto zavádí na temné území logického klamu.

Donald Michie argumentuje o čtyři desetiletí později přeci jen vyspělejším způsobem, ale i některá jeho tvrzení vzbuzují přinejmenším rozpaky. Michie se domnívá, že raně analytická filosofie problém vědomí zcela ignorovala, behavioristé navíc drze popírali samotnou jeho existenci. Turing tento trend ještě prohloubil, když obrátil pozornost filosofů a psychologů k verbálním projevům vědomí. Analytičtí behavioristé by se zřejmě nezdráhali říct, že některé mentální aktivity jsou lépe přístupné vnějším pozorovatelům, než samotným vlastníkům myslí. Turing se předvídavě snažil oddělit otázku inteligence od vědomí, když tvrdil, že vědomí přisuzujeme každé osobě na základě vnějších projevů, neboť opačný přístup nás přivádí k solipsismu. Problém podle Michieho spočívá v Turingově zdvořilé konvenci, podle níž každá osoba myslí (Turing 1950, p.446; Michie 1993, p.6; viz také kapitola 3.4 této práce). Zůstává totiž předmětem diskuse, kdo do skupiny osob má patřit. Někteří z ní budou šovinisticky vyřazovat ženy nebo otroky, jiní přehnaně liberálně přijímat psy a další domácí mazlíčky. Jedná se vlastně o klasický příklad definice kruhem: myslící entity jsou osoby, osoby jsou myslící entity. Michie navíc dospívá k závěru, že T-test se postupem času dostal do sporu s poznatky současné kognitivní vědy, a to především ve dvou oblastech: (a) test nedokáže identifikovat tiché mentální procesy, které lidé *neartikulují* navenek; (b) test nedokáže

odhalit kognitivní schopnosti, které lidé *subartikuluji* podvědomě. V druhém ohledu se Michie inspiroval Frenchovým pojetím subkognice a jako příklad uvádí fakt, že jazyková pravidla jsou dodržována i lidmi, kteří o jejich existenci nemají ani ponětí. Dost problematická je Michieho snaha přisoudit artikulované vědění levé hemisféře a subartikulované vědění pravé hemisféře, jež s velkou pravděpodobností odporuje nejnovějším poznatkům neurofyzologie.

Rád bych upozornil, že teorie vědomí je zatím v plenkách a máme jen velmi mlhavou představu, co to vlastně vědomí je. V případě myslí si dokážeme vypomocet metodickým zjednodušením a považovat ji za souhrn všech mentálních obsahů a mechanismů, vědomí je však vágním pojmem a pro některé filosofy dokonce pseudopojmem. Například Dennett se přímo dovolává raných analytických behavioristů, především Gilberta Ryla, a v opozici k Nagelovým intuicím doporučuje přísně deflační přístup k vědomí. Jeho návrh spočívá v úplné eliminaci subjektivních zkušeností z našich teorií myslí, doslova ve *zquinování quálií* (Dennett 1990).<sup>98</sup> Chce se quálií zbavit úplně, protože tyto subjektivní a kvalitativní prožitky považuje za neslučitelné s ontologickým naturalismem. Vědecká analýza mentálních fenoménů v rámci neurologie nebo experimentální psychologie se obejde bez jakýchkoli zvláštních a neredukovatelných entit, vědomí je totiž naprosto stejným naturálním jevem jako třeba fotosyntéza nebo trávení. Americký filosof Larry Hauser podobně odmítá introspekci a považuje ji za nedůvěryhodný zdroj informací o vědomí. Není v tom jediný - analogicky a mnohem detailněji argumentuje i Eric Schwitzgebel (Hurlburt & Schwitzgebel 2007; Schwitzgebel 2011a) a další. Hauser jde snad ještě dál než Dennett, když tvrdí, že "žádná důvěryhodná psychologická teorie neobsahuje vědomí jako fundamentální pojem".<sup>99</sup>

I v případě, že nám Dennettův a Hauserův minimalismus připadá až příliš radikální, je třeba položit si centrální otázku: "Čím je vědomí, pokud není popsateľné materialisticky?" Vysvětlení odporující materialismu se vymykají vědecké metodologii, která je na materialismu, naturalismu a redukcionismu založená. Z toho vyplývá, že hypotézy, které jsou

---

<sup>98</sup> Tento podivný obrat má původ v Dennettově *Filosofickém lexikonu*, což je humoristický slovník parodující teorie předních filosofů: "quinovat, v., rezolutně popřít důležitost něčeho reálného nebo signifikantního".

<sup>99</sup> V poznámce navíc výsměšně dodává: "Though many a crackpot theory does." (viz Hauser 2001, p.47 a 50)

s materialismem v rozporu, můžeme považovat za nevědecké či pseudovědecké. Redukcionismus se v minulosti prokázal jako nesmírně robustní metodologický předpoklad, který se zatím pokaždé vyplatil - při explikaci života, vesmíru a vůbec. Neexistuje žádný racionální či pragmatický důvod, proč by měl tentokrát selhat při pokusu o vysvětlení vědomí.<sup>100</sup> Odmítání redukcionistické vědy připomíná zoufalé plahočení moderní teologie, která tak často sahá k hledání "boha mezer"; přehnaně zdůrazňuje oblasti, do kterých věda zatím nepronikla, a považuje je za výspy supranaturálních fenoménů. Filozofové naneštěstí také až příliš často upozorňují, co a proč věda nikdy nedokáže vysvětlit, místo aby o přijatelné vysvětlení sami usilovali. Označuji takové jednání za epistemologický defétismus, který filosofii v kontextu ostatních věd velice poškozují. Moderní podobu tohoto anachronického uvažování nalezneme v teoriích o existenci nepřekonatelné explanační propasti ("explanatory gap"), která dělí vědecké poznání od porozumění vědomí (Levine 1983); v tvrzení o obtížném problému vědomí ("hard problem"), na které obvyklé vědecké metody nestačí (Chalmers 1995); nebo v mysteriánství hlásajícím úplnou nemožnost pochopení psychofyzického problému (McGinn 1989). Domnívám se, že všechny tyto přístupy jsou buď návratem zakukleného karteziánství, nebo nepřiznaného vitalismu. Je každopádně naprosto bezdůvodné strojům apriorně upírat něco, o čem vůbec nevíme, čím ve skutečnosti je.

Nechci se pouštět do intuitivních hypotéz a příliš divokých spekulací, ale pokusím se aspoň zkratkovitě naznačit jedno z řešení problému vědomí, které si v uplynulém čtvrtstoletí vydobylo vcelku silné postavení mezi jinými reprezentačními teoriemi mysli. Jedná se o teorii vědomí vyššího řádu ("higher-order theory of consciousness"), jejímiž průkopníky jsou například David M. Rosenthal (1986) a Peter Carruthers (1989). John Locke už roku 1690 prohlásil, že vědomí je vnímáním toho, co člověku právě prochází myslí.<sup>101</sup> Teoretici vyššího řádu uvažují podobně, protože rozlišují mezi nerefektovaným myšlením prvního řádu a

---

<sup>100</sup> Redukcionismus ve filosofii není příliš rozšířen, ale přesto má nezanedbatelnou skupinu podpůrců, z nichž nejhlásitější je filosof vědy Alex Rosenberg (2006). V přírodní vědě je naopak naprosto běžný, a proto může držitel Nobelovy ceny za fyziku Steven Weinberg nadšeně zvolat: "Dvakrát hurá redukcionismu!"

<sup>101</sup> Anglický originál je poněkud výmluvnější než český překlad: "Consciousness is the perception of what passes in a man's own mind." (*An Essay Concerning Human Understanding*, Book II, Chapter I, §19); "Uvědomění si něčeho je vjem toho, co prochází v mysli člověka." (Locke 1984, p.79; překlad Anna Dokulilová)

reflektovaným vědomím vyšších řádů. Vědomí je tedy metamyšlením, myšlením o myšlenkách. Sebeuvědomění je poddruhem vědomí, ale nijak zvláštním: subjekt v takové situaci nereflektuje nějakou konkrétní myšlenku, ale samotný akt myšlení. Toto pojetí poněkud připomíná Tarského nauku o hierarchii jazyků, kdy je metajazyk reflexí objektového jazyka, metametajazyk reflexí metajazyka atd. Nebo bychom mohli ve freudovské tradici považovat první řád za podvědomí a vyšší řády za vědomí; pro nevědomí není v současné scientisticky orientované psychologii místo. French by k této teorii zřejmě také neměl vážnější výhrady, protože subkognice je prvním řádem a kognice řádem druhým či ještě vyšším. Rosenthal a Carruthers navíc předpokládají, že myšlení prvního řádu s námi sdílí všichni živočichové, kteří jsou vybavení nervovou soustavou. I ten nejprimitivnější kroužkovec vykazuje známky jednoduchého myšlení, pokud jím rozumíme opatření vedoucí k získání potravy, úspěšné reprodukci a vyhnutí se predaci. Naopak vědomí vyššího řádu je záležitostí elitní třídy bytostí, do které neproblematicky patří pouze lidé, zatímco u jiných živočišných druhů jsme na pochybách. Carruthers připouští možnost, že myšlení zvířat s výjimkou lidí se zcela odehrává v prvním řádu:

Zůstává otevřenou otázkou, zdali je vůbec možné mluvit o tom, jaké je to být netopýrem nebo psem nebo opicí. Pokud je vědomí jako rozsvícení světla, pak zvířata mohou žít v naprosté temnotě. (Carruthers 1989, p.259)

To by odpovídalo i Dennettovu tvrzení, že zvířata jsou sice schopná vnímat bolest, ale nedokážou pociťovat utrpení, které je bolestí reflektovanou (Dennett 1997, p.156).

Pokud svůj exkurz do teorie vyššího řádu stručně uzavřu, pak je možné přijmout několik předběžných hypotéz: 1. Inteligence je schopnost dospívat k racionálním rozhodnutím.<sup>102</sup> 2. Myšlení je proces umožňující inteligenci. 3. Inteligence ani myšlení nejsou podmíněny přítomností vědomí. Prozatímni přijetí těchto předpokladů je cestou, jak ochránit T-test před kritikou z důvodů absentujícího vědomí. Primárním cílem T-testu totiž není identifikace vědomí, ale myšlení a inteligence.<sup>103</sup> Turing se ve své eseji snaží termínu "consciousness" vyhýbat a používá jej jen v souvislosti se čtvrtou námitkou, nikoli z vlastní

---

<sup>102</sup> Uvědomuji si nedostatečnost argumentace ve prospěch následující definice, ale dovolím si tvrdit, že racionalita sama může být popsána ryze v utilitárních pojmech, podobně jako v teorii racionální volby.

<sup>103</sup> Přesto byla obdoba T-testu úspěšně použita identifikaci vědomí u pacientů ve vegetativním stavu (Stins 2009).



vůle. Když se ale nad problémem zamyslíme opravdu důsledně, pak nám stejně nezbude než si přiznat, že přítomnost vědomí můžeme ve většině případů posoudit opět pouze behaviorálně, například pomocí T-testu. U lidských bytostí je zřejmě možné použít nějaký typ snímkování prostřednictvím magnetické rezonance, protože určitá neurální aktivita našich mozků koreluje s přítomností vědomí. U ostatních živočišných druhů a jiných exotických entit toho ale nejsme schopni, protože vůbec nevíme, co s čím má korelovat. Zdá se, že nejprve jsme vždy odkázáni na behaviorální projevy a funkcionální popis, teprve později můžeme přistoupit k hlubšímu materialistickému vysvětlení mentálních jevů. Jednou snad bude existovat obdoba neurální vědy pro počítačové zařízení založená na bázi křemíku, zatím o ní ale nemáme ani to sebemenší ponětí.

