Specifikace výpočetního modelu mysli vědomého vtěleného agenta

Jiří Wiedermann[1]

Hjakudžó svolal žáky a postavil před ně džbán: Kdo mi řekne, že to je džbán, aniž ho pojmenoval? Isan předstoupil a kopl do džbánu, až ho převrhl.

Abstrakt. Nastíníme jednoduchou, ale přesto kognitivně účinnou architekturu inteligentního agenta. Model využívá dvou komplementárních vnitřních modelů světa: jeden pro „syntax" poznaného světa a druhý pro jeho sémantiku. Tyto modely řeší problém porozumění konceptům a podporují algoritmické procesy, jejichž efekty se pro pozorovatele jeví jako projevy vyšších kognitivních funkcí, jakými jsou imitační učení, rozvoj komunikace, řeči, myšlení a vědomí.

1 Úvod

Citát z úvodu z práce předního českého japanologa Miroslava Nováka, *Každý sám svým pánembohem* [15] naznačuje, že mistři filosofie zen znali klíč k řešení problému, který se v umělé inteligenci skrývá pod názvem „*problém ukotvení symbolů*". Tento problém se týká otázky, jak slova získávají svůj význam a co to vlastně ten význam je [9]. Umělá inteligence se tímto problémem začala zabývat v osmdesátých letech dvacátého století, kdy se tento problém manifestoval v podobě „bajky" o tzv. čínské komnatě, kterou „vymyslel" filosof J. Searle [22]. Tato bajka ilustrovala problém, jestli počítač může rozumět slovům (tj. symbolům), pomocí kterých komunikuje s lidmi. Searle upozornil na skutečnost, že počítač nemůže se slovními symboly dělat nic jiného, než je podle nějakých pravidel transformovat na jiné symboly (v konečném důsledku na posloupnosti nul a jedniček), kterých sémantiku počítač také nemůže znát. Takže, alespoň podle Searla, „*tudy cesta nevede*". Tato bajka rozvířila snad dodnes neutuchající diskusi a vedla k rozvoji teorií, které tvrdily, že proto, aby uměle-inteligentní systémy „rozuměly" tomu, co dělají, potřebují mít tělo. Takovým systémům se někdy říká „vtělení agenti". Teprve tělo, obvykle v interakci s vnějším světem, dává těmto systémům prostředek, jak porozumět svým vlastním akcím i akcím ostatních vtělených agentů, a jak s nimi komunikovat. To ostatně využil i žák Isan z našeho citátu: nemít těla a nebýt džbánu, nemohl by kopnout do džbánu, ten by se nepřevrátil a nevydal by přitom charakteristické zvuky a nevyvolal by v přihlížejících soubor příslušných

[1] Ústav informatiky AV ČR, v.v.i., Pod Vodárenskou věží 2, 182 07 Praha 8, Česká republika, e-mail: jiri.wiedermann@cs.cas.cz

paměťových a smyslových vjemů, které jsou právě oním hledaným významem slova „džbán“. Poeticky řečeno, Isan tedy jednoduše u přihlížejících „zahrál na stejnou strunu“, kterou by v nich rozezvučelo nevyslovené slovo „džbán“.

Vraťme se ještě na chvíli k Searlovu počítači, který proto, aby komunikoval s lidmi „inteligentně“, musel zjevně mít nějakou informaci o tom, jak vypadá vnější svět, jaké asi jsou schopnosti a pocity lidí v různých situacích, atd. Zkrátka, musel mít jakýsi vnitřní model vnějšího světa, jakkoliv reprezentovaný. Tento model pak tvořil součást pravidel, pomocí kterých by Searlův počítač komunikoval s lidmi,

Myšlenka, že netriviální inteligentní systémy by měly využívat nějakou formu vnitřního modelu světa, sahá do prvopočátků umělé inteligence. Nicméně, pokusy o řízení chování pomocí formálních pravidel pracujících nad symbolickými modely světa selhaly. Následkem toho se hlavní proud výzkumu obrátil v poslední dekádě minulého století směrem k biologicky motivovaným modelům založeným na přímém napodobování chování živých tvorů, většinou hmyzu. Tento přístup zdůrazňoval nutnost ztělesnění a situovanosti pro využití chování jednoduchých robotů řízených senzorickými podněty (viz např. [4]). Toto paradigma fungovalo dobře zejména ve spojení s tzv. subsumpční architekturou, která využívala na sebe navazující vrstvy stále složitějšího chování, které bylo programováno na míru požadovaným stále náročnějším úkolům (viz např. [17]). Bohužel, po počátečních úspěších, převážně při konstrukci různých robotů reagujících na přímé podněty se narazilo na meze takového přístupu. Zjevné to bylo zejména v humanoidní robotice, kde se další postup směrem k vyšším kognitivním funkcím zdál nemožný bez zásadních inovací kognitivní architektury příslušných agentů. Možná, že zde je třeba hledat důvody útlumu konstrukčních aktivit v oblasti humanoidní robotiky, který se ale na druhé straně zdá být kompenzován nárůstem počtu teoretických prací v přilehlých oblastech, kladoucích si za cíl umělé vědomí (viz např. [10]).

