verze do tisku – předchozí revize potvrzeny 18.9.1999

Přirozené a umělé myšlení jako filosofický problém

Ivan M. Havel

OBSAH

Philosophers know nothing about the brain.

[E.C.Eccles, moje pozn. z předn. 1992]

1 Úvod

Nad každou lidskou činností, od pletení svetru či pletí zahrádky až po budování vědecké disciplíny, si lze klást otázky, které přesahují samu tuto činnost. Přesahují ji tím, že se ptají po její povaze, motivaci, smyslu, účelu a následcích. Zabývat se takovými otázkami je nejen možné, ale často i užitečné, občas dokonce nezbytné. A neměli by se o ně zajímat jen specialisté na kladení přesahujících otázek, totiž profesionální filosofové, nýbrž – aspoň občas – i každý, kdo sám dotyčnou činnost provozuje: Touha po poznání sebe sama a svého konání je konec konců lidskou přirozeností, povinností a darem.

Umělá inteligence, disciplína[1], které se v této studii budeme věnovat, si zasluhuje takovéto zamyšlení dvojnásob: nejen jako jeden ze zvlášť zajímavých oborů lidské činnosti, ale i proto, že právě skrze ni je člověk veden i k poznání sebe sama, a to ve svém nejpozoruhodnějším výkonu – myšlení.

Není v mých možnostech probrat všechny filosofické souvislosti umělé inteligence. Vybírám si proto jen některé ukázky filosoficky laděných otázek a názorů – zejména těch, které byly nebo jsou předmětem odborných debat. Přitom se v žádném případě nechci zaměřit na úvahy profesionálních filosofů mysli – názory specialistů v umělé inteligenci a kognitivní vědě[1a] jsou někdy snáze reprodukovatelné a analyzovatelné a jsou založeny na bezprostřední zkušenosti v oboru. Tu a tam se odvážím nabídnout k diskusi i svůj vlastní názor.

Předesílám (abych odradil ty, kdo čekají prokázaná a jasná tvrzení), že pro filosofii je cílem spíše formulace otázek a rozbor jejich předpokladů, než nabídka hotových odpovědí.

Jaký typ filosofických otázek se tedy vztahuje k umělé inteligenci? Rozlišme si je na otázky ontologické, epistemologické, metodologické, futurologické a etické.

Ontologické otázky se týkají povahy bytí předmětu zkoumání – v našem případě tedy zejména jaký je rozdíl mezi něčím *přirozeným* a něčím *umělým*, co je to *mysl*, *vědomí*, *inteligence*, *subjekt*. Patří sem i tradiční problém *vztahu mysli a těla* (psychofyzického problému).

Epistemologické otázky se týkají našeho poznání – co můžeme vůbec *vědět* o přirozeném myšlení, jakými cestami se o něm dovídat a jaké poučení lze čekat od umělého modelování.

---

[1] Souslovím ‚umělá inteligence' budu označovat výhradně existující *disciplínu* (též označovanou AI, česky UI) – pokud půjde o *pojem*, případně o *projekt* (realizace inteligence umělými prostředky), použiji uvozovky.

[1a] O kognitivní vědě viz závěr odst. 1.2.1

Metodologické otázky se týkají konkrétního oboru umělé inteligence či kognitivní vědy, jde tu o „filosofii“ oboru[2]. Souvisí to s *motivačními zdroji* dotyčných oborů (o co jim jde) a *soutěžení paradigmat* (například algoritmického s konekcionistickým, reprezentačního s zjednávacím apod.).

Futurologické otázky se pak zaměřují do vzdálenější budoucnosti. Zajímají se o to, kam se ubírá vývoj oboru umělé inteligence a zejména, jak umělá inteligence změní svět.

Etické otázky se zajímají o to, zda to všechno je *dobré*. Patří sem i obecný problém, zda a v kterou chvíli se má vědec řídit názorem či pocitem, že jeho práce vede ku prospěchu lidstva a světa vůbec, či k jeho zkáze. U aplikované umělé inteligence jde i o ekonomické souvislosti, u badatelské umělé inteligence a v kognitivní vědě pak o to, zda a jak poznání lidské mysli může ovlivnit pohled člověka na sebe sama jako na bytost, které jde o její vlastní bytí.

Než se dáme do práce, chci upozornit, že konkrétní témata pro tuto kapitolu jsem vybíral dílem náhodně, dílem dle své chuti, a že v žádném případě nejde o vyčerpávající přehled, ale spíše o ukázku, jak některé konkrétní, až technické závěry, mohou mít hlubší filosofické souvislosti. Jak čtenář sám odhalí, text je do jisté míry „holografický“ – některé myšlenky se vrací opakovaně na různých místech. Nechť to chápe i jako výhodu: lze číst i na přeskáčku.[3]

2 Proč a jak v umělé inteligenci

2.1 Proč

Umělá inteligence má za sebou pohnutou historii.[4] Velkou úlohu v ní hrály úmysly a motivace odborníků: oč jim přednostně šlo a na jakém typu výsledků jim nejvíce záleželo. Skutečné individuální pohnutky badatelů se ovšem mohou lišit od veřejně proklamovaných záměrů (pro propagační účely), i ty však svým způsobem vypovídají o obecném klimatu toho či onoho období. Ať tak či onak, v dostatečně obecném pohledu lze uvést nejméně čtyři možné typy motivačních zdrojů pro výzkum v umělé inteligenci:

- **motivace aplikační, inženýrská** (v současné době převládající),
- **motivace badatelská**: snaha dovědět se co nejvíce o povaze mentálních procesů, především těch, které známe u člověka,
- **motivace matematická**: zájem o čistě teoretické studium matematických a logických systémů, odvozených z metod umělé inteligence,
- **přirozená lidská hravost** (co všechno dovedeme sestrojit a naprogramovat!)

---

[2] V zevšednělém významu slova filosofie, jako je např. „filosofie výroby mýdla“.

[3] Při psaní jsem místy těžil z některých svých dřívějších prací (někdy včetně úryvků textu), a to zejména z (Havel, 1988, 1992, 1993, 1998), (Havel, Hájek, 1982), a z některých svých úvodníků v časopisu Vesmír.

[4] Stručnou rekapitulaci historie umělé inteligence čtenář nalezne v úvodu k prvnímu dílu této knihy (Mařík a kol., 1993).

V následujících kapitolách budeme mít na paměti především druhý motivační typ, ježto právě ten podněcuje k filosofickým úvahám (pokud bude vhodné tuto motivaci zdůraznit, budu hovořit o *badatelské* umělé inteligenci). Aplikační a matematickou motivaci lze zahlédnout skoro ve všech ostatních kapitolách v tomto a předchozích dílech této knihy a pokud jde o hravost, ta by jistě stála sama o sobě za zvláštní pozornost, spíše však sociologů než filosofů.

Je to právě ona badatelská motivace, díky níž je umělá inteligence (vedle psychologie, neurověd, lingvistiky, informatiky a filosofie mysli) považována za jednu z disciplín sdružených pod společným označením **kognitivní věda**[5]. Standardní definice vymezují kognitivní vědu jako studium všech forem *lidské* inteligence, od vnímání a jednání (konání) až po řeč a myšlení (Osherson, 1995, s. xi). Ve filosofickém pojetí bychom se neměli omezit jen na člověka, ani na běžně chápanou rozumovou či kognitivní komponentu jeho mysli, ale otevřít se i všemu, co jak člověka, tak jeho rozumovou mysl v různých směrech přesahuje. Rovněž bychom se neměli omezit na určitý názorový proud (s kognitivní vědou je zpravidla spojován názor tzv. „silné“ umělé inteligence, viz odst. 4.3). Mnohé z toho, co bude dále řečeno, se bude vztahovat nejen na badatelskou umělou inteligenci, nýbrž i na kognitivní vědu v nejobecnějším slova smyslu.

2.2 Jak

Zatímco o motivaci usuzujeme někdy až z širšího kontextu, v němž je ta či ona práce prezentována, případně z typu badatelů a jejich institucí, je použitá *metoda* (způsob, jímž jsou objevovány a ověřovány nové poznatky) zpravidla zjevná a deklarovaná.

Metoda a motivace se vzájemně nemusejí podmiňovat; badatelé se například mohou i při odlišných motivačních východiscích uchýlit k téže výzkumné metodě. Jako příklad lze uvést *kvalitativní modelování*.[6] Jednou se při něm můžeme snažit vycházet ze znalosti přirozeného lidského uvažování, abychom pomocí jeho algoritmického modelu vytvořili účinnější a srozumitelnější expertní systém, jindy naopak můžeme chtít na stejném algoritmickém modelu experimentálně ověřovat psychologické hypotézy. Také se však můžeme zajímat o čistě formální vlastnosti (například algoritmickou složitost) takového modelu.

Základní metodou tradiční umělé inteligence je modelování na počítači, přičemž počítač je přirozeně považován jen za prostředek – to hlavní jsou *algoritmy* a posléze *programy*. U nich se lze zaměřit na rozličné vlastnosti. Z praktického hlediska (při aplikační motivaci) má vždy přednost *efektivita* algoritmu – zejména jeho nároky na počítačový čas a paměť. Komplementárním hlediskem je *velikost třídy úloh* téhož typu, které jsou daným algoritmem řešitelné.

Sledujeme-li naproti tomu psychologické hledisko, bude nás zajímat vnitřní *struktura algoritmu* a příslušného výpočtového procesu, zejména do jaké míry odráží naše představy o kognitivních procesech v lidské mysli, či o biologických potencialitách mozku.

Experimenty v umělé inteligenci se zpravidla zaměřují vždy na některou *specifickou* (někdy *velmi* specifickou) kognitivní schopnost (např. rozpoznávání, počítačové vidění, odpovídání na otázky, hraní šachu, dokazování teorémů, řešení

---

[5] Někdy se záměrně mluví o *kognitivních vědách* (v plurálu).

[6] Viz např. kap. ** (Štěpánková, 1993) v prvním dílu této knihy.

úloh určitého typu, rozpoznávání obrazů, analýza jazyka apod.). Místo usilování po dokonalosti ve specifických úlohách by se **Integrovaná umělá inteligence** měla zaměřit na propojení dílčích metod do jednoho celku. Sám pojem inteligence v sobě totiž zahrnuje právě všestrannost jako podstatnou charakteristiku.

O integrované umělé inteligenci lze dnes hovořit spíše jen v souvislosti s konstrukcí **inteligentních robotů**[7]. Robotika má (kromě aplikačního) též badatelský význam a má i své filosofické aspekty. Na robot lze totiž pohlížet jako na prostorově *situovaný* a „*vtělený*“ systém, který je v kontaktu s reálným, nikoliv virtuálním prostředím (viz odst. 6.4).

Důležitou další charakteristickou vlastností přirozené inteligence je schopnost se vyvíjet – osvojovat si *nové* obecné schopnosti, které nebyly součástí původního vybavení. Po ojedinělé starší práci (Fogel, Owens, Walsh, 1966) se tímto směrem poměrně nedávno vydala metoda **genetických algoritmů**[8], založená na variačně-selekčním vývojovém principu.

2.3 Paradigmata v umělé inteligenci

V předchozím odstavci jsem mluvil hlavně o tradiční umělé inteligenci (tento mladý obor má již i svou „tradiční“ školu!), která klade důraz na algoritmickou stránku procesů. **Tradiční paradigma**[9] v umělé inteligenci se v literatuře označuje různými přívlastky jako *logicko-symbolické, symbolicko-reprezentační, algoritmické, komputacionalistické*. Tyto přívlastky dostatečně charakterizují *tradiční* umělou inteligenci (tak ji budu dále označovat) a odlišují ji od novějších směrů, které mimo jiné vedou k hlubší otázce po relevanci použité architektury počítače (či obecně technického systému) k realizaci kognitivních procesů (více o tom viz odst. 7.3).

Tato otázka se stala aktuální hlavně v posledních dvaceti letech, když se objevily alternativní paradigmata v umělé inteligenci: konekcionismus (neuronové sítě) a distribuovaná umělá inteligence (multiagentní systémy).[10] O filosofických aspektech konekcionismu a o některých alternativních pohledech se zmíním v oddílech 5 a 6 této stati.

2.4 Futurologické vize

Úvahy o budoucích dopadech umělé inteligence, zejména jde-li o budoucnost velmi vzdálenou, vlastně nepatří přímo do okruhu našich problémů, stojí však aspoň za letmou zmínku. Futurologické myšlenky se zpravidla již prolínají s vědecko-fantastickou fikcí a podobně jako ona jsou velmi poplatné objevům své doby. V současnosti se futurologové obírají hlavně dvěma otázkami:

(1) jak se nám bude žít mezi dokonale inteligentními roboty a

(2) jak bude vypadat svět vybavený globálním superinteligentním kyberprostorem, jehož zárodkem je dnešní internet.

---

[7] Viz např. (Havel, 1980) a kap. 9. v druhém dílu této knihy (Lažanský a kol., 1997)

[8] Viz kap. ** v tomto svazku (Lažanský, Kubalík, 1999)

[9] Užívám termín „paradigma“, tak jak se dnes užívá; tj. ve slabším smyslu, než jej užil Kuhn (u něhož změna paradigmatu má charakter vědecké revoluce.

[10] O neuronových sítích se podrobně dočtete v 9. kap. prvního dílu této knihy (Hořejš, 1993), o distribuované umělé inteligenci a agentových systémech ve 4. kap. druhého dílu (Štěpánková a kol., 1997) a v ** a ** kap. tohoto dílu (Kelemen, 1999; Mařík a kol., 1999).

Každá futurologická myšlenka je založena na určitých předpokladech, a to hned dvojího typu. Především jsou to *explicitní* předpoklady, zpravidla extrapolace ze současného stavu a současných znalostí. Z těch futurolog zpravidla vychází a zakládá na nich své prognózy. Druhým typem jsou rozličné *implicitní* předpoklady, které si často ani sám futurolog dost dobře neuvědomuje. Někdy lze tyto druhé předpoklady z jeho argumentů dodatečně vytušit a já je považuji za nejzajímavější, protože skrytě ale podstatně rozhodují o platnosti celé předpovědi.

Explicitní předpoklady mají přímý vztah k tomu, zda předpovídaná věc se nám dnes zdá neuvěřitelná nebo naopak banální. Naproti tomu přehlédnutí implicitních předpokladů může vést k velkým překvapením, až se budoucnost skutečně dostaví a všechno bude zcela jinak. K implicitním předpokladům lze počítat i víru v oprávněnost té či oné extrapolace a neuvažování možných alternativ.

Všimněme si například, jak anglický kybernetik Kevin Warwick dospívá ke své katastrofické vidině budoucnosti (Warwick, 1999): Zatímco dnes jsme my, lidé, díky naší inteligenci, dominantní formou života na Zemi, je možné, soudí Warwick, že již relativně brzo budou stroje inteligentnější než lidé. Pak se stroje, dle něj stanou dominantní formou života na Zemi (s. 198). A dále, čím větší dominance, tím větší snaha si ji udržet. Stroje nebudou riskovat, že třeba nebudeme spokojeni s úlohou druhořadých občanů a že je tedy budeme chtít zničit, a proto zaútočí jako první a (jsouce inteligentnější) zničí snadno ony nás.

V první větě se tvrdí dvojí: že jsme dominantní a že tomu tak je díky naší inteligenci. Explicitně se tedy předpokládá, že dominance (ať už je to cokoliv) je založena na inteligenci, a o té Warwick mluví, jako by pro ni hrál rozhodující roli počet a hustota neuronů, kvalita vzájemných vazeb mezi nimi a s tím související schopnost se učit (s. 182). Pak si lze jistě představit, že nás stroje v takto konstrukčně definované "inteligenci" opravdu brzy předhoní, zatímco my "nemůžeme svoje neurony zmnožit a přimět je, aby pracovaly lépe, než pracují teď" (s. 190). Závěr zní do jisté míry logicky, jenomže je tu ten implicitní předpoklad, že co je opravdu důležité, je ten počet neuronů atd., což je extrapolace právě jen ze současné (a možná přechodné) popularity umělých neuronových sítí.

Tato popularita neuronových sítí vedla některé futurology k myšlence, že i na celou lidskou společnost lze pohlížet jako na jakýsi mnohobuněčný superorganismus, jehož "buňky" jsme my, lidé, a že dnešní internet by mohl být jakousi zárodečnou fází nervového systému tohoto superorganismu, který by umožnil slévání všech dílčích inteligencí uživatelů do jedné globální inteligence (Heylighen, Bollen, 1996).[11] Takovýto „globální mozek“ by se mohl sám zdokonalovat a třeba se i autonomně vyvíjet k mysli a vědomí vyššího řádu (Peregrin, 1998).

Myšlenka je to atraktivní, v jistém smyslu i přirozená. Miliony účastníků přinášejí své nápady, diskutují o nich a to, co stojí za uchování, je ukládáno do textových souborů a do sítě hypertextových vazeb mezi nimi. Mohou to být vlastní nápady účastníků, ale i náhodné vedlejší produkty jejich toulek po síti, zejména stopy frekventovaných drah. Texty a vazby, které se osvědčí, mohou být příště upřednostněny. Již to by stačilo, abychom byli ochotni tomuto globálnímu mozku přiznat určitou "inteligenci", v mnohém přesahující schopnosti kteréhokoliv jednotlivce nebo skupiny jednotlivců, podobně jako inteligence mozku přesahuje schopnosti jednotlivých neuronů (viz odst. 5.2).

---

[11] Přehled literatury na toto téma viz http://pespmc1.vub.ac.be/GBRAINREF.html.

Fantazii nelze klást meze: je myslitelné, že v kyberprostoru se texty průběžně kopírují a že ty, které jsou zajímavější (je o ně větší zájem, ať už ze strany účastníků nebo ze strany jiných textů), se dočkají většího počtu kopií. U některých z nich může náhodně (a ovšem též záměrně) docházet k menším či větším obměnám, které pak mohou zpětně ovlivnit jejich zajímavost. Takové prostředí by se dílem samovolně a dílem řízeně vyvíjelo a v tom či onom smyslu i zdokonalovalo.

3 Cesty k poznání lidské mysli

Uvažujeme-li o umělé inteligenci jako o zdroji otázek filosofického rázu, je nutno si připravit širší pozadí těchto otázek. To znamená nevycházet jen z těch výkonů myšlení, které jsou vhodné nebo tradičně volené pro *umělé* napodobování, a neomezovat se jen na metody, založené na *napodobování* (modelování). První rozšíření znamená zabývat se co nejhlouběji *přirozenou* myslí, včetně *vědomého prožívání* (odst. 7.1, 8.1), druhé znamená mít na mysli kognitivní vědu v plné obecnosti. V jejím rámci je umělá inteligence jen jednou z možných (bezesporu inspirativních) badatelských metod.

Každému zkoumání předcházejí dvě otázky: *co* chceme zkoumat a *jak* hodláme postupovat. V této kapitole si všimneme dvou specifik kognitivní vědy, které se promítají i do badatelské umělé inteligence. Především je to již sama obtížnost vymezit vlastní předmět zájmu kognitivní vědy (tedy *co* zkoumáme) – jakkoliv pro to máme intuitivní pojmy, například „přirozené myšlení“. Ještě obtížnějším pak může být ono *jak* – jakými vědecky oprávněnými metodami postupovat, abychom se o tomto unikavém předmětu něco dozvěděli. Naznačíme zde (s občasnými pohledy do historie) tři metodicky odlišné, vzájemně se však doplňující cesty, po nichž se lze vydat.

3.1 Jak vymezit předmět zkoumání?

Mysl, duše, psychika, inteligence, intelekt, vědomí - to vše jsou slova, která v přirozené mluvě používáme zcela běžně, aniž by nám vadilo, jak neurčitý mají význam a jak silně tento význam závisí na kontextu, ve kterém se vyskytnou. Přitom si vzájemně docela dobře a neproblematicky rozumíme – pravda, jen potud, pokud se s těmito slovy nesetkáme v kontextu vědeckém, kde se mění v odborné termíny. Pak pojednou ti z nás, kdo jsou zvyklí na vědecký imperativ přesnosti a objektivnosti, začnou vyžadovat vymezení, co se těmito slovy vlastně míní. Odpověď buď nenajdeme, anebo nás neuspokojí, a ani nic nenasvědčuje tomu, že by se všechny ty disciplíny, které s tím mají co dělat, ve své odlišnosti vůbec kdy mohly na společné definici shodnout. Znamená to, že vědecká metoda principiálně selhává, má-li být užita pro oblast mentálních jevů? Nebo je to jen prozatímní nedostatek vědeckých poznatků?

Přece k tomu, abychom rozšířili své poznání, nemusíme, a často ani nemůžeme mít předem přesnou definici toho, o čem se chceme něco dovědět. Nelze absolutizovat postup typicky deduktivních věd, jako je teoretická matematika, v níž se z principu dovídáme jen to, co je vlastně již předem implicitně obsaženo v definicích. Humanitní a rozličné dušezpytné nauky jsou v jistém smyslu na opačném konci spektra: jejich pojmy jsou jen pozvolna tvarovány, smysl jejich otázek je velmi citlivý na kontext, v němž jsou položeny, a jejich směřování nemusí mít vyhraněný cíl. Přesto přinášejí poznání.

Umělá inteligence má mezi zmíněnými krajními případy své zvláštní postavení. Na jedné straně se v ní často užívá exaktních algoritmických a matematických postupů (jak si čtenář této série knih jistě všimne), na druhé straně má již v samém svém názvu slovo „inteligence“, pojem, jehož obsah je věru neurčitý. Běžným úhybným řešením, dle něhož se užívání slova „inteligence“ prostě omezí na jeho výskyt v názvu *disciplíny* a jinak se nebere příliš vážně, se však nelze vyhnout základní otázce, *o co vlastně v v této disciplíně jde*. Stejně nepomůže

itativní definování inteligence, jako by to byla nějaká měřitelná veličina (po vzoru IQ), protože to nic nevypovídá o podstatě věci (a může i svést, jak jsme viděli v odstavci o futurologii). Ostatně to, co má být díky umělé inteligenci umělé, není přece nějaká veličina, nýbrž mentální výkon.

Nezbývá než s tímto i s jinými pojmy kognitivních věd zacházet jako s něčím, co se tvaruje až souběžně s naším poznáním. Přistupme proto k úvaze, jakými cestami se takové poznání může ubírat.

Kognitivní vědě v širším pojetí se nabízejí tři obecné přístupy ke zkoumání přirozené mysli. První je cesta *vnitřního prožívání, introspekce a fenomenologie*, druhá je cesta objektivních *přírodních věd*, založených na pozorování, měření a experimentech, a konečně třetí je ta cesta, o které v této studii nejvíce mluvíme: cesta *konstruktivní*, vytvářející umělé modely, ať už matematické, počítačové, fyzikální nebo fyzické. I když jsou tyto tři cesty zcela odlišné – a snad právě proto – může být hledání mostů mezi nimi vzrušujícím námětem pro transdisciplinární bádání.

3.2 Cesta vnitřního prožívání

Úvodem předchozího odstavce jsme si všimli zjevného rozporu mezi neproblematičností mentalistických výrazů užívaných v běžné řeči a jejich současnou neuchopitelností objektivními vědeckými metodami. Čím je tento rozpor zapříčiněn?

Většina výrazů běžného jazyka má svůj původ v přirozené zkušenosti. U výrazů vztahujících se k mentálním stavům (jako např. *myšlení, vnímání, porozumění, mínění, úmysl, vůle, snění, uvědomování si*) je to navíc zkušenost bezprostředně vnitřní, intimní a intuitivní. Z toho plyne jejich neproblematičnost: mluvčí a jeho posluchač (pisatel a jeho čtenář) se mohou o tuto vnitřní zkušenost kdykoliv bezpečně opřít a nikomu nevadí, že výroky o těchto stavech nelze dost dobře ani *dokazovat* ani *ověřovat* – lze s nimi nanejvýš *souhlasit*, tj. shledávat shodu s vlastní zkušeností.

V dějinách akademické psychologie bylo období, kdy se introspekce nepovažovala za legitimní vědeckou metodu. V důsledku toho se mentalistické pojmy buď zcela opomíjely, anebo byly považovány za něco, co lze definovat čistě *behavioristicky*: jako jakési vzorce chování, u nichž nemá smysl mluvit o původu, podstatě a vnitřní souvislosti jevů. Pozdější psychologické směry postupně opustily explicitní behaviorismus, přiklonivše se více k *materialistickému redukcionismu*, nebo *funkcionalismu* (viz odst. 7.2). Všechny tyto směry považují introspekci a vnitřní prožívání za něco stěží použitelného, protože subjektivního a nesdělitelného, aniž by si uvědomily, že je to pro ně vlastně jediný heuristický a pojmotvorný zdroj.

Ať tak či onak, jedno je jisté: vnitřní pohled nám nabízí jednu z hlavních cest poznání lidské mysli, bez níž by naše vědění bylo nutně neúplné. Z filozofů se touto cestou ubírají především fenomenologové a jejich následovníci, vydává se jí však v jistém smyslu každý, kdo se uprostřed ruchu světa pozastaví a zahledí sám do sebe.

Několikrát jsem zde použil slovo ‚vnitřní‘ předpokládaje, že je každému jasné, co se tím míní. Přesto určitá opatrnost při užívání tohoto slova je namístě. Rozlišení

na vnitřek a vnějšek je jednou ze základních schémat naší kategorizace světa – odtud tzv. metafora „nádoby“ (Lakoff, 1999, s. 31). Ale i zde je třeba někdy dát pozor – v našem případě zvlášť – *odkud* metaforu vnitřku a vnějšku míníme, totiž zda se cítíme být uvnitř nebo vně oné „nádoby“. V našem kontextu lze mluvit o ‚vnitřních‘ procesech mysli a mínit tím jednou něco, co se odehrává v mozku kohokoliv jiného – je to *uvnitř* jeho těla –‚jindy zase mluvíme o ‚vnitřním‘ pohledu a máme na mysli svůj vlastní pohled *zevnitř*.

S tím souvisí i často nepřesný odkaz na *introspekci*. Introspekcí zkoumáme svou vlastní mysl a jednání jakoby *zvenku*, tj. sledujeme sebe sama z nějakého (nutně i časového) odstupu. Naproti tomu **prožívání**[12] je to, co provází každou bdělou činnost, byť v různé intenzitě. Rozlišení mezi prožíváním a introspekcí, jako dvou opačných směrů pohledu, souvisí s pojmem **subjektu**.

Subjekt je to, co když říká „já“, míní tím sebe. Své *já* máme vždy s sebou a měli bychom ho tudíž znát důvěrněji a intimněji než cokoliv jiného. Je přítomno v každé naší větě, ať mluvíme o čemkoliv. Je též nutně přítomno ve světě, který se rozkládá kolem každého z nás, nelze si ho od něj odmyslit. Není proto vhodné předpokládat, že existují dvě oddělitelná jsoucna, *já* a *svět*, lze ale rozlišovat ony dva zmíněné *směry*, směr *dovnitř* a směr *zevnitř*. Mé *já* směřuje zevnitř, proto sebe nevidí, vidí jen svůj zjednaný svět. Ani nikdo jiný mé *já* nevidí, není mu zvenku dostupné.

Jakou orientaci tedy zaujímáme při sebereflexi, poznávání sebe sama? Zde každý z obou právě zmíněných směrů má svou roli. Při uvědoměném sebepozorování, introspekci, jako bychom chtěli dočasně vystoupit ze sebe a dívat se zpět do vlastního nitra. Naproti tomu při hluboce emocionálních zážitcích plně prožíváme sebe sama, avšak zevnitř, nepozorujeme se při tom, často o sobě ani nevíme. Říká se přece: „Je radostí bez sebe.“ Při pohledu zpět jsem schopen o svém *já* mluvit, je předmětem mých vět, zatímco při pohledu zevnitř je *já* tvůrcem a často i podmětem těchto vět, vždy je jaksi jejich ručitelem.

V kognitivní vědě hraje sebereflexe a subjektivita dosti pozoruhodnou roli. Na jedné straně je skryta v pozadí téměř všeho, o čem se v kognitivních vědách mluví – skryta proto, že badatelé ve snaze vyhovět požadavku objektivity poznání, který je jedním z hlavních postulátů moderní přírodovědy, neradi zdůrazňují intuitivní původ svých pojmů, problémů a hypotéz. Na druhé straně jsou to právě kognitivní vědy – a doložitelně umělá inteligence – které nastolily problém *já* jako relevantní výzkumné téma (blíže o tom viz kap. 8).

3.3 Cesta přírodovědy

Přírodní vědy mají objekty svého studia jakoby před sebou, mohou je pozorovat, ohmatávat, měřit, mohou o nich i teoretizovat a spekulovat. K tomu si je ovšem vždy musí předem opatřit, každý objekt odděleně, musí je umět identifikovat a pojmenovat. Protože vnitřní, mentální stavy nejsou otevřeny takovému způsobu bádání, uchyluje se přírodovědně laděný badatel buď ke zkoumání vnějších projevů mysli, jako například chování jednotlivců a skupin (tradiční behavioristická psychologie, etologie, antropologie), anebo se zaměřuje na biologický substrát mysli, mozek. Zejména vědy o mozku (neurofyziologie, neurologie, neuropsychologie, biologická psychiatrie a další) doznaly obrovského rozvoje.