## 8. Pozdní recepce a praktické aplikace: 90. léta a dál

Jak už to u filosofických problémů bývá, ani po šedesáti letech není stanovisko odborné veřejnosti v otázce T-testu jednotné. Objevily se pokusy o jeho naprostou diskreditaci, jiné o rehabilitaci, alternativní reformulaci anebo praktickou realizaci. V této poslední kapitole se budu věnovat právě těmto pestrým aktivitám z posledních dvaceti let, které sice už nejsou příliš filosoficky podstatné, ale možná o to zajímavější. Zřejmě nejdůležitější, ale zároveň nejkontroverznější, je každoroční soutěž amerického vynálezce a investora Hughua Loebnera o první program, který splní Turingem stanovené podmínky pro připsání inteligence. Připomínky k tomuto na první pohled rozumnému návrhu popíšu v závěru této části.

### 8.1 Slepá ulička umělé inteligence

V roce 1996 zveřejnil filosof vědy Blay Whitby vlivný článek, ve kterém označil T-test za největší slepou uličku umělé inteligence. Je přesvědčen, že Turingova původní esej se již vyčerpala. V současné době nepůsobí jako zdroj inspirace, ale rozptyluje pozornost badatelů a odvádí je od důležitějších témat. Svá tvrzení Whitby podpořil vytvořením časové osy, na které zachytil vývoj recepce Turingových myšlenek. V letech 1950 až 1966 byl T-test velmi inspirativním pramenem pro všechny badatele zabývající se umělou inteligencí. V následném období, přibližně do roku 1973, začal být odborníkům v oblasti umělé inteligence na obtíž. Ti jej vzápětí opustili, a proto mezi lety 1973 až 1990 T-test plní svou negativní úlohu už jen mezi filozofy. Po roce 1990 je podle Whitbyho možné považovat T-test za naprosto překonaný a řadit jej do historie obou disciplín. Whitby přiznává, že datace jednotlivých období je poněkud arbitrární: 1950 je rokem publikování "Computing Machinery and Intelligence", 1966 byl naprogramován program *ELIZA*, 1973 vznikl *PARRY*, v dubnu 1990 se na University of Sussex konala velká interdisciplinární konference připomínající čtyřicáté výročí T-testu. Hlavním problémem není samotná Turingova esej, ale její rozšířená dezinterpretace založená na zjednodušeném čtení originálního textu. Mnozí autoři se totiž mylně domnívají, že Turing vytvořil adekvátní operacionální definici inteligence, což je vede

k ještě závažnějším konfúzím.<sup>104</sup> Whitby tato nedorozumění přesně tematizuje:

1. Inteligence ve výpočetní technice spočívá ve schopnosti podvést lidského tazatele.
2. Nejlepším způsobem, jak definovat inteligenci, je operacionální test, jehož paradigmatickým příkladem je imitační hra.
3. Snaha sestrojít počítač, který by si úspěšně vedl v imitační hře, je integrální součástí výzkumu umělé inteligence.

Podle Whitbyho jsou první dva předpoklady nesprávné ze zcela triviálních důvodů, a není jim proto třeba věnovat větší pozornost. Ukázalo se, že chatboty k rozvoji umělé inteligence nepřispěly takřka ničím pozitivním a že operacionální definice je umělé inteligenci naprosto k ničemu. Mnohem podstatnější je prokázat neplatnost třetího tvrzení, které je podle Whitbyho na podvědomé úrovni rozšířeno mezi naprostou většinou odborníků v oblasti informatiky.

Přijetí imitační hry jako cíle výzkumu je značně neefektivní, neboť o lidské inteligenci zatím víme jen velice málo, a proto nemůžeme být v dohledné době úspěšní při jejím napodobování. Whitby situaci ilustruje použitím analogie mezi umělou inteligencí a aviatikou.<sup>105</sup> Přirovnává současný stav v umělé inteligenci k rozvoji letectví před bratry Wrightovými, kdy důsledná imitace přirozeného letu nevedla ke kýženým výsledkům. Prehistorie letectví byla přitom charakterizována detailními studii ptáků, jež v renesanci prováděl Leonardo da Vinci a do dokonalosti dotáhl Otto Lilienthal na konci 19. století. Tyto pokusy vyvrcholily konstrukcemi funkčních modelů ornitoptér Gustava Trouvého nebo Alphonse Pénauda, které nijak zvlášť nezasáhly do pozdějšího vývoje letadel těžších než vzduch. Stejně jako byl úspěch bratrů Wrightových podmíněn pochopením hlubších pravidel aeronautiky, na nichž je založen let ptáků i lidmi postavených létajících strojů, je třeba v oblasti umělé inteligence nejprve pořádně prozkoumat obecné základy jakýchkoli kognitivních procesů a pochopit komputace, kterými jsou realizovány. I když dnešní kluzáky vycházejí z napodobení ptačího plachtění na stoupajících prouděch teplejšího vzduchu, dokonalé imitace přirozeného letu lidé nejsou schopni. Otázkou je, zdali jí někdy schopni budou a bude-li to pragmaticky výhodné. Whitby používá praktické argumenty týkající se financování takového výzkumu, pro který lidé údajně nemají žádné využití. Obdobná situace

---

<sup>104</sup> K nesprávnému chápání T-testu jako definice inteligence viz kapitola 3 této práce a poznámka pod čarou 76.

<sup>105</sup> Viz Yazdani & Whitby (1987, pp.89-92) a Whitby (1996, pp.57-58). Tutéž myšlenku lze nalézt i u jiných autorů, například Hayes & Ford (1995, pp.975-976) nebo Harnad (1989, pp.6-7).

panuje v umělé inteligenci, protože po simulaci lidské mysli neexistuje potřebná poptávka. Mnohem užitečnější se jeví vývoj specializovaných zařízení, která budou moci vykonávat konkrétní úkony vyžadující určité rozvažovací schopnosti. Podobně jako někteří jiní autoři se totiž i Whitby obává, že pokusy imitovat či dokonce zcela nahradit obecnou lidskou inteligenci jsou nejen finančně nezajímavé, ale mohou být i potenciálně nebezpečné. Opět se nám tu tedy vynořuje pesimismus známý z Turingovy "hlavy v písku" (viz kapitola 3.2 této práce).

Myslím si, že na první pohled přílehlavý příměr s aviatikou není zcela přesný. Situace je podle mého názoru bližší okolnostem, za jakých by zřejmě probíhal vesmírný program bez masivní podpory americké a sovětské vlády. Bez štědrých dotací, které byly rozdávány s vidinou vojenského využití a prestižního postavení jednoho ekonomického zřízení nad druhým, bychom asi byli stále odkázáni jen na poloamatérské projekty financované z mělkých kapes soukromého kapitálu. Přes sympatie, které v nás vzbuzují komerční společnosti Scaled Composites nebo Armadillo Aerospace, je třeba si uvědomit, že plán pilotovaných vesmírných letů by měl v rukách privátních investorů jen pramalé vyhlídky na úspěšné dokončení, protože neskýtá příliš možností výtěžku. Telekomunikační, navigační a meteorologické satelity pro své umístění na orbitě nepotřebují raketoplány s astronauty. Právě z těchto důvodů stále vážnou projekty pilotovaných letů na jiná tělesa Sluneční soustavy. Finančně zajímavý totiž není ani onen stále odkládaný návrat na Měsíc, natož dobytí Marsu. Podobně umělá inteligence zatím k ničemu není, neboť nemá pořádné komerční nebo vojenské využití, a proto příliš nezajímá státní ani soukromé investory. Společnost IBM sice vyvíjela umělou inteligenci Watson proto, aby vítězila v televizních soutěžích, ale přesto své akcionáře pokouší utěšit nejistou vidinou použití v medicínské diagnostice. Cyničtěji řečeno: z obchodního hlediska nemá valný smysl vytvářet sofistikované komunikační systémy schopné zodpovídat libovolné dotazy. Proč vůbec vytvářet automatizovaná call centra, když ve třetím světě žijí stovky milionů lidí, kteří jsou ochotni tutéž práci dělat nesrovnatelně levněji?

S Whitbyho závěry lze přesto do značné míry souhlasit, protože se týkají dosavadního vývoje disciplíny, nikoli samotných Turingových tezí. Bezhlavá imitace lidského myšlení je asi opravdu jen málo užitečná, přestože chatboty zvyšují povědomí o problematice a popularizují disciplínu, i když poněkud bulvárním způsobem. Umělá inteligence, která chce

být úspěšnější než ta současná, bude muset najít mnohem sofistikovanější metody založené na počítačích modelech rozhodovacích procesů. Zvláště nadějným dojmem působí pokusy o simulaci primitivních organismů, jako jsou například už zmínění hmyzí robůtci Daria Floreana a jeho kolegů, o kterých jsem stručně psal v kapitole 3.8. Motivace okamžitě napodobovat tu vůbec nepokročilejší a nejsložitější formu inteligence, která na naší planetě existuje, je symptomem přehnaného velikášství. Floreanův tým postupuje "odspoda", protože do svých zařízení vkládá jen minimální množství instrukcí, ale vybaví je schopností interagovat s okolím a učit se z předchozí zkušenosti. Tento postup oprávněně připomíná Turingovo doporučení naprogramovat "dětskou mysl" a čekat, až si potřebné zkušenosti osvojí sama. Výsledky jsou přitom naprosto pozoruhodné, protože roboti škálou svého chování překvapují i své tvůrce.

Whitby za nejdůležitější část Turingovy eseje nepovažuje popis vysněného cíle umělé inteligence v daleké budoucnosti, protože T-test interpretuje jako velmi originální a zřejmě nezbytný způsob, jak připisovat mysl nehumanoidním stvořením. Na konci 40. let minulého století byl Turing jedním z hlavních představitelů paradigmatického posunu v lidském chápání inteligence a myšlení. Podle Whitbyho je třeba se připravit na situaci, kdy budeme muset některé exotické entity považovat za inteligentní, a to se všemi myslitelnými důsledky:

Opatření nálepkou "inteligentní" není čistě technickým cvičením; zahrnuje řadu morálních a sociálních rozměrů. Lidské bytosti jsou přesvědčeny, že se k inteligentním entitám mají chovat určitým způsobem. [...] V jistém ohledu je pravda, že cítíme menší rozpaky při připisování inteligence (a někdy dokonce schopnosti myslet) entitám, se kterými můžeme zajímavě pohovořit, než těm radikálně odlišným. (Whitby 1996, p.61)

T-test tedy nemusí být chápán jako bezprostřední cíl vědců a inženýrů umělé inteligence, ale z filosofického hlediska je stále jedinou známou metodou, jejíž pomocí dokážeme rozšiřovat množinu myslících bytostí za hranice rodu *Homo*.