V současné době převažuje názor, že pro prolomení dříve zmíněné bariéry, na kterou narazila robotika přímými podněty řízených robotů, a k otevření cesty k vyšším mozkovým funkcím potřebujeme automatické mechanizmy rozšiřující dříve získané znalosti a dovednosti robotů [23]. Takové mechanizmy mohou využívat vhodné vnitřní modely světa. Dnes převažující trendy dávají přednost tzv. subsymbolickým modelům (především neuronovým sítím) před tzv. symbolickými modely, u kterých je přímočarý vztah mezi objekty a jejich reprezentací. Přehled současného stavu a diskusi o vnitřních modelech světa lze nalézt např. v pracích [5] a [11].

V práci [11] její autoři Holland a Goodman argumentují ve prospěch interního modelu světa sestávajícího ze dvou oddělených, avšak spolupracujících částí: modelu agenta a modelu jeho okolí. Nedávno byl v oblasti teoretické informatiky publikován podobný model [3] sloužící k úvahám o definici vědomí. V těchto a podobných pracích se jejich autoři domnívají, že klíč k pochopení vědomí je ukryt v definici a ve funkci výše zmíněného dvousložkového modelu. Cruse [5] přichází k podobnému závěru použitím vnitřního modelu, který zachycuje pouze agentovo vlastní tělo.

Tato práce navazuje na práce autorů zmiňovaných v předchozím odstavci. Metodologickým východiskem pro náš přístup bude postup používaný v softwarovém inženýrství při návrhu velkých systémů. Začneme nástinem architektury kognitivního systému a uvedeme neformální funkční specifikace jednotlivých jeho modulů. To znamená, že definujeme typ dat a jejich tok mezi jednotlivými moduly a také úkol jednotlivých modulů z hlediska zpracování příslušných dat. Dále již budeme argumentovat ve prospěch uvedeného modelu — jaké jsou důvody domnívat se, že model podporuje realizaci procesů, o kterých se lze oprávněně domnívat, že odpovídají vyšším kognitivním schopnostem, jakými jsou např. imitační učení a rozvoj komunikace, řeči, myšlení a vědomí.

Naše kognitivní architektura bude vycházet z myšlenky dvou spolupracujících vnitřních modelů. První model bude tvořen tzv. zrcadlovou neuronovou sítí, která se bude učit často se opakující „percepčně–behaviorální" jednotky. Tyto jednotky jsou reprezentovány v modelu pomocí tzv. multimodální informace, která je fúzí senzorických a motorických informací vztahujících se k „jednotce" situace. Návrh zrcadlové neuronové sítě, která je odpovědna za agentovu situovanost v jeho prostředí, byl inspirován zjištěnými vlastnostmi nedávno objevených biologických zrcadlových neuronů v mozku primátů [19], [20] a také nejnovějšími neurofysiologickými objevy [7], které naznačují, že zrcadlové neurony skutečně reprezentují kognitivní mechanizmus pro porozumění akcím, úmyslům a emocím, které jsou vybuzeny senzorickými stimuly. Zrcadlová neuronová síť reprezentuje v jistém smyslu jak agenta, tak i jeho okolí; zachycuje současně syntaxi i sémantiku korektního chování. V odpovídající multimodální informaci je svět reprezentován senzorickými vstupy a agentovy akce jsou zachyceny v odpovídajících motorických instrukcích a jeho „pocitech" danými zpětnovazební informací od jeho vnitřních senzorů. Takže, na jedné straně, zrcadlová síť zachycuje podobnou informaci jako Cruseho vnitřní model (což je vlastně také neuronová síť) agentova těla anebo model agenta v Hollandově a Good-

manově modelu. Na druhé straně, protože v zrcadlové síti jsou také prvky environmentální informace zprostředkované agentovými senzory, tato síť jistým fragmentovaným způsobem také reprezentuje okolí (na sub-symbolické úrovni) podobně jako druhá část Hollandova a Goodmanova modelu.

Druhý vnitřní model světa je tvořen agentovou řídící jednotkou. Tato jednotka kontinuálně zpracovává multimodální informaci dodávanou zrcadlovou sítí. Úkolem řídící jednotky je dolovat znalosti z toku multimodálních informací. V řídící jednotce jsou znalosti reprezentovány pomocí rekurentní sítě konceptů. Základní koncepty jsou tvořeny jednotkami multimodálních informací. Dále řídící jednotka také automaticky odvozuje koncepty, které neodpovídají žádné multimodální informaci, nýbrž reprezentují znalost odvozenou, abstrahovanou ze základních konceptů. Řídící jednotka odhaluje pomocí statistických mechanizmů často se vyskytující vzory v toku základních konceptů a na základě těchto vzorů formuje abstraktnější koncepty a učí se jejich časové či prostorové vztahy. To znamená, že příslušná síť konceptů se vlastně učí různým vzorcům chování. Vycházejíc z multimodální informace o současné situovanosti, řídící jednotka určí následující akci agenta. Také řídící jednotka je implementována pomocí rekurentní neuronové sítě. Je zřejmé, že řídící jednotka zachycuje dynamické aspekty agentovy interakce s jeho okolím a jako taková nemá protipól v modelech autorů zmiňovaných předtím.