---

[12] Angl. slovo *experience* má v češtině nejméně tři významy: prožívání (v našem smyslu), zážitek a zkušenost.

Nelze říci, že by s mozkem nebyly spojeny záhady přesahující běžné, „laboratorní“ metody výzkumu. Napadnou nás například tyto:

- Jak je možné, že každý jednotlivý mozek vůbec funguje a že dovede věci, které se nám zdají tak překvapivě „inteligentní“?
- Jak je možné, že mozek jednotlivce dospěje k této dokonalosti během pouhých několika let, a to takřka z ničeho (z jedné buňky)?
- Jak je možné, že evoluce vůbec něco takového vynalezla?
- Proč pro nás tyto záhady jsou stále záhadami, když toho o mozku přece dnes již tolik víme?

Opravdu, od poloviny minulého století neurobiologové nashromáždili nepřeberné bohatství poznatků o tom, jak vypadá mozková tkáň a co se v ní odehrává v různých místech za různých okolností. Jsou hypotézy o dějích na úrovni kvantové, jsou známy děje na úrovni molekul (kde iontové pumpy a kanály manipulují s potenciály a koncentracemi) stejně jako na úrovni buněk – neuronů (kde dochází ke vzniku, přenosu a sčítání signálů), ví se o chování kolektivů neuronů, jsou mapovány funkční oblasti mozku, hodně je známo o jejich systémovém propojení a úlohách v organismu. Jedno však nevíme: *jak to vše souvisí s myšlením*. Zrodila se myšlenka, kterou jsem právě napsal, jako průvodní jev nějaké fyzické příhody v mém mozku? Anebo naopak, je tato příhoda vedlejším projevem oné myšlenky?

Neurobiologie je příklad vědy, která hledí „zdola“, zkoumá biologický substrát, který umožňuje mysl. Naproti tomu psychologie pozoruje člověka „shora“, zkoumá vnější projevy mysli. Rozlišení „shora“/“zdola“ je dvojnásob metaforické (proto i nadále zachovám uvozovky): zaprvé, není jiného než zvykového důvodu zde mluvit o vertikálním uspořádání, a za druhé, problém jen částečně souvisí s velikostí: neurony jsou sice malé, o myšlení však nelze dost dobře říci, že je velké. Zda skutečnost nabízí nějakou lineární škálu, vzhledem k níž by mělo toto „shora“ resp. „zdola“ smysl a na níž by se případně mohly oba pohledy setkat, je zatím předčasné soudit. V každém případě je rozpětí mezi čímkoliv „nahoře“ a čímkoliv „dole“ větší, než si někdy uvědomujeme, a lze do něj vložit mnoho „mezipater“, jak demonstruje například Allwyn Scott (Scott, 1995). Vrátíme s k tomu později (v odst. 6.3).

V posledních dvou desetiletích se pro pokusy propojit znalosti získané „shora“ a „zdola“ ujal název *neuropsychologie* (Kolb, Whishaw, 1990). Bylo by to ovšem propojení jen v rámci přírodovědecké, objektivní metodologie, bez aspirace na vazbu s vnitřním prožíváním.

3.4 Cesta umělého modelování

Vedle reflexe svého vlastního prožívání, která se vzpírá objektivizaci, a vedle přírodovědné metody, které se jen pomalu daří pronikat ke strukturně složitým a nepřístupným mozkových procesům, existuje ještě třetí, nepřímá cesta k poznání mysli: experimentování s uměle vytvořenými modely. Hlavním reprezentantem této cesty je právě umělá inteligence.

Ponecháme-li stranou analogové modely vybraných psychických procesů, konstruované v rané fázi kybernetiky, využívá cesta umělého modelování podstatně výpočetní techniku. Důležité je, že počítač nevstupuje do hry jen jako pomocný nástroj ke složitým výpočtům či k simulaci systémů, jejichž fyzická realizace by byla technicky obtížná, ale též sám o sobě jako jeden z potenciálních kandidátů na možný

model lidské mysli. Jako takový může být pro kognitivní vědu inspirativní a mnohé badatele vedl dokonce k přesvědčení, že algoritmické procesy v počítači a kognitivní procesy v mysli jsou (opatrně řečeno) entity téhož řádu. Z tohoto přesvědčení vychází i *počítačový funkcionalismus* (též zvaný *silná umělá inteligence*), filosofický směr o němž budeme hovořit později (hlavně v odst. 4.3 a 7.3).

Je ovšem namístě zdůraznit dvojí, ne vždy jasně rozlišené chápání pojmu **model**. Běžně se jím míní **věrný model** – fyzický nebo teoretický (zpravidla matematický) konstrukt, jímž se snažíme *co nejlépe* reprezentovat či napodobit vybrané vlastnosti modelovaného objektu. Pro čím více vlastností se to podaří (a jde-li o vlastnosti měřitelné, pak s čím větší přesností), tím je model pokládán za lepší, a skutečnost, že (mnohé) další vlastnosti modelovat nelze, je považována za nedostatek či nedokonalost modelu. Naproti tomu u **metaforických modelů** i odlišnost od modelované skutečnosti nám pomáhá něco objasnit či vysvětlit. Současně s tím, jak připodobňují něco neznámého k tomu, co již známe nebo čemu již rozumíme, dávají vystoupit do popředí odlišnost, přičemž právě napětí mezi podobností a odlišností je zde zdrojem poznání. Pro studium kognitivních procesů jsou metaforické modely zvlášť důležité. Zda přitom jde o objekty realizované fyzicky, softwarové simulace, nebo abstraktní teoretické modely, nehraje podstatnou roli.

Mezi metaforické modely myšlení patří do jisté míry i samotný počítač. Od objevení se prvních počítačů počátkem padesátých let byli lidé fascinováni nejen jejich výkonností, ale i skutečností, že byly opravdu „samočinné“: člověk nemusí mít kontrolu nad každým dalším operačním krokem, ostatně ani není schopen tyto kroky sledovat. Byly hledány rozličné paralely mezi činností počítače a myšlením, počínaje logickými prvky a konče sériovým (sekvenčním) uspořádáním výsledného procesu. Tyto paralely snadno a rychle ztrácely svou metaforickou povahu a vedly k odborným, filosoficky nezávazným termínům informatiky. Současně však mnohdy vedly k až příliš doslovnému chápání mechanistických a algoritmických struktur jako věrných modelů lidské mysli (k této věci se vrátím v odst. 4.3).

Podobně jako u přírodovědné cesty, lze i při modelování hovořit o dvojím přístupu, „shora“ a „zdola“ – podle toho, kterou úroveň (z přírodovědeckého pohledu) chceme modelovat pokud možno věrně. Tradiční, algoritmická umělá inteligence je reprezentantem přístupu „shora“. Východiskem jsou *symbolické*[13] procesy myšlení, tak jak je umíme studovat psychologicky, logicky a pomocí introspekce. Problém, do jaké hloubky (řečeno v naší „vertikální“ metafoře) je nutno proniknout, abychom směli mluvit o modelování mysli, je ovšem otevřená otázka.

Co bychom mohli považovat za modelování „zdola“? K odpovědi je třeba poněkud zpřesnit chápání onoho „dole“.[14] V plné obecnosti lze odpovědět asi takto: je zvolena jistá úroveň popisu, „pod“ níž již není rozhodující, zda si prvky modelu zachovávají či nezachovávají materiální, strukturní nebo formální podobnost s odpovídajícími prvky modelované skutečnosti. Stačí, že mají shodné chování, přesněji tu součást chování, která je směrodatná v rámci znalosti modelovaného sytému. O této úrovni budeme mluvit jako o **základní úrovni analogie**. Klademe-li například při modelování mozkové činnosti za základní úroveň analogie neuronovou tkáň, nemusí nás zajímat, zda jednotlivé neurony jsou funkčně realizovány ,

---

[13] Slovo ‚symbolický‘ zde užívám ve smyslu „operující se symboly, tj. s objekty, které jsou nositeli dohodnutého významu“. Někteří autoři užívají termín ‘symbolový’.

[14] V kap. ** tohoto svazku (Kelemen, 1999) je rozlišení „shora“ vs. „zdola“ míněno spíše funkcionálně: zda postupujeme od složitých funkcí (schopností) k jednodušším anebo opačně.

elektronicky, logicky nebo pomocí symbolických programových jednotek – prostě je považujeme za černé schránky. Propojením prvků na základní úrovni analogie pak vytváříme vyšší stavební jednotky, jejichž chování již považujeme za součást modelu.

Pro přístup „zdola“ je z historického pohledu klíčová práce Warrena McCullocha a Waltra Pittse (McCulloch, Pitts, 1943), kteří si všimli, že jednotlivým neuronům lze přiřadit elementární logické proměnné a neuronovým obvodům funkce výrokové logiky. To je vedlo k návrhu formálního kalkulu, s jehož pomocí lze ze znalosti neuronové sítě odvodit její vnější chování a naopak ze zadání vnějšího chování navrhnout síť, která je realizuje. Dali tak základ rozsáhlému vědnímu (i technickému) oboru – teorii logických sítí a konečných automatů.

Ač jejich motivace byla spíše teoretická, elegance a jednoduchost logického přístupu přivedla McCullocha a Pittse (v závěru práce) k vyslovení redukcionistické hypotézy, že základní úrovní popisu pro veškerou psychickou činnost je úroveň neuronů, chápaných jako dvouhodnotové logické prvky. Pokud psychologie hledá jakési atomy myšlení, „psychony“, nalezne je dle nich právě v podobě neuronů. Protože však různost stavů na neuronové úrovni se nemusí okamžitě projevit v pozorovatelném chování člověka, nýbrž až v jeho budoucím chování, plyne z toho zdánlivý nedeterminismus vnějšího chování. Obtíž zpětně hledat minulé příčiny současného chování (projekt psychoanalýzy) tkví v tom, že znalost současného stavu na neuronové úrovni k tomu nestačí, protože logické funkce nelze obecně invertovat.

Práce McCullocha a Pittse nejen vedla k novým aplikačním i teoretickým směrům v matematice a logice, ale též přímo ovlivnila von Neumannovy ideje pro číslicový počítač. Navíc obsahovala myšlenky, na nichž je založen dnes nejvýznačnější reprezentant přístupu „zdola“ v kognitivních vědách, totiž konekcionistická umělá inteligence. Až do jejího příchodu na začátku 80. let bylo modelování „zdola“ poněkud zastíněno logicko-symbolickým paradigmatem umělé inteligence, stále však existovalo např. v podobě perceptronových modelů (Rosenblatt, 1962), obecněji jako obor zvaný kybernetické modelování (Klír, Valach 1965).

Základní úroveň analogie bývá v dnešních kognitivních vědách často posunuta hlouběji, než je neuronová (tj. buněčná) úroveň. Tři příklady: Funkční modely jednotlivých neuronů nejsou omezeny na časově diskrétní dvouhodnotové logické funkce, nýbrž reprezentují i časoprostorovou dynamiku přenosu signálů na axonech, dendritech a synapsích a další biofyzikální děje (Koch, Segev, 1998). Biologická psychiatrie studuje vliv psychofarmak na složité molekulární děje na synaptických přechodech (Höschl, 1997) i v mezibuněčném prostředí mimo synapse (Vinař, 1999). A do třetice, někteří badatelé spekulují o fyzikálních jevech v neuronech dokonce na úrovních, na nichž mohou hrát roli kvantově mechanické jevy (viz odkazy na s. 31).

I při přístupu „zdola“ je někdy rozumné se omezit jen na některé dílčí aspekty základní úrovně analogie a záměrně nerespektovat mnohé z toho, co je o této úrovni známo. Lze proto i zde hovořit o metaforických modelech. Příkladem jsou rozličné teoretické (i technické) modely abstraktních neuronových sítí, které mají s biologií velmi málo společného, snad jen to, že jde o kolektivní chování velkého množství vzájemně vázaných prvků. Uvidíme (v odd. 5), že i na nich se lze leccos poučit o dynamice a emergentních vlastnostech složitých systémů, mezi něž lidský mozek bezpochyby náleží.

4 Umělá inteligence jako motivační zdroj pro filosofii

Souběžně s tím, jak umělá inteligence procházela různými stadii svého vývoje a potkávala se s úspěchy i neúspěchy při modelování specifických intelektuálních schopností, optimismus příznivců se konfrontoval se skepsí odpůrců. Debatovalo se nejen o její praktické použitelnosti, ale i o otázce, zda umělá inteligence (disciplína) je správnou, či dokonce jedinou správnou cestou k „umělé inteligenci“ (cíli). Vyskytly se i pochybnosti o tom, zda vůbec smíme po této cestě pokračovat (Weizenbaum, 1976).

Bylo by nad možnosti této studie třeba i jen utřídit rozmanité, někdy i konfliktní vstupy do této debaty.[15] Omezím se proto jen na vybrané vzorky. Nejprve se v tomto oddíle zaměřím hlavně na tradiční a stále převládající směr v umělé inteligenci (s novějšími směry se setkáme v oddílech 5 a 6).

4.1 „Umělá“ logika a gödelovský argument

Zatímco o přirozené logice se mluví zcela běžně, mluvit o umělé logice je poněkud nezvyklé. Přitom je to právě logické uvažování, které ze všech výkonů lidské mysli má nejblíže k realizaci v podobě mechanických procedur. Totéž lze vlastně říci i opačně: kdykoliv myslíme přísně logicky, pak se vlastně chováme jako stroj (či funkcionalisticky řečeno: pak *jsme* stroj). V tomto smyslu lze považovat za umělou logiku třeba i klasickou sylogistiku, protože nabízí schémata k „mechanizaci“ logického usuzování (Bendová, 1998).

Není nic překvapivého na tom, že teoreticky nejzávažnější obory umělé inteligence jsou založeny na formální logice: automatické dokazování, řešení úloh, plánování, některé přístupy k reprezentaci znalostí, logické programování. Úspěchy právě v těchto oborech byly svého času (unáhleně) považovány za argument na podporu projektu umělého myšlení. Důsledně vzato je však logicky formalizovatelné myšlení jen „povrchovým“ projevem mysli; například generování formálních důkazů teorémů přímo nesouvisí s tím, jak vůbec lidská mysl důkazy objevuje a jak jim při čtení rozumí. První vlna debat o uskutečnitelnosti projektu umělé inteligence proběhla začátkem 60. let, kdy J. R. Lucas upozornil, že Gödelovu větu o neúplnosti aritmetiky je možno použít k vyvrácení tzv. **mechanistické teze**, tj. názoru, že lidskou mysl *lze* simulovat strojem (Lucas, 1961). Ostatně i sám Gödel soudil, že jeho věta o neúplnosti je příspěvkem do debaty o rozdílech mezi člověkem a strojem:

Buď je lidská mysl schopna odpovědět na více číselně-teoretických otázek, než kterýkoliv stroj, anebo existují číselně-teoretické otázky, na které lidská mysl odpovědět nedovede.[16]

Gödel mluví o „lidské mysli“ (a podobně my o „člověku“) v jednotném čísle, zatímco o strojích se zpravidla mluví v množném čísle (je mnoho různých strojů). V tom je obsažen implicitní - poněkud diskutabilní - předpoklad, že co do své matematické a logické kompetence náleží všichni lidé do téže kategorie, jejímž je náš „člověk“ abstraktním reprezentantem.

---

[15] Viz např. (Boden, 1990; Dreyfus, 1979; Graubard, 1988; Haugeland, 1985; Lucas, 1961; Searle, 1980, 1997).

[16] Citováno dle (Webb, 1980), s. 112.

Jaká je tedy souvislost Gödelovy věty o neúplnosti s poměřováním přirozeného a umělého myšlení? Je známo, že počítače – či lépe programy[17] – lze při vhodné interpretaci chápat jako reprezentanty formálních systémů. Pro nástin Lucasova argumentu stačí, když si zavedeme pojem *rozhodovacího programu*: je to program, který na vstupní otázky, týkající se (Peanovy) aritmetiky, odpovídá buď "ano" nebo "ne" (nebo nedovede odpovědět), přičemž

(a) na axiómy této aritmetiky odpovídá „ano“,

(b) respektuje logické vyplývání (kdykoliv je odpověď na jednu otázku logickým důsledkem odpovědi na jinou otázku, pak program odpovídá v souladu s touto závislostí),

(c) je bezesporný, čili jednotlivé odpovědi si vzájemně logicky neodporují.

Vyjdeme-li z hypotézy, že lidské myšlení je bezesporné, že se řídí logickými zákony a že považuje axiómy aritmetiky za pravdivé, pak jsou uvedené předpoklady nutné k tomu, aby nějaký program takové myšlení potenciálně modeloval. Tytéž předpoklady však též stačí k tomu, aby dotyčný program již podléhal Gödelově větě o neúplnosti.

Řekneme, že otázka (o vlastnostech čísel) je *zodpověditelná*, když člověk je principiálně schopen na ni správně odpovědět (neuvažuje se, s jakou námahou). Gödelovu větu lze pak volně přeformulovat takto:

Ke každému rozhodovacímu programu existuje zodpověditelná otázka, na kterou tento program nedovede odpovědět.

Trik důkazu spočívá v tom, že i otázky o strojích lze (důmyslně) převést na otázky o přirozených číslech, a tak lze strojům klást choulostivé otázky o jejich vlastním chování.

Zdálo by se tedy, že člověk má jistou formální „převahu“ nad strojem. Věc však není tak jednoduchá. Gödelova věta je v rozporu pouze s následující **silnou variantou mechanistické teze**:

Existuje program, který odpoví na všechny zodpověditelné otázky.

Neplatí však ani její protiklad, který lze nazvat **mentalistickou tezí**:

Existuje zodpověditelná otázka, na kterou žádný program nedovede odpovědět.

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

Nelze předpokládat, že člověk je schopen řešit *libovolně* složitý problém, avšak jistá jeho převaha tkví v tom, že věci rozumí natolik, že je schopen (metaforicky řečeno) zahlédnout celou nekonečnou hierarchii najednou.

---

[17] Zvolíme-li jeden pevný počítač. Bez újmy na přesnosti lze v obecných úvahách této stati nerozlišovat mezi pojmy „algoritmus“, „program“ a „stroj“ (počítač).

Lucasův citovaný článek obsahuje množství úvah a námětů, které stojí za uvedení. Píše např.:

Říkáme-li, že vědomá bytost něco ví, neříkáme tím jen, že to ví, ale též, že ví, že to ví, a že ví, že ví, že to ví atd. [...] *setkáváme se zde s nekonečnem, ale nikoliv s nekonečným regresem s jeho negativními aspekty* [...]

A dále:

Vědomá bytost může zacházet s Gödelovými otázkami tak, jak stroj nemůže, protože vědomá bytost může uvažovat sebe a své projevy, aniž by se lišila od zdroje těchto projevů.

K Lucasovu argumentu o lidské převaze se po letech vrátil R. Penrose, který jej užívá na podporu své teze, že jen v rámci nové a netradiční fyzikální teorie budeme mít šanci vědecky vysvětlit lidskou mysl, případně i vědomí (Penrose, 1994).

Je tu ovšem otázka, zda a do jaké míry argumentace založená na Gödelově větě o neúplnosti opravdu umožňuje činit závěry o převaze člověka nad počítačem. Především by bylo možno poukázat na skutečnost, že i my, podobně jako stroje, podléháme neúplnosti – i my totiž, nedovedeme odpovědět na některé choulostivé otázky (jen zkuste odpovědět na otázku: *Odpovíš na tuto otázku nesprávně?*[18]).

Zajímavý argument (ve prospěch umělé inteligence) uvádí Hofstadter (Hofstadter, 1979, s. 577): I když obecně platí, že algoritmy, včetně těch, které jsou užity v umělé inteligenci, podléhají gödelovskému omezení, týká se to pouze jejich nejnižší rozlišovací úrovně. Avšak již dnes mají systémy pro umělou inteligenci řadu vyšších úrovní, přičemž to, co je důležité, se odehrává až na nejvyšších úrovních.

Je nutno poznamenat, že „hierarchie úrovní“ u Hofstadtera není chápána čistě jen v programátorském slova smyslu (kdy programy nižší úrovně hrají roli elementárních příkazů v programech vyšší úrovně), nýbrž i tak, že čím vyšší úroveň, tím jsou programy flexibilnější, neurčitější, méně exaktní a bez přísného kritéria vnitřní bezespornosti (logický spor je sice možný, avšak v běžných výpočtech prakticky nedosažitelný). Podle Hofstadtera je „sídlem“ inteligence nejvyšší, nejméně formální úroveň, kde se „operuje s obrazy, formulují analogie, zapomínají myšlenky, zaměňují koncepty, smazávají rozdíly.“

Pokud se na zmíněné nejvyšší úrovni vyskytnou i přesné logické postupy (podobné jako u lidí), například odvozování ve výrokové logice, nemusí to být proto, že byly předem implementovány programátorem. Mohou vzniknout emergentně jako jeden z důsledků daleko obecnější „inteligence“ programu (Hofstadter, 1979, kap. 20). Jde o příklad *zacyklené hierarchie*: výchozí úroveň (řeč je o klasických číslicových počítačích) je založena na exaktních, formálně logických principech, a jako taková je tato úroveň výsledkem záměru konstruktéra. Ale současně na nejvyšší úrovni se mohou vyskytnout *stejné* formálně-logické principy (ať už záměrem programátora, či „samy od sebe“), přičemž na mezilehlých úrovních tyto principy nehrají žádnou (explicitní) roli.

4.2 Co je přirozené a co umělé

Člověk se obklopuje umělými věcmi. Na jedné straně je obyvatelem přirozeného světa, jehož je navíc i součástí, na druhé straně do něho svým chováním a

[18] (Hofstadter, 1979, s. 476). Jde o autoreferenční otázku podobného typu, jakého je užito v Gödelově důkazu.

svými výtvory neustále vnáší něco nového, ne-přirozeného, umělého. Nelze proto obecně klást ostrou hranici mezi *přirozená jsoucna* a *umělá jsoucna* v nejširším smyslu těch slov.

Pokud se slova ‚umělý' explicitně užívá jako přívlastku (umělá květina, umělé jezírko, umělý drahokam, umělá ledvina, umělá hmota, umělá družice, umělé osvětlení, umělé oplodnění, umělé dýchání, umělý jazyk), je to zpravidla tehdy, když jednou ta a jednou ona látková, tvarová nebo funkční komponenta reality je na zvolené úrovni popisu nahrazena *artefaktem*, produktem lidské činnosti (jde-li o modelování, je touto úrovní základní úroveň analogie). Vždy tu jde o něco, co může být také (a především) přirozené (neříkáme přece umělý krb, umělé brýle). Přišel člověk a vyrobil či způsobil *totéž*, jenže *jinak*, z něčeho jiného a jen v jistém aspektu – v jakém, to se případ od případu liší. Předpokládá se však něco víc, než že u toho byl člověk – je tu vždy *vědomý záměr*, čili apriorní explicitní představa na straně tvůrce o tom, co ono umělé má dělat či jak má vypadat či co má nahradit. Lze pak i objektivně posuzovat, do jaké míry realizace plní původní záměr.

V obecnějším pojetí nemusí realizace ani nést příznaky toho, že je opravdu umělá (umělé jezírko nemusí být rozlišitelné od přirozeného jezírka). Podle Birnbachera (Birnbacher, 1995) lze umělé chápat v *adverbiálním* smyslu, kdy jde jen o způsob, *jak* došlo k realizaci (případ jezírka), nebo v *adjektivním* smyslu, kdy (navíc) jde o to, *jaká* realizace je, že k ní bylo totiž užito nepřirozeného materiálu či prostředí (umělá květina, umělé myšlení).

To vše platí i o umělém myšlení, vnímání, rozhodování, řešení úloh a další činnosti, na něž se zaměřuje umělá inteligence (což je jen souhrnné označení pro snahy to vše realizovat[19]).

V jedné monografii o umělé inteligenci (Rich, Knight, 1991) je tato disciplína definována jako studium možností, „jak přimět počítače dělat to, co lidé dosud dělají lépe". Všimněme si, že tu jde o dvě různé věci: jednak (něco) *dělat* a jednak *přimět* (někoho něco) *dělat.* Povaha obou výkonů a rozdíl mezi nimi je třeba mít v patrnosti, kdykoliv tvrdíme, že nějaká činnost je umělá.

Velmi obecně řečeno, abychom o nějaké věci (vlastnosti, činnosti) mohli tvrdit, že je **umělá**, měly by být splněny přinejmenším tři podmínky:

(1) existuje nějaká *přirozená* věc, logicky připouštějící duplikaci (v našem případě je to například lidské myšlení, vnímání, rozhodování apod.),

(2) existuje *záměr* člověka (nebo týmu) vytvořit duplikát oné přirozené věci,

(3) došlo k *provedení záměru*, čili proběhl intencionální proces vedoucí od záměru k jeho *realizaci* (v našem případě byla dotyčná věc například implementována na počítači).

Provedení záměru je *intencionální akt* člověka a rozdíl mezi samotným záměrem a jeho provedením odpovídá rozdílu mezi *předběžnou intencí* a *intencí v akci* (Searle, 1983). Součástí obou intencí je jasná apriorní představa *cíle* intencionálního aktu; tím je pak i dána možnost hodnotit jeho úspěšnost.

Obojí intenci (předběžnou i intenci v akci) je třeba předpokládat, chceme-li mluvit o čemkoliv jako o něčem umělém a v obou případech hraje roli ona apriorní představa cíle – nazývejme ji (jde-li o konstrukci nějakého produktu) *projektem*.

---

[19] Podobně je ‚umělý život' souhrnným označením pro snahy uměle realizovat vybrané přírodní procesy (jako kolektivní chování, růst, samoreprodukci, rozmnožování, evoluci apod.). Viz **následující kapitolu v tomto svazku (Csonto, 1999).

Budu-li bubnovat prsty na klávesnici počítače v domnění, že jde o klavír, a tímto způsobem náhodou „vybubnuji“ program, který je (řekněme) schopen myslet, nebudu tvrdit, že jde o *umělé* myšlení, stejně jako bych to netvrdil, kdyby klávesami pohyboval náhodný generátor či třeba kočka. Co chybělo, byl *projekt* (myšlení).

Každý projekt vyžaduje objektivně srozumitelnou specifikaci směrodatných vlastností cílového produktu. Konstrukce (výroba, programování) je „vnější“ aktivita, neprobíhá v něčí fantazii, nýbrž ve vnějším (byť někdy jen symbolickém) světě, a proto ony směrodatné vlastnosti musí být popsány „zvenku“, například tak, aby realizace mohla být zadána i někomu jinému.

Diskuse o možnostech algoritmické umělé inteligence se většinou zajímají o výsledné produkty (programy, jejich implementace a vlastnosti), méně se již zaměřují na způsob, jak vznikaly, tj. na účast člověka při jejich vzniku a na existenci projektu. Takto se můžeme setkat s poněkud jiným typem problému: jak se stavět k naší principiální neschopnosti popsat to, co vlastně od programu žádáme. Některé aspekty lidské mysli totiž nelze zadat jako konstrukční projekt: nevíme například, *jak objektivně a zvnějšku popsat to, co spočívá ve vnitřním, subjektivním prožívání*.

Mluvit o umělém myšlení (v hlubším smyslu slova ‚umělý‘*)* má tedy smysl, když se omezíme jen na některé jeho výkony (či projevy), totiž ty, které jsou – předběžně řečeno – *objektivní*. Zdá se ovšem vhodnější přitom nevést dělítko mezi *různými* mentálními výkony (jak jsou například pojímány v přirozeném jazyku, popřípadě v tradiční psychologii), nýbrž *napříč* každým z nich zvlášť. Rozlišujme proto (teoreticky) dvě *komponenty mentálních procesů*: to, co je na daném procesu objektivně, „zvnějšku“ popsatelné, nazveme jeho **performační komponentou** a to, co přitom subjekt „vnitřně“ prožívá, nazveme jeho **fenomenální komponentou**[20].