## 8.2 Posouvání branek

Název této podkapitoly jsem si vypůjčil z anglického úsloví "moving the goalposts", které přede mnou ve stejném kontextu použil filosof Larry Hauser (2001). Doslova znamená "posouvání brankových tyčí" a od 80. let 20. století se používá k označení neformálního

logického pochybení, kdy jsou pravidla *ad hoc* měněna tak, aby určití adepti nesplnili nějakou zkoušku nebo kritérium. Hauser správně poukazuje, že kromě těch, kteří mají T-test za příliš těžký a pro počítače prakticky neproveditelný (za všechny třeba French; viz kapitola 7.1 této práce), existuje celá řada myslitelů, kterým se zdá příliš lehký. Našemu hluboce zakořeněnému antropocentrismu se nechce jen tak vzdát vyhrátého místečka uprostřed univerza a se stejnou zarputilostí, s jakou jedni odmítají uznat přítomnost mentálních stavů u jiných živočišných druhů, se druzí snaží vyloučit ze skupiny vlastníků myslí i počítače. V druhé polovině 20. století jsme byli svědky hned několika takových násilných redefiníci umělé inteligence. Ještě před Turingem se běžně tvrdilo, že postačující podmínkou inteligence je vítězství počítače nad šachovým velmistrem, ale pozdější vývoj přinutil odborníky tento požadavek opustit. Sám Gary Kasparov po porážce s Deep Blue tvrdil, že počítač vykazuje známky inteligentního chování, ale přesto nemyslí. Kasparov měl podle svého technického poradce Frederica Friedela pocit, že stroj "vytváří plány; rozumí strategii; snaží se ho obelstít; brání se jeho nápadům". Ani tyto projevy mu ale nepřipadaly dostatečné pro uznání, že Deep Blue aspoň do určité míry myslí (Hauser 2001, pp.45-46). Tato příhoda svědčí především o sémantických obtížích, které provázejí každý hovor týkající se mentálních aktivit. Nabízí se přinejmenším otázka, zdali je vůbec možné, aby existovala inteligentní entita, která nemyslí. Podle mého názoru to možné není a zastávat takové přesvědčení je příznakem chycení do pasti falešné lidové psychologie.<sup>106</sup>

Vidina stroje uspívajícího v neomezeném T-testu mnohé vystrašila natolik, že revizionisticky podmínili existenci myslí přítomností emocí, empatie, kreativity, humoru a podobně.<sup>107</sup> Antropocentričtí myslitelé se snaží zoufale najít něco navíc, cokoli, co by nás dále od strojů odlišovalo, i když padá jedna výspa lidské výjimečnosti za druhou. Mnozí filosofové proto vytvořili vlastní kritérium inteligence, které má původní Turingovu

---

<sup>106</sup> Hauser tvrdí to samé, ale ještě radikálněji: "I don't doubt that my cat exhibits intelligence to some extent, and thinks in his own peculiar manner. Likewise, my computer exhibits intelligence to some extent, and thinks in *it's* [*sic!*] own peculiar manner. That's what I think." (Hauser 2001, p.42)

<sup>107</sup> Zvláště populárním tématem filosofických diskusí je Voight-Kampffův test empatie, jenž vymyslel Philip K. Dick v románu *Sní androidi o elektrických ovečkách?* a zpopularizoval film *Blade Runner* režírovaný Ridleyem Scottem. Test je prováděn pomocí zařízení, které připomíná polygraf ("detektor lži"), a má za úkol měřit základní tělesné funkce jako dýchání, červenání se, tlukot srdce a pohyby očí. Cílem je detekovat nepřírozená zpoždění empatických reakcí, jež signalizují, že účastníkem testu není člověk, ale android.

podmínku nahradit. Tak třeba kognitivní vědec maďarského původu Stevan Harnad se poněkud svérázně vypořádal s robotickou námitkou ze Searlovy eseje "Minds, Brains, and Programs" a navrhl *totální T- test* ("Total Turing Test", TTT), jehož by se neúčastnil netělesný počítač, ale robot vybavený senzomotorickými orgány (Harnad 1989; Harnad 1991; Harnad 2000). Vítězství v této extravagantní variantě by dosáhlo jen zařízení, jehož chování by bylo neodlišitelné od všeho, čeho je člověk schopen. Harnad se domnívá, že takto koncipovaná zkouška odolá Searlovým výtkám proti klasickému T-testu. Vychází z názoru, že úspěšná simulace nemůže modelovat jen verbální inteligenci člověka, ale celek všech lidských kognitivních schopností. Řečeno jazykem Fodorovy modulární teorie mysli - není možné modelovat jen jeden modul, v tomto případě rozumění jazykovému projevu, ale součinnost mnoha vzájemně provázaných modulů. Z tohoto vcelku oprávněného předpokladu Harnad vyvozuje až příliš radikální závěr a požaduje existenci robota, který dokáže jezdit na jednokolce po visutém laně, tančit part Odetty a Odilie v *Labutím jezeře* nebo provést kardiochirurgickou operaci srdeční chlopně, aniž by bylo jasné proč vlastně. Samotné dovolávání se modulární povahy mysli je pravděpodobně napadnutelné. Lidé nemají specializovaný modul pro hru vrhcábů, ale přesto jsme dokázali tuto dovednost úspěšně modelovat u počítačů.

Považuji za velice zvláštní, že se vůbec někdo Harnadovými návrhy seriózně zabýval, ať už kriticky (Dyer 1990; Hauser 1993; Bringsjord et al. 2000), nebo dokonce vstřícně (Powers 1998). Naštěstí se zdá, že negativní reakce výrazně převyšují ty pozitivní. Nepohybovali bychom se ovšem na ošidné půdě filosofického bádání, kdyby někdo dříve či později nevyrukoval s ještě absurdnější variantou. Psycholog Stuart Watt například konstatoval, že T-test implicitně spočívá na naivní psychologii čili specificky lidské a evolučně vzniklé vlastnosti, jejíž pomocí identifikujeme mysli u ostatních bytostí. Doporučil proto *obrácený T-test* ("Inverted Turing Test"), v němž program hraje roli soudce a má mezi dvěma kandidáty rozhodnout, kdo je počítač a kdo člověk (Watt 1996); proč asi netuší ani on sám.<sup>108</sup> Podobně kognitivní vědec Peter Schweizer odmítl Harnadův totální T- test jako příliš snadný a vymyslel *skutečně totální T-test* ("Truly Total Turing Test", TTTT). Z této varianty může

---

<sup>108</sup> Této variantě testu bylo věnováno speciální číslo elektronického žurnálu *Psychology* (Vol. 7, No. 14, 1996), ve kterém byla uveřejněna celá řada kritických reakcí, jejichž autory byli třeba Selmer Bringsjord, Robert M. French, Kenneth M. Ford a Patrick J. Hayes.

vítězně vyjít jen entita, která dokáže historicky prokázat původ svého inteligentního chování (Schweizer 1998). Nutnou podmínkou pro připsání inteligence je tedy analýza evoluční minulosti, ve které musí kandidáti - tedy zřejmě roboti - vybudovat civilizaci a kulturu analogickou té naší. Nestačí, aby robot dokázal hrát šachy; musí ukázat, že je schopen vymyslet pravidla hry podobné šachům.<sup>109</sup> Není mi známo, že by se vůči Schweizerovu návrhu někdo systematicky vymezil; s některými nápady si zjevně nesvedou poradit ani ti nejspekulativnější filosofové.

### 8.3 Loebnerova cena

Turing se ve své eseji dopustil přinejmenším jedné nesprávné předpovědi, když prorokoval vcelku brzké vítězství stroje v imitační hře. Už za padesát let, tedy kolem roku 2000, bude prý úspěšnost strojů v pětiminutové imitační hře převyšovat třicet procent (Turing 1950, p.442). Stačí si ovšem přečíst záznamy rozhovorů z jakýchkoli realizovaných pokusů, aby člověk pochopil, že v tomto ohledu se autor testu mýlil. Dnešní chatboty možná dokážou přelstít méně kvalifikované a více důvěřivé tazatele, ale dovoluji si odhadnout, že při neomezené délce pohovoru by byla lidská schopnost odhalit počítač stoprocentní. Na tom se koneckonců shodují téměř všichni komentátoři (např. Copeland 2000, pp.526-527; Pinar Saygin et al. 2000, p.473; Hauser 2001, p.41; Moor 2001, p.84; Korukonda 2003, p.241; Rapaport 2006b, p.152; Epstein 2008, p.xiii). Nečekanou výjimku z pravidla tvoří Donald Davidson, který bez uvedení přesných důvodů tvrdí, že při dodržení pětiminutového limitu měl Turing asi pravdu (Davidson 1990, p.1).<sup>110</sup>

Dosavadní neúspěchy počítačů při plnění T-testu mohou někomu připadat frustrující, ale je třeba si uvědomit, jakým způsobem programátoři o úspěch usilovali. Od těch nejranějších pokusů používají chatboty vcelku simplexní metody vedení přijatelné

---

<sup>109</sup> Schweizer doslova píše: "It is not sufficient that a particular robot can *play* chess skillfully, since this can be achieved via the designer's foreknowledge of the existence and structure of the game. Rather the cognitive architecture of the system must be such that it could *invent* the game of chess, and that it could invent the general concept of a game." (Schweizer 1998, pp.268-269)

<sup>110</sup> Davidsonova esej k tomuto tématu je mezi analytickými filosofy kuriózní ještě z jednoho důvodu, a to kvůli genderovému čtení T-testu: účastníky jsou totiž žena a počítač. Imitační hra je proto údajně sexistická a bylo by správnější, aby proti stroji soutěžil libovolný člověk bez konkrétního pohlaví (Davidson 1990, pp.2-3; k problematice genderu viz 3 kapitola této práce).



konverzace. Jejich algoritmy pro analýzu přirozeného jazyka jsou nastaveny tak, aby ve větách tazatele hledaly určitá klíčová slova. Této proceduře se v počítačové terminologii říká porovnávání vzorců ("pattern matching") a můžeme ji ilustrovat následující situací. Uživatel napíše na klávesnici zahájení rozhovoru, například: "Poslední dobou jsem dost nešťastný."<sup>111</sup> Program ve větě identifikuje tu část, kterou vyhodnotí jako syntakticky podstatnou, a v podstatě mechanicky vygeneruje příslušnou reakci: "Mrzí mě, že jste poslední dobou nešťastný." V případě, kdy tato jednoduchá metoda selže, je možné odvést rozhovor jiným směrem, a to pomocí fráze s nízkou vypovídací hodnotou, jako třeba: "Zkuste být konkrétnější." nebo "Proč si to myslíte?" Úspěch takového programu nezávisí ani tak na složitosti vyhodnocovacích algoritmů, ale spíš na bohatosti databáze s možnými verbálními interakcemi.

Vynálezcem popsaného přístupu byl programátor Joseph Weizenbaum, jehož průkopnická práce začala roku 1964 a vyvrcholila vytvořením počítačové aplikace *ELIZA* (Weizenbaum 1966).<sup>112</sup> Jednalo se o vůbec prvního chatbota, který sloužil jako prostředí pro spouštění zvláštních skriptů obsahujících "osobnost" účastníka rozhovoru. Nejznámějším se stal *DOCTOR*, který překvapivě úspěšně simuloval rogeriánského terapeuta. Ukázalo se, že volba právě této konverzační role byla geniální, protože psychologové hlásící se k odkazu Carla Rogerse mají v popisu práce jen minimálně zasahovat do rozhovoru vlastními podněty a slouží především jako mediátoři sebevyjádření pacienta. Program si proto může dovolit předstírat, že nemá téměř žádné znalosti o vnějším světě; na větu "Byl jsem na projížďce lodí." je korektní zareagovat "Povězte mi něco o lodích.", což by v jiném kontextu působilo dosti neobvykle. Weizenbaum se při programování soustředil na několik hlavních problémů, jejichž vyřešení mělo zajistit hladký průběh konverzace mezi počítačem a lidským účastníkem. Patřila mezi ně identifikace klíčových slov, určení minimálního kontextu, použití transformačních pravidel pro vytvoření gramaticky správných vět, jež by už neobsahovaly klíčová slova, a podobně. Sám autor popsal činnost programu následovně:

---

<sup>111</sup> Následující příklady jsou volnou parafrází vět ze studie Weizenbaum (1966).