Náš model umožní věrohodné vysvětlení výpočetních mechanizmů stojících za jevy, které se ve svých důsledcích podobají vyšší mozkové činnosti, včetně vědomí. V našem modelu je výpočetní vědomí chápané jako poslední fáze posloupnosti postupně stále více kognitivně náročnějších schopností systému rozvíjejících se ve stále více stimulujícím prostředí. Odpovídající posloupnost začíná na úrovni schopnosti učení pomocí imitace, pokračuje přes schopnost naučit se a porozumět řeči těla, posunkům a artikulované komunikaci mezi příslušníky stejného druhu, a dále vede přes schopnost mluvení sama k sobě k myšlení. Ve finále tento vývoj vede do stavu, ve kterém jsou kognitivní entity schopné popsat ve vyšším jazyku libovolnou minulou, přítomnou anebo očekávanou událost a přemýšlet o nich („*generovat na míru šité příběhy*“, jak to nazývá Blum ve své práci [3]). Také v našem modelu se tento stav považuje za příznak vědomí. To ostatně dobře odpovídá Minského poznámce o tom, že „*vědomí je velký kufr*“, obsahující mnoho různých mentálních schopností [14].

Myšlenka, že zrcadlové neurony jsou klíčem k imitačnímu učení a že hrají důležitou roli při rozvoji přirozeného jazyka, počala klíčit začátkem tohoto století (viz. např. [1], [8], [13], [18], [19]). Jeden

z prvních výpočetních modelů, založených na zrcadlových neuronech, byl publikován v práci [27]. V nynější práci je rozpracována myšlenka chápání zrcadlových neuronů jako vnitřního „syntaktického" modelu světa, který společně s vnitřním sémantickým modelem světa skýtá rámec pro řešení problému ukotvení symbolů a pro rozvoj výpočetního vědomí; dále je zde prezentován nástin příslušných kognitivních algoritmů včetně definice výpočetního vědomí. Práce představuje rozpracování autorových myšlenek, prezentovaných v [26], [27], [28] a [29]. Naše výsledky potvrzují konstruktivním způsobem intuici dřívějších badatelů, že totiž mání těla a vnitřní modely světa představují základ pro rozvoj vyšších mentálních funkcí, včetně vědomí.

Struktura článku je následující: ve 2. části představíme podrobněji náš model. Ve 3. části popíšeme jeho fungování vedoucí ke vzniku výpočetního vědomí.

2 Model

Struktura modelu je znázorněna na obr. 1. Model se skládá ze 4 hlavních součástí: senzomotorické jednotky, senzomotorického modelu světa reprezentovaného zrcadlovou neuronovou sítí, řídící jednotky a těla. Tok (digitálních) dat mezi jednotlivými moduly je zobrazen šipkami.

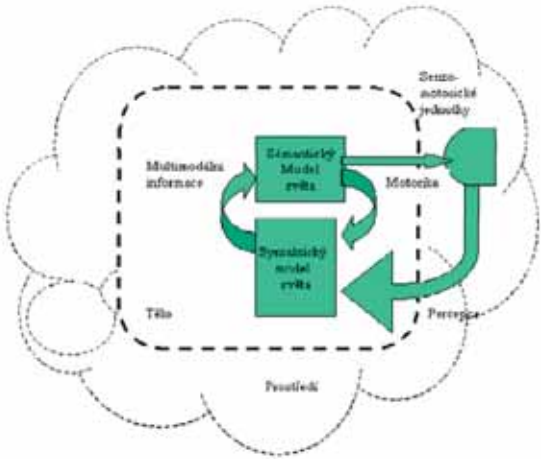
Senzomotorické jednotky dostávají od řídící jednotky *motorické instrukce*. Nejsou to pouze instrukce pro agentovy lokomoční orgány, ale také instrukce určené senzorickým orgánům sloužící k nastavení jejich parametrů: zacílení na určitý cíl, zaostření apod. Kopie těchto instrukcí jsou současně nasměrovány do zrcadlové sítě. Senzomotorické jednotky posílají do zrcadlové sítě dva druhy dat.

Prvním druhem jsou tzv. *extracepční data* nesoucí informaci od senzorických jednotek, které zkoumají agentovo okolí. V tomto případě reagují senzory na fyzikální vstupy (elektromagnetické vlnění, zvuky, tlak, teplota, atd.) a transformují je do digitální podoby. Druhý typ dat jsou tzv. *propriocepční data* pocházející od senzomotorických jednotek agenta rozmístněných uvnitř jeho těla. U člověka je příkladem takové informace vnitřní teplota, krevní tlak, svalový tonus, apod.