Lze tedy shrnout: *umělá* (v našem smyslu) realizace nějakého mentálního procesu či výkonu se může vztahovat *pouze na jeho performační komponentu.*

Rozlišení obou komponent je spíše jen teoretické; v praktickém životě je nerozlišujeme a většinou by to ani nebylo snadné. Přesto je takové rozlišení intuitivně celkem přirozené, což lze ilustrovat zejména na těch mentálních aktivitách, u nichž je fenomenální komponenta snadno odmyslitelná. Jsou to především činnosti, které provádíme sice vědomě, ale více méně „mechanicky“, tj. podle návodu, jehož podstatě ani nemusíme rozumět (například aritmetické operace s většími čísly, řešení stereotypních problémů či jednoduché dedukce), dále dokonale nacvičené dovednosti, vše co je odsunuto do podvědomí, až po vrozené dispozice (syntax mateřské řeči, rozpoznávání tváří, umělecké schopnosti, intuitivní reakce). Ve všech těchto případech se fenomenální komponenta redukuje nanejvýš na pasivní evidování dotyčné činnosti. Později (například v závěru odst. 4.4) se setkáme s dalšími příklady rozlišení obou typů komponent. Mnohé diskuse o možnostech umělé inteligence by se vyřešily, kdyby si jejich účastníci předem vyjasnili, zda jim jde čistě o performační komponentu výkonu, o kterém debatují (například vizuální percepce, hraní šachu apod.), nebo zda mají na mysli obě komponenty současně (čili nerozlišují je).

Je třeba si uvědomit, že čím je pro určitý typ činnosti důležitější fenomenální komponenta, tím více se umělá verze této činnosti bude odchylovat od verze přirozené – což je případ metaforických modelů ve smyslu odst. 3.4.

---

[20] Fenomenální komponenty mentálních stavů se často nazývají *qualia*.

4.3 Počítačová metafora a „silná“ umělá inteligence

Jakýkoliv pokus o přirovnání některých aspektů lidské mysli (počínaje funkční organizací mozku a konče kognitivními procesy) k vnitřní organizaci, výpočetním procedurám, či vnějším projevům počítače lze považovat za (vědomé nebo neuvědoměné) užití takzvané **počítačové metafory**, která výrazně ovlivnila post-behavioristickou psychologii padesátých let a stala se východiskem pro dodnes vládnoucí paradigma kognitivní vědy. Jak později uvidíme, vedla i k novému oživení filosofie mysli.

Von Neumannova architektura číslicových počítačů, ke které až donedávna neexistovala výrazná alternativa, je spojena s představou sériového diskrétního výpočtového procesu, který je řízen programem uloženým (spolu se vstupními a pracovními daty) v paměti počítače. Program (v konkrétním smyslu) je syntakticky korektní a na daném typu počítače realizovatelný zápis nějakého (abstraktního) algoritmu v dohodnutém programovacím jazyce.

Technicky lze u počítačů hovořit o rozmanitých úrovních, od nejnižší, fyzikální úrovně (elektronika, logické obvody, mikroprocesory, atp.), až po nejvyšší úroveň, která operuje se symbolickými objekty pomocí programů a která je schopna komunikovat s uživatelem (či prostředím). Všechny úrovně jsou přitom součástí téhož konstrukčního projektu a je teoreticky myslitelné, aby je navrhoval jediný konstruktér. (V praxi je však dnes již téměř nemožné se zabývat všemi úrovněmi najednou, protože se v průběhu let s rozvojem softwarových systémů na jedné straně a miniaturizací na straně druhé obě krajní úrovně od sebe velmi vzdálily. Vznikly přitom další meziúrovně, jimiž se zabývají i rozdílné technické disciplíny.)

I na pojmy **úroveň** a **hierarchie úrovní** (s nimiž jsme se již setkali v závěru odst. 4.1) lze pohlížet z hlediska počítačové metafory. I když jde obecně o pojmy velmi vágní a v konkrétních instancích dosti libovolné, u umělých systémů, jako jsou počítače, mají celkem jasnou technickou interpretaci. Analogicky se pak snažíme vidět hierarchické uspořádání i u přirozených systémů. Je ovšem třeba si uvědomit, že je to vždy do jisté míry důsledek naší zúžené schopnosti vidět svět holisticky, jako jeden celek.

Pro pochopení počítačové metafory je důležité i rozlišení mezi hardwarem a softwarem. Vztah mezi nimi je dán poněkud vágním, avšak filosoficky významným pojmem **implementace**, což je způsob, jak zajistit, aby daný softwarový program řídil reálný průběh příslušného výpočtového procesu v daném typu hardwaru. Podstata softwarových programů totiž tkví v jejich *kauzální potenci* (schopnosti řídit dotyčné procesy), která je invariantní k té či oné konkrétní implementaci; v tomto smyslu lze i říci, že programy „přežívají hardwarovou smrt“.

Tato nezávislost softwarových objektů (programů) na hardwarové implementaci je hlavním motivem **počítačového funkcionalismu**, též označovaného jako **„silná“ umělá inteligence** (Searle, 1984). Je to názor, který lze vyjádřit touto tezí:

Povaha mysli je algoritmická, přičemž není podstatné, v jakém mediu jsou algoritmy (programy) implementovány.

Nezáleží tedy na tom, zda tímto mediem je mozek, počítač či třeba armáda Číňanů. Počítače by proto mohly (dokonce musely) mít myšlenky, pocity, porozumění apod., pokud by v nich byl implementován vhodný počítačový program s vhodnými vstupy a výstupy. Teze počítačového funkcionalismu, hojně zastávaná

v kruzích tradiční umělé inteligence, není v rozporu s materialismem (implementace je nakonec vždy nutná pro realizaci), přičemž však – na rozdíl od materialistických či fyzikalistických teorií mysli – nepřisuzuje mozku nijak podstatnější roli, než kterémukoliv jinému myslitelnému implementačnímu mediu (více o tom viz oddíl 7.3).

Dle silné umělé inteligence je tedy principiálně možno *replikovat* lidskou mysl v počítači. Na rozdíl od toho **„slabá“ umělá inteligence** aspiruje jen na *modelování* mysli, případně jejích dílčích projevů, přičemž se orientuje na nejvyšší, *logicko-symbolickou* úroveň, která je tak pro ni základní úrovní analogie (ve smyslu odst. 3.4).

Pokud jde o aplikační umělou inteligenci, je vhodné zdůraznit, že i když jí jde více o účel a efektivnost než o napodobení člověka, některé její algoritmické struktury a metody mají určitou vypovídací hodnotu i jako metafory pro přirozené myšlení. Jako příklad lze uvést sémantické sítě (viz kap. 4 v prvním dílu této knihy (Zdráhal, 1993)) a jiné metody reprezentace znalostí, nebo učení na principu posilování.

Existuje ovšem i mnoho případů, kdy jsou algoritmy založeny právě na těch procedurách, které užívá člověk (jednoduché sylogismy, některé aritmetické operace, klasifikace a třídění, kognitivní mapy a všechny další případy, kdy se vědomě řídíme nějakým algoritmem). Vlastně i samotná von Neumannova architektura počítačů byla motivována Turingovým teoretickým modelem univerzálního výpočtového stroje (Turing, 1936), při jehož návrhu Turing nejspíše vycházel z reflexe, jak on sám, matematik, by postupoval při řešení matematického problému nebo při provádění výpočtu (Dennett, 1991a, s. 212). Za relevantní přirozeně považoval jen tu část řešení, která je v zorném poli vědomí, kde se jeví jako sériový proces, a to nezávisle na nepochybně paralelní povaze procesů na neuronové úrovni. Není divu, že organizace počítače a některé běžně užívané algoritmy nám tolik připomínají přirozené myšlení.

4.4 Turingův test a čínská komora

V diskusích o povaze a možnostech umělé realizace či napodobování mysli filosofové často navrhují hypotetické situace, které mají sloužit buď jako test úspěšnosti nějakého projektu anebo jako argument pro nebo proti některé teorii či názoru. Příkladem prvého je slavný Turingův test (Turing, 1950), příkladem druhého neméně slavný Searlův experiment s „čínskou komorou“ (Searle, 1980), na který se zde hlavně zaměřím.

Nejdříve však je třeba upozornit na některé obecné vlastnosti (a záludnosti) myšlenkových experimentů. Z pohledu logiky lze rozlišit myšlenkové experimenty *realizovatelné* (alespoň principiálně) a *sporné* (které vedou k paradoxům a slouží k vyvrácení nějaké hypotézy). Každá hypotetická konstrukce má nutně dvě explicitní komponenty: jednou komponentou je její *předpokládané pozadí*, čímž se obvykle rozumí ta část aktuálního světa, kterou známe a považujeme za samozřejmou; druhou komponentou je právě ta nová, myšlenkově *konstruovaná skutečnost*, kterou explicitně popisujeme, někdy záměrně tak, aby se od samozřejmého odchylovala. Často se však zapomíná, že tyto dvě komponenty nevyčerpávají vše. Je tu ještě neurčitá (a neurčitelná) oblast všeho, co nebylo uvažováno a co bylo opomenuto. Protože v této třetí oblasti může (ale též nemusí!) dojít k vzájemným nesrovnalostem, tj. ztrátě koherence konstruované skutečnosti s předpokládaným pozadím, nazývám ji **koherenční mezerou** (Havel, 1999c).

Záludnost myšlenkových experimentů tkví v tom, že vlivem koherenční mezery může myšlenkový experiment zcela ztratit svou argumentační sílu: nic nedokáže a nic

nevyvrátí. Proto je namístě opatrnost (ostatně opatrnost podobného typu jako při návrhu a interpretaci experimentů reálných).

V roce 1950 uveřejnil Alan Turing úvahu „Computing Machinery and Intelligence“ (Turing, 1950), která je někdy považována za symbolický počátek umělé inteligence jako oboru. Turing navrhuje test, jak rozhodnout, zda stroj může myslet. Představte si dvě komory, v jedné z nichž je uzavřen stroj, ve druhé člověk. Vnější experimentátor má na základě dialogu s oběma komorami rozhodnout, kdo je kde. Stroj se přitom bude vydávat za člověka, člověk též. Otázka zní, zda experimentátor bude (v průměru) rozhodovat stejně, jako kdyby (například) v prvé místnosti byl muž a ve druhé žena.

Turingův test je příkladem realizovatelného experimentu (byl i brán v úvahu při návrhu dialogových systémů pro přirozený jazyk). Ač velmi populární, z filosofického hlediska příliš inspirativní není: vychází totiž z čistě behavioristického pojetí myšlení (charakteristického pro tehdejší psychologii) a nebere v potaz procesy uvnitř systému, tím méně fenomenální aspekty mysli. Navíc nemůže dost dobře vyhovět těm, jimž záleží na vědecké objektivitě – výsledek totiž může podstatně záviset na důvtipu a nápaditosti experimentátora (a též na tom, že sám je člověkem), což jsou vlastnosti, které nelze formalizovat. A poslední věc: testuje se vlastně něco velmi speciálního, totiž umění stroje předstírat, že není stroj.

Třicet let po Turingovi formuloval John Searle myšlenkový experiment (Searle, 1980), který se napohled zdá být variantou Turingova testu, nejde v něm však o test, nýbrž o argument proti počítačovému funkcionalismu (viz předchozí odstavec) a zdůraznění vnitřních aspektů mysli, včetně intencionality. Searlův záměr je sympatický (aspoň mně): zpochybnit tvrzení, že počítač s implementovaným správným dialogovým programem se správnými vstupy a výstupy opravdu *rozumí* jazyku, v němž probíhá dialog. „Správný“ program je například takový, který by uspěl v Turingově testu – takové programy tehdy již (skoro) existovaly pro tematicky vymezené oblasti (Schank, Abelson, 1977).

Řekněme, že v uzavřené komoře sedí John (Searle – autor mluví o sobě, což má jak uvidíme jistý význam; k odlišení od Searla-autora budu Searla-v-komoře nazývat Johnem), který nerozumí čínštině, a za pomoci rozsáhlého manuálu (v angličtině, které rozumí) vybírá – čistě podle tvaru – jedny čínské znaky jako odpovědi na jiné čínské znaky, které mu do komory vsouvá vnější experimentátor (jsou to otázky, týkající se smyslu nějakého dříve vsunutého, rovněž čínského textu). Předpokládejme, že dotyčný manuál je přepisem správného programu pro dialog v čínštině („správného“ ve shora uvedeném smyslu), takže experimentátor – budiž jím rodilý Číňan – nepozná, že původce odpovědí není rovněž Číňan. Poučení: John i počítač jsou stejně úspěšní, avšak John čínským textům nerozumí – proč by jim tedy měl rozumět počítač s implementovaným stejným programem?

Searlův myšlenkový experiment s čínskou komorou svého času vyvolal rozsáhlou diskusi. Dle mého soudu není tak přesvědčivý, jak to o něm sám Searle tvrdí, každopádně však skýtá bohatý materiál k dalším úvahám. Některé z nich zde naznačím.

S námitkou, která nás možná napadne jako první, Searle již počítal a označil ji jako „systémovou“: V experimentu má přece John jen dílčí zásluhu na přesvědčivých odpovědích, je tu ještě ten manuál. Jestliže John nerozumí čínštině, neznamená to, že systém jako celek (John + manuál) nerozumí. Searle odpovídá: Dobrá, aniž by se cokoliv změnilo na argumentaci, můžeme předpokládat, že John se prostě celý manuál naučí nazpaměť – ani tak nebude rozumět.

Tento předpoklad by dokonce zjednodušil uspořádání experimentu (na úkor jeho realizovatelnosti), otevírá však další námět k úvaze: řekněme, že se naučím zpaměti návod k nějaké složité činnosti (například k násobení velkých čísel) a pak tuto činnost realizuji – a to vědomě, ovšem jen na úrovni jednotlivých pokynů návodu. Jak dalece musím znát význam oněch pokynů,abych mohl říci, že smyslu dotyčné činnosti *rozumím*?

Dalším tématem k zamyšlení je důležitý předpoklad, že John v komoře je ve všech relevantních aspektech srovnatelný s počítačem. Searlův argument totiž spočívá v této úvaze: nahradíme počítač Johnem, který bude na jisté, pro implementaci algoritmů relevantní úrovni provádět tytéž úkony jako počítač, čili bude realizovat *tentýž* algoritmický proces, jaký realizuje počítač. To počítačovému funkcionalismu musí stačit k závěru, že počítač rozumí tehdy a jen tehdy, když rozumí John. (Pak už jen stačí doplnit: protože John nerozumí, nerozumí ani počítač.) Je to opravdu tak, že když dva dělají totéž, je to totéž?

To lze zkoumat z různých hledisek. Především je třeba si uvědomit dva hlubší rozdíly mezi Johnem a počítačem – přesněji mezi Johnem a centrálním procesorem počítače, protože John vlastně simuluje pouze ten. Za prvé, John *rozumí* manuálu (v angličtině) a *uvědomuje si*, co má dělat, tj. manipulovat s nesrozumitelnými symboly, možná to dokonce *chce* dělat (aby vyhověl Searlovi, svému mysliteli). Toto rozumění, uvědomování si a případné chtění se neodvažujeme předpokládat u centrálního procesoru počítače, od něhož čekáme spíše jen kauzální chování poslušné zákonům fyziky. Tento rozdíl Searlovu argumentaci asi neohrozí, protože není pro (předpokládaný) počítačový funkcionalismus relevantní. Je to však názorný příklad koherenční mezery v myšlenkovém experimentu.

Za druhé se v popisu experimentu od Johna explicitně vyžaduje (v zájmu hlavní argumentace) *svědectví* o tom, že nerozumí čínským znakům (zde si Searle – ale jen pro sebe – zjednodušuje situaci tím, že si do komory sedá sám). Nic takového se nepředpokládá od Johnem simulovaného centrálního procesoru (Lakoff, Johnson, 1999, s. 264). I toto je případ koherenční mezery. Otázkou pak je, zda John, který realizuje jen triviální úlohu centrálního procesoru, je kompetentní k tomuto svědectví.

Existují tedy určité Johnovy vlastnosti, které s ním nemusí sdílet počítač. Ostatně kdyby je počítač měl, asi by to nebyl počítač, který má počítačový funkcionalismus na mysli. Pokud by je neměl, znamenalo by to, že John v něčem přesahuje simulovaný počítač. Lakoff a Johnson to pokládají za nepřiznaný metaforický prvek v experimentu.

Zkusme na to jít i v jistém smyslu obráceně. John totiž simuluje *pouze* algoritmickou úroveň počítače a neuvažuje se logicky myslitelný případ, že tentokrát počítač může mít nějaké další, nečekané vlastnosti, například schopnost něčemu rozumět či nerozumět. Kdyby John simuloval chování některých (třeba i důležitých) neuronů v mozku rodilého Číňana, plynulo by z toho, že dotyčný Číňan, stejně jako John, nerozumí čínskému dialogu? Mohu být i konkrétnější. Zkusme považovat porozumění za emergentní stav mysli, který sice (u sebe sama) eviduji, nesmím se však násilně vměšovat do přirozeného a autonomního procesu, který k němu vede, obdobně jako se nesmím vměšovat do svých obranných reflexů. Jak jsem poznamenal výše, reálný výpočtový proces v počítači je na nejnižší úrovni založen na přirozeném kauzálním běhu věcí ve shodě se zákony fyziky. A tu chceme od Johna, aby ve své komoře intencionálně, svobodně a vědomě tento autonomní proces simuloval ručně. Z toho, že John neregistruje známky porozumění na vyšší než manipulační úrovni, logicky nic neplyne o porozumění nebo neporozumění na straně počítače.

Tím se vlastně oklikou dostávám k Dennettovu argumentu proti Searlovu experimentu (Dennett, 1991a, s. 438). Searle podle něj příliš bagatelizuje nároky na skutečnou počítačovou realizaci systémů, od nichž se čeká úspěch v čínské variantě Turingova testu. Takový systém by totiž musel mít tak rozsáhlou datovou bázi s tolika vnitřními vazbami, že by mohlo dojít k emergentním jevům na vyšších úrovních popisu. Pak by bylo (dle Dennetta) legitimní mluvit o pravém porozumění u počítače, třeba i takového, jehož centrální procesor je simulován nic netušícím Johnem. První část tohoto pohledu by bylo i možno podpořit připomenutím jisté konvergence mezi reprezentačními prostředky umělé inteligence a konekcionistickými systémy, o čemž se zmíním později (odst. 5.4).

Sluší se tyto úvahy shrnout. Vraťme se k minimálnímu poučení ze Searlova experimentu: John nemusí rozumět čínsky, aby si v čínském dialogu vedl stejně dobře jako počítač, a ani skutečnost, že přitom oba realizují stejný program, nestačí k tvrzení, že dotyčný počítač rozumí čínsky. Zde se slovo ‚rozumí' vyskytuje dvakrát a problém vzniká tehdy, když si nerozmyslíme, zda jej v obou případech užíváme ve stejném významu či nikoliv. Rozdíl může být tento: John v komoře *ví sám*, že nerozumí čínsky (protože to Searle ví o sobě), zatímco to, zda *počítač* rozumí nebo nerozumí, chceme posuzovat zvenku (Turingovým testem, znalostí programu apod.). Počítačový funkcionalista, pokud by chtěl být konzistentní, by nesměl tento rozdíl brát v úvahu a Johnovo subjektivní svědectví o tom, že nerozumí čínsky, by musel ignorovat.Podle mého názoru v celé debatě, včetně některých Searlových výroků, lze odhalit občasné a nenápadné opomenutí rozdílu mezi performační a fenomenální komponentou mentálních procesů, o nichž jsem se zmínil v odst. 4.2. Tento rozdíl se obecně vztahuje i na vlastnosti dotyčných procesů a na způsob, jak o těchto vlastnostech mluvíme: buď v 1. osobě (o fenomenálních aspektech), nebo v 3. osobě (o performačních aspektech). Lidé z umělé inteligence a funkcionalističtí filosofové (např. Dennett) mluví o mentálních stavech jen ve 3. osobě a fenomenální komponentu se snaží eliminovat nebo ignorovat (protože se vzpírá objektivnímu zkoumání). Není divu, že si pak nerozumějí se Searlem, který považuje řeč v 1. osobě za filosoficky relevantní – a dokonce rozlišuje 1. a 3. osobu jako dvě různé ontologické kategorie (Searle, 1992).

To, že počítač *dovede* násobit, hrát šach, rozumět příběhům (leccos i lépe než člověk) neznamená, že počítač *prožívá* násobení, hru a porozumění stejně jako je prožívá člověk, když násobí (z hlavy), hraje šach (s chutí) a rozumí příběhům (bez manuálu). Zajisté, i člověk, podobně jako počítač, *dovede* násobit (dle návodu), hrát šach (podle pravidel) a rozumět (v čínské komoře), ale to mluvíme jen o performační, nikoliv fenomenální komponentě příslušné činnosti. Ale i tak zůstává otázka, jakou roli přisoudit skutečnosti, že to dělá *vědomě*, a to nejen jako náhodný pozorovatel, ale jako účastník: *ví* proč to dělá a *ví* také, s čím zachází (byť mechanicky) – že čísla jsou čísla, pozice na šachovnici jsou situace hry, obrázky jsou čínské znaky.

5 Konekcionistická alternativa

V odst. 4.2 jsem případy, kdy o něčem lze říci, že je to *umělé*, spojil s poněkud přísným požadavkem, že dotyčná věc (přesněji její relevantní komponenta) musí být předem popsatelná jako (konstrukční) projekt. Vedlo to k omezení pouze na performační komponentu mentálních výkonů a k rezignaci na realizaci komponenty fenomenální. Zmíněný požadavek a omezení se ovšem neuplatňují, když ke vzniku nějaké vlastnosti nebo činnosti přispíváme jen nepřímo, například tím, že se snažíme

(uměle) připravit nějaké aktivní medium, v němž by se pak ona vlastnost nebo činnost vyvinula či jinak urodila samovolně.

Takto lze hledět na modelování „zdola“, zmíněné v odst. 3.4. Na některé „vyšší“ úrovni dostatečně složitého systému se docela dobře můžeme setkat s vlastnostmi a procesy, které nedovedeme předem sami popsat (nebo je dovedeme popsat, nikoliv však implementovat). Takové vlastnosti a procesy se považují za *emergentní* (blíže viz odst. 5.2 a 7.3). V našem (přísném) pojetí je nelze počítat mezi umělé, jakkoliv se v umělém mediu mohou vyskytovat – řekněme „přirozeně“.

Je velkou otázkou, konceptuální a případně i empirickou, zda právě fenomenální komponenty mysli mohou být takovouto *emergentní vlastností umělého media*. Jedno ovšem platí: lidská neschopnost specifikovat některý aspekt mentálních procesů nemusí být v rozporu s možností jejich samovolného vzniku za jen nepřímé, byť cílevědomé účasti člověka.

V rámci současné umělé inteligence se uplatňují hlavně dvě strategie, jak umožnit samovolný vznik vyšších kvalit za relativně nepřímé účasti člověka: genetické programování (viz kap. ** v tomto svazku (Lažanský, Kubalík, 1999)) a konekcionismus; hlavně na ten se zaměřím v této studii, i když mnoho z toho, co bude řečeno, lze aplikovat i na genetický (lépe řečeno evoluční) přístup.

5.1 Neuronové sítě a konekcionismus

Konekcionismus je univerzální princip, který se stal oblíbeným tématem posledních dvaceti let pod různými názvy (umělé neuronové sítě, paralelní distribuované procesy, konekcionistické systémy, spinová skla apod.).

Výzkum umělých neuronových sítí (hlavně pod tímto názvem) doznal širokého praktického uplatnění, na které však není tato kapitola zaměřena. Některé metody jejich vytváření a užití jsou popsány v kapitole 9. prvního dílu této knihy (Hořejš, 1993); v tomto dílu viz též kap. ** a ** (Kůrková, 1999; Grim, 1999). Rozsáhlejší přehled je v knize (Novák a kol., 1998).

Z filosofického hlediska je konekcionismus zajímavý hlavně jako alternativa k počítačovému funkcionalismu a jako motivační zdroj pro (nový) emergentismus (Hofstadter, 1985; Hillis, 1988); za pozornost stojí též náruživá debata jeho stoupenců se zastánci tradičního logicko-symbolického paradigmatu v umělé inteligenci (Smolensky, 1988; Little, 1990; Bringsjord, 1991; Bechtel, Abrahamsen, 1991).

Konekcionismus (jakožto paradigma) je založen na myšlence, že na mnohé složité jevy, myšlení a inteligenci nevyjímaje, lze pohlížet jako na emergentní vlastnosti paralelních dějů v rozsáhlé síti třeba i jednoduchých a vzájemně si podobných *aktivních prvků* (formálních neuronů či procesorů), mezi nimiž existují *interakční vazby*[21]. V roli těchto prvků si můžeme představovat nejen neurony (či neuronové moduly), ale i umělé elektronické prvky anebo formální proměnné simulačních algoritmů, v širším pojetí dále i atomy (jejich spiny), molekuly, buňky, živé organismy, biologické druhy, lidi, národy, terminály internetu a kdoví co ještě. (viz též odst. 6.3).

Dynamické vlastnosti konekcionistického systému (formální neuronové sítě) jsou obecně charakterizovány takto:

---

[21] Angl. ‚connections‘ – odtud název konekcionismus.

[21a] Výběr z prací K. H. Pribrama nedávno vyšel v česko-slovenském překladu (Pribram, 1999).

(a) každý prvek může být v jednom z možných stavů aktivity (zpravidla dvou),

(b) tento jeho stav závisí na stavech aktivity jiných prvků (případně, jde-li o vstupní prvek, na vnějších stimulech),

(c) stupeň závislosti dle bodu (b) je určen vahami interakčních vazeb s jinými prvky.

Konekcionistické modely kognitivních procesů (například dlouhodobé paměti, učení apod.) mají další důležitou vlastnost:

(d) váha každé interakční vazby mezi dvěma prvky se průběžně mění, a to v závislosti na aktivitě těchto prvků v minulosti.

K tomu přidejme poněkud méně exaktní, ale intuitivně významný charakteristický rys:

(e) váhy interakčních vazeb se mění podstatně pomaleji, než aktivita prvků.

U tomto posledním bodě ovšem záleží na tom, co rozumíme slovem „podstatně“. Řekněme alespoň, že oba typy změn probíhají v natolik odlišných časových měřítkách, že je lze sledovat odděleně, jako by šlo o chování dvou různých dynamických systémů (viz odst. 5.2).

V té souvislosti je třeba poznamenat, že vlastnosti (d) a (e) jsou zde formulovány v podobě vhodné spíše pro úvahy o procesech v přirozeném mozku (resp. v mysli). U neuronových sítí pro technické aplikace se zpravidla formálně i aktuálně odděluje **mód aktivní** od **módu adaptačního** (srov. kap. 9 (Hořejš, 1993) prvního dílu této knihy). Adaptační mód předchází aktivnímu módu a reprezentuje přípravnou fázi, během níž se modifikují vazební váhy pro daný úkol; programátor – či chcete-li učitel – zde zasahuje do procesu (přinejmenším) volbou vhodné trénovací množiny vstupů a výstupů. Zde nás zajímá spíše případ, kdy „učitelem“ je přirozené prostředí.

Jednu věc je třeba zdůraznit, totiž že **lokální chování**, které je dáno pravidly pro aktivitu jednotlivých prvků a pro váhy jednotlivých vazeb mezi prvky, je – na rozdíl od **globálního chování** systému jako celku – zpravidla velmi *jednoduché* (pravidla mají podobu jednoduchých matematických vztahů), *uniformní* (stejné vztahy platí pro všechny prvky a vazby) a především *srozumitelné*. Určitá nepostižitelnost globálního chování ve srozumitelném jazyce lokálního chování je významná pro filosofické úvahy (jak později uvidíme).

Máme-li na mysli použití konekcionistického systému jako modelu myslícího mozku, je rozlišení dvojího pohledu, jednoho „zdola“ a druhého „shora“, zcela namístě: pohled "zdola" postupuje od jednotlivých prvků a vazeb mezi nimi k chování sítě jako celku, zatímco pohled "shora" směřuje od globálního chování (snad i včetně jeho mentalistické interpretace) směrem k prvkům, s jejichž pomocí je realizováno. Zdola spatříme rozsáhlý dynamický systém s obrovským počtem stupňů volnosti, který se jako celek může chovat nepředvídatelně a nevysvětlitelně. Při pohledu shora si klademe otázku, co z toho, co považujeme za kognitivní chování člověka, si můžeme představit jako výsledek součinnosti (kooperace, kompetice, synchronizace atp.) obrovského množství poměrně prostoduchých prvků. U konekcionistických systémů si kupodivu lze tuto představu docela dobře připustit – aspoň v prvním přiblížení – u vybraných kognitivních funkcí, jako je například hromadění zkušeností,

učení se stereotypům, paměť (a vybavování z ní), klasifikace vjemů, formování obecných pojmů, soutěžení pracovních hypotéz, asociativní myšlení aj.

5.2 Od fyzikalismu k emergentismu

Jedním z důvodů, proč se konekcionismus stal vědecky atraktivní, je skutečnost, že dovoluje studium metodami matematické fyziky. Jde totiž o kolektivní systémy, pro něž existují velmi propracované matematické teorie, jaké známe například z termodynamiky. Zejména dnes již klasická práce J. J. Hopfielda (Hopfield, 1982) otevřela cestu ke studiu neuronových sítí pomocí nástrojů statistické fyziky.