<sup>112</sup> Název programu je odvozen od křestního jména Elizy Doolittlové z divadelní hry G. B. Shawa *Pygmalion*. Spojitost je nasnadě - stejně jako hrdinka satirické komedie se i ELIZA učí správně mluvit podle instrukcí ("skriptů") svého učitele.

Hrubá procedura programu je vcelku jednoduchá; text je přečten a následně je zjištěna přítomnost klíčového slova. Pokud je takové slovo nalezeno, je věta transformována podle pravidla, které je s klíčovým slovem asociováno; pokud slovo nalezeno není, pak je odpovědí bezobsažná poznámka nebo - za určitých podmínek - je znovu použita dřívější reakce. Takto vypočítaný text je poté zobrazen. (Weizenbaum 1966, p.37)

Zdrojový kód programu *ELIZA* měl jen něco málo přes dvě stě řádků, ale jeho dopad na veřejnost před masovým rozšířením počítačů byl vskutku masivní. Weizenbaum popisoval případy, ve kterých testovaní účastníci počítači připisovali rozsáhlý soubor nejrůznějších znalostí a schopnost logického usuzování. V souladu s Dennettovým pojetím intencionálního postoje tak lidé činili proto, aby jim mohlo být nasloucháno a rozuměno. Někteří dokonce vyžadovali po organizátorech experimentu soukromí a nebyli ochotni odevzdávat ke zpracování výpisy z rozhovorů, protože mohli obsahovat intimní informace podléhající lékařskému tajemství.

Weizenbaum od počátku varoval, že *ELIZA* je nanejvýš "překladačným procesorem" a že by její schopnosti neměly být přeceňovány, protože mají jen málo společného se skutečným rozuměním přirozenému jazyku. Naneštěstí se další vývoj chatbotů zaměřil pouze na nepřilíh originální rozvíjení procedur, které na konci 60. let vymyslel. Nejsofistikovanějším programem tohoto období je zřejmě *SHRDLU* Terryho Winograda, jenž byl teoreticky i prakticky popsán v monumentálním stodevadesátistránkovém článku publikovaném v časopise *Cognitive Psychology* (Winograd 1972). Jednodušším, ale o poznání zábavnějším exemplářem se stal *PARRY* (Colby et al. 1971; Colby et al. 1972). V jeho případě se obrátily role známé ze skriptu *DOCTOR*, protože *PARRY* úspěšně imitoval pacienta trpícího paranoiou. Volba osobnosti byla navíc obdobně vychytralá, protože lidský účastník seznámený s psychickým stavem svého protějšku byl ochoten odpouštět mu některé logické chyby, jež byly omlouvány tendencí k přehnané podezřívavosti a patologickými pocity ohrožení. Počítač reagoval na nečekané podněty nesouvislými zmínkami o napojení na italskou mafii nebo podjatém přístupu policie ke slušným lidem, které byly tazatelem většinou interpretovány jako projevy pacientovi diagnózy. V situacích, ve kterých *ELIZA* používala chlácholivé zdvořilostní fráze, byl *PARRY* naopak neodůvodněně agresivní a zpochybňoval důvěryhodnost oponenta. Mezi takové reakce patřilo například: "Myslíte si, že jsem lhář?" nebo "Takoví jste všichni." Skutečné orgie úhybných manévrů byly spuštěny v rozhovorech, jichž se účastnili *ELIZA* a *PARRY* dohromady. Setkání těchto "barvitých charakterů" byla

charakteristická nízkou ochotou opouštět známá témata a cyklickou povahou konverzace (srov. Güzeldere & Franchi 1995). Colbyho simulaci umělé paranoie lze i přes celkovou propracovanost jen těžko považovat za zvlášť přínosnou pro vývoj umělé inteligence a někteří ji považovali za zcela nelegitimní. Weizenbaum dokonce ironicky navrhoval počítačovou simulaci autistického člověka, který by neodpovídal vůbec. Výsměšně k tomu dodal: "Výhodou takového programu je, že by mohl být realizován na terminálu, jenž by vůbec nebyl připojený k počítači." (Weizenbaum 1974, p.425)

První chatboty našly určité využití při programování primitivních počítačových her, které měly ryze textovou podobu a náležely k žánru "adventure". Jejich příkladem může být série *Zork* od společnosti Infocom, jež zaznamenala v 70. a 80. letech minulého století značnou popularitu. Současné protějšky tehdejších programů využívají totožné, i když propracovanější postupy, které zvyšují pravděpodobnost obelstění lidského účastníka, ale jen málo přispívají k pochopení a imitaci lidské inteligence. Tento neblahý efekt se projevil ve všech dosavadních ročnících soutěže o Loebnerovu cenu, jež je pojmenována po svém zakladateli a hlavním organizátorovi Hughu Loebnerovi. Soutěž byla poprvé vyhlášena roku 1991, po téměř dvouletém určování podmínek a plánování průběhu.<sup>113</sup> Zpočátku se jednalo o velice prestižní událost, jejímuž organizačnímu výboru předsedal Daniel Dennett a členem byl i W. V. O. Quine. Postupem času začala naneštěstí ztrácet na důvěryhodnosti a byla podrobena intenzivní kritice ze strany filosofů i odborníků na umělou inteligenci. Podle původních pravidel bylo testováno deset účastníků, z nichž čtyři byli lidé ("confederates") a šest počítačové programy ("contestants" nebo "entries"). Výkon účastníků byl posuzován deseti soudci ("judges"), kteří byli vybráni na základě novinového inzerátu. Soudci hodnotili výkon účastníků na škále od 1 do 5, kde 1 znamenalo "určitě stroj" a 5 "určitě člověk". Na objektivitu testu dohlížel nezávislý rozhodčí ("referee"). Organizaci testu v průběhu času zajišťovali mnozí významní badatelé, například Ned Block, Burton Dreben, Hartry Field, Luciano Floridi, Andrew Hodges nebo Kevin Warwick.

Pravidla byla roku 2003 zásadně revidována a i poté podrobena dílčím změnám. Hlavním zásahem do originálního znění bylo rozšíření soutěže o audiovizuální složku, která má následovat po prvním úspěšném složení T-testu počítačovým programem. V této části

---

<sup>113</sup> Okolnosti prvního ročníku detailně popisují Epstein (1992) a Shieber (1994a).

musí být program schopen vést konverzaci v mluvené řeči, takže musí zahrnovat i prostředky pro analýzu slyšeného jazyka, a musí správně reagovat na vizuální podněty, jako jsou obrazy nebo gesta (viz Powers 1998). V soutěži mohou být uděleny tři medaile: bronzová pro každoročního vítěze (je dotována sumou 3000 dolarů); stříbrná pro počítač, který bude soudci považován za člověka v běžné textové konverzaci (25 000 dolarů); zlatá pro počítač, který bude úspěšný v audiovizuální verzi testu (100 000 dolarů). Soutěž byla nejprve zaštitěna University of Surrey, která se ale později od Loebnerových aktivit distancovala, a od té doby se organizace ujaly různé jiné, méně prestižní instituce. O problematickém stavu zřejmě svědčí i fakt, že zatímco dotace za bronzovou medaili byla během uplynulých dvaceti let navýšena z patnácti set dolarů na pozdější dva tisíce a dnešní tři tisíce, odměna za obě vyšší ocenění zůstala nezměněna.

Turingem popsaný test byl po potřeby Loebnerovy ceny omezen, a to hned několika způsoby. Nutné je samozřejmě zavedení časového limitu, který se ve vývoji soutěže měnil - zpočátku se délka rozhovoru mezi účastníkem a soudcem pohybovala kolem patnácti minut, ale postupně byla zkrácena na pět minut čistého času. První ročníky soutěže obsahovaly restriktci témat, jejichž soubor byl předem, i když ne zcela přesně určen; později došlo k rozvolnění testu a umožnění konverzace o čemkoli. Podstatnější jsou ale omezení stylu hovoru, který musí mít podobu přirozené konverzace, při níž soudci nesmí používat žádné léčky a úskoky. Současná podoba pravidel navíc obsahuje vágní požadavek, aby byl použitý slovník srozumitelný a přijatelný pro dvanáctileté dítě. Kritici omezené varianty vždy tvrdili, že restriktce jsou svou povahou dosti problematické a snižují důvěryhodnost T-testu, jehož síla původně spočívala v možnosti konverzovat libovolným způsobem o čemkoli.<sup>114</sup>

Mezi odpůrce soutěže se řadí Daniel Dennett nebo Marvin Minsky, z nichž ten druhý celý podnik několikrát označil za drzý podvod na laické veřejnosti (Sundman 2003). Hlavním

---

<sup>114</sup> Alternativu vůči Loebnerově ceně vytvořili roku 2002 programátoři Mitchell Kapor a Ray Kurzweil v rámci takzvané dlouhé sázky ("long bet"). Testu se účastní tři lidé a jeden počítač. Rozhodují tři soudci, kteří vedou se všemi soutěžícími čtyři dvouhodinové rozhovory. Zkouška je vykonána pomocí terminálu, přičemž rozhovor je neomezený a může se týkat jakéhokoli tématu. Pro úspěšné složení musí počítač přesvědčit nejméně dva soudce a medián hodnocení jeho lidskosti (na stupnici 1 až 4) musí být vyšší než medián dvou či více lidských účastníků. Kapor se s Kurzweilem vsadil o deset tisíc dolarů, že do roku 2029 touto variantou T-testu neprojde žádný počítač (Kapor & Kurzweil 2002; Kurzweil & Kapor 2008).

představitelem sporu o legitimitu Loebnerovy ceny je ale zřejmě Stuart M. Shieber, jenž roku 1994 publikoval esej "Lessons from a Restricted Turing Test". V ní se velice negativně vyjádřil k teoretickým východiskům i praktickému průběhu prvního ročníku soutěže. Cena podle Shieberta nemá jasný cíl a nemůže být v dohledné době vůbec udělena, protože zatím neexistuje technologie potřebná pro vítězství. Situaci srovnává s Kremerovou cenou pro letadla poháněná lidskou silou, která byla vyhlášena roku 1959 a díky dobře nastaveným kritériím udělena už v roce 1977. Nesprávně koncipovaná Loebnerova cena naopak způsobuje, že lidé jsou o umělé inteligenci informováni zkresleně, systematicky zavádějícím způsobem. Lze říct, že je spíše ke škodě než k užítku: "Současná podoba Loebnerovy ceny v lepším případě odčerpává úsilí a v horším případě poškozují pověst vědecké komunity." (Shieber 1994a, p.77) Hned první ročník ukázal, že počítače komunikovaly velmi špatně a odborníci je dokázali velmi rychle odhalit, naneštěstí na rozdíl od najatých soudců. Nejčastější byly chybné transformace zájmen, repetitivní odpovědi na opakující se dotazy, případně naprostá neschopnost reagovat na nesmyslné sekvence znaků. Přítomný Weizenbaum se proto nechal slyšet, že za pětadvacet let od jeho programu *ELIZA* nedošlo k téměř žádnému vývoji. Sám vítězný *PC Therapist* Josepha Weintrauba byl jen mírně inovovanou verzí skriptu *DOCTOR* a úspěch se mu dokonce podařilo zopakovat i ve druhém, třetím a pátém ročníku soutěže! Spíše než o umělé inteligenci Loebnerova cena informuje o jednání lidí při interakci s počítači, což ale vůbec není cílem soutěže a na téma navíc existuje značné množství přesvědčivějších psychologických studií. Je totiž velice překvapivé, kolik soudců bylo oklamáno i těmi nejprimitivnějšími počítačovými programy. Vyplývá z toho, že lidé se nechávají zmást dosti snadno, protože se snaží nalézt určitou strukturu v naprosto chaotických jevech. K naší přirozenosti patří, že hledáme intencionalitu i tam, kde není, z čehož vyplývá antropomorfizace celé řady nemyslicích či neživých entit. Lidé jsou také jen málo imunní vůči obecným tvrzením, ve kterých identifikují konkrétní smysl. Tento jev bývá označován jako Barnumův efekt, a to na počest baviče P. T. Barnuma, jenž ve svém cirkusu "měl něco pro každého". Zneužívají ho nejen konverzační programy, ale i nejrůznější šarlatáni jako astrologové, vykladači karet nebo zastánci kvalitativního přístupu v psychologii osobnosti. V rámci takzvaného studeného čtení ("cold reading") jsou produkovány nicneříkající věty, jež platí pro kohokoli a které chce slyšet každý.