Další součástí modelu je *zrcadlová síť*. Je to síť umělých neuronů, které modelují chování reálných zrcadlových neuronů. V každé jednotce této sítě (skládající se z několika neuronů) se setkávají extracepční a propriocepční data od senzomotorických jednotek s motorickými instrukcemi z řídící jednotky. Jednotka tvoří jejich konjunkci, která se nazývá *multimodální informace*. Zrcadlová síť plní tři hlavní úkoly:

- *Učení:* pamatuje si často se vyskytující multimodální informace;
- *Identifikace:* k dané vstupní multimodální informaci síť najde „nejpodobnější" multimodální informaci zapamatovanou v síti;

• *Asociativita:* v případě, že do sítě vstoupí neúplná či poškozená multimodální informace, síť doplní její chybějící či poškozené části.



Obrázek 1. Vtělený kognitivní agent se dvěma vnitřními modely světa

Aby mohla síť tímto způsobem pracovat, musíme zařídit, aby si pamatovala pouze konečné množství „důležité" informace. To lze dosáhnout vhodnou parametrizací percepčních dat a motorických instrukcí pomocí konečné množiny hodnot parametrů. Jinou možností je využití fuzzy přístupů, kterých efekt je hrubá klasifikace dat do konečného počtu shluků podobných multimodálních dat. Pro správnou práci asociativního mechanizmu zrcadlové sítě je nezbytné, aby multimodální informace obsahovala co nejvíce redundantních údajů, které umožní zrekonstruovat celou informaci v případě, že některé její části chybějí. Proto je důležité, aby agent byl vybaven dostatečným množstvím senzorů, poskytujících k jedné „události" informace různého typu. Speciálně budeme předpokládat, že motorické údaje samotné stačí pro rekonstrukci zbytku multimodální informace. Tento předpoklad budeme potřebovat pro práci algoritmu myšlení.

Každá jednotka zrcadlové sítě se specializuje na naučení a poté i rozeznání specifické multimodální informace, která odpovídá „*percepčně-behaviorální jednotce*". Učení se děje průběžně v případě, kdy do sítě vstupují „nepoškozené" kompletní multimodální informace. Taková situace se nazývá *standardním učícím režimem*. Učení postupuje pomocí tzv. hebbovských principů, tj. posilováním vah neuronů

které reprezentují danou multimodální informaci vždy, když se taková informace rozezná.

Pokud funguje zrcadlová síť naznačeným způsobem, tak jejím výstupem je proud kompletních multimodálních informací bez ohledu na to, jestli informace vstupovala původně do sítě poškozena anebo neúplná. Tento proud teče do řídící jednotky; zde se jednotky multimodální informace nazývají *koncepty*. Úkolem řídící jednotky je dolovat znalosti ze vstupujícího proudu konceptů a aktivovat jiné koncepty, zapamatované v řídící jednotce. Motorická část aktivovaných konceptů je pak zaslána senzomotorickým jednotkám a její kopie do zrcadlové sítě. Řídící jednotka tedy určuje další akci agenta.

V řídící jednotce existují koncepty odpovídající každému výskytu multimodální informace vystupující ze zrcadlové sítě. Navíc se v řídící jednotce formují nové tzv. *abstraktní koncepty*, které bezprostředně neodpovídají žádné konkrétní multimodální informaci. Koncepty jsou v řídící jednotce propojeny přes vazby zvané *asociace*, které jsou ohodnoceny *vahami* různé velikosti. Koncepty společně s asociacemi a jejich vahami tvoří agentovu paměť.

Pravidla pro formování nových konceptů, vznik asociací a velkost jejich vah jsou založena na principech, kterých původ sahá až k anglickému filozofovi 18. století, D. Humovi [12]:

- *Současnost*: dva koncepty budou asociovány (anebo, pokud již jsou asociovány, tak váha asociace vzroste), pokud se často vyskytují současně; souběžně vznikne nový koncept, který je sjednocením obou konceptů. Tento nový koncept je *konkretizací* každého z původních dvou konceptů,
- *Souslednost*: dva koncepty budou asociovány (anebo, pokud již jsou, tak váha asociace vzroste), pokud se často vyskytují jeden po druhém.
- *Podobnost*: koncept bude asociován s jiným konceptem, pokud jsou si podobny; míra podobnosti je definována velikostí překryvu příslušné multimodální informace.
- *Abstrakce*: společná část dvou konceptů tvoří koncept, který je abstrakcí obou konceptů; tento koncept se přidá k množině původních konceptů.