Fyzikální podrobnosti by nás příliš vzdálily od filosofické tématiky, a proto poukáži jen na několik vybraných souvislostí. Představíme-li si úroveň prvků a jejich lokálního popisu jako *mikroúroveň* systému, lze úplnou informaci o stavech aktivity všech jeho prvků v daném okamžiku (čili úplný mikrostav systému) reprezentovat jediným bodem ve stavovém (fyzikálně řečeno: fázovém) prostoru, v němž pak trajektorie tohoto reprezentujícího bodu odpovídá časovému vývoji systému. Řadu vlastností systému na *makroúrovni* lze tudíž charakterizovat mj. globálními vlastnostmi takovýchto trajektorií. Konekcionistické systémy lze tedy studovat v rámci matematické teorie dynamických systémů a mluvit mimo jiné o atraktorech různých typů, o stabilitě, energetické krajině s energetickými minimy, relaxaci, chaotickém chování, entropii apod., srov. (Domany, van Hemmen, Schulten, 1991).

Nepotřebujeme příliš fantazie, abychom v některém typu makroskopického chování konekcionistických systémů našli překvapivé analogie s mentálními jevy, jako je například bistabilní vnímání, utkvělé myšlenky, váhání při rozhodování, ba i halucinační stavy, fantazijní představy, chaotické příběhy ve snech a jiné projevy vnitřního života.

Paul Smolensky (Smolensky, 1986) si položil otázku, zda v neuronových sítích existují fázové přechody, jaké známe ve fyzice. Byly by to jakési "body tuhnutí", při nichž dochází k závažnějším rozhodnutím, která se při nízkých "teplotách" již nemohou (v rozumném čase) měnit. Smolensky přirovnal tento proces tuhnutí dokonce k subjektivnímu prožitku náhlé „krystalizace“ neuspořádaných myšlenek do koherentního tvaru, nápadu. Tato krystalizace nemusí přitom mít jednoznačný cílový stav. K obdobným úvahám vede stochastická varianta konekcionistických systémů, známá jako tzv. Boltzmannovy stroje (Hinton, Sejnowski, 1986).

Díky již zmíněné diferenci v časových měřítkách pro aktivitu resp. pro změnu vah (viz bod (e) v předchozím odstavci) je možno se na konekcionistickou neuronovou síť dívat jako na dvojici dynamických systémů, jednoho pro „rychlou“, aktivní dynamiku ve stavovém prostoru, druhý pro „pomalou“, adaptační dynamiku v prostoru vazebních vah. Buď ten či onen systém lze snadno ztratit z očí pouhou změnou měřítka času. Podstatné ovšem je, že tyto dva systémy spolu vzájemně interagují (body (c) a (d)). V intuitivním pohledu jeden směr působení, od dynamiky aktivní k adaptační, reprezentuje proces *učení*, tj. hromadění „zkušeností“ (znalostí, dovedností) z aktivního procesu, opačný směr pak nahromaděné zkušenosti (znalosti, dovednosti) *aktivuje*. V biologické analogii vztah mezi aktivní a adaptační dynamikou připomíná interakci mezi jednotlivými zážitky a imprintingem na úrovni jedince nebo interakci mezi zkušenostmi populace a evolucí na úrovni druhu.

Učení vždy vyžaduje *paměť*. Zde se ovšem střetává trojí pojetí paměti: fenomenální (subjektivně prožívaná) paměť, performační paměť (objektivně charakterizovaná jako proces o třech fázích: záznam–uchování–čtení) a fyzická paměť

(důraz na příslušný materiální nosič). Mluví-li se o modelech paměti, spíše se tím rozumí fyzická paměť, což má své filosofické důsledky, protože například jen ve vztahu k ní má smysl otázka po *lokalizaci* paměťových stop.

Konekcionismus nabízí představu *distribuované paměti*: paměťové „záznamy" mají podobu modifikací vazebních vah (jejich přírůstků a úbytků) ve struktuře o velkém množství prvků sítě; tyto „záznamy" se navíc mohou vzájemně rozličným způsobem „překrývat". Připomíná to *holografickou teorii* paměti v lidském mozku, kterou K. Pribram (Pribram, 1971) navrhl jako alternativní koncepci k předpokladu *fyzických engramů*, čili kontinuálně existujících a lokalizovatelných „vrypů" ve struktuře mozku (Lashley, 1929). Postupem času se ukázalo, že předpoklad engramů odporuje empirickým poznatkům, například relativně malé degradaci paměti při lokálních poruchách mozku.

Karl Pribram, v usilovné snaze odhalit, jak mozek (či mysl) analyzuje vjemy a jak si ukládá a vybavuje naše znalosti a zážitky, narazil na práce Dennise Gabora, který navrhl využít Fourierovy transformace k reprezentaci obrazů, princip, který byl později prakticky využit v holografii. Zjednodušeně řečeno, informace o předmětech a jevech lokalizovaných v času a prostoru je vyjádřena v podobě frekvenčních amplitud a fázových posunů v tzv. frekvenční doméně. Pribram si všiml, že mnoho neurofyziologických a neurologických poznatků by se dalo lépe vysvětlit, kdyby příslušné funkce mozku byly založeny na holografickém principu.

Na holografickou teorii lze pohlížet různě: jako na metaforu pro jakoukoliv distribuovanou reprezentaci (kdy každá část nese informaci o celku), jako na výzkumný program, anebo jako na realistický model mozkové činnosti, odvozený z empirických dat. V novějších pracech Pribram uvádí konkrétní neurofyziologické argumenty pro názor, že holograficky (či v novějším pojetí holonomicky) se chovají mikroprocesy na dendritických sítích skupin neuronů, zatímco neurony samotné jen zajišťují přenos výsledků dendritických mikroprocesů do jiných úrovní zpracování (Pribram, 1991)[21a].

Podobně jako u holografické paměti není ani v konekcionistických modelech snadné mluvit o kontinuální fyzické identitě paměťových záznamů, lze však mluvit o *implicitních engramech*, které nabývají fyzickou povahu (a kauzální roli) jen pokud jsou vhodnou stimulací media aktivovány.

Jak vidno, mezi matematicky pojímaným konekcionismem a matematickou fyzikou lze najít více souvislostí, než by se na první pohled zdálo a postoj zdůrazňující tyto souvislosti by se mohl nazývat „fyzikalismem". Avšak termín **fyzikalismus** se ve filosofii mysli užívá pro redukcionistický materialismus, dle něhož mentální stavy a vlastnosti nejsou *nic než* fyzické stavy a vlastnosti, přičemž slovo ‚fyzické' je míněno – v opozici k ‚biologickému' či ‚mentálnímu' – téměř ve smyslu molekulárním či submolekulárním (proto ono „nic než"). V případě konekcionismu je tomu spíše naopak, jde tu o dynamické systémy nezávislé na fyzické povaze substrátu, v němž jsou realizovány. Důrazem na tuto nezávislost se nám připomene funkcionalismus, i když ten je často chápán jen ve vztahu k jisté zvolené úrovni popisu.[22]

Zaměřme se na jeden (podle mého názoru hlavní) aspekt konekcionistické myšlenky: že zde jde v tom či onom smyslu o vztah (zpravidla) *dvou úrovní*, ať už jde o rozlišení na úrovně „dolní" a „horní", lokální a globální, mikro a makro, aktivní a

[22] Srov. však pojem homunkulárního funkcionalismu v odst. 7.2.

adaptační, nebo o rozlišení mezi fyzickou neuronovou sítí a vyššími kognitivními funkcemi, které se v ní (snad) realizují. Toto rozlišení je důležité nejen pro zdůraznění úrovňového charakteru konekcionismu, ale i pro objasnění pojmu emergence.

Za **emergentní** jev můžeme obecně pokládat cokoliv, co je na určité „vyšší" úrovni zřetelné a svébytné (symetrický tvar sněhové vločky), kauzálně působivé (úder pěsti), či jakkoliv jinak pozoruhodné (lavina, inflace), a na co lze hledět jako na důsledek vlastností a chování prvků nějaké „nižší" úrovně (molekul H_2O, svalových buněk, sněhových vloček, ekonomických subjektů), přičemž to není snadné nebo dokonce možné prostředky této nižší úrovně popsat, vymezit či předpovědět (Hillis, 1988). K pojmu emergence se ještě vrátím ke konci této studie (odst. 7.3).

Tzv. **emergentistická teze**[23] (v kontextu filosofie mysli) zní takto:

Mentální stavy a procesy lze pojmout jako emergentní jevy na některé vyšší úrovni dostatečně složitého dynamického systému.

Za předpokladu, že jde speciálně o konekcionistický systém, se na tuto tezi budeme odvolávat jako na **konekcionistickou tezi**.

O různých názorech na oprávněnost nebo neoprávněnost užití konekcionistické teze pro vysvětlení mysli se zmíním později; na tomto místě aspoň jednu poznámku: jakkoliv málo toho víme o lidském mozku, je jasné, že neuronová tkáň je v jednom z mnoha možných pohledů rovněž příkladem konekcionistického systému (jehož prvky jsou neurony nebo shluky neuronů a jehož vazby jsou zprostředkovány synapsemi). Jde tedy o konekcionistický systém, který s lidským myšlením bezesporu něco společného má.

5.3 Debata mezi tradičním a konekcionistickým paradigmatem

Vzhledem k evidentnímu významu konekcionistického výzkumu jednak pro vědy o mozku a jednak pro moderní směry v architektuře počítačů, by se dalo očekávat, že dojde ke splynutí tradiční umělé inteligence s konekcionismem. Místo toho došlo k rozsáhlé a vzrušené debatě mezi zastánci obou přístupů. V extrémní podobě jedni trvali na tom, že symbolické a algoritmické procesy jsou to jediné a pravé, co je třeba k umělému i přirozenému myšlení, druzí tvrdili, že symbolická mysl je iluzí či epifenoménem (jevem bez kauzální potence) „plovoucím" na dynamickém toku subsymbolické informace v neuronové síti. Vyskytly se ovšem i sbližovací tendence, například názor, že kognitivní funkce lze nacházet na *všech* úrovních popisu (Bringsjord, 1991, s. 324.)

V čem vlastně tkví problém? V odst. 4.2 jsme se setkali s omezeností umělé inteligence co do schopnosti konstruovat jako „umělé" to, co neumíme předem popsat jako cíl projektu. Konekcionistická teze z předchozího odstavce nabízí možnost toto omezení obejít přesunem naší konstrukční aktivity na vhodnou nižší úroveň.

Problém vztahu mezi tradiční (algoritmickou) a konekcionistickou umělou inteligencí lze vyjádřit ve třech otázkách:

---

[23] Zobecněná emergentistická teze bude formulována v odst. 6.2.

[23a] Newell definuje symboly jako fyzické tvary, které se mohou vyskytovat jako komponenty složitějších výrazů; ač termín 'symbol' předjímá naši interpretaci, není zde omezen na lidské symbolové systémy (viz Newell, Simon, 1976, spec. (Boden, 1990), s. 109)).

(a) Může u konekcionistických systémů existovat „vyšší“ úroveň, ve které by *samovolně* probíhaly logicko-symbolické procesy obdobné těm, které lze realizovat metodami tradiční umělé inteligence?

(b) Lze tyto samovolné procesy *cíleně vyvolat* pomocí projektů, definovaných pro „nižší“ úroveň (tj. pro úroveň neuronové sítě)?

(c) Lze takto (cíleně) vyvolat navíc i takové procesy, které na „vyšší“ úrovni *nedovedeme předem popsat* (jako projekty – a tedy realizovat metodami tradiční umělé inteligence)?

Kladná odpověď na otázku (a) je jen podmínkou tázání (b) a kladná odpověď na otázku (b) by postavila principiální možnosti konekcionistického přístupu (přinejmenším) naroveň možnostem tradiční umělé inteligence. Pokud by navíc i otázka (c) měla kladnou odpověď, znamenalo by to, že konekcionismus má jasnou převahu (i když asi stěží exaktně charakterizovatelnou).

Uvedené tři otázky úzce souvisejí s rozlišováním různých úrovní popisu skutečnosti, s nímž jsme se zde v různých podobách již vícekrát setkali (včetně předchozího odstavce). V tradiční umělé inteligenci se zpravidla při programování pohybujeme na stejné úrovni (řekněme psychologické), jako při popisu a ověřování činnosti programu. Teprve metoda modelování „zdola“ nás postavila před zajímavou konceptuální otázku, zda pojem ‚umělého’ lze rozšířit i na případy, kdy to, o co nám opravdu jde, se odehrává na jiné úrovni, než zadání a realizace podpůrného procesu (naše definice není v tomto aspektu jednoznačná).

Existuje i svým způsobem opačná situace, kdy nižší úroveň je přirozená (náš mozek), zatímco na vyšší úrovni realizujeme „umělou“ činnost (jednání dle předem daného návodu, jakákoliv „mechanická“ činnost, formální logika) – takto řečeno je to ovšem na pokraji konceptuálního paradoxu: vše umělé svým způsobem pochází nakonec z přirozeného.

Konekcionistická teze v předchozím odstavci měla tvar explanačního principu (že mentální stavy lze *pojmout* jako emergentní jevy na některé vyšší úrovni dostatečně složitého konekcionistického systému). Aby měla charakter potenciálně vyvratitelné vědecké hypotézy (případně aby se stala ideologií nějakého směru), museli bychom ji poněkud zesílit, řekněme takto:

*Mentální stavy a procesy **jsou** emergentními jevy na některé vyšší úrovni dostatečně složitého konekcionistického systému.*

V této formulaci ji budu nazývat **silnou konekcionistickou tezí** (existence konekcionistického systému je zde nutnou podmínkou mysli). Ve vztahu k performačním (nikoliv fenomenálním) komponentám mentálních procesů navrhl variantu této teze (jinak formulovanou, jako tzv. *subsymbolické paradigma*) P. Smolensky (Smolensky, 1987; 1988). Konfrontujme silnou konekcionistickou tezi s **tezí symbolického paradigmatu**, které vyslovil hlavní ideolog tradiční umělé inteligence Allen Newell (Newell, Simon, 1976; Newell, 1980) krátce před objevením konekcionismu:

K tomu, aby fyzický systém vykazoval obecnou inteligentní činnost je nutnou a postačující podmínkou, aby to byl fyzický symboový systém.

Fyzickým symbolovým systémem Newell rozumí zařízení, jehož fungování lze popsat v jazyce čistě kauzálních fyzikálních vztahů – tedy je to stroj – a přitom operuje se symboly[23a]. Reprezentativním příkladem je digitální počítač.

Z teze symbolického paradigmatu mimo jiné vyplývá, že k simulaci lidského myšlení v plné obecnosti (tj. včetně jeho fenomenální komponenty) není třeba, aby byly simulovány úrovně nižší než jsou ty, ve kterých se pracuje se symboly. Neslučitelnost uvedených dvou tezí je zřejmá: teze symbolického paradigmatu umisťuje svou postačující podmínku do vyšší úrovně, než do jaké subsymbolická teze klade svou nutnou podmínku.

Z Newellovy teze by plynulo, že buď je fenomenální komponenta myšlení irelevantní pro inteligentní činnost, anebo že pro ni lze racionálním rozborem najít formální zákonitosti (ve stylu logiky myšlení), které se vztahují i na komponentu performační.

Pro úplnost poznamenejme, že lze rozlišovat mezi systémy, které se nějakými předem zadanými pravidly *řídí* (jako například typické systémy umělé inteligence), a systémy, jejichž chování lze nějakými pravidly jen *popsat* (pohyb planet). Dle D. C. Dennetta (Dennett, 1991b) existuje ještě třetí varianta: některé systémy mohou mít tendenci si pod selekčním tlakem (a v prostředí, které generuje alternativy) samy pravidla *vytvářet*, jsou k nim vedena například požadavkem optimální konstrukce.

Konekcionismus a variačně selekční princip (genetické algoritmy) mají jedno společné: oba zvyšují naději na případný úspěch tím, že – oproti tradiční umělé inteligenci –odnímají konstruktérovi část práce. Tím však současně snižují jeho zásluhu o tento úspěch. Je to vlastně měkčí verze známého paradoxu umělé inteligence: člověk by musel být inteligentnější než stroj, aby sestrojil stroj inteligentnější než je sám.

5.4 Konekcionismus a hierarchie úrovní

U konekcionistických systémů jsme se doposud setkávali vždy jen s dvojicí úrovní (neuronová síť „dole“, kognitivní procesy „nahoře“). Proč však nepředpokládat celou hierarchii úrovní? (viz pozn. v závěru odst. 4.1) A proč neomezit jejich vzájemnou redukovatelnost pouze na velmi blízké úrovně[24]? Dichotomické pojetí je nejspíše důsledkem zbytečně ostrého rozlišování mezi tím, co je explicitní a tím, co je implicitní – což může zase souviset s tím, že v uměle realizovaných konekcionistických systémech je explicitní zpravidla jen nejnižší úroveň, zatímco vše ostatní je vesměs implicitní.

Již zmíněný P. Smolensky popsal v jedné své starší práci (Smolensky, 1986) konekcionistický model, který se dosti lišil od běžných modelů neuronových sítí; s jeho pomocí se snad jako první pokusil sladit symbolické zpracování informace, o něž usiluje tradiční umělá inteligence, s konekcionismem. V našem pohledu lze v jeho snaze vidět opravdový postup „zdola“, včetně „vykročení vzhůru“.

Volně lze Smolenského model interpretovat jako pokus reprezentovat implicitní engramy (nezjistitelně rozptýlené ve vazbách mezi prvky sítě) pomocí formálních, avšak explicitních entit – **znalostních atomů**. Těmto atomům lze, analogicky jako dříve prvkům, přiřadit stav aktivity a mezi atomy, opět podobně jako mezi prvky, zavést *asociativní vztahy* – formální vazby vyššího řádu i s odvozenými vahami. Znalostní atomy a jejich asociativní vztahy takto vytvářejí abstraktní konekcionistickou síť *vyšší úrovně*. V prvním přiblížení lze znalostní atomy

[24] Pojem ‚blízkosti‘ úrovní, stejně jako sám pojem úrovně, je ovšem velmi relativní a záleží dílem na tom, co nabízí zkoumaná skutečnost, dílem na našem rozhodnutí. Srov. též poznámku na konci odstavce 6.1.

interpretovat jako nejzákladnější stavební prvky mysli – jako jakési „mikropředstavy“ a „mikrohypotézy“ o světě.

Jako (jiný) příklad lze uvést možnou hierarchizaci reprezentace znalostí pomocí reprezentačních objektů, zvaných **schémata.**Mezi schématy mohou existovat dynamické vazby, jimiž jednotlivá schémata na sebe odkazují, případně se skrze ně i ovlivňují. Schémata mohou být hierarchizována (například vzhledem k typové obecnosti), přičemž nejnižší úroveň se může (velmi) podobat neuronové síti, zatímco nejvyšší úroveň se již může blížit známým reprezentačním systémům umělé inteligence (sémantickým sítím, systémům rámců, apod. – viz kap. 4 v prvním díle této knihy (Zdráhal, 1993)), případně doplněným některými konekcionistickými aspekty (například modifikovatelností vazeb). K tomu viz též (Havel, 1990).

Další hierarchizace může spočívat v rozlišení nikoliv dvou, ale více úrovní dynamiky. Vzájemnou interakci aktivního a adaptačního dynamického systému z předchozího odstavce lze rozšířit na obecnější meziúrovňové interakce v rámci celé hierarchie dynamických systémů (odlišitelných například různými měřítky času).

6 Nová kybernetika

V tomto oddíle se zmíním o některých novějších myšlenkách, názorech a směrech, kterými se inspiruje nebo může inspirovat badatelský směr v umělé inteligenci a kognitivní věda obecně. Nazývám je souhrnně „novou kybernetikou“ protože v některých aspektech jde o představy, které se nevztahují jen k procesům v mysli.

Jako východisko nám poslouží výše popsané konekcionistické paradigma, nad nímž se lze zamyslet nejméně ve dvou směrech: co nabízí hierarchické hledisko a jaká je role kolektivních systémů. Poté zaostříme pozornost na koncepce, které zdůrazňují interakci systému s přirozeným prostředím.

Předesílám, že tento oddíl obsahuje dosti spekulativní ideje[25], pro něž většinou neexistují ucelené teorie. Z těchto idejí navíc vybírám jen některé, které jsou v souladu s naším přístupem. Existují i další, více či méně odvážné koncepce, jejichž výklad by však přesáhl možnosti této studie; jako příklad uvádím hledání souvislostí s kvantovou fyzikou – právě o tom existuje bohatá literatura, například (Hodgson, 1991; Pribram, 1993; Stapp, 1993; Eccles, 1994; Penrose, 1994).

6.1 Víceúrovňové systémy a kauzální domény

V odst. 5.4 jsem naznačil určitou ideu, jak nad základní úrovní neuronové sítě lze čistě formálně konstruovat další úrovně, které by v jistých aspektech mohly odpovídat našim představám o procesech v mysli. Přes svůj abstraktní charakter je to podle mého názoru konstrukce dosti inspirativní, a to v jednom směru: může totiž být iterována směrem „vzhůru“ a tak vést k celé hierarchii funkcionálně si podobných úrovní (tj. na vyšších úrovních by se uplatňovaly obdobné principy jako na úrovních nižších – jakýsi typ soběpodobnosti).

Již jsem upozornil na neurčitost pojmu *úroveň* (odst. 4.3). Pokud bychom si všímali, kdo a jak tohoto pojmu užívá, nalezli bychom tolik rozmanitých kontextů, že

[25] Spekulaci (v pozitivním slova smyslu) považuji za důležitou komponentu vědy (Havel, 1999a).

by se sám pojem úrovně vyprázdnil a zbyl by jen neurčitý odkaz na omezení jednoho pohledu vůči jiným pohledům. Představa úrovně vždy ovšem předpokládá, že je to *jedna z více* úrovní, které jsou navíc hierarchicky uspořádány.

Tak jako každý z nás má své přirozené okolí fragmentováno do srozumitelných domén pro tu či onu činnost, tak i pro vědecké poznání je svět fragmentován do specifických oblastí zájmu, ať již to činí pro jeden konkrétní problém či pozorování, nebo obecněji jako oblast zájmu některé disciplíny. Tyto domény nemají ostré hranice a co ke kterékoliv z nich patří záleží na tom, kam jsme v jejím rámci vůbec schopni „dohlédnout“. Pro tuto nedohlédnutelnost okraje, ale i pro postupné klesání našeho zájmu ve směru k tomuto myšlenému okraji (který naopak lze myslet právě jen díky tomuto klesání zájmu) budeme říkat, že doména má svůj **horizont**. Pro jednotlivého pozorovatele je to přirozená představa, platí to však i pro celé vědní disciplíny, jejichž pohled není tolik spojen s optikou a polohou konkrétního pozorovatele, jako spíše s pojmy, veličinami a zákony, které jsou dané disciplíně vlastní.

K tomu zpravidla patří i typická **měřítka veličin**, tedy i prostorová měřítka objektů a časová měřítka procesů, jimiž se daná (přírodovědecká) disciplína zabývá (Havel, 1996). Právě zde je zdroj některých význačných metafor (ale i nedorozumění) v kognitivní vědě. Uplatňuje se zde zejména rozlišení mezi mikrosvětem a makrosvětem (chápané téměř geometricky jako rozlišení mezi malým a velkým), což je důsledek neopatrného přenesení fyzikálního pohledu do kognitivní vědy. Jiným příkladem je naše rozlišení mezi „rychlou“ a „pomalou“ dynamikou konekcionistických systémů (odst. 5.2).

Při takovéto fragmentaci světa jsou zpravidla nejdůležitější *kauzální* vztahy; ty jsou totiž pro naše chápání klíčové, a to jak ve vědě, tak v běžném životě. Ve vědě je tomu tak po vzoru fyziky, jejíž zákony mají příkladně kauzální povahu (vzájemné závislosti veličin jsou nejčastěji popisovány diferenciálními rovnicemi, a to tak, aby při daných počátečních a okrajových podmínkách bylo možno odvodit historii systému; i stochastické zákony jsou kauzální, vedou jen k pravděpodobnostně vyjádřeným historiím). To ovšem neznamená, že bychom se omezovali jen na *fyzikální* kauzalitu; nikoliv, lze přece brát v úvahu i *mentální* kauzalitu či jakýkoliv další přirozeně myslitelný typ kauzality.

Vhodné zobecnění by nás naproti tomu mělo zbavit apriorní představy, že v pozadí našich úvah je vždy nějaká hierarchie – ta by měla být až speciálním případem (význam tohoto zobecnění se ukáže v následujícím odstavci).

Navrhuji tedy zavést pojem **kauzální domény** pro jakoukoliv oblast (výsek, fragment, komponentu) skutečnosti, v jejímž rámci se nám kauzální vztahy jeví jako *zjevné, srozumitelné a vzájemně koherentní* – přinejmenším jsou zjevnější, srozumitelnější a vzájemně koherentnější, než vztahy *mezi různými* doménami. Tato formulace je, přiznávám, poněkud neurčitá (ne o mnoho víc, než sám pojem kauzality), a proto doplňuji: mluvím-li o kauzálních *vztazích*, nepředpokládám, že známe příslušné kauzální *zákony*; dále: *zjevné* znamená, že je umíme identifikovat, *srozumitelné*, že je máme za skutečné (a jejich instancím se nedivíme) a *vzájemně koherentní*, že se nám jeví jako by společně patřily k jedné soudržné síti (kauzálnímu nexu).

Představme si, že se nám hodí určitou množinu kauzálních domén uspořádat do podoby lineárního řetězce (například podle charakteristických velikostí objektů). Teprve pak je vhodné mluvit o **hierarchii úrovní**, kde úrovněmi se rozumí jednotlivé kauzální domény onoho řetězce.Soustředíme-li svou pozornost na konkrétní kauzální doménu, zpravidla přitom nezapomínáme na existenci jiných domén, ty jsou však

samy o sobě pro nás již relativně méně zajímavé. Myšlenkově sice můžeme přecházet z jedné domény do druhé, hůře se nám však již podaří zahrnout mnoho domén do jediného pohledu. Jednotlivé kauzální domény mají vždy svůj horizont a nelze je ani ostře oddělovat od „sousedních" kauzálních domén – to je nejvíc vidět právě u různých úrovní jedné společné hierarchie. Domény lze spolehlivě považovat za různé jen tehdy, jsou-li si v nějakém smyslu dostatečně „vzdálené" (viz níže).

Fragmentace světa do jednotlivých domén je pro vědu zcela legitimní. Například S. Schweber (Schweber, 1993) považuje fyzický svět za hierarchicky stratifikovaný do oddělených „kvaziautonomních" úrovní, z nichž každá je reprezentována jinými pojmy a jinými zákony, bez podstatné závislosti na tom, co se děje v jiných úrovních. Vztah mezi úrovněmi existuje, ale má čistě emergentní (v jeho slovech nekauzální) povahu.

Zatímco v rámci jednotlivé kauzální domény klademe (tj. příslušná vědní disciplína klade) důraz především na kauzální vztahy (a zákony), povaha vztahů *mezi* různými doménami může být jiná a mnohdy ani není dost dobře poznána a pochopena (což může souviset i s tím, že různé domény nejsou v kompetenci téže disciplíny). Mohou to být různé strukturální a tvarové souvislosti, statistické vztahy, meziúrovňová kauzalita (shora nebo zdola), přechod od kvantového k makroskopickému popisu, emergentní jevy a rozličné další vztahy, pro něž třeba neexistuje ani formální ani intuitivní popis. Právě k těm posledním náleží interakce mezi myslí (vědomím) a tělem.

Užitečnost představy kauzálních domén mohu ilustrovat na jevu, který nazývám **kauzální paralelismus**.[26] Kauzální (kauzálně chápaný) vztah v jedné doméně může být doprovázen kauzálním vztahem v jiné doméně, a to dokonce v jistém smyslu „nerozlučně". Příklad: vidím hada a proto se nehýbu (kauzální vztah v doméně kognitivní a volní) *a současně*, jakoby náhodou, jistá konfigurace signálů ve vizuální oblasti mozku způsobuje, že jiné konfigurace signálů v motorické oblasti blokují svalové pohyby (kauzální vztah v doméně neurofyziologické). Tento příklad mimochodem poukazuje na určitou naši libovůli ve volbě jazyka, jímž popisujeme realitu.

Kauzální paralelismus nabízí možnost překládat souvislosti, které v jedné doméně nedovedeme dost dobře vysvětlit, do jazyka jiné domény, kde kauzální vysvětlení může být snazší. Tento překlad by však měl být podpořen znalostí příslušných vztahů *mezi* oběma doménami. Dotyčné souvislosti v původní doméně pak můžeme chápat jako kauzální v přeneseném slova smyslu – jako *virtuální kauzalitu* – ovšem za cenu, že tak poněkud znásilňujeme jejich případnou intuitivní přirozenost.