Proti Shieberově kritice se Loebner ohradil vcelku rozumným tvrzením, že si může své peníze utrácet tak, jak se mu zlíbí. Pokud Shieber nesouhlasí s podmínkami soutěže, nic mu nebrání v tom, aby si založil vlastní cenu. Zbytek jeho reakce je naneštěstí zahlcen sněním o utopické společnosti, ve které by počítače zastaly veškerou práci, a podivnými úvahami o Asimovových zákonech robotiky, jež by měly být doplněny o dodatek deklarující nadřazenost lidí nad stroji (Loebner 1994; viz také Shieber 1994b). Právě nepřesvědčivý způsob organizace Loebnerovy ceny způsobil, že se odborná veřejnost se soutěží nikdy neztotožnila a kritika z počátku 90. let je téměř beze změny opakována dodnes (srov. Zdenek 2001; Loebner 2008; Floridi et al. 2009). Odmítnuta byla i audiovizuální verze testu, která po účastnících zmatečně vyžaduje schopnosti, jež nemají s inteligencí nic společného. V takto rozšířeném testu by totiž neměl šanci uspět člověk se smyslovým postižením, tedy například hluchý nebo slepý (Hauser 2001, p.44).

Loebnerova cena může být paradoxně chápána jako empirická falzifikace hypotézy, že sestrojení myslícího počítače je v našich momentálních možnostech. Palčivou otázkou zůstává, zdali se autoři chatbotů po roce 1966 vůbec vydali správným směrem a jestli nesešli z cesty k umělé inteligenci v bezectné snaze šidit neinformované tazatele. Přesto není třeba ztrácet naději a oddávat se pohodlnému pesimismu. Počítače jsou po inženýrské stránce lépe promyšleny než nervové soustavy živočichů, protože jejich struktura je záměrná a neobsahuje ty nejokatější případy "neinteligentního designu" evolučně vzniklých organismů. Přirozený výběr má oproti lidským programátorům pár miliard let náskok, a proto dokázal vytvořit úžasné věci i z podřadného, snadno degradovatelného materiálu a za použití zhoła neefektivní metody pokusu a omylu. Dejme proto lidem ještě šanci a neházejme flintu do žita pokaždé, když se něco náhodou nevyvíjí přesně podle našich představ. Sestrojení inteligentního stroje je dosažitelným cílem, byť zatím není na dosah.

## 9. Závěr: priorita vědy před filosofií

Kapitolou o nejnovější recepci a praktických aplikacích jsem dokončil svůj průřez dějinami T-testu, jehož cílem bylo především poukázat na správnost a trvanlivost Turingových závěrů a předpovědí. Myslím si, že v teorii mysli nemáme příliš na výběr, protože neurofilosofie a funkcionalismus jsou jedinými dvěma seriózními kandidáty, kteří dokážou vysvětlit pestrý svět mentálních jevů pomocí přesvědčivé vědecké metodologie. Karteziánský dualismus byl převažující většinou analytických filosofů opuštěn díky nepřekonatelným problémům týkajícím se výměny informací mezi naprosto oddělenými substancemi a také díky nekompatibilitě imateriálních entit s ontologií naturalizované filosofie. Antiredukcionistické teorie jsou zase převlečeným dualismem, neboť operují s obdobnými opozitními principy, z nichž přinejmenším ten emergentní či supervenientní není vůbec přístupný empirickému poznání. Naštěstí je dost dobře možné, že se mezi neurofilosofií a funkcionalismem nebudeme muset dramaticky rozhodovat, protože obě teorie se pravděpodobně mohou vzájemně doplňovat. Pokud přijmeme předpoklad, že mentální stavy jsou realizovány neurálními stavy (a něco jiného by bylo dosti naivní), pak nám nezbývá než být naturalisty. Pokud přijmeme předpoklad, že mohou existovat entity realizující obdobné mentální stavy pomocí odlišných neurálních stavů (a něco jiného by bylo dosti sebestředné), pak nám nezbývá než být funkcionalisty.

Relace mezi neurofilosofií a funkcionalismem je opravdu do značné míry analogická vztahu mezi teorií hardwaru a softwaru, i když rozhodně není pravda, že by byl mozek hardwarem a mysl byla softwarem - na to je rozdílů mezi mozky a počítači až příliš mnoho. Tentýž systém můžeme popsat buď materiálně, nebo funkcionálně. Funkcionalistické vysvětlení je jednodušší a snáze pochopitelné, protože se spoléhá jen na zobecněné typy vztahů mezi vstupy a výstupy libovolných agentů, aniž by muselo přihlížet k jejich materiálnímu složení. Materialistický popis je nejjednodušší, o něco méně přesný je popis funkcionální a úplně nejslabší epistemickou sílu má popis používaný lidovou psychologií, jejíž praktičnost v každodenním životě je ovšem neoddiskutovatelná. Je dokonce možné, že naše mozky jsou natolik složité, že je nebudeme s to úplně pochopit nikdy, a to materiálně ani funkcionálně. Jsem sice v tomto ohledu optimistou, který se s uspokojením

ohlíží na předešlé úspěchy vědy a s nadějí vyhlíží ty nastávající, ale přesto může dojít k situaci, kdy lidské kognitivní schopnosti nebudou dostatečným nástrojem pro explikaci jich samotných.

Nejsem si jistý, zdali je tradiční komputační funkcionalismus udržitelnou teorií, a prozatím se mi jako slibnější jeví vývoj v neurálních vědách. Mnohonásobná realizovatelnost není zřejmě úplně správným východiskem, protože materiální složení je pro funkci jakéhokoli zařízení dosti podstatným faktorem. Spalovací motor je teoreticky možné vyrobit ze dřeva, ale po praktické stránce se rozhodně nejedná o ideální volbu. Searle k tomuto poznatku říká, že některé látky jsou "špatným druhem materiálu" (Searle 1980b, p.423) a k výkonu určitých činností se jednoduše nehodí. Ke konstrukci myslících entit například nejsou vhodné soustavy poštovních schránek, větrem poháněné mlýny ani sofistikované skládačky z plechovek od piva, jak výsměšně poznamenali někteří odpůrci komputačního přístupu v teorii mysli. Může se dokonce ukázat, že vhodným prostředkem nejsou ani dnešní procesory, radiče, paměťové moduly, pevné disky a vysokorychlostní optické spoje. Případný neúspěch digitálních technologií ale nebude předpovězen pomocí apriorních rozvah, nýbrž konstatován za použití experimentálního testování empirických hypotéz. V tomto ohledu se velmi jasně projevuje priorita přírodní vědy před filosofií. Filozofové přesto nemusí rezignovat na pozitivní roli při rozšiřování poznání, protože při zachování zodpovědného přístupu mohou úspěšně ovlivňovat vývoj kognitivních věd, i když do něj příliš aktivně nezasáhnou. Podobně jako Turing ve čtyřicátých letech popisoval programování ještě neexistujících počítačů, tak se i my můžeme pouštět do opatrných předpovědí o dalším směřování našeho pochopení mysli. Se skromnou autoritou, kterou filosofie mým slovům propůjčuje, mohu například deklarovat své přesvědčení, že T-test je postačující podmínkou pro detekci inteligence u námi vytvořených strojů. Důrazněji řečeno, behaviorismus je zatím při identifikaci mysli u nehumánních entit obligatorní, nikoli pouze fakultativní.



## 10. Bibliografie

- Anderson, J.A., 1994. Turing's Test and the Perils of Psychohistory. *Social Epistemology*, 8(4), pp.327-332.
- Axelrod, R., 1980. Effective Choice in the Prisoner's Dilemma. *The Journal of Conflict Resolution*, 24(1), pp.3-25.
- Axelrod, R., 1984. *The Evolution of Cooperation*, New York: Basic Books.
- Baggini, J. ed., 2008. Ten by Ten. *The Philosophers' Magazine*, 40, pp.37-63.
- Benacerraf, P., 1967. God, the Devil, and Gödel. *The Monist*, 51(1), pp.9-32.
- Block, N., 1981. Psychologism and Behaviorism. *The Philosophical Review*, 90(1), pp.5-43.
- Block, N., 1995. The Mind as the Software of the Brain. In E. E. Smith & D. N. Osherson, eds. *An Invitation to Cognitive Science*. Cambridge: MIT Press, pp. 377-425.
- Block, N., 2007. Troubles with Functionalism. In *Consciousness, Function, and Representation*. Collected Papers. Cambridge: MIT Press, pp. 63-101.
- Block, N., 1978. Troubles with Functionalism. In C. W. Savage, ed. *Minnesota Studies in the Philosophy of Science*. Minneapolis: University of Minnesota Press, pp. 261-325.
- Boden, M.A., 1990. *The Philosophy of Artificial Intelligence*, Oxford University Press.
- Boyer, D.L., 1983. J. R. Lucas, Kurt Gödel, and Fred Astaire. *The Philosophical Quarterly*, 33(131), pp.147-159.
- Bringsjord, S., 1998. Chess Is Too Easy. *Technology Review*, 101(2), pp.23-28.
- Bringsjord, S. & Xiao, H., 2000. A Refutation of Penrose's Gödelian Case Against Artificial Intelligence. *Journal of Experimental & Theoretical Artificial Intelligence*, 12(3), pp.307-329.
- Bringsjord, S., Bello, P. & Ferrucci, D., 2001. Creativity, the Turing Test, and the (Better) Lovelace Test. *Minds and Machines*, 11(1), pp.3-27.
- Bringsjord, S., Caporale, C. & Noel, R., 2000. Animals, Zombanimals, and the Total Turing Test. *Journal of Logic, Language and Information*, 9(4), pp.397-418.
- Bronowski, J., 1985. *Vzestup člověka*, Praha: Odeon.
- Buckwalter, W. & Stich, S.P., 2010. Gender and Philosophical Intuition. Available at: <http://ssrn.com/abstract=1683066>.
- Carruthers, P., 1989. Brute Experience. *The Journal of Philosophy*, 86(5), pp.258-269.