Kontrolní jednotka pracuje dle následujících pravidel. V každém okamžiku některé její koncepty jsou v aktivním stavu. Tyto koncepty reprezentují současný *„mentální stav"* agenta. Vstupující multimodální informace aktivuje další koncepty. Nový mentální stav se počítá pomocí asociací, které jsou mezi aktivními a ostatními koncepty. Aktivní koncepty excitují koncepty, se kterými jsou asociovány silou

úměrnou vahám jednotlivých asociací. Tato excitace samotná mírně posiluje váhy příslušných asociací. Váhy ostatních asociací jsou naopak mírně sníženy — to modeluje proces zapomínání. V množině excitovaných konceptů se vyberou nejvíce excitované koncepty a tyto jsou aktivovány; ostatní koncepty jsou deaktivovány. Množina nově aktivovaných konceptů určuje nový mentální stav a také další motorické akce agenta. Na množinu aktivních konceptů lze nahlížet jako na *krátkodobou (operační) paměť*. Množina všech konceptů, asociací a jejich vah odpovídá *dlouhodobé paměti agenta*.

Je zřejmé, že mechanizmus výpočtu nového mentálního stavu připomíná mechanizmus výpočtu konečného automatu a že jej také lze implementovat podobně jako model zrcadlových neuronů — tj. pomocí neuronové sítě.

Na základě předchozích principů je řídící jednotka schopna realizovat jednoduché kognitivní úkoly, jako např. učení současně se vyskytujících konceptů (pomocí současnosti), jejich posloupnost (tzv. jednoduché podmiňování pomocí souslednosti), podobné chování a počítat jejich abstrakce. Mechanizmus je dokonce schopen realizovat *pavlovské reflexy* (viz např. [25], str. 217), kdy je agent podmíněn produkovat jisté chování v odpovědi na zdánlivě nesouvisející podnět.

Pokud chceme pokročit směrem k realizaci náročnějších kognitivních úkolů, musíme zavést speciální koncepty zvané *afekty*. Tzv. základní afekty tvoří podmnožina konceptů aktivovaných prostřednictvím senzorů. Existují dva druhy afektů: *pozitivní*, odpovídající příjemným či žádoucím pocitům, a *negativní*, odpovídající negativním pocitům. Asociace vycházející z afektů nesou „znaménko“ afektů a excitace přes takové asociace modulují excitaci cílového konceptu — mohou ji posílit anebo utlumí. Pomocí afektů lze simulovat *učení pomocí odměn anebo trestů*, které následují buď bezprostředně, anebo až s jistým zpožděním po vykonání akce. Zdá se, že schopnost pavlovských reflexů a učení pomocí odměny a trestu představují kognitivní minimum, které musí agent splnit, pokud aspiruje na netriviální chování.

Během agentovy interakce se stimulujícím prostředím se koncepty v rámci řídící jednotky začnou samoorganizovat do *klastrů* definovaných na základě podobnosti. Jádra těchto klastrů jsou tvořena koncepty, které jsou největší společnou abstrakcí ostatních konceptů v klastru a stále si zachovávají jistou netriviální míru podobnosti s ostatními koncepty v klastru. Pomocí časové souslednosti se jádra klastrů řetězí do tzv. *zvyků.* Jakmile se zvyky dostatečně upevní, chování agentů se jimi začne řídit. Takové chování se ve většině případů odvíjí bez jakéhokoliv „mentálního“ úsilí. Pouze v situacích, kdy se zvyk

v nějakém kontextu „kříží“ s jiným zvykem, si řídící jednotka vyžádá další informace (např. v tzv. off–line režimu od zrcadlové sítě, anebo v on—line režimu od senzomotorických orgánů — viz v dalším), které upřesní kontext a umožní výběr odpovídajícího chování. Více detailů o práci řídící jednotky a jejích kognitivních schopnostech viz v práci [26], [28] a tam uvedených odkazech na literaturu.

Poslední složkou našeho modelu je *tělo*. Tělo je nositelem senzomotorických orgánů agenta a současně tvoří ochrannou schránku všech jeho dalších součástí.

Nyní se vraťme k otázce vnitřních modelů. Na zrcadlovou síť můžeme zjevně pohlížet jako na syntaktický model světa. V tomto modelu je svět reprezentován tak, jak je poznán prostřednictvím agentových smyslů a jeho motorických akcí. Tento model můžeme nazvat i *senzomotorickým modelem*, protože popisuje „syntaxi“ agentova světa. V zrcadlové síti jsou uloženy ty kombinace exterocepčních a propriocepčních vstupů a motorických instrukcí, které se „hodí k sobě“. Všimněme si, že součástí multimodální informace jsou vždy i propriocepční informace, které se samozřejmě bezprostředně a výlučně týkají agenta samotného. To znamená, že v zrcadlové síti jsou přítomny i prvky modelu samotného agenta.

Na druhé straně, řídící jednotka je specifickým modelem světa, který zahrnuje „sémantiku“ světa. Asociace a koncepty uložené v tomto modelu zřejmě odpovídají relacím a jevům ve skutečném světě tak, jak je agent vnímal pomocí svých senzorů a akcí. To vše vlastně reprezentuje *vnitřní dynamický model světa*. Na tento model lze také nahlížet jako skladiště „vzorců“ chování, která jsou smysluplná v daných situacích.

V další části popíšeme, jak interakce obou částí modelu vede k složitějšímu chování.