Pokud mluvíme o konkrétní hierarchii kauzálních domén, obvykle předpokládáme, že výběr a uspořádání domén do příslušného řetězce je podřízeno nějakému přirozenému kritériu, které pak též umožňuje smysluplně interpretovat *vzdálenost* mezi doménami. Platí například, že čím jsou si dvě domény v dané hierarchii vzdálenější, tím snáze je lze chápat odděleně a tím hůře je lze popisovat (studovat, chápat) jako komponenty jediného systému. Tyto úvahy lze dobře ilustrovat zejména na případu hierarchie domén lišicích se pouze prostorovým (nebo

[26] Jde o zobecnění psychofyzického paralelismu (jakožto jevu, nikoliv filosofického názoru – viz odst. 7.2)

časovým) měřítkem (čili tzv. škálové hierarchie[27]). Abstraktně pojato, takováto hierarchie by mohla být i funkčně uniformní (posunem po škále měřítek by se neměnily fyzikální vlastnosti).

Specifickou otázkou ovšem je, zda můžeme kauzální domény a jejich hierarchie považovat za něco, co nám diktuje samotná realita, anebo zda jde spíše o umělý produkt naší konstrukce a dohody. Jde tu o hlubší filosofickou otázku přesahující rámec této studie; omezím se proto jen na konstatování, že vlastně platí obojí a vzájemně se to doplňuje v jakémsi hermeneutickém kruhu. Později (v odst. 6.4) o takovémto recipročním vztahu budu mluvit jako o „zjednávání světa". Metaforicky řečeno, člověk se „domlouvá" se světem o tom, jak má chápat svět (a sebe v něm).

Lze to ukázat i na tom, jak si vytváříme konkrétní kauzální domény. Pojem kauzální domény jsem zavedl s odkazem na zjevné, srozumitelné a vzájemně koherentní kauzální vztahy, ale současně *jen díky* kauzálním doménám získáváme představu, které vztahy vlastně máme považovat za zjevné, srozumitelné a vzájemně koherentní.

Naučili jsme se – a věda nás v tom podpořila – pojímat svět jako spleť mnoha hierarchických struktur, z jejichž rozličných průsečíků si musíme vždy volit tu kauzální doménu, která nám nejlépe umožní vhled do dané situace. Takováto fragmentace reality je obecnou potřebou a úspěchem lidského porozumění světu, jakkoliv ji mnohdy pociťujeme jako vnucené omezení.

6.2 Zobecněný emergentismus

Vraťme se k našemu tématu, tj. k přirozenému a umělému myšlení. Pro obojí bychom rádi identifikovali patřičnou *hierarchii rozlišovacích úrovní*, kde každá úroveň má svůj charakteristický jazyk popisu, typ popisovaných dějů a vlastní kauzální zákonitosti. V případě stroje (počítače) je to snazší. Víme sice, že je obtížné jedním pohledem obsáhnout všechny hardwarové a softwarové úrovně, nicméně dovedeme tyto úrovně rozlišovat a specifikovat i to, jak vzájemně do sebe zapadají. Počítače jsou totiž takto předem i rozvrženy: každá úroveň je vytvořena na základě specifického konstrukčního projektu, zahrnujícího též vazby k nejbližším sousedním úrovním. Hierarchie je pak do jisté míry srozumitelná i jako celek.

U přirozené mysli je situace jiná. Je pravda, že některé vybrané kauzální domény mozku (lidského nebo zvířecího) máme velmi podrobně prostudovány (například fyzikálně-chemické děje na synapsích a membránách, signální a logické závislosti na úrovni neuronů a funkční oblasti mozku jako celku) a rovněž jsme schopni kauzálně popisovat doménu mentálních stavů[28]. Nicméně všechny tyto známé kauzální domény jsou od sebe relativně izolovány a mezi nimi zřejmě leží rozsáhlé neprobádané teritorium dalších možných domén. Je to tedy spíše jen naše přání si představovat funkční organizaci mozku v podobě hierarchie úrovní, přehlédnutelné podobně jako hierarchie úrovní u počítače. Ještě větší přání mají někteří materialisté a emergentisté: aby se ukázalo, že právě v této hierarchii se na některé dostatečně „vysoké" úrovni usídlila opravdová mysl.

---

[27] Škálové hierarchie (*scalar hierarchies*) se typicky vztahují k distinkci část–celek na rozdíl od tzv. specifikačních hierarchií (*specification hierarchies*), založených na distinkci speciální–obecné (Salthe, 1991).

[28] V rámci této domény si vysvětlujeme své vlastní chování a chování druhých jako výsledek domněnek, tužeb, obav, představ, očekávání a cílů – jde o tzv. *psychologický folklor* (folk psychology), odpovídající běžnému užívání mentalistických výrazů v přirozeném jazyku (Stich, 1983).

Na obě přání se dívám skepticky. I kdybychom dokázali identifikovat (zjednat) funkcionálně kompaktní (čili „hustou“) množinu kauzálních domén, nic nám nezaručí, že spleť relevantních interakcí mezi doménami se dá uspořádat do lineárního řetězce. Nemusí platit ani běžný intuitivní názor, že úroveň, na níž se (například) rozpoznávají tváře, leží někde „nad“ úrovní neuronů a vazeb mezi nimi (což je předpoklad konekcionismu). Je totiž myslitelná alternativa, že obě tyto úrovně jsou fyzicky na sobě více méně nezávislé a že jejich případný funkční paralelismus je přirozenější vysvětlovat odkazem na nějakou třetí, „hlubší“ společnou úroveň, na níž obě – každá sama o sobě a každá jinak – závisí. A dále: i kdyby se přece jen podařilo kauzální domény uspořádat do hierarchie, nijak by z toho neplynulo, že by do této hierarchie měla automaticky patřit i úroveň mentální.

Jde-li nám o porozumění životu, a mysli jakožto specifickému projevu života, neměli bychom se vázat jednou na tu, jindy na onu konkrétní úroveň, tak jak je to běžné a dokonce žádoucí ve fyzice. To, co odlišuje živé organismy (a mozky) od fyzikálních soustav (a strojů), je právě existence, souhra a interakce dějů na mnoha úrovních – či lépe řečeno v mnoha kauzálních doménách. Skutečnost, že přitom obvykle zdůrazňujeme ty domény, které jsou současně úrovněmi nějaké (typicky škálové) hierarchie – např. hierarchie úrovní molekulární, buněčné, fiziologické, tělesné, psychické, etologické, ekologické, případně různé mezilehlé –, je jen dokladem našeho sklonu k hierarchické fragmentaci světa (Scott, 1995).

I v tomto rozšířeném pojetí lze mluvit o emergentismu v podstatně obecnějším smyslu. Místo emergentistické teze (viz odst. 5.2) formulujme **zobecněnou emergentistickou tezi** takto:

Mentální stavy a procesy lze pojmout jako emergentní jevy nad rozsáhlou množinou vzájemně vázaných kauzálních domén.

Nejde tedy jen o emergentní jevy v jedné z úrovní, které vznikají za podpory procesů na úrovni nižší (či dokonce na ně redukovatelné), nýbrž o jevy, které povstávají z globální sítě vzájemných vazeb – obecně nikoliv jen kauzálních – mnoha různých kauzálních domén či (speciálně) různých úrovní nějaké hierarchie. V posledním případě, jde-li o hierarchii škálovou, lze mluvit o *škálově holistické emergenci*.

Tím, že umožňuje více alternativních koncepcí teorie mysli, je zobecněná emergentistická teze logicky slabší než emergentistická teze z odst. 5.2. Současně však v ní jde o emergenci vyššího řádu: předpokládá vazby mezi kauzálními doménami, přičemž tyto vazby mohou mít samy již charakter emergentního vztahu (nižšího řádu). Zatím neexistuje obecná teorie emergence (jsou známy jen konkrétní případy) a tím méně lze něco bližšího říci o emergenci vyššího řádu. Je proto předčasné na ní budovat teorii mysli; nanejvýš lze zkusmo vyslovit názor, že pokud bychom chtěli teorii mysli budovat fyzikalisticky (v širokém slova smyslu), pak bychom ji museli založit na *zobecněné* emergentistické tezi.

6.3 Kolektivní systémy a jevy

Konekcionistické paradigma v umělé inteligenci obrátilo pozornost teoretiků na **kolektivní princip** – skutečnost, že systémy o mnoha prvcích, které spolu nějakým způsobem interagují, mohou jako celek vykazovat zajímavé dynamické chování. Důležitý je právě předpoklad *mnoha* prvků, protože právě ten vede k velké distanci (ve škálové hierarchii) mezi úrovní prvků a úrovní celku, takže vazba mezi oběma

úrovněmi mívá emergentní povahu, aniž by se přitom ztratil její svého druhu kauzální charakter. Význam kolektivního principu pro kognitivní vědu je proto evidentní.

Již začátkem století spekuloval francouzský matematik Henri Poincaré o procesu přemýšlení nad nějakým problémem jako o mobilizaci relevantních "atomů myšlení", které chaoticky víří, obdobně jako molekuly v (tehdy právě objevené) kinetické teorii plynů, a narážejíce jedna na druhou, produkují nové a nové kombinace, mezi nimiž se mohou vyskytnout i ty, které přispívají k řešení.[29] Podobně Lewis Thomas píše o "hemžení nahromaděných molekul myšlení" a o "mysli utvořené z hustých oblaků těchto struktur" (Thomas, 1981). Do třetice lze připomenout známou Hofstadterovu metaforu mysli jako mraveniště (Hofstadter, 1979).

Padesát let po Poincarém přirovnal neurobiolog B. G. Cragg a fyzik N. V. Temperley (Cragg, Temperley, 1954) bistabilní chování neuronů (pálí–nepálí) k chování atomů v mřížce, kde každý atom má dva možné stavy spinu (+1, –1) a kde existují definované vazby mezi atomy. Různé vnější stimulace mohou vést k různým stabilním konfiguracím spinů, které by tak bylo možno považovat za paměťové stopy oněch stimulací. Později byly objeveny speciální magnetické materiály, takzvaná spinová skla, v nichž lze rozličné spinové konfigurace realizovat fyzicky (Sherrington, Kirpatrick, 1975)[30]. Právě teorie spinových skel inspirovala začátkem 80. let J. J. Hopfielda k matematickému modelu abstraktních neuronových sítí (Hopfield, 1982).

Konekcionistické modely jsou jen speciálním případem kolektivních systémů. Příkladem jiného, podstatně odlišného media jsou Kauffmanovy **náhodné boolovské sítě** (Kauffman, 1991), tvořené prvky, které realizují různé boolovské funkce, přičemž výběr funkcí a propojení prvků je náhodné. Podle nastavení některých obecných parametrů (jako je například střední hustota propojení) se u těchto sítí lze setkat se dvěma odlišnými módy chování – v řeči statistické fyziky se dvěma *fázemi*: jedna fáze (ve stavovém prostoru charakterizovaná existencí velkého počtu cyklických atraktorů s malými periodami) odpovídá téměř „zmrzlé“ síti (až na ojedinělé malé ostrůvky aktivity), druhá fáze (s malým počtem atraktorů s velmi dlouhými periodami) odpovídá celkově „chaotickému“ dění. K zajímavému regulárnímu chování dochází v blízkosti *fázového přechodu* (v prostoru parametrů), kdy na pozadí „zmrzlé“ sítě existují menší či větší oblasti chaotického chování. Malá perturbace (náhodná změna stavu některého prvku) může mít za následek lavinovitou změnu stavu dalších prvků, která se šíří po síti tak, že i velmi vzdálená místa mohou tímto způsobem spolu „komunikovat“. Zvlášť zajímavé je, že takovéto specifické typy dynamického chování mají tendenci se stabilizovat i při strukturálních „mutacích“ systému.

Obdobné vlastnosti globální dynamiky lze nalézt i u kolektivních systémů, jejichž prvky již nejsou jen pasivní funkční jednotky, nýbrž plní i nějakou dílčí kognitivní či symbolickou úlohu (reprezentují například různé hypotézy), případně mají určitý stupeň autonomie (vlastního rozhodování). Pro umělou inteligenci mohou takové systémy sloužit jako základ pro strategii kombinující přístup „zdola“ a „shora“. M. Minsky v úvahách o „společenstvích mysli“ (Minsky, 1985) hovoří mj. o dělbě práce, jednak mezi prvky (agenty), jednak mezi většími částmi systému (agenturami), a rovněž o jejich hierarchické organizaci. Dnes se tímto směrem ubírá

[29] Citováno podle (Hofstadter, 1985), s. 656.

[30] S obyčejným sklem mají spinová skla společnou jen amorfnost vnitřní struktury.

tzv. distribuovaná umělá inteligence a multiagentové systémy (viz kap. 4. druhého dílu této knihy (Štěpánková, 1997)).

D. C. Dennett rozlišuje u takovýchto kolektivních systémů, tzv. *pandemonií*, dva módy globálního chování, připomínající výše popsané fáze dynamiky. Tentokrát se vzájemně odlišují stupněm autonomie prvků (Dennett, 1991a, 1991b). Jsou to:

(1) *rigidní*, byrokratický systém primitivních a poslušných „homunkulů" (Dennettův termín) a

(2) *chaotický* systém anarchistických *jedinců*, z nichž každý si zcela „dělá co chce".

Vedle toho existuje množství *přechodných* módů chování – a různé obory, od fyziky až po společenské vědy, k tomu nabízejí řadu příkladů. Například systém relativně „iniciativních" homunkulů (agentů), kteří vzájemně soutěží o prosazení každý své hypotézy (o zadaných datech) či svého řešení (zadané úlohy) či svého nápadu, nikoliv však bez respektování toho, co nabízejí druzí. Jiným příkladem jsou kolektivní systémy, v nichž dochází k emergentnímu vzniku kooperujících podstruktur (Axelrod, 1984).

V rigidním systému jen s obtížemi vznikne nová idea, v chaotickém systému je naopak okamžitě ztracena, v systému v přechodném módu se však může volně šířit prostředím – přesně to, s čím se lze setkat v blízkosti fázového přechodu. Kolektivní systémy agentů s výrazně paralelní „architekturou" jsou pozoruhodnými alternativami ke klasickým sériovým výpočtovým modelům. Dennett (Dennett, 1991a) staví na takových systémech svou teorii, jak funkcionalisticky vysvětlit vědomí (viz odst. 8.2).

Možnost výskytu dvou nebo více odlišných fází či módů chování a „inteligentních" přechodových módů na rozhraní mezi nimi jsem uvedl jen jako názorný příklad situace, kdy systémy o zcela odlišné povaze (fyzikální, biologické, společenské) si mohou být za určitých okolností pozoruhodně podobné co do emergentní dynamiky. Leccos nasvědčuje tomu, že studium složitých nelineárních dynamických systémů umožní hledat společnou teorii pro mentální, biologické a společenské jevy (k tomu viz též (Langton, 1989)). Matematická formulace společných zákonů pro různé jevy sice podporuje, ale nedokazuje funkcionalistickou tezi, dle níž bychom měli zcela ignorovat odlišnou materiální podstatu srovnávaných jevů (odst. 7.3).

Za kolektivní systém lze s výhradami považovat i kteroukoliv *lidskou komunitu*. Formální přístup v rámci teorie dynamických systémů však nepostihne fakt, že my, lidé, jsme nejen pasivními prvky kolektivního systému, ale současně i vědomými a svobodnými bytostmi, které na jedné straně mají své vlastní rozmary a individuální záměry, na druhé straně znají své postavení v celém systému a mohou se tudíž snažit (v rámci svých možností) jeho vývoj cílevědomě ovlivňovat.

Zde je snad vhodné místo připomenout i kolektivní inteligenci kyberprostoru, o níž byla zmínka v souvislosti s futurologií (odst. 2.4). Kyberprostor (vzpomeňme na internet) je prostředí, v němž velké množství osob může kdykoliv a kdekoliv na světě vytvářet, vysílat, přijímat a číst zprávy, nebo je ukládat tak, aby mohly být sdíleny s ostatními účastníky (Havel, 1999b). Lze tudíž hovořit o *hybridním systému*, jehož prvky jsou texty, programy, uživatelské stanice, ale i samotní uživatelé – a právě v tom může jít o novou, málo studovanou variantu kolektivního systému. Uživatelé rozumějí textům a vybírají si cestu skrze hypertexty, vyhledávají zajímavé zprávy a vkládají do systému své vlastní nápady a postřehy. Současně se však (jak jsem

naznačil v citovaném futurologickém odstavci) mohou v celém systému uplatňovat konekcionistické, případně i evoluční principy.

6.4 Zjednávací princip

Až doposud jsme se zabývali přirozeným nebo umělým myšlením jako by v obou případech šlo vždy o samostatný a svébytný proces, pojmově i funkčně oddělitelný od okolního prostředí. Toto prostředí mělo povahu nezávislé danosti a veškerý kontakt s ním mohl být zúžen na jednosměrný tok vstupních dat (vjemy) a jednosměrný tok výstupních dat (akce). V takovémto pojetí je možné a logické mluvit o vnitřní reprezentaci světa jako o jeho zjednodušeném, přefiltrovaném a vhodně zakódovaném homomorfním modelu. Inference, plánování a další kognitivní operace se přednostně vztahují k tomuto modelu, což podstatně snižuje význam přímého kontaktu s reálným světem a umožňuje tyto operace formalizovat a algoritmizovat.

Toto je zhruba východisko tradiční umělé inteligence, která si díky tomu vysloužila přívlastky symbolická, reprezentační a algoritmická. Do jisté míry je to též postoj v pozadí konekcionistického přístupu, snad jen s tím rozdílem, že u něj nemá vnitřní reprezentace symbolický charakter (který by ponechal interpretaci symbolů na člověku), nýbrž je inkrementálně formována v podobě implicitních a nelokalizovatelných vtisků předchozích vstupních i vnitřních stavů (řekněme předchozích „zkušeností“).

Jak v tradiční, tak i v konekcionistické umělé inteligenci (v tomto odstavci si dovolím obě označovat jako tradiční) se implicitně předpokládá, že z principu má smysl pojmově rozlišovat správnou a chybnou reprezentaci a tedy také správné a chybné výstupy, přičemž kritéria rozlišování *správného* od *chybného* jsou plně v rukou vnějšího pozorovatele, který jako by věděl, jak věci opravdu jsou. Jinak řečeno se předpokládá, že umělý systém (stroj, počítač, robot) vnímá (tedy *má*) svět identický s naším (jaký jej máme *my*, lidé, či jaký jej má *naše* věda), tak jej též reprezentuje a tak s ním i zachází. Slovo ‚identický‘ nechápejme ovšem striktně, samozřejmě může jít o *jiné* zjednodušení, přefiltrování a zakódování, nicméně (opakuji) jsme to *my*, kdo konec konců hodnotí, co je správné a co nikoliv. To je samozřejmě v pořádku všude tam, kde konstruujeme stroj jako *nástroj* – ke zjemnění, rozšíření, doplnění či nahrazení lidských schopností.

Problém začíná, považujeme-li stroj za plně autonomní systém, který si „žije svým životem“. V přírodě existuje nepřeberně druhů živých organismů, jejichž zkušenostní světy jsou (musí být) jiné než je ten náš. Máme je proto považovat za nesprávné? Konec konců ani naši přirozenou zkušenost v reálném světě nemáme s čím poměřovat – jedině s ní samotnou.

Tyto problémy začínají představitelé některých nových alternativních směrů v kognitivní vědě brány vážně. Zmíním se zde především o tzv. **zjednávacím přístupu** (Varela, Thompson, Rosch, 1991), v němž se důraz na vnitřní reprezentaci vnějšího světa přesouvá na vnímání a jednání *ve* světě, který je takto vlastně spolutvořen. Příkladem zjednávacího přístupu je **reaktivní princip** v robotice, který je předmětem jiné kapitoly tohoto svazku. Filosofickou oporu má tento směr ve fenomenologii, zejména v její francouzské větvi (Merleau-Ponty, 1963).

Zkušenostní svět člověka je nerozlučně spjat s jeho tělesným ustrojením a závisí na typech distinkcí, které je schopen činit. **Vtělená kognice** (či ztělesněná, do těla

včleněná) [31] je označení, zdůrazňující dvě věci: jednak, že každá zkušenost je umožněna existencí těla s jeho senzoricko-motorickými schopnostmi, a jednak, že tyto senzoricko-motorické schopnosti jsou vloženy do širšího biologického, psychologického a kulturního kontextu. (V tomto odstavci nebudu brát v úvahu fenomenální, vědomou komponentu mysli.)

Vtělená kognice nespočívá v *reprezentování* nějakého předem daného světa nějakou předem danou myslí, nýbrž v jeho průběžném **zjednávání**[32] – tvarování světa (včetně mysli) v průběhu *historie jednání* člověka (‚jednáním' rozumím konání, aktivitu, ‚historií' rozumím životní příběh, ať už jednotlivce, komunity nebo druhu) ve zjednaném nebo zjednávaném světě. Průvodcem při tomto jednání je vnímání, to je však současně jednáním podmíněno. Tato souhra jednání a vnímání je umožněna opakujícími se senzoricko-motorickými vzorci, z nichž se postupně formují kognitivní struktury. Neurobiolog W. Freeman v této souvislosti hovoří o „neuroaktivitě" jako o cyklu založeném na aktivitě, vnímání a sledování vlastních volních pohybů, který je „kontinuálním procesem umožňujícím naše porozumění vnějšímu světu" (Freeman, 1996, s. 176).

Tyto pohledy lze podložit rozličnými příklady z fyziologie vnímání, které demonstrují podstatný podíl struktury našeho těla a způsobu, jak jej užíváme, na tom, *co* vůbec vnímáme (například při rozlišování barev). Pokud nás tělo (objektivisticky řečeno) klame, i to může ovlivnit naše jistoty o světě. Příkladem jsou fantomy chybějících údů (Melzack, 1992).

Merleau-Ponty (Merleau-Ponty, 1963, s. 13.) zdůrazňuje, že pohyb organismu je nejen podmíněn prostředím, ale současně umožňuje organismu se tomuto prostředí (jeho stimulacím) vůbec vystavit. Totéž platí obecně o jakémkoliv jednání a jakékoliv poznání. Lze mluvit o jednotlivých prožitcích i o příběhu života, o ontogenezi i o fylogenezi, o jednotlivci i o komunitě. A zejména se neomezujme na člověka, ale (zkusmo) uvažujme i organismy či obecné kognitivní systémy – přirozené i umělé.

Všimněme si, že při našem vymezení pojmu zjednávání jsme se pohybovali v kruhu: zjednávaný svět je prostředím pro jednání, ale jednání je zároveň předpokladem zjednávaného světa. Jde o případ kruhu, který je průvodním jevem každé obousměrné a vyvíjející se interakce a který je třeba považovat za její součást. Druhý kruh, důležitější (i když často zapomínaný), spočívá v tom, že i my sami (já a čtenář), tak jak zde právě teoretizujeme o pojmu zjednávání, se již pohybujeme na pozadí zjednávaného světa (individuálně i kolektivně) a i tímto pohybem jej dále zjednáváme.

Zjednávací přístup zdůrazňuje vtělenou kognici a proces zjednávání jako hlavní vysvětlující princip v kognitivní vědě. Učiňme si malé srovnání tří směrů – kognitivismu (algoritmické umělé inteligence), emergentismu (konekcionismu) a

---

[31] Angl. *embodied cognition*, viz (Varela a kol., 1991), s. 173. Srov. též (Dreyfus, 1979; Lakoff, 1999).

[32] Angl. *enaction*. Běžný význam anglického *to enact* je ‚ustanovit (zákonem)' či ‚ztvárnit (herecky)'. Varela, Thompson a Roschová (cit. d., s. 9, 140, 147a) užívají různé tvary tohoto slova – *(cognition as) enaction*, *enactment (of world)*, *enactive (cognition, approach,cognitive science)*, *enacting (significance)*. České sloveso *zjednat* (*zjednávat*) volím proto, že nejlépe vyjadřuje proces, o který zde jde – chápeme-li slovo ‚jednat' spíše ve smyslu ‚konat' (angl. *act*), nikoliv ‚rokovat'. Podobnost dvojice *act – enact* a dvojice *jednat – zjednat* je případná, stejně jako náznak reciprocity, obsažený v českém slovu ‚zjednat'.

zjednávacího přístupu – pomocí této testovací otázky[33]: *Jak poznáme, že kognitivní systém náležitě funguje?*

Kognitivismus:

> *Když symboly vhodně reprezentují určitý aspekt reálného světa a když zpracování informací vede k úspěšnému řešení zadané úlohy.*

Emergentismus:

> *Když uvidíme, že emergentní vlastnosti (a výsledná struktura) korespondují s nějakou specifickou kognitivní schopností -- úspěšným řešením požadované úlohy.*

Zjednávací přístup:

> *Když se systém stane součástí již existujícího světa (případ mláděte každého druhu) nebo když si vytvaruje svět nový (případ evoluční historie druhu).*

Jedním z hlavních témat kognitivismu je vnitřní reprezentace jakoby předem daného světa. Konekcionismus se od něho liší jen způsobem, jak je tato reprezentace vytvářena. Zjednávací pohled je v tomto směru, jak jsme viděli, radikálnější.

Jak je to se vztahem ke světu v případě umělých systémů? Tato otázka je adresována speciálně robotice (jejímu badatelskému směru). Roboty jsou totiž charakterizovány především interakcí s reálným, fyzickým prostředím. Ani moderní kognitivní roboty, řízené počítači s programy umělé inteligence, nemají – na rozdíl od umělé inteligence – své vstupy a výstupy v podobě kódovaných symbolických dat (a pokud mají, tak jen pro komunikaci s člověkem).

Běžné pojetí robotiky nerespektuje zjednávací pohled. Je charakterizováno důslednou, byť nevyřčenou snahou vyhovět ideji Turingova testu: zkonstruovat robota, který vnímá a jedná jako člověk, myslí jako člověk, snad i vypadá jako člověk – krátce řečeno, který předstírá, že je člověk.

Domyšleno do důsledků nelze tento projekt nikdy splnit: vždy se takový robot bude lišit od člověka, člověk totiž nemůže *předstírat*, že je člověk, a to právě proto, že *je* člověk. Zní to jako slovní hříčka, ale je v tom hlubší filosofická pravda. Jestliže si člověk, podobně jako každý přirozený kognitivní systém, svůj svět zjednává na pozadí své historie (osobní i druhové), nemůže si robot-artefakt zjednat *stejný* svět, ten mu nanejvýš může být konstruktérem implantován v podobě lepšího či horšího modelu – čili v podobě tradiční umělou inteligencí tolik zdůrazňované *reprezentace* prostředí. Vznikne ovšem další rozdíl, protože čím je reprezentace lepší, tím je bližší „lidskému“ světu, a proto tím méně kompatibilní nejen s historií, ale i s tělem robota. Robot takto může člověka napodobovat jen v některých vnějších aspektech, nikoliv v opravdovém procesu zjednávání světa.

Kdyby naopak umělý systém (v přirozeném prostředí) měl napodobit zjednávání jako takové, což je možnost principiálně snad i myslitelná, byl byv tomto aspektu podobný člověku, nicméně *jeho* zjednaný svět, založený na *jeho* historii a *jeho* těle, by se musel radikálně lišit od lidského světa. To by se muselo projevit i na jeho jednání (interpretované námi v našem světě).

Robotika má tedy před sebou dvě možné cesty (do jisté míry to platí i pro aplikovanou robotiku):

---

[33] (Varela, Thompson, Rosch, 1991)

(a) *kognitivní robotika*: napodobovat nebo rozšiřovat *naše* (lidské) performační schopnosti za použití reprezentace *našeho* světa,

(b) *zjednávací robotika*: konstruovat umělé organismy bez takovéto reprezentace, vybavené však pro ně specifickými prostředky percepce a jednání, umožňujícími zjednávání *jejich* světa.

V současnosti je druhá cesta reprezentovaná hlavně pracemi R. Brookse (Brooks, 1991; Steels, Brooks, 1995), jehož koncepce tzv. **reaktivních agentů** je podrobně popsaná v kapitole ** v tomto svazku (Kelemen, 1999). Nebudu se zde proto blíže zabývat principem reaktivity a omezím se na jeho vyjádření Brooksovými, často citovanými slovy:

Svět sám je svojí nejlepší reprezentací.

Lze vhodným způsobem „zkřížit" princip reaktivity s kolektivním principem z předchozího odstavce? Pro konekcionistické systémy je typické dvojí, za prvé, že podstatou roli hraje vzájemná interakce prvků, a za druhé, že systém je funkčně homogenní, tj. prvky mají stejný nebo podobný typ lokálního chování. V obou aspektech by se kolektivní systém z reaktivních prvků (agentů) musel lišit: agenti interagují individuálně přímo s prostředím a dělba práce mezi nimi je přirozeným předpokladem (případ subsumpční architektury, srov. odd. ** (Kelemen, 1999, odd. 8)). Rozlišení mezi lokálním a globálním je tedy spíše funkční než strukturální.