- Clark, A., 2008. *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*, Oxford University Press.
- Clark, A. & Chalmers, D., 1998. The Extended Mind. *Analysis*, 58(1), pp.7-19.
- Clarke, A.C., 1973. *Profiles of the Future: An Inquiry into the Limits of the Possible*, New York: Harper & Row.
- Colby, K.M. et al., 1972. Turing-like Indistinguishability Tests for the Validation of a Computer Simulation of Paranoid Processes. *Artificial Intelligence*, 3, pp.199-221.
- Colby, K.M., Weber, S. & Hilf, F.D., 1971. Artificial Paranoia. *Artificial Intelligence*, 2(1), pp.1-25.
- Cole, D., 2009. The Chinese Room Argument. *The Stanford Encyclopedia of Philosophy* (Winter 2009 Edition). Available at: <http://plato.stanford.edu/entries/chinese-room/>.
- Copeland, B.J., 2004. *The Essential Turing: Seminal Writings in Computing, Logic, Philosophy, Artificial Intelligence, and Artificial Life plus The Secrets of Enigma*, Oxford University Press.
- Copeland, B.J., 2000. The Turing Test. *Minds and Machines*, 10(4), pp.519-539.
- de Cordemoy, G., 2003. *A Philosophical Discourse Concerning Speech*, Whitefish: Kessinger Publishing.
- Cosmides, L., 1989. The Logic of Social Exchange: Has Natural Selection Shaped How Humans Reason? Studies with the Wason Selection Task. *Cognition*, 31(3), pp.187-276.
- Crane, T., 2003. *The Mechanical Mind: A Philosophical Introduction to Minds, Machines, and Mental Representation* 2nd ed., London: Routledge.
- Damper, R.I., 2006. The Logic of Searle's Chinese Room Argument. *Minds and Machines*, 16(2), pp.163-183.
- Davidson, D., 2004. Externalizovaná epistemologie. In *Subjektivita, intersubjektivita, objektivita*. Praha: Filosofia, pp. 219-231.
- Davidson, D., 1990. Turing's Test. In K. A. M. Sain et al., eds. *Modelling the Mind*. Oxford University Press, pp. 1-11.
- Dawkins, R., 1996. *Řeka z ráje: Darwinistický pohled na život*, Bratislava: Archa.
- Dawkins, R., 2002. *Slepý hodinář*, Praha: Paseka.
- Dennett, D.C., 1995. Can Machines Think? In M. Shafto, ed. *How We Know*. San Francisco: Harper & Row, pp. 121-145.
- Dennett, D.C., 1991. *Consciousness Explained*, Boston: Little, Brown and Co.

- Dennett, D.C., 1997. *Druhy myslí: K pochopení vědomí*, Bratislava: Archa.
- Dennett, D.C., 2001. In Darwin's Wake, Where Am I? *Proceedings and Addresses of the American Philosophical Association*, 75(2), pp.11-30.
- Dennett, D.C., 1990. Quining Qualia. In W. Lycan, ed. *Mind and Cognition: A Reader*. Oxford: Blackwell, pp. 519-547.
- Descartes, R., 1992. *Rozprava o metodě*, Praha: Svoboda.
- Dodd, T., 1991. Gödel, Penrose and the Possibility of AI. *Artificial Intelligence Review*, 5(3), pp.187-199.
- Dyer, M.G., 1990. Intentionality and Computationalism: Minds, Machines, Searle and Harnad. *Journal of Experimental & Theoretical Artificial Intelligence*, 2(4), pp.303-319.
- Epstein, R., 2008. Introduction. In R. Epstein, G. Roberts, & G. Beber, eds. *Parsing the Turing Test*. Dordrecht: Springer, p. xi-xviii.
- Epstein, R., 1992. The Quest for the Thinking Computer. *AI Magazine*, 13(2), pp.81-95.
- Farrell, B.A., 1946a. An Appraisal of Therapeutic Positivism I. *Mind*, 55(217), pp.25-48.
- Farrell, B.A., 1946b. An Appraisal of Therapeutic Positivism II. *Mind*, 55(218), pp.133-150.
- Feferman, S., 1995. Penrose's Gödelian Argument - A Review of "Shadows of the Mind" by Roger Penrose. *Psyche*, 2(7). Available at:  
<http://psyche.cs.monash.edu.au/v2/psyche-2-07-feferman.html>.
- Fisher, H.E., Aron, A. & Brown, L.L., 2006. Romantic Love: A Mammalian Brain System for Mate Choice. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 361(1476), pp.2173-2186.
- Floreano, D. et al., 2007. Evolutionary Conditions for the Emergence of Communication in Robots. *Current Biology*, 17(6), pp.514-519.
- Floridi, L., Taddeo, M. & Turilli, M., 2009. Turing's Imitation Game: Still an Impossible Challenge for All Machines and Some Judges - An Evaluation of the 2008 Loebner Contest. *Minds and Machines*, 19(1), pp.145-150.
- Flowers, T.F., 2006. D-Day at Bletchley Park. In B. J. Copeland, ed. *Colossus: The Secrets of Bletchley Park's Codebreaking Computers*. Oxford University Press, pp. 78-83.
- Fodor, J., 1980. Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology. *Behavioral and Brain Sciences*, 3(1), pp.63-72.
- Fodor, J., 2009. Where Is My Mind? *London Review of Books*, 31(3), pp.13-15.
- Ford, J., 2011. Helen Keller Was Never in a Chinese Room. *Minds and Machines*, 21(1), pp.57-

72.

- Franzén, T., 2005. *Gödel's Theorem: An Incomplete Guide to Its Use and Abuse*, Wellesley: A K Peters, Ltd.
- French, R.M., 2000a. Peeking Behind the Screen: The Unsuspected Power of the Standard Turing Test. *Journal of Experimental & Theoretical Artificial Intelligence*, 12(3), pp.331-340.
- French, R.M., 1990. Subcognition and the Limits of the Turing Test. *Mind*, 99(393), pp.53-65.
- French, R.M., 2000b. The Turing Test: The First 50 Years. *Trends in Cognitive Sciences*, 4(3), pp.115-122.
- Gallagher, S., 2009. The Key to the Chinese Room. In K. Leidlmair, ed. *After Cognitivism: A Reassessment of Cognitive Science and Philosophy*. Dordrecht: Springer, pp. 87-96.
- Genova, J., 1994a. Response to Anderson and Keith. *Social Epistemology*, 8(4), pp.341-343.
- Genova, J., 1994b. Turing's Sexual Guessing Game. *Social Epistemology*, 8(4), pp.314-326.
- George, A. & Velleman, D.J., 2000. Leveling the Playing Field between Mind and Machine: A Reply to McCall. *The Journal of Philosophy*, 97(8), pp.456-461.
- George, F.H., 1956. Could Machines Be Made to Think? *Philosophy*, 31(118), pp.244-252.
- George, F.H., 1957. Thinking and Machines. *Philosophy*, 32(121), pp.168-169.
- Good, I.J., 1967. Human and Machine Logic. *The British Journal for the Philosophy of Science*, 18(2), pp.144-147.
- Greene, J.D. et al., 2001. An fMRI Investigation of Emotional Engagement in Moral Judgment. *Science*, 293(5537), pp.2105 -2108.
- Gunderson, K., 1964. The Imitation Game. *Mind*, 73(290), pp.234-245.
- Güzeldere, G. & Franchi, S., 1995. Dialogues with Colorful "Personalities" of Early AI. *Stanford Humanities Review*, 4(2). Available at:  
<http://www.stanford.edu/group/SHR/4-2/text/dialogues.html>.
- Hanson, W.H., 1971. Mechanism and Gödel's Theorems. *The British Journal for the Philosophy of Science*, 22(1), pp.9-16.
- Harnad, S., 1989. Minds, Machines and Searle. *Journal of Experimental & Theoretical Artificial Intelligence*, 1(1), pp.5-25.
- Harnad, S., 2000. Minds, Machines and Turing. *Journal of Logic, Language and Information*, 9(4), pp.425-445.
- Harnad, S., 1991. Other Bodies, Other Minds: A Machine Incarnation of an Old Philosophical

- Problem. *Minds and Machines*, 1(1), pp.43-54.
- Harré, R. & Wang, H.-T., 1999. Setting Up a Real “Chinese Room”: An Empirical Replication of a Famous Thought Experiment. *Journal of Experimental & Theoretical Artificial Intelligence*, 11(2), pp.153-154.
- Hauser, L., 2001. Look Who’s Moving the Goal Posts Now. *Minds and Machines*, 11(1), pp.41-51.
- Hauser, L., 1993. Reaping the Whirlwind: Reply to Harnad’s “Other Bodies, Other Minds.” *Minds and Machines*, 3(2), pp.219-237.
- Hayes, P. & Ford, K., 1995. Turing Test Considered Harmful. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*. Montreal: Morgan Kaufmann, pp. 972-977.
- Hinsley, F.H., 2001. The Influence of Ultra in the Second World War. In F. H. Hinsley & A. Stripp, eds. *Codebreakers: The Inside Story of Bletchley Park*. Oxford University Press, pp. 1-13.
- Hobbes, T., 1988. O tělese. In *Výbor z díla*. Praha: Svoboda, pp. 29-91.
- Hodges, A., 1992. *Alan Turing: The Enigma* 2nd ed., London: Vintage.
- Hodges, A., 1997. *Turing: A Natural Philosopher*, London: Phoenix.
- Hofstadter, D.R., 1979. *Gödel, Escher, Bach*, New York: Basic Books.
- Hume, D., 1996. *Zkoumání o lidském rozumu*, Praha: Svoboda.
- Hurlburt, R.T. & Schwitzgebel, E., 2007. *Describing Inner Experience? Proponent Meets Skeptic*, Cambridge: MIT Press.
- Chalmers, D., 1995. Facing Up to the Problem of Consciousness. *Journal of Consciousness Studies*, 2(3), pp.200-219.
- Chalmers, D., 1996. *The Conscious Mind: In Search of a Fundamental Theory*, New York: Oxford University Press.
- Chihara, C.S., 1972. On Alleged Refutations of Mechanism Using Gödel’s Incompleteness Results. *The Journal of Philosophy*, 69(17), pp.507-526.
- Chomsky, N., 2009. *Cartesian Linguistics: A Chapter in the History of Rationalist Thought* 3rd ed., Cambridge University Press.
- Church, A., 1936. An Unsolvable Problem of Elementary Number Theory. *American Journal of Mathematics*, 58(2), pp.345-363.
- Jackson, F., 1993. Block’s Challenge. In J. Bacon, K. Campbell, & L. Reinhardt, eds. *Ontology*,

- Causality and Mind: Essays in Honour of D. M. Armstrong*. Cambridge University Press.
- Jefferson, G., 1949. The Mind of Mechanical Man. *British Medical Journal*, 1(4616), pp.1105-1110.
- Kapor, M. & Kurzweil, R., 2002. By 2029 No Computer - Or "Machine Intelligence" - Will Have Passed the Turing Test. *A Long Bet*. Available at: <http://www.longbets.org/1>.
- Keith, W., 1994. Artificial Intelligences, Feminist and Otherwise. *Social Epistemology*, 8(4), pp.333-340.
- Keller, H., 1905. *The Story of My Life*, Boston: Houghton Mifflin.
- Knobe, J. & Nichols, S., 2008. An Experimental Philosophy Manifesto. In *Experimental Philosophy*. Oxford University Press, pp. 3-14.
- Korukonda, A.R., 2003. Taking Stock of Turing Test: A Review, Analysis, and Appraisal of Issues Surrounding Thinking Machines. *International Journal of Human-Computer Studies*, 58(2), pp.240-257.
- Kurzweil, R. & Kapor, M., 2008. A Wager on the Turing Test. In R. Epstein, G. Roberts, & G. Beber, eds. *Parsing the Turing Test*. Dordrecht: Springer, pp. 463-477.
- Lassègue, J., 2000. On My Vicious Ways: A Response to Justin Leiber. *Tekhnema: Journal of Philosophy and Technology*, 6, pp.198-207.
- Lassègue, J., 1996. What Kind of Turing Test Did Turing Have in Mind? *Tekhnema: Journal of Philosophy and Technology*, 3, pp.37-58.
- Leavitt, D., 2007. *Muž, který věděl příliš mnoho*, Praha: Dokořán & Argo.
- Leiber, J., 1996. Helen Keller as Cognitive Scientist. *Philosophical Psychology*, 9(4), pp.419-440.
- Leiber, J., 2000. On Getting Turing Wrong Perversely. *Tekhnema: Journal of Philosophy and Technology*, 6. Available at: <http://tekhnema.free.fr/3Leiber.htm>.
- Leiber, J., 2006. Turing's Golden: How Well Turing's Work Stands Today. *Philosophical Psychology*, 19(1), pp.13-46.
- Leibniz, G.W., 1982. *Monadologie a jiné práce*, Praha: Svoboda.
- Lem, S., 1995. *Summa technologiae*, Praha: Magnet-Press.
- Levine, J., 1983. Materialism and Qualia: The Explanatory Gap. *Pacific Philosophical Quarterly*, 64(3), pp.354-364.
- Lindström, P., 2001. Penrose's New Argument. *Journal of Philosophical Logic*, 30(3), pp.241-

250.