3 Vyšší kognitivní funkce

Nejprve podáme nástin mechanizmu imitačního učení, které představuje základ vyšších mentálních schopností (viz např. [1], [11]). Představme si následující situaci: agent *A* pozoruje agenta *B,* který koná jistý dobře rozlišitelný specifický úkol (např. napřahuje ruku po banánu). Pokud *A* má ve svém repertoáru percepčně-behaviorálních jednotek chování situaci, která se dobře shoduje s pozorováním zprostředkovaným senzory agenta *A*, tak jeho zrcadlová síť tuto jednotku identifikuje pomocí asociačního mechanizmu, pouze na základě senzorické informace. Současně, protože v odpovídající percepční informaci chybí propriocepční složka, asociativní mechanizmus ji doplní (na základě podobnosti s minulou agentovou zkušeností)

a navíc přidá i jednobitový příznak znamenajícím „*toto není má vlastní akce*“. Asociativní mechanismus také doplní i chybějící motorickou informaci, tj. „ukotví“ vnímané vjemy do tělesných akcí (ale tyto akce, samozřejmě, nerealizuje). Takto zkompletovaná multimodální informace postoupí do řídící jednotky, kde je adekvátním způsobem zpracovaná. Řídící jednotka agenta *A* má v tuto chvíli k dispozici informaci o tom, v jaké je *B* situaci, a z ní může odvodit (jako by sám *A* byl v situaci, ve které je *B*), co pravděpodobně udělá *B*. Toto „předjímání“ je realizováno jednoduše sledováním asociací, které začínají v současném mentálním stavu agenta *A*. Agent *A* dokonce má k dispozici informaci o „pocitech“ agenta *B* — jednoduše jsou to propriocepční informace (resp. příslušné afekty) agenta *A*, které doplnil asociativní mechanizmus, a samozřejmě také zmíněný příznak, který v tomto kontextu znamená „*toto nejsou mé pocity*“. Toto můžeme považovat za model „*vcítění*“ (empatie) v našem zjednodušeném modelu. Navíc, pokud nadáme naše agenty schopností zapamatovat si krátké nedávně pozorované posloupnosti akcí, tak *A* může zopakovat akce, které pozoruje u agenta *B*. Za strany agenta *A* to ovšem není nic jiného než *imitace* chování agenta *B*.

Stejný mechanizmus pomáhá formaci detailnějšího modelu agenta samotného — tzv. koncept *self*. Funguje to takto. Pozorování aktivit agenta stejného druhu jiným agentem umožňuje pozorovateli „doplňovat si“ mezery v jeho vlastním dynamickém modelu světa, protože již od samého počátku má pozorovatel k dispozici informaci o tom, „*jaké to je, vnímat svou vlastní motoriku*“, a teď ji doplňuje tím, „*jaké to je, když vnímám tutéž motoriku u jiného*“. V této fázi jsme již u *primitivní komunikace* prováděné pomocí *gest*, anebo obecněji, pomocí *řeči těla*. Naznačením nějaké akce agent vysílá vizuální informaci, která je doplněna pozorovatelovou zrcadlovou sítí na úplnou multimodální informaci. To znamená, že pomocí jediného gesta může být sdělena komplexní informace. V tomto případě tedy gesto zastupuje prvek jakéhosi (proto)jazyka vyšší úrovně. Mimochodem, zde mohou do komunikace vstoupit i *výpočetní emoce* jako jedna ze složek komunikace. Jejich účelem je modulovat agentovo chování. Samozřejmě, že za tím účelem musí být agenti vhodným způsobem vybavení (např. specifickou mimikou, možností změny barvy, apod.). Jakmile má agent schopnost artikulace, může doplnit gesta, a později je dokonce nahradit, *artikulovanými zvuky*. To lze chápat jako zrození mluvené řeči. Někdy v této době začíná proces stratifikace abstraktních konceptů od vtělených a agent začíná „rozumět“ gestům (tj. jazyku těla), což jsou vlastně ztělesněné abstraktní koncepty. Je dobré si uvědomit, že agenti rozumějí gestům a řeči prostřednictvím „vcítění

se“ do takové komunikace, v termínech ukotvení obsahu komunikace ve stejné senzomotorice [8][9] a ve složitějších případech, ve stejných zvycích (viz např. [9]). Dále je třeba si uvědomit, že agent, který by alespoň ve fázi poznávání světa a učení porozumění svým vlastním akcím a vjemům neměl tělo, by se nemohl naučit rozumět komunikaci (viz např. [16]).

Pokud se vrátíme k procesu stratifikace abstraktních a vtělených konceptů, tak na tyto dvě třídy konceptů lze nahlížet jako na koncepty na symbolické a sub-symbolické úrovni. Tento náhled dává odpověď na často zmiňovaný problém, jestli mysl pracuje s jednou či druhou třídou konceptů: náš model pracuje s oběma třídami a přechází plynule od jedné ke druhé. Sub-symbolická úroveň vtělených konceptů je nutná pro porozumění abstraktním konceptům. Jazyk je potom vlastně nadstavbou nad vtělenými koncepty. Přechod od gest a případně řeči těla k artikulované řeči neznamená pouze to, že se gesta „naváží“ na příslušné zvuky, ale především na motoriku mluvidel. To dále umožní „samomluvu“ (hovoření sama k sobě) a později umožní přechod k myšlení (viz dále).