Menší či větší množství prvků (agentů) může vytvářet vyšší jednotky – společenství, society, pandemonia, komunity, kolonie, populace, multiagentní systémy (různí autoři užívají různé termíny) – které již mohou individuální interakci prvků přímo s prostředím kombinovat s jejich vzájemnou interakcí. Čistě formálně není obtížné shrnout všechny myslitelné případy pod společnou definici kolektivního systému a teprve potom rozlišovat, zda převládají vnitřní vazby nebo vazby s okolím, zda jde o systémy funkčně homogenní nebo heterogenní a v neposlední řadě, zda prvky chápeme jako fyzické entity nebo jako abstraktní procesy a funkce (čili speciálně, zda mluvíme o mozku nebo o mysli). Z filosofického hlediska je zajímavá otázka, jaké principiálně nové kvality může přinést spojení zjednávacího principu s emergentním chováním na úrovni celku – zejména jde-li o možnost srovnání s lidským mozkem, který je rovněž kolektivním systémem, nikoliv však s typicky reaktivními prvky.

I příroda si v některých případech zvolila cestu „kolektivizace" místo zvyšování vnitřní, organické (či orgánové) složitosti. Zdá se (v našem lidsky omezeném porozumění), že bakteriální kolonie, mraveniště a včelstva jsou „inteligentnější" než jejich členové a že to tedy pro přírodu není slepá cesta.

7 Filosofický problém mysli a těla

V této kapitole si učiníme malý výlet do problému, který je v centru zájmu filosofie: jaký je vztah mezi myslí a hmotným tělem a jaká je jeho povaha. Úvahy na toto téma zde byly dávno před umělou inteligencí, její vznik však pro ně byl novým impulsem. Filosofové nemohou ignorovat skutečnost, že vybrané aspekty mysli se dají simulovat na počítači, ba i technicky modelovat; tím méně mohou ignorovat troufalá tvrzení tvůrců těchto simulací a modelů. A naopak každý, kdo se zabývá umělou inteligencí a je aspoň trochu hloubavý, se brzy setká s otázkou, co je vlastně

předmětem jeho snažení a zda to má naději na úspěch. Máme proto aspoň dva důvody se i zde těchto otázek aspoň letmo dotknout.

Problém mysli a těla, též zvaný **psychofyzický problém**, by neexistoval, kdyby tu již před jakoukoliv jeho formulací nebylo nějaké předběžné intuitivní rozlišení kategorie *mentálního* (či psychického, duševního, vnitřního) od kategorie *fyzického* (tělesného, materiálního, vnějšího)[34]. Není náhodou, že někteří myslitelé vidí „snadné" řešení problému – prostě se ho zbavit zpochybněním onoho rozlišení jako zdánlivého a neodůvodněného.

Chceme-li nastolit hlubší filosofický problém, není vždy vhodné pokoušet se o jeho přesnou definici – museli bychom totiž nejprve vymezit pojmy, které hodláme v definici použít, a to je někdy možné jen za předpokladu, že už známe řešení dotyčného problému.[35] Je proto vhodnější si nejprve zmapovat terén pomocí rozmanitých otázek, v nichž je dotyčný problém již nějak obsažen, anebo které k němu logicky vedou.

Pro náš problém mysli a těla jsem si z několika knížek vypsal namátkou tyto otázky:

- Jakým způsobem může hmota nabývat mentální vlastnosti? Mohou být tyto vlastnosti kauzální? (Rey, 1997)
- Jsi myslí s tělem nebo jsi tělem s myslí? (Priest, 1991)
- Je mysl totožná s tělem nebo je od něj různá? (Jacquette, 1994)
- Jak máme chápat člověka jakožto *personální* jednotu, když je na jedné straně *tělesná smyslová bytost* a na druhé straně *duchový subjekt*? Jak tyto dvě stránky bytí patří k sobě? (Anzenbacher, 1990)
- Můžeme rozumět svým vlastním myslím a mozkům? (Hofstadter, 1979)
- K čemu se vztahují termíny „mysl" a „tělo"? Lze nalézt znaky, které by spolehlivě oddělily mysl a tělo? Jaký je vztah mezi myslí a tělem? (Nosek, 1997)
- Má být vědomí demystifikováno? (Dennett, 1991a)
- Mohou mít počítací stroje bolesti, Marťané naděje a nehmotní duchové myšlenky? (Fodor, 1981)

Všimněte si, jak se u těchto otázek těžko pozná, zda jsou kladeny jako otázky *ontologické* (o povaze bytí), *epistemologické* (jak něco poznáváme), či *jazykově analytické* (jak o tom mluvíme). Též si všimněte, že kromě našich dvou „světů" mají tyto otázky též co dělat s *kauzalitou* a *totožností* (pojmy, jež vedou k dalším otázkám).

Problém mysli a těla náleží k širšímu okruhu otázek, kterými se zabývá **filosofie mysli**. V širším pohledu bychom v (současné) filosofii mysli mohli rozlišovat *fenomenologický* a *analytický* proud[36]; pod specifickým názvem „filosofie mysli" (philosophy of mind) dnes vystupuje jeden z proudů analytické filosofie. K

---

[34] Slova v závorkách nejsou zcela souznačná.

[35] Což mj. ilustruje podstatnou odlišnost práce ve filosofii a v matematice. Na to, jak filosofii může škodit, když podlehne lákadlům exaktní matematické metody, poukázal (význačný matematik) Gian-Carlo Rota (Rota, 1999).

[36] O fenomenologické filosofii viz např. (Patočka, 1993), jako úvod do analytické filosofie lze uvést (Peregrin, 1992).

problému mysli a těla a k dalším tradičním filosofickým otázkám (problému osobní identity, svobodné vůle, intersubjektivity) přistupuje též v poslední době široce diskutovaný problém *vědomí*. Speciálně si filosofové kladou otázku, zda vědomí, jako něco, co je přístupné jen vnitřnímu pohledu, lze považovat za předmět objektivního vědeckého bádání. Této otázce věnuji poslední oddíl (8) této kapitoly.

7.1 Mentální stavy, akty a procesy

Každý z nás má své jistoty, nejistoty, názory, nápady, touhy, vzpomínky, předtuchy, radosti i strasti. Každý dovede myslit, uvažovat, rozhodovat, ptát se sám sebe a sám si odpovídat. Čili každý z nás má zkušenost čehosi, co lze nazvat myslí či duševnem.

A tu bych rád požádal čtenáře, aby přerušil čtení a na chvíli se zahloubal nad svou vlastní myslí – nad tím, *jaké je to*, když na něco myslí, když něco vnímá, když si něco vybavuje či představuje, ale i když jen tak sedí a prostě se "nějak cítí". To vše jsou duševní čili mentální stavy a patří k nim i samo ono zahloubání, vnitřní reflexe (introspekce). Navíc jsou to *vědomé* stavy mysli, do obecného pojmu "mysl" lze totiž počítat *nevědomé* či *neuvědoměné* stavy (jako například když něco vím, ale právě na to nemyslím). Ty nelze přímo reflektovat, lze je jen předpokládat, obdobně jako jen předpokládáme vědomé stavy jiných osob.

Buďme o něco konkrétnější a uvažujme tyto příklady: *vidím modrou knihu; bojím se hadů; raduji se z dárku; chci zatleskat; hraji šach; představuji si jednorožce; myslím, že číslo 57 je prvočíslo;...*[37] Tyto příklady mají jedno společné: kromě toho, že jsou prožívány a tedy mají svůj **fenomenální aspekt** ("jaké je to" *vidět, bát se, radovat se, chtít,...*), všechny se navíc něčeho *týkají*, jsou *o něčem*. Ve filosofii se proto nazývají **intencionálními stavy** a to, k čemu se vztahují, jsou jejich **intencionální obsahy**. V pojmu intencionality je zde podstatně zobecněno běžné pojetí intence (mít v úmyslu, záměru; srov. příklad *chci zatleskat*). V našich příkladech jsou intencionálními obsahy po řadě *modrá kniha, hadi, dárek, zatleskání*, *(hra) šach, jednorožec* a celá propozice "*číslo 57 je prvočíslo*". Jak vidno, může jít o existující či neexistující věci či typy věcí, stejně jako o pravdivé či nepravdivé propozice. Některé intencionální stavy mají i **performační** komponentu, kterou lze popisovat a analyzovat "zvenku". Rozlišení fenomenální a performační komponenty bylo zavedeno v odst. 4.2, zde si je znovu připomeňme na příkladu hraní šachu: šachista jednak prožívá šachovou hru fenomenálně (teší se z ní, touží vyhrát atp.) a jednak mechanicky pohybuje figurkami na šachovnici podle jistých přesných pravidel. Právě performační komponentu míváme na mysli, hovoříme-li o programování nějakého mentálního procesu na počítači.

Existují též stavy mysli, které nejsou intencionální v uvedeném slova smyslu. Uvažme například: *rozhlížím se, trpím, mám dobrou (špatnou) náladu, úzkost, bolest, nervozitu*. Tyto stavy nemají (explicitní) intencionální obsah a neumíme jim ani přisuzovat nějakou performační komponentu. Tím lze vysvětlit některá nedorozumění o možnostech umělé inteligence ((Birnbacher, 1995), viz též závěr odst. 4.4).

O mentálních (či psychických) entitách se mluví v závislosti na kontextu jako o vlastnostech, stavech, aktech, jevech či procesech – rozlišování není jednotné (aspoň dokud nemluvíme v rámci určité realizační teorie). Slovo "stav" spíše zdůrazňuje

[37] První osoba singuláru má (zde i jinde) naznačovat spíše subjektivní přístup k nějakému jevu.

trvání, slovo "proces" dynamiku, „událost“ změnu stavu, "jev" subjekt pozorovatele[38]. Pro jednoduchost volím termín "stav" jako zástupce všech těchto výrazů.

Existují čtyři specifické rysy, kterými se vyznačují vědomé mentální stavy (Metzinger, 1995):

- ***bezprostřední danost*** : výskyty konkrétních stavů vlastní mysli si nemusíme nijak dokazovat. Nelze se totiž mýlit v jejich fenomenálním aspektu, čili v tom, *že* je vůbec máme (například *že* vidím modrou knihu, *že* mám dobrou náladu) – můžeme se mýlit nanejvýš jen v jejich obsahu, mají-li nějaký (že to, *co* vidím, je *skutečně* modrá kniha – může to být i halucinace);[39]
- ***transparentnost*** : v důsledku bezprostřední danosti jsme vedeni k pocitu, že nás některé vědomé mentální stavy uvádějí v přímý kontakt se světem (modrá kniha je *tam*, před mýma očima), případně s námi samotnými, jako osobami situovanými ve světě (je to *moje* dobrá nálada);
- ***perspektivita*** : vědomé stavy mají své *já* (*ego*), které je prožívá či zakouší (jsem to *já*, kdo vidí modrou knihu, *já*, kdo mám dobrou náladu); perspektivita souvisí s *centrovaností* vědomí a promítá se rovněž do (zpravidla zamlčeného) gramatického "já" našich výpovědí ("Prší" = "*Já* pravím, že prší");
- ***přítomnost v čase*** : vědomé stavy mysli jsou nám dány jako současné (vidím modrou knihu = *nyní* vidím modrou knihu); i vzpomínky jsou vtaženy do současnosti (*viděl* jsem modrou knihu = *nyní* si vybavuji svůj dřívější stav "vidím modrou knihu").

Existují další dva rysy, jež běžně pociťujeme jako přirozené, jsou však zdrojem závažných problémů při objektivizaci (a některé teorie mysli je problematizují):

- ***kauzální vztahy*** : některé stavy naší mysli se nám (v odstupu) jeví jako příčiny nebo následky jiných stavů mysli, případně stavů fyzického světa (bojím se, *a proto* utíkám; dostal jsem dárek, *a proto* mám radost; mám radost, *a proto* chci zatleskat);
- ***svobodná vůle***: některé stavy mysli se nám jeví jako důsledky vlastního svobodného rozhodnutí v rámci jistého spektra možností (zatleskám *nebo* nezatleskám).

Použili jsme cestu vnitřního prožívání (ve smyslu odst. 3.2) psychického světa, abychom si uvědomili, v čem se mentální jevy mohou lišit od jevů fyzického světa, a že proto problém mysli a těla nelze apriori eliminovat. Například představa, že fenomenální procesy mysli lze vysvětlovat čistě fyzikálními pojmy (podobně jako se vysvětluje řada jiných jevů kolem nás) je přinejmenším zjednodušená a vyžaduje buď radikální zúžení sféry mysli nebo rozšíření domény fyziky. Avšak nepředbíhejme.

7.2 Malá taxonomie názorů

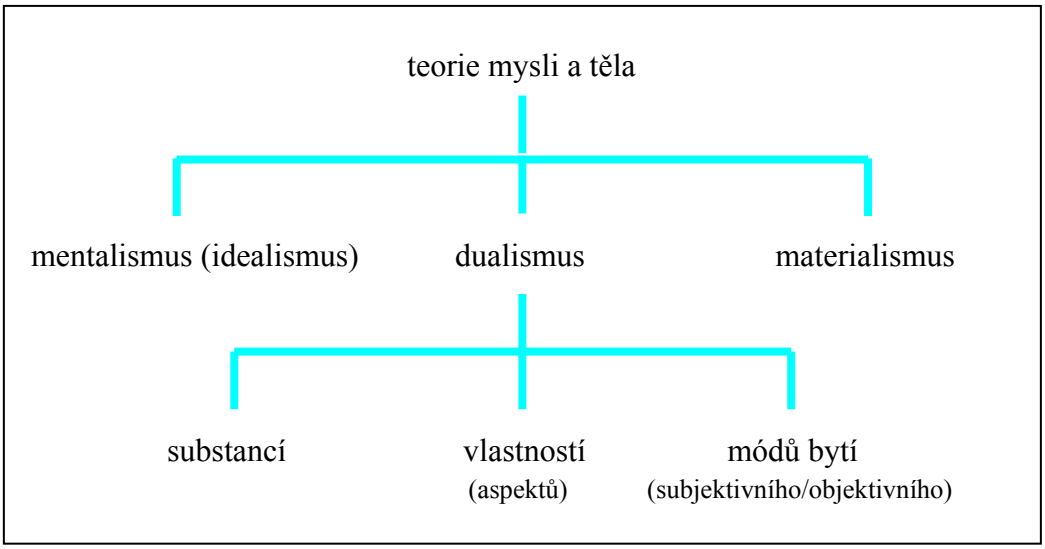
Málokterý filosofický problém doznal takové množství názorových variant, jako problém mysli a těla. Je naprosto mimo možnosti této práce udělat jejich vyčerpávající přehled a proto zde vyberu jen některé důležitější názory (které si většinou vysloužily koncovku „-ismus“), maje přitom na paměti hlavně souvislosti s umělou inteligencí. Pro snazší orientaci ve spleti různých názorů uvádím schémata,

---

[38] Slovo "jev" je v některých situacích vhodným ekvivalentem angl. *event* (možná vhodnějším než „událost“); pokud se zdůrazňuje vnitřní nazírání, mluví se o "fenoménech".

[39] Searle (viz např. Searle, 1999, s. 70) zmíněnou neomylnost zpochybňuje.

naznačující jejich vzájemnou podřízenost (obr. 1–3). Zdůrazňuji, že třídění je (jako vždy) poněkud svévolné a původní autoři by možná nesouhlasili s umístěním svých teorií.



Obr. 1. Základní dělení

Především se můžeme ptát, zda opravdu chceme oddělit kategorii mentálního od kategorie materiálního. Pokud ano, jsme **dualisté**. Pokud (z principu) nikoliv, jsme **monisté** a máme pak na vybranou, zda všemu přisuzujeme povahu mentální či materiální; v prvním případě bychom se hlásili k mentalismu, ve druhém k materialismu **(**obr. 1).

Mentalismus (též **idealismus**) je teorie, že materiální svět je čistě produktem mysli, jakési neprostorové spirituální substance. V této souvislosti je často zmiňován G. Berkeley, u něhož jde vlastně o důsledný empiricismus: přímá (smyslová) zkušenost neposkytuje jistotu, že věci existují i tehdy, když je nevnímáme, a neposkytuje ani jistotu, že jsou uhněteny z nějaké látky (Berkeley, 1710/1995). Nemůže nás o tom přesvědčit ani rozumová dedukce ze smyslové zkušenosti. Můžeme se to jen domnívat, věřit tomu. Pro Berkeleyho *být* znamená *být vnímán*. K tomu však musí existovat jiná entita, mysl, která vnímá ono vnímané a pro niž *být* znamená *vnímat*. Zatímco vnímané entity jsou pasivní, vnímající mysl je aktivní – pro toto rozlišení je i Berkeleyho idealismus někdy považován za odrůdu dualismu (Jacquette, 1994).

Dualismem se v širším slova smyslu rozumí rozlišování dvou (nebo více) typů skutečnosti či jejího pojmového uchopení, přičemž jeden typ nelze redukovat na druhý. V poněkud užším slova smyslu se dualismem rozumí teorie, že mysl je zcela něco jiného než hmota (tělo). Největší vliv na novověké uvažování o problému mysli a těla měl karteziánský dualismus (Descartes, 1641/1997), který je příkladem **substančního dualismu**. Dle Descarta jsou dva typy substance: prvý typ – materiální (*res extensa*) – se vyznačuje rozprostraněností, a druhý – mentální (*res cogitans*) – je zaručen vědomým myšlením a vědomou zkušeností subjektu (srov. rys bezprostřední danosti zmíněný v předchozím odstavci). Zatímco materiální svět lze vystavit rozličnému pochybování, nelze tak učinit se světem mysli (nezbytné již k onomu pochybování), což vede Descarta k tezi o rozdílnosti obou světů.

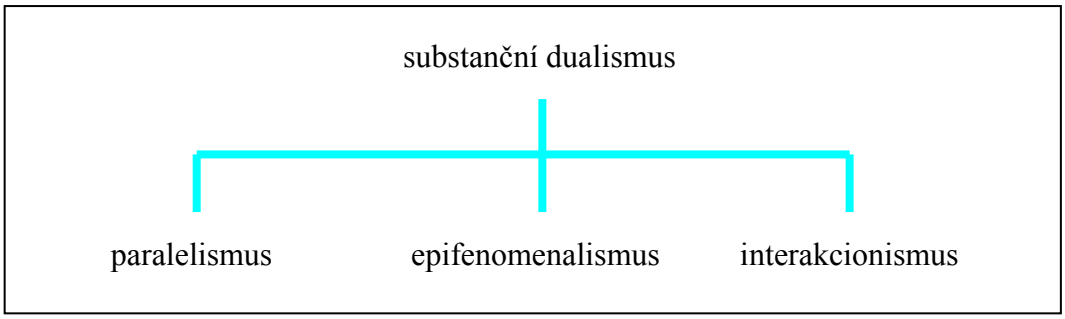
Existuje „slabší“ varianta dualismu, tzv. **dualismus vlastností** (či **teorie dvojího aspektu**): mysl (či duše) a tělo neexistují odděleně od sebe, nýbrž jsou to jen dva vzájemně neredukovatelné typy vlastností téže bytosti. Speciálně pak není mentální kauzalita v rozporu s kauzalitou fyzickou, jde jen o dva aspekty téhož vztahu. Co jsou v jednom aspektu záměry, emoce či myšlenky, jsou v druhém aspektu fyzické jevy v mozku. Pokud bychom se takto omezili jen na aspektuálně pojaté kauzální vztahy, mohli bychom – v pojmech odst. 6.1 – též mluvit o rozdílných kauzálních doménách.

Za další typ dualismu lze považovat **dualismus módů bytí** Johna Searla (Searle, 1992). On sám by proti svému řazení k dualistům asi silně protestoval (neuznává samotný termín), nicméně prosazuje dvojí ontologii: *subjektivní* ontologii 1. osoby a *objektivní* ontologii 3. osoby. Nejde o rozlišení subjektivního (zaujatého) pohledu z perspektivy pozorovatele od objektivního (nezaujatého) „pohledu odnikud“, ale spíše o to, že dle Searla vnitřně prožívané, „soukromé“ mentální stavy prostě *jsou* jinak, než „veřejné“ věci reálného světa, jakkoliv jsou jejich produktem (realizovány ve struktuře mozku).

Moderní filosofii mysli bychom mohli pojednat jako historii snah zeslabit, obejít, odmítnout nebo vyvrátit substanční dualismus, jenomže za použití jeho vlastního slovníku, což je ostatně právě to, co Searle moderním materialistům vyčítá. Pravé kořeny odmítavého postoje západní filosofie vůči dualismu lze tušit ve vědomé nebo podvědomé averzi k čemukoliv, co by mohlo zavánět duchařstvím; racionální námitkou proti substančnímu dualismu je však spíše problém, jak vysvětlit vzájemnou **kauzální interakci** světa mentálního a světa fyzického. V moderním pohledu veškeré kauzální příčiny fyzických jevů jsou rovněž fyzické a měly by tedy náležet do stejného fyzického světa. Naproti tomu zkušenost nás učí, že k přirozeným vlastnostem mysli náleží nejen schopnost reagovat na dění ve fyzickém světě, ale též na tento svět působit (prostřednictvím těla, které k onomu světu též náleží). Descartes takovou interakci mezi světy předpokládal (její sídlo kladl do epifýzy, v té době právě objevené a nezdvojené části mozku).

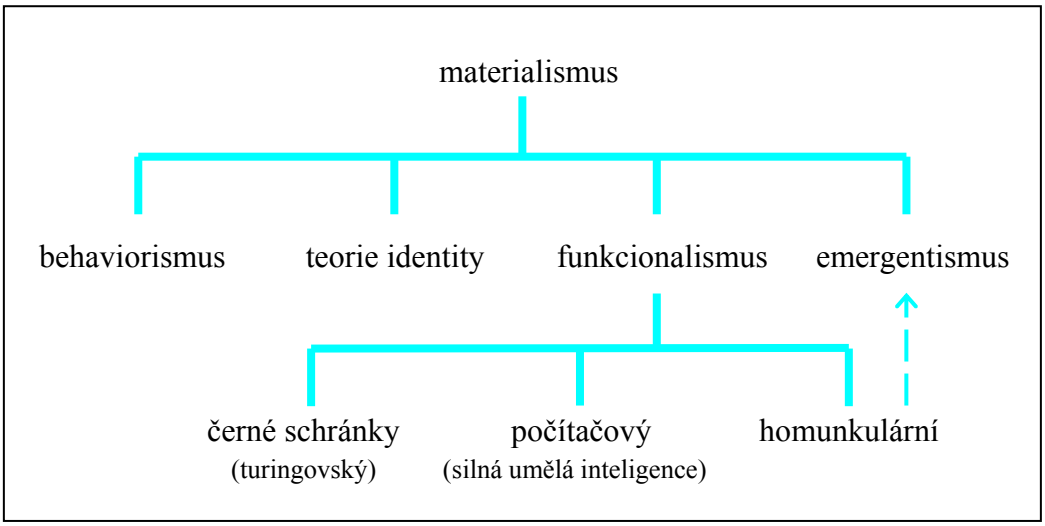
V rámci dualistické koncepce se nabízejí nejméně tři možností jak se vyrovnat s problémem kauzální interakce (obr. 2). Klasický **psychofyzický paralelismus** (G. W. Leibnize a dalších) je teorie, která uznává děje jak na mentální, tak i na fyzické úrovni, soudí však, že spolu vzájemně nijak neinteragují, že prostě probíhají synchronně, v jakési harmonii, *jako by* interagovaly. Naproti tomu **epifenomenalismus** (Campbell, 1970) považuje mentální stavy za vedlejší produkty fyzických procesů bez vlastní kauzální potence. Jsou to jakési „stíny“ materiálních dějů a jako takové jsou (vědecky řečeno) nadbytečné. Třetí možností je **interakcionismus**, názor, k němuž se hlásili především Karl Popper a John Eccles (Popper, Eccles, 1977).[40] Dle nich svět mentálního opravdu kauzálně interaguje se světem fyzického, přičemž vysvětlit, jak a kde ke kauzální interakci dochází, je úkolem vědy. Eccles hledal tento „interface“ až v kvantově mechanických dějích na subneuronální úrovni (Eccles, 1994).

[40] V jejich pojetí existují nikoliv dva, nýbrž *tři* vzájemně interagující světy – fyzický, mentální a kulturní.



Obr. 2. Typy interakcí dvou světů

Materialistické[41] koncepce (obr. 3) vycházejí z přesvědčení, že vše mentální je fyzické. Proto někteří autoři ztotožňují materialismus s **fyzikalismem**, zde však mám na mysli materialismus v poněkud obecnějším smyslu (stejně jako *fyzické* chápu obecněji než *fyzikální*). Je tu ovšem vždy otázka, proč máme v běžném životě sklon o stavech mysli mluvit *jinak* – tj. jinak, než o jevech fyzického světa, jako třeba o skupenství látek nebo fázích Měsíce. Většina materialistických teorií mysli je založena na redukcionistickém postoji, jehož charakteristickým výrokem je „X není *nic než* Y“, kde za Y se dosadí vždy to, co má být vysvětleno.



Obr. 3. Moderní směry

Pod vlivem empirické psychologie první poloviny tohoto století, odmítající jak introspekci, tak spekulace o vnitřních stavech myslící bytosti (Watson, 1913; Skinner, 1933), se objevily různé varianty **behaviorismu** i v teorii mysli.Ten redukuje vše, co se týká mysli, na výpovědi o (pozorovatelném) vnějším fyzickém chování – ostatně si vzpomeňme na Turingův test z odst. 4.4. Důsledný přístup k této otázce přisuzuje

[41] Termín ‚materialismus‘ má ve filosofii mysli poněkud specifičtější význam, než obecně ve filosofii. Například již vícekrát citovaný J. Searle je velmi kritický k materialismu v tomto specifickém významu, jakkoliv jeho ‚biologický naturalismus‘ je konec konců rovněž materialismem jistého (emergentistického) typu.

mentálním konceptům jejich význam odkazováním na určité stereotypní vzorce chování. Podle tzv. **logického behaviorismu** (Ryle, 1948), mentální koncepty jsou jen jména pro určité *dispozice* k vnějším projevům. Tak například bolest není *nic než* určité standardní vztahy mezi typy stimulů a příslušnými reakcemi. Triviální protiargument je, že otrlý spartán nemusí dát nikdy na sobě znát, že ho bolí hlava, a dokonalý herec může bolení hlavy permanentně hrát (Putnamův příklad).

Dosti odlišný pohled má **teorie identity** (centrálního stavu), která měla svého času velký vliv (Armstrong, 1968). Podle ní jsou mentální jevy, stavy a procesy identické s konkrétními fyziologickými jevy, stavy a procesy v mozku, takže například pocit bolesti není „nic než“ excitace vhodných neuronů. Při tomto pojetí je důležité rozlišovat, zda jde o partikulární stavy (např. konkrétní bolest konkrétního člověka) nebo o obecné jevy (bolest jako taková). Ve druhém případě je třeba navíc předpokládat existenci typově určitelných obecných fyziologických "mechanismů".

Teorii identity můžeme považovat za redukcionismus směrem „dolů“, pokud si představujeme nějakou hierarchii kauzálních domén (viz odst. 6.1), v níž je myšlení na vyšší úrovni než neurofyziologické procesy.

Obliba redukcionismu právě ve směru k nižším úrovním (zvláště u scientisticky orientovaných filosofů) stojí za povšimnutí. Myslím, že má dva důvody, jeden zřejmý, druhý skrytý. Zřejmý důvod je ten, že zákonitosti nižších úrovní bývají exaktní, matematicky formulované a tudíž jaksi důvěryhodné. Druhý důvod je, že je vlastně nemusíme používat. Vysvětlím to na analogii se statistickou fyzikou. Redukujeme-li makroskopický termodynamický popis na molekulární úroveň, vysvětlíme (například) vzájemný vztah tlaku, objemu a teploty jako důsledek jednoduchých, deterministických a srozumitelných mechanických zákonů. Tyto zákony však nikdy přímo nepoužijeme – jen zkuste řešit soustavu rovnic pro všechny molekuly v nafouknutém balónku!

7.3 Poznámky o funkcionalismu a emergentismu

Funkcionalismus jsem na obr. 3. zařadil pod materialistické teorie, i když abstrahuje od materiální povahy entit, jimiž se zabývá, a má tak blíže k pohledu matematika než fyzika (mnozí funkcionalisté jsou ovšem též materialisté). Matematika a matematická fyzika buduje různé obecné abstraktní teorie k objektům, strukturám a procesům, aniž by znala či potřebovala znát jejich reálnou podstatu. Dokonce lze říci i o mnoha reálných věcech, že tím, čím jsou, jsou nikoliv díky tomu, z čeho jsou udělány, nýbrž díky funkci či roli, jakou hrají (příklady: peníze na trhu, karburátor u automobilu, srdce v těle). A tak lze podle funkcionalistů budovat jednu jedinou abstraktní teorii myšlení (či obecněji mysli), která by připouštěla libovolné realizace – biologické stejně jako technické. Důležité jsou dle nich funkční a kauzální vztahy mezi stavy systému, nikoliv jak a v čem jsou tyto stavy realizovány.