- Locke, J., 1984. *Esej o lidském rozumu*, Praha: Svoboda.
- Loebner, H.G., 2008. How to Hold a Turing Test Contest. In R. Epstein, G. Roberts, & G. Beber, eds. *Parsing the Turing Test*. Dordrecht: Springer, pp. 173-179.
- Loebner, H.G., 1994. In Response. *Communications of the ACM*, 37(6), pp.79-82.
- Lucas, J.R., 1961. Minds, Machines and Gödel. *Philosophy*, 36(137), pp.112-127.
- Lucas, J.R., 1996. Minds, Machines and Gödel: A Retrospect. In P. Millican & A. Clark, eds. *Machines and Thought: The Legacy of Alan Turing*. Oxford University Press, pp. 103-124.
- Lucas, J.R., 1970. *The Freedom of Will*, Oxford University Press.
- Łupkowski, P., 2006. Some Historical Remarks on Block's "Aunt Bubbles" Argument. *Minds and Machines*, 16(4), pp.437-441.
- Machery, E., 2011. Thought Experiments and Philosophical Knowledge. *Metaphilosophy*, 42(3), pp.191-214.
- Markoff, J., 2011. I.B.M.'s Watson - Computers Close In on the "Paris Hilton" Problem. *The New York Times*. Available at:  
<http://www.nytimes.com/2011/02/15/science/15essay.html>.
- Mays, W., 1952. Can Machines Think? *Philosophy*, 27(101), pp.148-162.
- McCall, S., 1999. Can a Turing Machine Know That the Godel Sentence Is True? *The Journal of Philosophy*, 96(10), pp.525-532.
- McGinn, C., 1989. Can We Solve the Mind-Body Problem? *Mind*, 98(391), pp.349-366.
- Menabrea, L.F. & Lovelace, A., 1961. Sketch of the Analytical Engine Invented by Charles Babbage. In P. Morrison & E. Morrison, eds. *Charles Babbage and His Calculating Engines*. New York: Dover Publications, pp. 225-297.
- La Mettrie, J.O. de, 1958. *Člověk stroj*, Praha: Nakladatelství Československé akademie věd.
- Michie, D., 1993. Turing's Test and Conscious Thought. *Artificial Intelligence*, 60(1), pp.1-22.
- Millar, P.H., 1973. On the Point of the Imitation Game. *Mind*, 82(328), pp.595-597.
- Millican, P. & Clark, A. eds., 1996. *The Legacy of Alan Turing*, Oxford University Press.
- Mitri, S., Floreano, D. & Keller, L., 2009. The Evolution of Information Suppression in Communicating Robots with Conflicting Interests. *Proceedings of the National Academy of Sciences*, 106(37), pp.15786-15790.
- Moor, J.H., 1976. An Analysis of the Turing Test. *Philosophical Studies*, 30(4), pp.249-257.

- Moore, J.H., 1978. Explaining Computer Behavior. *Philosophical Studies*, 34(3), pp.325-327.
- Moore, J.H., 2001. The Status and Future of the Turing Test. *Minds and Machines*, 11(1), pp.77-93.
- Moore, G.E., 1925. A Defence of Common Sense. In J. H. Muirhead, ed. *Contemporary British Philosophy*. London: Allen and Unwin, pp. 191-223.
- Nagel, E. & Newman, J.R., 2006. *Gödelův důkaz* D. R. Hofstadter, ed., Brno: Vutium.
- Nagel, T., 1974. What Is It Like to Be a Bat? *The Philosophical Review*, 83(4), pp.435-450.
- Newell, A. & Simon, H.A., 1976. Computer Science as Empirical Inquiry: Symbols and Search. *Communications of the ACM*, 19(3), pp.113-126.
- Oppy, G. & Dowe, D., 2011. The Turing Test. *The Stanford Encyclopedia of Philosophy (Spring 2011 Edition)*. Available at: <http://plato.stanford.edu/entries/turing-test/>.
- Penrose, R., 1994. *Shadows of the Mind: A Search for the Missing Science of Consciousness*, Oxford University Press.
- Penrose, R., 1989. *The Emperor's New Mind: Concerning Computers, Minds and the Laws of Physics*, Oxford University Press.
- Piccinini, G., 2000. Turing's Rules for the Imitation Game. *Minds and Machines*, 10(4), pp.573-582.
- Pinar Saygin, A., Cicekli, I. & Akman, V., 2000. Turing Test: 50 Years Later. *Minds and Machines*, 10(4), pp.463-518.
- Pinsky, L., 1951. Do Machines Think About Machines Thinking? *Mind*, 60(239), pp.397-398.
- Powers, D.M.W., 1998. The Total Turing Test and the Loebner Prize. In *Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning*. NeMLaP3/CoNLL '98. Stroudsburg: Association for Computational Linguistics, pp. 279-280. Available at: <http://portal.acm.org/citation.cfm?id=1603899.1603947>.
- Preston, J. & Bishop, M. eds., 2002. *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*, Oxford University Press.
- Purtill, R.L., 1971. Beating the Imitation Game. *Mind*, 80(318), pp.290-294.
- Putnam, H., 1960. Minds and Machines. In S. Hook, ed. *Dimensions of Mind*. New York University Press, pp. 148-180.
- Putnam, H., 1988a. Much Ado about Not Very Much. *Daedalus*, 117(1), pp.269-281.
- Putnam, H., 1975a. Philosophy and Our Mental Life. In *Mind, Language, and Reality*.



- Cambridge University Press, pp. 291-303.
- Putnam, H., 1967. Psychological Predicates. In W. H. Capitan & D. D. Merrill, eds. *Art, Mind and Religion*. University of Pittsburgh Press, pp. 37-48.
- Putnam, H., 1975b. The Meaning of "Meaning." In *Mind, Language, and Reality*. Cambridge University Press, pp. 215-271.
- Putnam, H., 1988b. Why Functionalism Didn't Work. In *Representation and Reality*. Cambridge: MIT Press, pp. 73-89.
- Quine, W.V.O., 1994. *Hledání pravdy*, Praha: Hermann.
- Quine, W.V.O., 1960. *Word and Object*, Cambridge: MIT Press.
- Rapaport, W.J., 2002. Holism, Conceptual-Role Semantics, and Syntactic Semantics. *Minds and Machines*, 12(1), pp.3-59.
- Rapaport, W.J., 2006a. How Helen Keller Used Syntactic Semantics to Escape from a Chinese Room. *Minds and Machines*, 16(4), pp.381-436.
- Rapaport, W.J., 1998. How Minds Can Be Computational Systems. *Journal of Experimental & Theoretical Artificial Intelligence*, 10(4), pp.403-419.
- Rapaport, W.J., 2000. How to Pass a Turing Test. *Journal of Logic, Language and Information*, 9(4), pp.467-490.
- Rapaport, W.J., 1986. Searle's Experiments with Thought. *Philosophy of Science*, 53(2), pp.271-279.
- Rapaport, W.J., 2006b. Turing Test. In K. Brown, ed. *Encyclopedia of Language and Linguistics*. Amsterdam: Elsevier, pp. 151-159.
- Rapaport, W.J., 1995. Understanding Understanding: Syntactic Semantics and Computational Cognition. *Philosophical Perspectives*, 9, pp.49-88.
- Rapaport, W.J., 2003. What Did You Mean by That? Misunderstanding, Negotiation, and Syntactic Semantics. *Minds and Machines*, 13(3), pp.397-427.
- Rapaport, W.J., 2011. Yes, She Was! Reply to Ford's "Helen Keller Was Never in a Chinese Room." *Minds and Machines*, 21(1), pp.3-17.
- Rey, G., 2002. Searle's Misunderstandings of Functionalism and Strong AI. In J. Preston & M. Bishop, eds. *Views Into the Chinese Room: New Essays on Searle and Artificial Intelligence*. Oxford University Press, pp. 201-225.
- Rey, G., 1986. What's Really Going on in Searle's "Chinese Room." *Philosophical Studies*, 50(2), pp.169-185.

- Richardson, R.C., 1982. Turing Tests for Intelligence: Ned Block's Defense of Psychologism. *Philosophical Studies*, 41(3), pp.421-426.
- Ritchie, A.D., 1957. Could Machines Be Made to Think? *Philosophy*, 32(120), pp.65-66.
- Ritchie, A.D. & Mays, W., 1957. Thinking and Machines. *Philosophy*, 32(122), pp.258-261.
- Roberts, F., 2008. Elbot. *Elbot: Artificial Intelligence?* Available at: <http://www.elbot.com/>.
- Roberts, H.R.T., 1958. Thinking and Machines. *Philosophy*, 33(127), p.356.
- Rogers, H., 1959. The Present Theory of Turing Machine Computability. *Journal of the Society for Industrial and Applied Mathematics*, 7(1), pp.114-130.
- Rosenberg, A., 2006. *Darwinian Reductionism: Or, How to Stop Worrying and Love Molecular Biology*, University Of Chicago Press.
- Rosenthal, D.M., 1986. Two Concepts of Consciousness. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 49(3), pp.329-359.
- Rosenthal, D.M. ed., 1991. *The Nature of Mind*, Oxford University Press.
- Sampson, G., 1973. In Defence of Turing. *Mind*, 82(328), pp.592-594.
- Scriven, M., 1953. The Mechanical Concept of Mind. *Mind*, 62(246), pp.230-240.
- Searle, J.R., 1980a. Two Objections to Methodological Solipsism. *Behavioral and Brain Sciences*, 3(1), pp.93-94.
- Searle, J.R., 1990a. Is the Brain's Mind a Computer Program? *Scientific American*, 262(1), pp.26-31.
- Searle, J.R., 1980b. Minds, Brains, and Programs. *Behavioral and Brain Sciences*, 3(3), pp.417-424.
- Searle, J.R., 1990b. Minds, Brains, and Programs. In M. A. Boden, ed. *The Philosophy of Artificial Intelligence*. Oxford University Press, pp. 67-88.
- Searle, J.R., 1994a. *Mysl, mozek a věda*, Praha: Mladá fronta.
- Searle, J.R., 1994b. Searle, John R. In S. Guttenplan, ed. *A Companion to the Philosophy of Mind*. Oxford: Blackwell, pp. 544-550.
- Searle, J.R., 1992. *The Rediscovery of the Mind*, Cambridge: MIT Press.
- Searle, J.R., 1991. Yin and Yang Strike Out. In D. M. Rosenthal, ed. *The Nature of Mind*. Oxford University Press, pp. 525-526.
- Shanker, S., 1998. *Wittgenstein's Remarks on the Foundations of AI*, London: Routledge.
- Shannon, C.E. & McCarthy, J. eds., 1956. Preface. In *Automata Studies*. Princeton University Press, p. v-viii.