Vycházejíc ze struktury a funkčnosti řídící jednotky a zrcadlové sítě, náš model realizuje řešení problému ukotvení symbolů v podobném duchu, jaký byl naznačen (ale nedořešen do takové úrovně, jako v našem modelu) v práci [23]. Výše naznačená realizace vyšších kognitivních funkcí je v dobrém souladu s tzv. teorií intencionálního souznění, která vychází z předpokladu, že sdílené vzorce neurální aktivity a doprovodní vtělené simulace tvoří biologický základ pro porozumění záměrům jiných agentů [7].

Agent schopný komunikace v našem smyslu je jenom krůček od myšlení. V našem modelu je *myšlení realizováno jako mluvení se sebou samým*. Agent mluvící sám se sebou spustí mechanizmus rozlišující mezi externími stimuly („*poslouchám mluvenou řeč*“) a interními („*jsem to já, kdo mluví*“). Zde je základ *uvědomění* v našem modelu. Další malou modifikací (z hlediska inteligentního designéra agenta) můžeme dosáhnout, že agent do samomluvy ani nemusí zapojovat svá mluvidla. V tomto případě příslušné instrukce se nedostanou k tomuto orgánu, pouze jsou přímo usměrněny do zrcadlové sítě (viz obr. 1). Zde vybudí stejnou multimodální informaci, jako v případě, kdyby agent slyšel příslušné artikulované zvuky anebo vnímal řeč svého těla prostřednictvím propriocepce. (Zde využíváme naši poznámku o tom, že motorická část multimodální informace podmiňuje její zbytek). Zřejmě během myšlení agent „odpojí“ jakoukoliv interakci s vnějším světem (tj. percepci a motorické akce). Agenta v *režimu myšlení* znázorňují tmavé šipky na obr. 1 představující cyklus

z kontrolní jednotky do zrcadlové sítě a zpět. Všimněme si, že z hlediska interního mechanizmu pracuje agent v režimu myšlení podobně jako ve standardním režimu učení. Rozdíl je v tom, že v posledně zmíněném případě agent pracuje s reálnou percepcí a vykonává všechny motorické instrukce, kdežto v předchozím případě běží stejné „mentální" procesy. Tyto procesy vycházejí z virtuálních, nikoliv skutečných, dat uložených v zrcadlové síti; příslušné motorické instrukce se nerealizují. Řečeno počítačovou terminologií, v režimu myšlení agent pracuje *off–line*, kdežto ve standardním režimu pracuje *on—line*. Ještě si všimněme, že pokud má agent schopnost odpojit se od reality v režimu myšlení, pak agent rozlišuje mezi myšlením a realitou. To se považuje za základ vědomí [21].

V našem modelu budeme definovat výpočetní vědomí v duchu Minského ideje, že „*vědomí je velký kufr*" obsahující mnoho různých mentálních schopností. Prologem k vědomí je komunikace a myšlení. Následující „definice" výpočetního vědomí předpokládá, že agent je schopen komunikovat v abstraktním vyšším jazyku. Vyšší jazyk je „abstraktní" jazyk používající slovní výraz či gesto k označení relativně složité akce (odpovídající posloupnosti mentálních stavů) anebo abstraktnímu konceptu. Úroveň jazyka je tím vyšší, čím bohatší je jazyk, tj. čím větší a abstraktnější je množina věcí a událostí, o kterých lze v jazyce komunikovat. Budeme říkat, že *agenti mají vědomí*, pokud jejich jazykové schopnosti dosahují takové úrovně, že agenti jsou schopni fabulovat na dané téma. Přesněji, agenti jsou schopni:

- Mluvit, přemýšlet a vysvětlit z hlediska 1. anebo 3. osoby minulé, přítomné anebo očekávané zážitky, pocity, záměry a pozorované jevy;
- Napodobit pozorované akce jiných agentů, verbálně je popsat, a naopak realizovat akce na základě jejich verbálního popisu ve vyšším jazyku;
- Rozšiřovat svůj jazyk vyšší úrovně o nová slova, učit se novému jazyku.

Zdá se tedy, že stav vědomí nelze dosáhnout bez toho, že by agent měl k dispozici vnitřní model světa doplněný znalostmi o tom, jak svět funguje a jak funguje agent sám (a jemu podobní) v tomto světě. To je nemyslitelné bez toho, aby agent měl schopnost učení. Nutným požadavkem pro vznik vědomí je sociální interakce agentů ve vyšším jazyce se stejnou či podobnou sémantikou. Zřejmě vědomí není vlastnost, kterou entita buď má, anebo nemá. Tuto vlastnost může agent mít v různé míře. Např. z předchozích požadavků vyplývá, že vědomý agent by měl být schopen rozeznat sám sebe v zrcadle; ještě „vědo-

mější“ agent by měl být schopen rozumět pohádkám, ale „vědět“, že to jsou smyšlené historky. Ještě více vědomý agent by měl umět lhát (nikoliv např. pomocí mimikrů, ale ve vyšším jazyku!) a být si toho vědom, atd. A jak zjistíme, jestli vědomý agent „rozumí“ tomu, co říká? Zřejmě nejjednodušší bude, zeptat se ho. Schopnost odpovídat na takové otázky je vlastně jádrem naší definice vědomí.