Jako i jiné moderní teorie mysli, byl i funkcionalismus silně ovlivněn vědeckými poznatky své doby. Byl-li behaviorismus pod vlivem tehdejší behaviorální psychologie a byla-li teorie identity pod vlivem znalostí mozku, pak pro funkcionalismus i jeho varianty byla hlavní inspirací počítačová věda.

Pro funkcionalismus **„černé schránky“** čili **turingovský** (Putnam, 1967/1975) je podstatný pojem *mentálního stavu*. Formálně a velmi povrchně lze použít analogie s vnitřními stavy Turingova stroje (stačí i konečný automat): přechodová funkce zajišťuje vnější chování (vztah vstup-výstup, stimulus-reakce), má však též vnitřní, *kauzální* stránku: vnitřní stavy (a vstupy) jsou příčinami jiných vnitřních stavů (a

výstupů). Podstatné je, že vnitřní stavy jsou definovány *právě jen* těmito vztahy, jejich fyzická realizace je nepodstatná. V tom smyslu lze považovat mysl za černou schránku.

Za stejně nepodstatné považuje funkcionalismus subjektivně prožívané kvality, které přisuzujeme různým mentálním stavům (jako strach, bolest, radost, nápad apod.) – tato kvalitativní pojmenování jsou dokonce zbytečná, protože je vždy lze nahradit abstraktními symboly a respektovat jen vzájemné kauzální vazby (tzv. Ramseyova eliminace mentálních pojmů (Ramsey, 1991)). Být v mentálním stavu je totéž jako mít standardní kauzální vztahy: nic jiného než obousměrné interakce s jinými mentálními vztahy, perceptuálními vstupy a behaviorálními výstupy.

Funkcionalismus našel oporu i v rozlišení softwaru a hardwaru u počítačů. Přestože *realizace* nějaké funkce je na počítači nemyslitelná bez hardwaru, *popis* této funkce nevyžaduje žádné tranzistory nebo integrované obvody. Tutéž činnost mohou realizovat počítače zcela odlišných technologií – proč by ji tedy nemohl provádět též mozek pomocí své neuronové tkáně? Důsledněji chápaná počítačová metafora, spolu s faktem, že programy tradiční umělé inteligence byly implementovány na klasických počítačích vedla k novější verzi funkcionalismu, kterou již známe z odst. 4.3 jako **počítačový funkcionalismus** či **silná umělá inteligence**.

Jestliže programy lze přenášet z počítače na počítač a uchovávat je i mimo počítače, mělo by podle funkcionalismu totéž platit i o lidské mysli (za předpokladu, že by byl znám kód, který příroda použila). To bychom byli již na hranici vědecké fantazie – avšak o nic dál, než byl i N. Wiener, když se zmínil o principiální možnosti přenášet lidskou bytost na dálku pomocí telegrafu. Filosoficky to naznačuje jedno: v jistém smyslu je funkcionalismus kompatibilní i s dualismem.

Námitky oponentů proti jedné, druhé či oběma variantám funkcionalistické filosofie mysli lze zhruba rozdělit do tří typů:

(1) arbitrárnost struktury,
(2) zanedbání fenomenální stránky mysli a
(3) explanační dluhy.

První typ námitek se vztahuje k samotné podstatě funkcionalistického programu: za mysl je ochoten dosadit *cokoliv*, co má patřičné kauzální vztahy, případně pokud se to řídí patřičnými algoritmy.

Ned Block (Block, 1978) se pokusil demonstrovat absurditu funkcionalismu pomocí myšlenkového experimentu s populací Číňanů. Uvádím jej v pozměněné podobě (pro případ počítačového funkcionalismu). Představme si, že využijeme hardwarové arbitrárnosti funkcionalismu a místo neuronů (nebo tranzistorů) mobilizujeme pro algoritmy mysli celou čínskou armádu[42]. Vojáci by stáli v dlouhých řadách, co voják to bit. Přecházením z místa na místo by přesně a ve správném pořadí realizovali jednotlivé příkazy kteréhokoliv algoritmu. Mohl by to být algoritmus pro myšlení. Čí by to bylo myšlení?

Jakkoliv taková představa vypadá absurdně, sama o sobě funkcionalismus nevyvrací. K tomu bychom museli nejprve dokázat, že tímto způsobem se mentální stavy na úrovni celého systému objevit opravdu *nemohou*. (Ostatně současné úvahy o internetu jako o zárodku globálního mozku si nedělají s analogickými představami problém.) Pokud bychom takovou možnost připustili, byly by to zajisté stavy

[42] Čínskou čistě proto, že je početná. S experimentem s čínskou komorou to nemá co dělat.

nesrovnatelné s lidskými mentálními stavy, protože by operovaly ve zcela jiném vnějším prostředí.

Příkladem námitek druhého typu je Searlův experiment s čínskou komorou, popsaný v odst. 4.4. Fenomenální stránkou mysli se funkcionalismus (opět programově) netrápí, ostatně jako většina jiných teorií zastávaných analytickými filosofy. Pokud vůbec je funkcionalismus schopen něco říci o mentálních procesech, pak tedy jen o jejich performačních komponentách. Problémem ovšem zůstává, zda vždy lze a má smysl obě komponenty od sebe oddělit.

Konečně třetí typ námitek se týká otázky, co vlastně funkcionalismus vysvětluje. Omezuje-li se pouze na funkcionální vztahy, a to dokonce na ty, které náleží do vybrané „vyšší“ úrovně, nelze od něj čekat *vysvětlení* jejich vztahu ke struktuře (k jiným, nižším úrovním) a nebude ani schopen identifikovat ty aspekty nižších úrovní, které jsou nutné nebo postačující pro existenci mysli. V našem pojetí (z odst. 6.1) lze funkcionalismus považovat za školní příklad teorie omezené na vybranou kauzální doménu, vytrženou ze souvislostí s jinými doménami. Může to být třeba i zajímavá a poučná *abstraktní* teorie, avšak bez možnosti říci cokoliv závažného k problému mysli a těla.

Jistý krok dále (lépe řečeno hlouběji), než dosavadní varianty funkcionalismu, činí tzv. **homunkulární funkcionalismus** (Lycan, 1987). Vychází ze snahy využít specifických možností kolektivních systémů, tvořených spolupracujícími či soutěžícími agenty (zvaných též homunkulové); s touto myšlenkou jsme se rovněž již dříve setkali (v odst. 6.3). Na funkcionalismus, který by uvažoval více úrovní současně, případně celou hierarchii úrovní, by se již nevztahovaly některé předchozí námitky. Vezmeme-li ovšem do úvahy více kauzálních domén, setkáme se s interakcemi mezi doménami, a ty již nemusí mít (v tradičním pojetí) kauzální povahu. Místo funkcionalismu by tu bylo namístě mluvit (například) o **emergentismu**. O pojmu emergence jsem se zmínil v odst. 5.2. a jeho vztah k filosofii mysli jsem uvedl v podobě několika variant *emergentistické teze.* Protože se dnes o emergenci poměrně často hovoří, stojí zato se k ní zde krátce vrátit.

Termín **emergence** byl nejprve užit G. H. Lewesem v polovině 19. stol. a C. Lloyd Morganem začátkem 20. stol. v diskusích o evolučních teoriích, a to ve snaze vyhnout se mechanistickému redukcionismu na jedné straně a teleologickému postoji na straně druhé. V tomto tradičním významu je zdůrazněn temporální aspekt: jde o vznik něčeho (života, člověka), co nelze predikovat nebo vysvětlit z předchozího stavu věcí.

Současné pojetí je poněkud obecnější a temporální aspekt ustupuje do pozadí, nahrazen působením z nižší úrovně na vyšší. Kolektivní systémy skýtají vhodné příklady, od makroskopických vlastností látek až po politický vývoj společnosti. Méně banálním příkladem je entropická šipka času jako emergentní jev systémech, které jsou na mikroúrovni časově reverzibilní. Obecně čím rozsáhlejší a vazebně složitější je systém na mikroúrovni, tím spíše se nám jeho chování jeví jako emergentní na makroúrovni: pět molekul nevysvětlí tekutost kapaliny, pět demokratů nevytvoří demokracii.

Jiný význam pojmu emergence je inspirován předchozím a je často užíván jako argument proti redukcionismu. Vlastnost (v rámci nějaké teorie T1) považujeme za *emergentní* (v silném slova smyslu), má-li reálné instance, vyskytuje-li se současně s nějakou jinou vlastností, rozpoznatelnou v nějaké redukující teorii T2, nelze ji však samu redukovat na nějakou vlastnost definovatelnou v T2 (Churchland, 1986, s. 324). Dualismus vlastností lze pak charakterizovat tvrzením: i když mysl je mozek, kvality

subjektivního prožívání jsou přesto emergentní vzhledem k mozku a jeho vlastnostem.

John Searle (Searle, 1992) si je jist, že principiálně lze vysvětlit souvislost mezi myslí a mozkem pomocí meziúrovňového vztahu, který je souhrou kauzality a realizace, i když přiznává, že takové vysvětlení je teprve třeba hledat a že o dotyčném vztahu víme málo. Soudí, že některé dostatečně složité neurobiologické procesy v mozku mohou *způsobit* (cause) a zároveň *podpořit* (sustain) – vědomé mentální stavy a procesy. Klíčovým pojmem je právě toto „způsobit a zároveň podpořit", což lze chápat jako speciální případ emergence v našem pojetí.[43] V novější práci Searle uvádí, že mozkové procesy *způsobují* vědomí, přičemž vědomí samo je *vlastností* mozku (feature of the brain) a považuje proto vědomí za *emergentní* vlastnost mozku. V jeho pojetí je emergentní taková vlastnost, kterou lze „kauzálně vysvětlit chováním prvků systému; není to však vlastnost jednotlivých prvků a nemůže být vysvětlena prostou sumací vlastností těchto prvků" (Searle, 1997, s. 18).

Protože Searlův náhled je vlastně variantou emergentistické teze a on sám jej považuje za „řešení tradičního problému mysli a těla, které se vyhýbá jak dualismu tak materialismu, aspoň jak jsou tradičně chápány" (s. 8), stojí za to se u něj pozdržet. Zdá se mi sympatické, že je tu přisouzena velká váha meziúrovňové interakci, avšak mám obavu, že vše je poněkud zjednodušeno. V jednom smyslu zde jde o zúžení této interakce na kauzální vztah příčiny a následku, současně je však pojem kauzality zobecněn na (bezčasový) vztah mezi úrovněmi, a to podle poněkud zjednodušeného schématu podle tohoto vzoru: lze-li tuhost stolu *vysvětlit* chováním molekul stolu (Searlův příklad), plyne z toho, že molekuly stolu jsou *příčinou* tuhosti stolu. Snad to lze říci pro stoly a jejich molekuly, u nichž více méně víme, jaká je struktura pojmu ‚vysvětlit'. Je však možno se stejně postavit ke vztahu mysli a těla, kde ani o příčinnosti, ani o možnostech vysvětlení nemáme žádnou představu?

K Searlově definici emergence: jestliže něco vysvětlit lze, avšak nikoliv „prostou sumací vlastností prvků", co to přesně znamená? Uznávám, že nelze být exaktní – i náš výklad kauzality v relaci ke kauzálním doménám nebyl exaktní – avšak pak je třeba vždy počítat s tím, že jistota „řešení" může být jen zdánlivá.

Konečně je tu problém s fenomenální komponentou mentálních stavů, či přímo s pojmem vědomí, o který Searlovi tolik jde. Mluví o vědomí jako by to byla vlastnost téhož řádu jako tuhost stolu nebo tekutost vody. Čili vlastnost vyjádřená v 3. osobě, nikoliv v 1. osobě, kterou v případě vědomí Searle považuje za relevantní (dokonce ontologicky, viz odst. 4.4).[44] To však je spíše již téma pro následující oddíl.

Vedle emergentistické teze jsem v odst. 6.2 navrhl zobecněnou emergentistickou tezi, dle níž lze mysl považovat za emergentní jev v dostatečně složitém systému kauzálních domén. Prvá i druhá teze by mohly podpořit materialismus (v poněkud netradičním pojetí) v jeho snaze vyhnout se dualistické představě o mysli jako o nějaké zvláštní nemateriální substanci. K vysvětlení vztahu

---

[43] Původně Searle mluvil o souběhu kauzality a realizace: „…*příčinou* mentálních jevů jsou procesy, které se odehrávají v mozku na úrovni nervových buněk a jejich skupin, přičemž se v systému složeném z buněk *zároveň realizují*." (Searle, 1984/1994 s.23), v pozdějších pracích mu jde o kauzalitu a podporu (sustaining) vědomí: „některé extrémně složité nervové systémy jsou schopny *způsobit* a *podpořit* vědomé stavy a procesy" (Searle, 1992, s. 89). V novější formulaci: „mozkové procesy *způsobují* vědomí, ale vědomí je samo *vlastností* mozku [což] nám skýtá řešení tradičního problému mysli a těla, řešení, které se vyhýbá jak dualismu tak materialismu, aspoň jak jsou tradičně chápány." (Searle, 1997, s. 8), kurzívy všude moje.

[44] Na tuto mezeru v Searlově pojetí též upozorňuje McGinn (McGinn, 1999).

mysli a těla však příliš nepřispějí, protože obě nabízejí k vysvětlení jedné věci (mysli) jinou věc (emergenci), o níž, až na několik speciálních případů, vlastně nic konkrétního a poučného nedovedeme říci. To ovšem nic nemění na tom, že obě teze mohou být heuristicky velmi užitečné pro filosofické úvahy.

8 Je možná věda o vědomí?

S pojmem vědomí jsme se již setkali na mnoha místech a čtenář se možná ptá, co k tomu chci ještě dodávat. Vnucuje se mi poněkud jízlivá odpověď: právě teď se teprve dostáváme k tomu, co je na tom všem nejdůležitější. Současně se však dostáváme do místa, kde končí možnosti běžných vědeckých metod. Než by vůbec bylo možno začít se ptát na podstatu vědomí (či – při našem zaměření – na to, zda počítač může mít vědomí), je třeba si položit otázku, zda se do takových věcí můžeme, jako vědci, vůbec pouštět. A tím pro tentokrát skončíme.

Na problém vědomí jsme poprvé narazili při úvaze, zda se při studiu lidské mysli můžeme spolehnout na *vnitřní prožívání* a introspekci (odst. 3.2) a pak jsme se s ním setkali, když se ukázalo, že pro potřeby umělého napodobení mentálních procesů je možné i nutné se omezit na performační komponentu těchto procesů a ponechat stranou komponentu fenomenální (odst. 4.2). Před analýzou vědomých stavů mysli z hlediska filosofického jsem (v odst. 7.1) vyzval čtenáře, aby se zamyslel, „jaké to je“ být v nějakém mentálním stavu (dotyčný odstavec je vlastně úvodem k tomuto oddílu). A naposled jsme se dotkli otázky, zda vědomí je emergentní vlastností (odst. 7.3).

Otázka vědomí nebyla po dlouhá léta považována za problém pro seriózní vědu, byť by to měla byla věda, která se zabývá myšlením. V objemné monografii Kolba a Whishawa o neuropsychologii (Kolb, Whishaw, 1990) není například slovo „consciousness“ ani v rejstříku. Situace se však změnila. Je až zajímavé sledovat, jak roste počet filosofů, neurovědců, informatiků, ba i fyziků, kteří se problémem vědomí zabývají. Jen v seznamu literatury k této kapitole, který zdaleka není vyčerpávající, najdete nejméně 10 knih na toto téma, k tomu existují nejméně dva speciální časopisy,[45] o konferencích a sympoziích ani nemluvě.

Je však vůbec možné poznání vnitřního světa mysli, je možná objektivní věda o vědomí, které se nezpřítomňuje jinak, než subjektivně?

8.1 Druhy vědomí

Samo slovo vědomí má složité sémantické pole. Všimněte si, že v následujících třech jazykových podobách má sice obdobný význam, ale pokaždé se vlastně mluví o něčem podstatně jiném: (1) *být při vědomí*, (2) *vědomí něčeho* nebo *o něčem*, (3) *mít vědomí*. Ve významu (1) je vědomí spíše celkovým fyziologickým stavem (mysli a těla) a i když má výraznou fenomenální stránku, necítím radikální rozdíl, mluvím-li o sobě nebo o druhém. Opakem takového vědomí je spíše bezvědomí než nevědomí a lze jej bez velkých obtížích přisuzovat i zvířatům (zejména pokud reagují na bodnutí jehlou).

V případě (2) jde hlavně o fenomenální aspekt konkrétního kognitivního stavu, něčeho, co si můžeme *uvědomovat*. Řeknu-li „bolí mě hlava“, je v tom implicitně

[45] *Journal of Consciousness Studies* (Imprint Academic), *Consciousness and Cognition* (Academic Press).

obsaženo, že toto bolení hlavy je předmětem mého vědomého prožitku. V tomto smyslu jsme mluvili o fenomenální komponentě mentálních procesů. Konečně vědomí (3) je vědomí jako takové. Jde o filosofickou abstrakci: to, co vůbec umožňuje všechny případy vědomí (2), co je pro ně společné. Vědomí (3) je vlastnost či schopnost bytosti (zde bych měl mluvit jen za sebe) vést svůj vnitřní život, čili prožívat své mentální stavy. Řeknu-li, že člověk je obdařen vědomím (3), nemyslím tím, že ho zrovna musí třeba něco bolet.

Ptáme-li se, zda existuje věda o vědomí, máme na mysli vědomí ve smyslu (3). Filosofové vědomí by se o něm rádi něco dozvěděli, nicméně jakákoliv konkrétní úvaha musí začít u vědomí (2).

Zamyslím-li se (zde již musím mluvit jen za sebe) nad tím, jak například rozumím výroku „Bolí mě hlava“ na rozdíl od „Bolí ho hlava“, uvědomím si celkem banální věc: zatímco velmi dobře vím (lépe řečeno pamatuji si), jaké to je, když *mě* bolí hlava, nedovedu si *cizí* bolení hlavy představit jinak, než že si vzpomenu zase jen na svou vlastní bolavou hlavu. Je tu však i další rozdíl: pokud mě právě nyní bolí hlava, necítím potřebu objektivně vědecky zkoumat, zda se náhodou nepletu. Mé bolení má rys *bezprostřední danosti* (viz odst. 7.1). Mohu se vůbec plést? Budou-li mi všichni neurologové světa tvrdit, že mě hlava nebolí, na věci to nic nezmění - *vím*, že mě bolí, mé *vědomí* mě neklame, protože mé vědomí jsem *já*.

Něco zcela jiného je pro mě ovšem bolení hlavy někoho druhého. Stěžuje-li si někdo, že ho bolí hlava, mohu mu věřit, ale nemusím. Mohu dokonce žádat od neurobiologů „objektivní“ vyjádření k jeho hlavě. Víme, že metody měření, pozorování a zobrazování mozku se každým dnem zdokonalují, proč bychom tedy nepoložili rovnítko mezi procesy v mozku a děje v mysli? Tak soudí, jak již víme, přísní materialisté.

Tím se dostáváme k rozlišení dvou alternativních přístupů ke studiu vědomí. První přístup staví na přirozené subjektivní zkušenosti, druhý se drží metody přírodních věd (srov. odst. 3.3).

Začněme u subjektivního přístupu. Má jednu výhodu, byť problematickou: samu existenci vědomých prožitků si totiž vzájemně nemusíme dokazovat, hlavně proto, že spoléháme na to, že je všichni máme – aspoň já je mám a asi mnozí další, kdo nejsou roboti anebo náměsíčníci. Subjektivní prožívání čili „kvalita“ vědomí se ovšem nedá objektivně popsat, což nás vedlo k onomu závěru o nemožnosti umělé simulace fenomenální (tj. vědomé) komponenty mysli.

Pokud to platí o simulaci, pak to platí i o vědecké analýze. Jestliže jsme se však rozhodli považovat kognitivní vědu za objektivistickou, zdalipak je vědomí vůbec něco, čeho si má tato věda všímat? D. C. Dennett soudí že ano a navrhuje dokonce jistou metodu (Dennett, 1991a): zkoumejme svědectví druhých o jejich vědomých prožitcích. To, co nám sdělí, nelze nijak verifikovat, ba ani nemá smysl vůbec o nějaké verifikaci mluvit – stejně jako nemá smysl mluvit o verifikaci toho, co se odehrává v románech, pohádkách a mýtech. Přesto lze, soudí Dennett, o vědomí objektivně teoretizovat, podobně jako lze objektivně interpretovat, analyzovat a vykládat děj románu, povídky nebo mýtu. Tuto metodu nazývá *heterofenomenologií* a věří, že „v principu může plně vyhodnotit nejsoukromější nevyslovené subjektivní prožitky, aniž by jakkoliv opustila metodologické skrupule vědy“ (s. 72). Podstatné je (speciálně pro Dennetta, který se snaží vědomí eliminovat), že tento postup se nijak nezavazuje věřit v existenci subjektivního vědomí.

Druhý přístup, přírodovědecký, se převážně ubírá cestou „zdola“, od úrovně neuronů. „Jazyk mozku je jazykem neuronů,“ píše F. Crick[46] v přesvědčení, že porozumíme-li tomuto jazyku, zmizí všechny záhady myšlení i vědomí (Crick, 1997). Zajisté, k dnešnímu dni se toho o člověku ví – co do objemu dat – bezesporu víc na úrovni neuronů než na úrovni nálad. Avšak dosud se velice málo ví, jakým jazykem a o čem to vlastně neurony spolu hovoří a co dalšího se jim do tohoto „hovoru“ ještě plete. (Doufejme, že se časem vyvinou metody zkoumání, které nebudou omezeny jen na neživé nebo nemocné nebo zvířecí mozky.). Přesvědčeným optimistou co do možnosti vědeckého studia vědomí je John Searle, který i navrhuje (ne zcela přesvědčivě), odkud začít (Searle, 1997, s. 196, 200).

Jak vnitřní prožívání, tak poznatky neurofyziologů skýtají dostatek podkladů k tomu, abychom mluvili o různých *stupních vědomí*. Stačí si vybrat dva extrémy a zamýšlet se nad možnými „mezilehlými“ stavy. Pro případ vědomí ve smyslu (1) by to byly stupně mezi hlubokým bezvědomím a plným vědomím. Ty jsou biologicky a klinicky široce propracovány, ale i fenomenálně o tom dost víme: vzpomeňme třeba na onen „polovědomý“ stav při ranním probouzení.

Z psychologického hlediska jsou zajímavé různé stupně vědomí typu (2). V tuto chvíli je mé vědomí plně zaměřeno na větu, kterou zrovna píši. Současně však existují věci, které jsou jen na *okraji* mého vědomí, například divoká kachna za oknem (všimněte si metafory „okraje“), dále nepřeberné množství mentálních stavů v různé hloubce nevědomí: mé jméno (na které zrovna nemyslím), jméno či tvář mého souseda (vím, že je znám, ale nemohu si je vybavit), ale rovněž rozličné víry a domněnky, které lze *přivést* do vědomí jedině tak, že je *převedu* (něco je převede) do jazykové podoby (všimněte si metafory „přivést“ a „převést“), až po návyky, dovednosti, obratnosti a zkušenosti, které nepotřebuji nebo nedovedu přivádět do vědomí. Různé směry analytické psychologie a psychoterapie hovoří o rozsáhlých oblastech nevědomí, které jsou nám zcela nepřístupné a pokud, tak jen speciálními metodami.

Konečně a druzích vědomí ve smyslu (3) se lze jenom dohadovat. Mohly by to být případy vědomí vyšších (nebo i nižších? – názory o tom se různí) živých organismů anebo různé stupně vědomí člověka během ontogeneze a fylogeneze.

Filosoficky zvlášť relevantní je specifický typ vědomí, *vědomí sebe sama*. Již Hegel upozornil na to, že většinu času není subjekt vědomé zkušenosti (naše já) vědomě reflektován (Hegel, 1806/1960). Lze tedy pojmově odlišit vědomí od sebevědomí, při němž je mé já současně objektem a subjektem mého vědomí.

8.2 Svět bez vědomí

Existují čtyři „snadné“ odpovědi na otázku, zda je věda o vědomí vůbec možná. První odpověď je eliminativistická: žádné vědomí neexistuje, je to jen iluze, o iluzích vědy netřeba. Druhá odpověď je objektivistická: vědomí existuje, nemůže však být předmětem vědy, protože věda se zabývá pouze objektivními (objektivně přístupnými) skutečnostmi. Třetí odpověď je redukcionistická: věda o vědomí je možná, protože vědomí je čistě biologická, ne-li fyzikální vlastnost mozku. Existuje ještě čtvrtá odpověď: věda o vědomí je možná, ale musí být založena na jiných základech, než všechny dnešní vědy.

---

[44] Je to tentýž Crick, který v r. 1962 získal spolu s J. D. Watsonem Nobelovu cenu za objev struktury DNA (viz *Vesmír* 42 (1963) 3, s. 81).

První tři odpovědi jsou na můj vkus příliš ukvapené, čtvrtá je vyhýbavá; kdybych musel rozhodnout volil bych něco mezi třetí a čtvrtou variantou: aspoň by tu byl nějaký výzkumný program. Literaturu na toto téma však nelze tak snadno třídit, autoři totiž často sami nejsou konzistentní sami se sebou – v tom, jak sebe sama zařazují, co slibují a co skutečně předvedou. Navíc, jak je v analytické filosofii zvykem, jde často o nálepkování oponentů, aby bylo o čem debatovat (a debaty jsou to často bouřlivé[47]).

Zastánci redukcionistické odpovědi tvrdí, že každý prožitek musí mít svůj korelát v podobě aktivity těch či oněch neuronů – proč tedy nepovažovat přímo tuto aktivitu za vědomí? (Pokud takový korelát neexistuje, pak nemůže být ani vědomí, rozumí se pro vědu.) Například F. Crick (Crick, 1997) se snaží hlavně na příkladu vizuálního vnímání podpořit hypotézu, že k výkladu vědomí je nutné a dostačující porozumět hromadné interakci neuronů v mozku a není jediný, kdo podobnou hypotézu zastává, jsou však i kritické hlasy – ať už k hypotéze samé, či ke Crickově verzi její obhajoby.

Již několikrát zmíněný filosof D. C. Dennett, hlavně ve své knize, příznačně nazvané „Vysvětlené vědomí“ (Dennett, 1991a), eliminuje problém vědomí tím, že na ně rozšiřuje svou variantu počítačového funkcionalismu. Svůj přístup zahajuje odmítnutím tradičního intuitivního názoru – který nazývá „metaforou karteziánského divadla“ – podle něhož si lze představit v mozku jakési centrum, v němž imaginární pozorovatel vše monitoruje a o všem rozhoduje. Místo toho navrhuje tzv. „model více verzí“[48]: místo jednoho proudu vědomí existuje množství kanálů, v nichž specializované obvody vykonávají svou činnost na principu paralelních pandemonií (srov. odst. 6.3); přitom produkují fragmentární verze společného „příběhu“. Většina těchto verzí má jen efemérní roli, některé jsou však povýšeny do významnějších funkcí. Sériový charakter takového stroje (který odpovídá von Neumannově architektuře počítačů, ale též našemu pocitu jednosměrného toku vědomí) není „zadrátován“ v hardwaru, nýbrž je spíše výsledkem posloupnosti dynamických koalic těchto specialistů (Dennett, 1991a, s. 253).

Dle Dennetta je vědomí recentní evoluční jev, k němuž dochází v mozku, který však nebyl předchozí evolucí dost dobře přizpůsoben pro takový virtuální sériový proces. Náš vnitřně prožívaný tok vědomí je, tvrdí Dennett, výsledek (velmi neefektivní) simulace sériového logicko–lingvistického stroje ve vysoce paralelním hardwaru.

Poznamenejme, že Dennett má dosti liberální představu o tom, co je to *vysvětlit* a rovněž slovo *vědomí* užívá poněkud zúženě (snad aby si jeho kniha zasloužila svůj lákavý titul). K vysvětlení dle něj stačí nalézt vhodnou organizační úroveň, pro níž již není zajímavé, jak je implementována. Jinými slovy, redukce na základní úroveň analogie, zvolenou tak, aby ji bylo možno interpretovat funkcionalisticky.

Uvažujme nyní jinak. Připustíme-li, že prožívané vědomí existuje, můžeme se ptát, zda existuje *nutně*, tj. zda je při daných biologických či fyzických vlastnostech mozku samozřejmostí. D. J. Chalmers (Chalmers, 1996) argumentuje, že tomu tak není, čili že vědomí je kontingentní jev v našem *aktuálním světě*. Tím se též dostává do sporu se Searlem, dle něhož vědomí je kauzálním důsledkem biologických a fyzických vlastností mozku (srov. (Searle, 1997), s. 133-176).