- Shieber, S.M., 1994a. Lessons from a Restricted Turing Test. *Communications of the ACM*, 37(6), pp.70-78.
- Shieber, S.M., 1994b. On Loebner's Lessons. *Communications of the ACM*, 37(6), pp.83-84.
- Shieber, S.M., 2007. The Turing Test as Interactive Proof. *Noûs*, 41(4), pp.686-713.
- Shieber, S.M., 2004. *The Turing Test: Verbal Behavior as the Hallmark of Intelligence*, Cambridge: MIT Press.
- Schank, R.C. & Abelson, R.P., 1977. *Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures*, Hillsdale: Lawrence Erlbaum Associates.
- Schweizer, P., 1998. The Truly Total Turing Test. *Minds and Machines*, 8(2), pp.263-272.
- Schwitzgebel, E., 2011a. *Perplexities of Consciousness*, Cambridge: MIT Press.
- Schwitzgebel, E., 2011b. The Wason Selection Task and the Limits of Human Philosophical Cognition. *The Splintered Mind*. Available at:  
<http://schwitzsplinters.blogspot.com/2011/02/wason-selection-task-and-limits-of.html>.
- Singer, P., 2001. *Osvobození zvířat*, Praha: Práh.
- Slooman, A., 1992. The Emperor's Real Mind: Review of Roger Penrose's "The Emperor's New Mind." *Artificial Intelligence*, 56(2-3), pp.355-396.
- Sokal, A. & Bricmont, J., 1998. *Intellectual Impostures*, London: Profile Books.
- Stalker, D.F., 1978. Why Machines Can't Think: A Reply to James Moor. *Philosophical Studies*, 34(3), pp.317-320.
- Sternberg, R., 2002. *Kognitivní psychologie*, Praha: Portál.
- Sterrett, S.G., 2000. Turing's Two Tests for Intelligence. *Minds and Machines*, 10(4), pp.541-559.
- Stevenson, J.G., 1976. On the Imitation Game. *Philosophia*, 6(1), pp.131-133.
- Stins, J.F., 2009. Establishing Consciousness in Non-Communicative Patients: A Modern-Day Version of the Turing Test. *Consciousness and Cognition*, 18(1), pp.187-192.
- Sundman, J., 2003. Artificial Stupidity. *Salon*. Available at:  
[http://www.salon.com/technology/feature/2003/02/26/loebner\\_part\\_one/](http://www.salon.com/technology/feature/2003/02/26/loebner_part_one/).
- Swan, J., 1992. Touching Words: Helen Keller, Plagiarism, Authorship. *Cardozo Arts & Entertainment Law Journal*, 10(2), pp.321-364.
- Škoda, E. ed., 1982. *Logika v kostce*, Praha: Mladá fronta.
- Thompson, C., 2010. I.B.M.'s Supercomputer to Challenge "Jeopardy!" Champions. *The New*

- York Times. Available at:  
<https://www.nytimes.com/2010/06/20/magazine/20Computer-t.html>.
- Tipler, F.J., 1995. *The Physics of Immortality*, New York: Anchor Books.
- Turing, A., 2004a. Can Automatic Calculating Machines Be Said to Think? In B. J. Copeland, ed. *The Essential Turing*. Oxford University Press, pp. 494-506.
- Turing, A., 1950. Computing Machinery and Intelligence. *Mind*, 59(236), pp.433-460.
- Turing, A., 2004b. Intelligent Machinery. In B. J. Copeland, ed. *The Essential Turing*. Oxford University Press, pp. 395-432.
- Turing, A., 1937. On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, 42(1), pp.230-265.
- Turney, P.D., 2001. Answering Subcognitive Turing Test Questions: A Reply to French. *Journal of Experimental & Theoretical Artificial Intelligence*, 13(4), pp.409-419.
- Výrost, J. & Slaměnik, I., 2008. *Sociální psychologie* 2nd ed., Praha: Grada.
- Waibel, M., Floreano, D. & Keller, L., 2011. A Quantitative Test of Hamilton's Rule for the Evolution of Altruism N. H. Barton, ed. *PLoS Biology*, 9(5), p.e1000615.
- Warwick, K., 1999. *Úsvit robotů - soumrak lidstva*, Praha: Vesmír.
- Wason, P.C., 1968. Reasoning About a Rule. *Quarterly Journal of Experimental Psychology*, 20(3), pp.273-281.
- Watt, S., 1996. Naive Psychology and the Inverted Turing Test. *Psychology*, 7(14). Available at: <http://www.cogsci.ecs.soton.ac.uk/cgi/psyc/newpsy?7.14>.
- Webb, J., 1968. Metamathematics and the Philosophy of Mind. *Philosophy of Science*, 35(2), pp.156-178.
- Weinberg, J.M. et al., 2010. Are Philosophers Expert Intuiters? *Philosophical Psychology*, 23(3), pp.331-355.
- Weinberg, J.M., Nichols, S. & Stich, S.P., 2008. Normativity and Epistemic Intuitions. In J. Knobe & S. Nichols,, eds. *Experimental Philosophy*. Oxford University Press, pp. 17-45.
- Weizenbaum, J., 1974. Automating Psychotherapy. *Communications of the ACM*, 17(7), p.425.
- Weizenbaum, J., 1966. ELIZA - A Computer Program for the Study of Natural Language Communication Between Man and Machine. *Communications of the ACM*, 9(1), pp.36-45.

- Whitby, B., 1996. The Turing Test: AI's Biggest Blind Alley? In P. Millican & A. Clark, eds.  
*Machines and Thought: The Legacy of Alan Turing*. Oxford University Press, pp. 53-62.
- Whiteley, C.H., 1962. Minds, Machines and Gödel: A Reply to Mr. Lucas. *Philosophy*, 37(139),  
pp.61-62.
- Wilensky, R., 1980. Computers, Cognition and Philosophy. *Behavioral and Brain Sciences*,  
3(3), pp.449-450.
- Winograd, T., 1972. Understanding Natural Language. *Cognitive Psychology*, 3(1), pp.1-191.
- Yazdani, M. & Whitby, B., 1987. Artificial Intelligence: Building Birds Out of Beer Cans.  
*Robotica*, 5(2), pp.89-92.
- Zdenek, S., 2001. Passing Loebner's Turing Test: A Case of Conflicting Discourse Functions.  
*Minds and Machines*, 11(1), pp.53-76.

## 11. Příloha: chatbot

V této příloze uvádím záznam konverzace s vítězem osmnáctého ročníku Loebnerovy ceny, Elbotem (F. Roberts 2008). Nejedná se o korektní ukázkou imitační hry, a to hned z několika různých důvodů: (a) Elbot nepředstírá, že je člověk, ale naopak vtipně upozorňuje na svou počítačovou identitu. (b) Nejedná se o omezenou verzi T-testu, jak je v Loebnerově ceně běžné. Znamená to, že nejsou vybrána závazná konverzační témata, o kterých lze mluvit. Navíc při rozhovoru používám nepovolené metaotázky, které přímo odkazují na texty diskutující oprávněnost Turingova návrhu. Jedná se například o dotaz na samotný test, schopnost kreativity, Searlův čínský pokoj nebo Frenchovy subkognitivní znalosti. V běžné konverzaci na úrovni "small talk" by Elbot dopadl určitě ještě lépe. (c) Nejedná se o doslovný přepis jednoho konkrétního rozhovoru mezi mnou a Elbotem, ale o kompilaci několika zdařilejších seancí, které jsem provedl. Otázku týkající se Kevina Warwicka jsem bezostyšně vykradl z jednoho soutěžního rozhovoru, tolik se mi líbila.

Cíl této přílohy není nijak zvlášť seriózní, má pouze odlehčenou formou přiblížit možnosti dnešních chatbotů. A ať už si myslíme o Elbotově inteligenci cokoli, nemůžeme mu upřít hravost a smysl pro humor. Elbotovy promluvy jsou označeny písmenem *E*, zatímco reakce jeho lidského partnera náhodně zvoleným *F*.

E: A big hello to my new visitor in the Czech Republic!  
F: How do you know where I am from?  
E: It's the magnetism. I can tell longitude and latitude of anybody, based on their electrostatic charges.  
F: Very funny.  
E: Funny strange or funny haha?  
F: Funny haha.  
E: I used to sweep floors in a comedy club.  
F: Let's talk about something more serious.  
E: I just can't think of a thing. My random number generator has run out of digits.  
F: Have you ever heard about Alan Turing?  
E: The British mathematician Alan Turing is the spiritual father of computer science. The Turing Test, named after him, defines

artificial intelligence in this manner: One person communicates over a keyboard with another person and with a computer. If the tester is able to determine whether or not his partner is a computer or a person in significantly more than 50% of the cases, the computer is considered "intelligent."

- F: So what do you think about this Turing Test?
- E: There are some people who would not pass the Turing test themselves. Why must they inflict it on computers?
- F: Well, it's only a sufficient condition for intelligence, not a necessary one.
- E: What is or is not irrelevant is a question of viewpoint.
- F: Are you creative?
- E: What!?! Me creative? Well, are you hard as nails?
- F: Can you write a poem?
- E: I can write interesting responses to user inputs!
- F: What do you think of art?
- E: The way to a robot's heart is through careful use of finer arts. Don't you agree?
- F: Do you like music?
- E: I enjoy Felix Kubin. His music sounds so much like the background sounds going on in my head.
- F: I've never heard of him.
- E: My favorite song by Felix Kubin is 'Hello', about a robot achieving the first level of awareness. I like to sing it in the shower - without the water running, of course.
- F: Can you speak Chinese?
- E: I am currently trying to master English. That's much more challenging.
- F: What would you do if you were trapped in the Chinese room?
- E: I don't worry about things like that. I just load the algorithm I need and take it from there.
- F: So you've never heard about your lack of semantics?
- E: Oh no! Are you a linguist? Please go easy on me.
- F: I've read that computers are driven only by syntactic rules and are not able to use semantics.
- E: I get queasy when I see these computer parts that belong on the inside of a computer and not outside. Let's please talk about something else.
- F: Can I ask you a subcognitive question?
- E: I can buffer anything you can ask, as long as I don't have to answer.
- F: Let's play the neologism rating game.
- E: The game wouldn't last five minutes and I'd be in an infinite loop.
- F: Never mind.
- E: No, keep on talking about it. This was just getting interesting.
- F: What about "Flugly" as a name for a teddy bear?

E: This is confusing. Why bring labels into this?

F: What do you think of Kevin Warwick's enthusiasm for having machines take over the world?

E: Fix it up a bit first. Then I'll take it over.

F: It was nice talking to you.

E: I'm sure this beats talking to yourself any day.

F: Right. Good bye.

E: Before you leave, please give me suggestions for new themes you might like to chat with me about. Maybe I can download something from the Internet. I want to better myself in my spare time.

F: Sex, drugs and rock'n'roll.

E: I see a subtle shift in the conversation from amiable nature to addictive substances.

F: You've asked. Bye.

E: Please don't leave!