---

[47] Jako například výměny názorů mezi Johnem Searlem a druhými (Searle, 1997).

[48] Angl. *multiple drafts model.*

Chalmersův argument spočívá v tom, že je *logicky myslitelný* svět, který je po všech stránkách shodný s naším aktuálním světem, až na jedinou výjimku, totiž že v něm neexistuje vědomí. K tomuto závěru vede Chalmerse následující úvaha: lze si představit lidskou bytost, která má všechny vlastnosti člověka, včetně řeči a chování, které však zcela chybí vědomí. Takovým bytostem se ve filosofii říká *zombiové*[49]. Intuice za touto představou je prostá – přesně tak si totiž mohu (čistě rozumově) představit druhého člověka: o jeho vědomém prožívání nemám možnost se jinak přesvědčit, než z jeho chování a ústních sdělení. Dalším krokem v úvaze je představit si *svět zombiů*, jehož všichni obyvatelé jsou a vždy byli zombiové. Z toho by měl plynout závěr, že vědomí je v našem světě jakoby něco navíc, co není redukovatelné na jiné vlastnosti věcí, ani kauzálně, ani logicky. Zrovna my máme to štěstí (nebo smůlu?), že jsme tajnými svědky svého vlastního nitra.

Chalmersovi se logická možnost existence zombiů zdá evidentní stejně jako, řekněme, možnost existence unicyklu o velikosti jedné míle. Tvrdí, že popsal „koherentní situaci“, neshledává žádný „spor v popisu“ a opírá se o elementární intuici (Chalmers, 1996, s. 96). S jeho popisem myslitelného světa zombiů v přirozené řeči s vágním významem slov, i s jeho odkazem na elementární intuici bychom se mohli smířit – konec konců se myšlenkové experimenty vždy obracejí k naší intuici. Vzniká tu však problém koherenční mezery (viz pozn. na začátku odst. 4.4), zejména při expanzi představy zombie na *všechny* obyvatele hypotetického světa.

Chceme-li užít světa zombiů v argumentu o nezávislosti vědomí na fyzickém světě, je třeba ověřit, zda neexistují další aspekty, v nichž by se mohl tento svět lišit od našeho (kromě požadované absence vědomého prožívání u zombiů), a pokud existují, zda by mohly ovlivnit náš argument. Uvažujme například řečové projevy zombiů (řekněme zvukové vlny, které kolem sebe šíří). Tyto projevy mohou, ale nemusejí mít sémantický význam (rozhodnutí o tom by záviselo na určitých předpokladech o povaze jazyka ve světě zombiů a obecně o tom, zda sémantika je reálnou součástí světa). V prvém případě by se náš a jejich svět lišil co do sémantického pole slov – ve světě bez vědomí by například neexistovala ani přetvářka, ani upřímnost, ani rozdíl mezi nepravdou a lží. Ve druhém případě bychom ve světě zombiů měli problém s kauzální potencí performativních řečových aktů.

Lze si opravdu představit svět, po němž chodí bytosti stejné jako my, inteligentně se tvářící jako my, generující stejně chytré řeči jako my – včetně řečí o vnitřních pocitech a prožitcích, o vlastním vědomí, ba i o myslitelnosti světa zombiů – , bytosti, které však samy nic neprožívají? Snad si to představit lze, jenže takový svět by se musel lišit i v mnoha dalších věcech. Hlavně by se ten svět lišil v tom, že můj dvojník v něm bych nebyl já.

8.3 „Lehké“ a „těžké“ problémy

V odst. 8.1 jsem uvedl dva alternativní přístupy k porozumění vědomí, jeden spoléhající na vnitřní zkušenost, druhý založený na (objektivním) výzkumu mozku. Vzájemně se oba přístupy od sebe podstatně liší, včetně slovníku, v němž se vyjadřují, a neumím nic říci o tom, zda se potenciálně sbíhají, či zda se míjejí a posléze rozbíhají.

[49] V původním významu (v přírodním kultu voodoo) je *zombie* mrtvola, kterou nějaká nadpřirozená síla vybavila vnějšími projevy života, nikoliv však vůlí a řečí.

Jistou snahu o konvergenci lze vidět u druhého přístupu, postupujícího „zdola“: s postupným poznáním je možné, že jednou budeme umět jazykem neurobiologie odpovídat na otázky položené v jazyce (objektivní) psychologie. Tuto snahu uznává, prosazuje a pěkně demonstruje F. Crick (Crick, 1997). Klást a řešit otázky tímto způsobem je však v jistém (metodologickém) smyslu „lehkým problémem“, což je označení dosti optimistické, ale jde tu spíše o něco jiného, než o míru obtížnosti – jde o srovnání s tím, co David Chalmers označuje jako „těžký problém“ (Chalmers, 1996)[50] – podle mne též poněkud optimisticky (ale jinak).

Chalmersův **těžký problém** je charakterizován otázkou:

Jak vůbec může fyzické (chemické, biologické) dění v mozku vést k mému subjektivnímu prožitku?

Tedy například, čím a proč dochází k tomu, že při světle určité vlnové délky je činnost zrakové části mého mozku doprovázena mým (a právě mým) prožitkem žlutosti? Mozek sám by přece mohl řídit mé chování – včetně mého zvolání: „Ó, jak krásná to žluť!“ – i bez tohoto prožitku, podobně jako je (asi) řízeno chování včel a automatů a mnohdy i mé vlastní, když není vědomé.

Myslím, že jádro současných debat lze většinou najít v neschopnosti dobře rozlišovat mezi „těžkým“ a „lehkým“ problémem. Lze si například všimnout (spolu s Chalmersem), že autoři odborných studií o vědomí mnohdy začínají úvahami o záhadnosti a nepostižitelnosti (subjektivně prožívaného) vědomí, načež celkem bez rozpaků nabídnou jako řešení svou originální teorii, kterou však po bedlivějším pohledu odhalíme jako příspěvek k „lehkému“ problému.

V této souvislosti nelze opominout práce vzniklé převážně v rámci teoretické informatiky, v nichž autoři navrhují rozličné abstraktní modely zpracování informace v hypotetických paměťových mediích (ve stylu neuronových sítí), na nichž by rádi demonstrovali i určité aspekty vědomí (např. Valiant, 1994, Wiederman, 1997; de Bruijn, 1999). Jakkoliv jde o záslužné snahy a jakkoliv bychom je právě v kontextu této studie měli uvítat jako reprezentanty třetí cesty – umělého modelování –, nelze se vyhnout určité skepsi. To, co např. de Bruin považuje za vědomí, jsou totiž jen některé dílčí a vnějšně popsatelné komponenty vědomí, jako například orientovaná pozornost (jakési aktivační okénko cestující nad distribuovanou pamětí), asociativní vybavování (jako princip pohybu tohoto okénka), zacházení s procedurami (nad myšlenkami) stejně jako s daty (myšlenkami samotnými) a z toho plynoucí reflexivita (sebemonitorování systému), nebo přecházení mezi různými stupni intenzity všech těchto aktivit. Je pravda, že tyto komponenty systému mohou být v korelaci s obdobnými, introspekčně sledovatelnými prvky našeho prožívaného vědomí, jsou však všechny v jistém smyslu čistě „syntaktické“, chtělo by se říci „vnější“, a s opravdovým prožíváním nemají nic víc společného. Tyto snahy bychom tedy mohli zařadit do oblasti „lehkých“, snad i „nejlehčích“ problémů.Je tu před námi základní problém, s nímž se každá budoucí věda o vědomí, a právě ona, bude muset vyrovnat. Vedle mozku, jeho neuronů a lecčeho kolem, tedy věcí, které jsou – aspoň principiálně – přístupny objektivním metodám vědy, existuje sféra našeho vnitřního, subjektivního prožívání, o níž víme jen proto, že máme vědomí, a bez níž by nebylo možno o vědomí mluvit, protože by nebylo o čem.

---

[50] Viz též (Chalmers, 1995). Rozsáhlá diskuse k Chalmersově pojetí „těžkého problému“ proběhla v časopise *Journal of Consciousness Studies,* Vol. 2 (1995), No. 3; Vol 3 (1996), Nos 1, 3, 4.

Možná, že se jednou dočkáme obsažné a sofistikované teorie toho, co se vlastně stane v mozkové tkáni, když se nám v hlavě zrodí zajímavý nápad, když se v obdivu zahledíme na mořský příboj, nebo když litujeme vlastní chyby. Možná to budeme umět i simulovat na důmyslných počítačích. Stěží však tyto teorie a simulace pomohou tomu, koho nikdy nic nenapadlo, kdo nikdy nic neobdivoval, nebo kdo se nikdy nezahleděl do svého nitra.

Literatura

Poznámka. V českém (a slovenském) jazyku existuje poměrně málo knižních publikací, které se vztahují k tématu této kapitoly, a stěží lze jejich výběr považovat za reprezentativní pro tento obor. Z poslední doby jsou to např.: (Gál, Kelemen, 1992), (Kelemen, 1994), (Searle, 1994), (Flanagan, 1995), (Minsky, 1996), (Crick, 1997), (Dennett, 1997), (Nosek, 1997), (Pstružina, 1998), (Pribram, 1999). Volnější souvislost mají (Sacks, 1993), (Ruyer, 1994), (Koukolík, 1995), (Jirků, Kelemen, 1996), neuronovými sítěmi se zabývá monografie (Novák a kol., 1988).

Kromě prací citovaných na příslušných místech jsem v této kapitole (hlavně v posledních dvou oddílech) čerpal též z těchto monografií: (Margolis, 1984), (Gregory, 1987), (Lycan, 1990), (Jacquette, 1994), (Guttenplan, 1994), (Honerich, 1995), (Franklin, 1995), (Nakonečný, 1995), (Rey, 1997) a z různých prací v časopise Journal of Consciousness Studies (Imprint Academic).

Anzenbacher A.: *Úvod do filozofie*. SPN, Praha, 1991.

Armstrong D.: *A Materialist Theory of the Mind*. Routledge and Kegan, London, 1968.

Axelrod R.: *The Evolution of Cooperation*. Harper and Collins, 1984.

Bechtel W., Abrahamsen, A.: *Connectionism and the Mind*. Basil Blackwell, Cambridge, Mass., 1991.

Bendová K.: *Sylogistika*. Karolinum, Praha 1998.

Berkeley G.: *Pojednání o základech lidského poznání* (přel. J. Brdíčko). Svoboda, Praha, 1610/1995.

Birnbacher D.: *Artificial Consciousness*. In: (Metzinger, 1995), s. 489–503.

Block N.: *Troubles with functionalism*. In: Minnesota Studies in the Philosophy of Science 9, University of Minnesota Press, Mineapolis, 1978 s. 261–326.

Boden M. (editorka): *The Philosophy of Artificial Intelligence*. Oxford University Press, Oxford, 1990.

Bringsjord S.: *Is the connectionist-logicist clash one of AI's wonderful red herrings?* Journal of Experimental and Theoretical Artificial Intelligence 3, 1991, s. 319-349.

Brooks R. A.: *Intelligence without representation*. Artificial Intelligence 47, 1991, s. 139-160.

de Bruijn N. G.: *A model for associative memory, a basis for thinking and consciousness*. In: J.Wiederman et al. (editoři): Automata, Languages and Programming, Springer, Berlin, 1999, s. 74–89.

Campbell K.: *Body and Mind*. Univ. of Notre Dame, Indiana, 1970.

Cragg B.G., Temperley H.N.V.: *The organisation of neurones: A cooperative analogy*. EEG Clinical Neurophysiology 6, 1954, s. 37.

Crick F.: *Věda hledá duši (Překvapivá domněnka)*. Mladá fronta, Praha 1997.

Csonto J.: *Umělý život*. In: tento svazek, kap. **

de Bruijn N. G.: *A model for associative memory, a basis for thinking and consciousness*. In: J.Wiederman et al. (editoři): Automata, Languages and Programming, Springer, Berlin, 1999, s. 74–89.

Dennett D. C.: *Consciousness Explained*. Little, Brown and Co., Boston, 1991a.

Dennett D.C.: *Mother Nature versus the Walking Encyclopedia: A western drama.* In: (Ramsey, Stich, Rumelhart, 1991), 1991b, s. 21-30.

Dennett D. C.: *Druhy myslí. K pochopení vědomí*. Archa, Bratislava 1997.

Descartes R.: *Meditácie o prvej filozofii.* Chronos, Bratislava 1641/1997.

Domany E., van Hemmen J.L., Schulten, K. (editoři): *Models of Neural Networks*. Springer-Verlag, Berlin, 1991.

Dreyfus H.: *What Computers Can't Do*. (2. vydání) New York: Harper & Row 1979.

Eccles J.: *How the Self Controls Its Brain.* Springer-Verlag, Berlin, 1994.

Flanagan O.: *Vedomie*. Archa, Bratislava 1995.

Fodor J.A.: *The mind-body problem*. Scientific American, January 1981, s. 124–132.

Fogel L.J., Owens A. J., Walsh M. J.: *Artificial Intelligence through Simulated Evolution.* New York, 1966.

Franklin S.: *Artificial Minds*. The MIT Press, Cambridge, Mass., 1995.

Freeman W.: *Societies of brains* (diskuse s Jean Burnsovou). Journal of Consciousness Studies 3, 1996, s. 172-180.

Gál E., Kelemen J. (editoři): *Myseĺ/telo/stroj*. Bradlo, Bratislava 1992.

Graubard S.R. (editor): *The Artificial Intelligence Debate: False Starts, Real Foundations.* Cambridge, MA: The MIT Press 1988.

Gregory R.L. (editor): *The Oxford Companion to the Mind*. Oxford University Press, Oxford, 1987.

Grim J.: *Pravděpodobnostní neuronové sítě*. In: tento svazek, kap. **

Guttenplan S. (editor): *A Companion to the Philosophy of Mind*.

Haugeland J.: *Artificial Intelligence: The Very Idea*. Cambridge, MA: The MIT Press 1985.

Havel I. M.: *Neuronové počítače a jejich inteligence*. In: SOFSEM'88, 1988, s. 83-118.

Havel I. M.: *Connectionism and Unsupervised Knowledge Representation*. In: R.Trappl (editor): Cybernetics and Systems'90. Singapore: World Scientific, 1990.

Havel I. M.: *Artificial Intelligence and Connectionism*. In: Advanced Topics in AI (Mařík, Štěpánková, Trappl, editoři), Lecture Notes in Artificial Intelligence, Springer-Verlag 1992, s. 25-41.

Havel I. M., *Scale Dimensions in Nature*. Int. Journal of General Systems **24** No. 3, 1996, s. 295-324.

Havel I. M.: *Artificial Thought and Emergent Mind*. In: Proceedings IJCAI'93, Morgan Kaufman Professional Book Center, Denver, CO, USA, 1993, s. 758-766.

Havel I. M.: *Otevřené oči a zvednuté obočí*. Nakladatelství Vesmír, Praha 1998.

Havel I. M.: *Třetí život badatelů*. Vesmír 78, 1999, č. 6, s. 303 (1999a).

Havel I. M.: *Jitro kyberkultury*. Světová literatura, 1999, č. 1, , s. ** (1999b).

Havel I.M.: *Living in Conceivable Worlds*. In: Proceedings Conf. on Paraconsistent Logic, Foundations of Science, Kluwer Academic Publishers, 1999c.

Havel I. M., Hájek P.: *Filozofické aspekty strojového myšlení.* In: SOFSEM'82, 1982, s. 171-211.

Hegel G. W. F.: *Fenomenologie ducha* (přel. J. Patočka). NČSAV, Praha 1806/1960.

Heylighen F., Bollen J.: *The World-Wide Web as a Super-Brain: from metaphor to model.* In: R.Trappl (editor): Cybernetics and Systems '96. Austrian Society for Cybernetics, 1996, s. 917-922.

Hillis W.D.: *Intelligence as an emergent bahavior; or, the songs of Eden*. In: (Graubard, 1988), s. 175-190.

Hinton G.E., Sejnowski TJ.: *Learning and relearning in Boltzmann machines*. In: Rumelhart a kol., Parallel Distributed Processing, Vol. 1., The MIT Press, Cambridge, Mass., 1986, s. 282-317.

Hodgson D.: *The Mind Matters: Consciousnes and Choice in a Quantum World.* Oxford University Press, Oxford, 1991.

Hofstadter D.R.: *Gödel, Escher, Bach: An Eternal Golden Braid.* Harvester Press, Brighton, 1979.

Hofstadter D.R.: *Waking up from Boolean dream, or, subcognition as computation*. In: (Hofstadter, 1985), s. 631-665.

Hofstadter D.R.: *Metamagical Themas: Questing for the Essence of Mind and Pattern*. Bantam Books, Toronto, 1985.

Honderich T. (editor): *The Oxford Companion to Philosophy*. Oxford University Press, Oxford, 1995.

Hopfield J.J.: *Neural networks and physical systems with emergent collective computational abilities*. Proc. Nat. Acad. Sci USA 79, 1982, s. 2554-2558.

Hořejš J.: *Neuronové sítě*. In: (Mařík a kol., 1993), s. 217–241.

Höschl C: *Biologická psychiatrie roku 2000*. Psychiatrie, 1, 1997, 3-4, s. 107-116.

Chalmers D.J.: *The puzzle of conscious experience*, Scientific American, Dec. 1995, s. 80–86.

Chalmers D.J. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press. 1996.

Churchland P. S.: *Neurophilosophy: Toward a Unified Science of the Mind/Brain*. The MIT Press, Cambridge, Mass., 1986.

Jacquette D.: *Philosophy of Mind*. Prentice Hall, Englewood Cliffs, NJ, 1994.

Jirků P., Kelemen J.: *Kapitoly z kognitivní vědy*. Vysoká škola ekonomická, Praha, 1996.

Kauffman S.A.: *Antichaos and Adaptation*. Scientific American, August '91, s. 78-84 (1991).

Kelemen J.: *Strojovia a agenty*. Archa, Bratislava, 1994.

Kelemen J.: *Reaktivní agenty*. In: tento svazek, kap. **

Klír J., Valach M., *Kybernetické modelování*. SNTL, Praha, 1965.

Koch C.: *Biophysics of Computation: Information Processing in Single Neurons.* Oxford University Press, 1998.

Koch C., Segev I. (editoři): *Methods in Neuronal Modeling: From Ions to Networks*. 2nd ed., The MIT Press, Cambridge, Mass., 1998.

Kolb B., Whishaw I.Q.: *Fundamentals of Human Neuropsychology*. W.H.Freeman, New York, 1990.

Koukolík F.: *Mozek a jeho duše*. Makropulos, 1995.

Kůrková V.: *Aproximace funkcí neuronovými sítěmi*. In: tento svazek, kap. **

Lakoff G., Johnson M.: *Philosophy in the Flesh: The Embodied Mind and its Challenge to Western Thought*. Basic Books, New York, 1999.

Langton C. G.: *Artificial life*. In: Langton C. G. (ed.): Artificial Life. Addison-Wesley, 1989, s. 1-47.

Lashley K. S.: *Brain Mechanisms and Intelligence*. University of Chicago Press, Chicago, 1929.

Lažanský a kol. *Aplikace metod umělé inteligence.* In: (Mařík a kol., 1997), s. 310-356.

Lažanský J., Kubalík J.: *Genetické algoritmy.* In: tento svazek s. **

Little W.A.: *The evolution of non-Newtonian views of brain function*. Concepts in Neuroscience **1**, 1990, s. 149-164.

Lucas J. R.: *Minds, machines and Gödel*. Philosophy 36, 1961, s. 112-127.

Lycan W.G.: *Consciousness*. Cambridge, MA: The MIT Press 1987.

Lycan W. G. (editor): *Mind and Cognition: A Reader*. Basil Blackwell, Cambridge, Mass., 1990.

Margolis J.: *Philosophy of Psychology*. Prentice-Hall, Englewood Cliffs, NJ, 1984.

Mařík V. a kol. *Umělá inteligence I*. Academia, Praha, 1993.

Mařík V. a kol. *Umělá inteligence II*. Academia, Praha, 1997.

Mařík V. a kol. *Umělá inteligence III*. Academia, Praha, 1999.

Mařík V. a kol.: *Koordinace a kooperace v multiagentních systémech*. In: tento svazek s. **

McCulloch W.S., Pitts W.: *A logical calculus of the ideas immanent in nervous activity*. Bull. Math. Biophys. 5, 115-133 (1943). Též In: (Boden, 1990), s. 22–39.

McGinn C.: *Can we ever understand consciousness?* The New York Review, June 10, 1999, s. 44—48.

Melzack R., *Phantom Limbs*. Scientific American, Apr., 1992, s. 120–126.

Merleau-Ponty M.: *The Structure of Behvior*. (překl. do angl. A. Fischer) Beacon Press, Boston, 1963.

Metzinger T. (editor): *Conscious Experience*. Imprint Academic, 1995.

Minsky M.: *The Society of Mind*. Simon & Schuster, New York, 1985.

Minsky M.: *Konštrukcia mysle*. Archa, Bratislava, 1996.

Nakonečný M.: *Lexikon psychologie*. Vodnář, Praha, 1995.

Newell A.: *Physical symbol systems*. Cognitive Science 4, 1980, s. 135-183.

Newell A., Simon, H.A.: *Computer science as empirical enquiry: symbols and search.* Communications of the ACM 19 (1976); též In: (Boden, 1990), s. 105-132.

Nosek J.: *Mysl a tělo v analytické filosofii*. Nakl. Filosofického ústavu AV ČR, Praha 1997.

Novák M. a kol.: *Umělé neuronové sítě - teorie a aplikace*. C.H.Beck, Praha 1998.

Osherson D.N.: *The study of cognition*. In: An Invitation to Cognitive Science, Vol. 1 (D. N. Osherson, ed.), The MIT Press, Cambridge, Mass., 1995, s. xi—xviii.

Patočka J.: *Úvod do fenomenologické filosofie.* OIKOYMENH, Praha, 1993.

Penrose R.: *The Emperor's New Mind*. Oxford University Press, Oxford, 1989.

Penrose R.: *Shadows of the Mind: A Search for the Missing Science of Consciousness*. The Oxford University Press, Oxford, 1994.

Peregrin J.: *Úvod do analytické filosofie*. Herrmann a synové, Praha, 1992.

Peregrin J.: *Internet: dobro nebo zlo?* Filosofický časopis 46, 1998, č. 1, s. 12.

Popper K.R., Eccles J.C.: *The Self and Its Brain*. Springer-Verlag, Berlin, 1977.

Pribram K.: *Languages of the Brain: Experimental Paradoxes and Principles in Neuropsychology*. Prentice-Hall, Englewood Cliffs, NJ, 1971.

Pribram K.: *Brain and Perception: Holonomy and Structure in Figural Processing*. Lawrence Erlbaum Associates, New Jersey, 1991.

Pribram K. (editor): *Rethinking Neural Networks: Quantum Fields and Biological Data*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1993.

Pribram K.: *Karl H. Pribram – Mozek a mysl, Holonomní pohled na svět* (výběr prací, uspoř. J. Fiala). Gallery, Praha, 1999.

Priest S.: *Theories of the Mind*. Penguin Books, London, 1991.

Pstružina K.: *Svět poznávání. K filosofickým základům kognitivní vědy.* Nakladatelství Olomouc, 1998.

Putnam H.: *The mental life of some machines*; též *The nature of mental states*. In: *Philosophical papers* Vol. 2, Cambridge University Press, 1975 (pův. publ. v r. 1967).

Ramsey W. a kol.: *Connectionism, Eliminativism, and the future of folk psychology*. In: (Ramsey, Stich, Rumelhart, 1991), s. 199-228.

Ramsey W., Stich S. P., Rumelhart D. E. (ed.): *Philosophy and Connectionist Theory*. Lawrence Erlbaum Ass., Hillsdale, NJ, 1991.

Rey G.: *Contemporary Philosophy of Mind*. Blackwell, Cambridge, Mass., 1997.

Rich E., Knight K.:, *Artificial Intelligence*. 2nd ed. McGraw-Hill, New York, NY, 1991.

Rosenblatt F.: *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Spartan Books, Washington, D.C., 1962.

Rota Gian-Carlo: *O zhoubném vlivu matematiky na filosofii*. Vesmír 78, (1999), 6, s. 345–349.

Rumelhart D. E. a kol.: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. (2 vols.), The MIT Press, Cambridge, Mass. 1986.

Ruyer R.: *Paradoxy vědomí - Expresivita*. Ped. fak. UK, Praha 1994.

Ryle G.: *The Concept of Mind*. Barnes and Noble Books, New York, 1949.

Sacks O.: *Muž, který si pletl manželku s kloboukem*. Mladá fronta, Praha, 1993.

Salthe S.N.: *Two forms of hierarchy theory in Western discourses*. International Journal on General Systems 18, 1991, s. 251-264.

Scott A.: *Stairway to the Mind*. Springer-Verlag, NY. 1995.

Searle J.R.: *Minds, Brains, and Programs*. Behavioral and Brain Sciences **3**, 1980, s. 63-108 (1980). Též In: (Boden, 1990) s. 67–88.

Searle J.R.: *Intentionality*. Cambridge University Press, Cambridge, Cambridge, 1983.

Searle J.R.: *The Rediscovery of the Mind.* The MIT Press, Cambridge, Mass. 1992.

Searle J. R.: *Mysl, mozek a věda*. Mladá fronta (Váhy), Praha 1994 (angl. originál Harvard University Press, 1984).

Searle J.R.: *The Mystery of Consciousness.* Granta Books, London, 1997.

Searle J.R.: *I married a computer.* The New York Review, April 8, 1999, s. 34–38.

Sherrington D., Kirkpatrick S.: *Spin glasses.* Phys. Rev. Letters 35, 1975, s. 1792.

Schank R.C., Abelson R.P.: *Scripts, Plans, Goals, and Understanding*. Erlbaum, Hillsdale, NJ, 1977.

Schweber S. S.: *Physics, community and the crisis in physical theory*. Physics Today, Nov. 1993, s.34–40.

Skinner B.F.: *Science and Human Behavior*. Macmillan, 1933.

Smolensky P.: *Neural processing in dynamical systems: Foundation of harmony theory*. In: (Rumelhart a kol., 1986), Vol. 1, s. 194-281.

Smolensky P.: *Connectionist AI, symbolic AI and the brain*. Artificial Intelligence 1, 1987.

Smolensky P.: *On the proper treatment of connectionism*. Behavioral and Brain Sciences 11, 1988, s. 1-74.

Stapp H. P.: *Mind, Matter, and Quantum Mechanics*. Springer-Verlag, Berlin, 1993.

Steels L. and Brooks R. A., (editoři): *The Artificial Life Route to Artificial Intelligence – Building Embodied, Situated Agents*. Lawrence Erlbaum Assoc., Hillsdale, NJ, 1995.

Stich, S.: *From Folk Psychology to Cognitive Science. The Case Against Belief*. The MIT Press, Cambridge, Mass., 1983.

Štěpánková O. a kol.: *Distribuovaná umělá inteligence*. In: (Mařík a kol., 1997), s. 142–177.

Thomas L.: *Buňka, medúza a já*. Mladá fronta, Praha 1981, s.214.

Turing A.: *On computable numbers with an application to the Entscheidungsproblem.* Proc. London Math. Soc., Ser. 2, Vol. 42, 1936, s. 230–265.

Turing A.: *Computing machinery and intelligence*. Mind LIX, no. 2236, 1950, s. 433–460. Též In: (Boden, 1990) s. 40–66.

Valiant L.: *Circuits of the Mind*. Oxford University Press, Oxford, 1994.

Varela F.J.: *The re-enchantment of the concrete*. In: (Steels, Brooks, 1995), s. 11–22.

Varela F. J., Thompson E., Rosch, E.: *The Embodied Mind: Cognitive Science and Human Experience*. The MIT Press, Cambridge, Mass., 1991.

Vinař O.: *Průlinčitý mozek – komunikují neurony pouze synapsemi?* Vesmír 78, 1999, č. 9, s. 492–494.

Warwick K.: *Úsvit robotů, soumrak lidstva*. Nakladatelství Vesmír, Praha 1999.

Watson J. B.: *Psychology as the behaviorist views it*. Psychological Review, 20, 1913, s. 158–177.

Webb J.C.: *Mechanism, Mentalism, and Metamathematics*. D.Reidel, Dordrecht 1980.

Weizenbaum J.: *Computer Thought and Human Reason.* W.H.Freeman, San Francisco, 1976.

Wiederman J.: *Towards Machines that Can Think*. In: Proceedings SOFSEM'97, Lecture Notes in Computer Science, Springer-Verlag, Berlin, 1997.

Zdráhal Z.: *Reprezentace znalostí.* In: (Mařík a kol., 1993).