Právě popsaná definice vědomí je vlastně testem, který může aplikovat entita, která si myslí, že má vědomí, na jinou entitu, aby rozhodla, jestli jiná entita má také vědomí. Dle této definice postupujeme i my, lidé. Nicméně, uvědomme si, že v článku jsme dokázali něco více, než že jsme dospěli k definici výpočetního vědomí. Zdůvodnili jsme, že kognitivní agent s navrhovanou architekturou a právě popsanou funkčností jednotlivých jejích modulů v principu splňuje všechny předpoklady pro to, aby byl vědomý. Jestli se vědomí rozvine je pak otázkou agentova „správného ztělesnění“, vhodných technických parametrů jeho orgánů (paměťové kapacity a efektivity, operační rychlosti, vlastností senzomotorických jednotek, atd.), a také, samozřejmě, je to otázka správné výchovy agenta. Tento přístup připomíná poměry v teorii výpočtů: každý správně navržený počítač (Turingův stroj, anebo osobní počítač, řekněme), může být v principu univerzálním počítačem; nicméně, proto, aby fungoval, musí být patřičným způsobem sestrojen a naprogramován. Stejné závěry platí i pro náš model vzhledem k myšlení a vědomí. Podobná myšlenka, totiž, že „*zjistit, jestli nějaký mechanizmus může podporovat rozvoj vědomí v nějakém organizmu, lze pouze analýzou tohoto mechanizmu*“ byla zmíněna např. v práci [1], avšak příslušný mechanismus nebyl naznačen.

Věříme, že náš nástin modelu vědomého kognitivního agenta představuje první krok směrem k teorii kognitivních systémů, o kterých nebude třeba rozhodovat pouze pomocí testů, jestli mají nějaké kognitivní schopnosti, ale na základě zkoumání jejich architektury bude možné rozhodnout, jestli alespoň v principu takové schopnosti mohou mít.

Poděkováni: Tato práce vznikla v rámci výzkumného záměru AV0Z10300504 s částečnou podporou grantu 1ET100300419

Literatura

[1] I. Aleksander, B. Dummall: Axioms and Tests for the Presence of Minimal Consciousness in Agents. *Journal of Consciousness Studies*, Volume 10, No. 4-5, 2003.

[2] M. A. Arbib: The Mirror System Hypothesis: How did protolanguage evolve? In: Maggie Tallerman, editor, *Language Origins: Perspectives on Evolution*. Oxford University Press, 2005.

[3] M. Blum, R. Williams, B. Juba, M. Humphrey: Toward a High-level Definition of Consciousness, Invited Talk to the *Annual IEEE Computational Complexity Conferenc*e, San Jose CA, (June 2005).
[4] R.A. Brooks: Intelligence without reason. *Proceedings of the 12th Intl. Conference on Artificial Intelligence* (IJCAI-91), 1991, pp. 569-595
[5] H. Cruse: The evolution of cognition— a hypothesis. *Cognitive Science* Vol. 27 No. 1, 2003, pp. 135-155.
[6] D. Dennett: Consciousness Explained, The Penguin Press, 1991.
[7] J. Feldman: From Molecule to Metaphor. MIT Press, Cambridge, MA, June 2006.
[8] V. Gallese, M. E. Eagle, P. Migone: Intentional attunement: Mirror neurons and the neural underpinnings of interpersonal relations. J. *of the American Psychoanalytic Association*, 55: 131-176, 2007.
[9] S. Harnad: The Symbol Grounding Problem. *Physica D* 42: 335-346, 1990.
[10] O. Holland (Editor): Journal of Consciousness Studies, Special Issue: *Machine Consciousness*. Volume 10, No. 4-5, April-May 2003.
[11] O. Holland, R. Goodman: Robots With Internal Models: A Route to Machine Consciousness? *Journal of Consciousness Studies* Volume 10, No. 4-5, April-May 2003.
[12] D. Hume: Enquiry Concerning Human Understanding, in *Enquiries concerning Human Understanding and concerning the Principles of Morals,* edited by L. A. Selby-Bigge, 3rd edition revised by P. H. Nidditch, Oxford: Clarendon Press, 1975.
[13] J. R. Hurford: Language beyond our grasp: what mirror neurons can, and cannot, do for language evolution. In: O. Kimbrough, U. Griebel, K. Plunkett (eds.): *The Evolution of Communication systems: A Comparative Approach.* The Vienna Series in Theoretical Biology, MIT Press Cambridge, MA, 2002.
[14] M. Minsky: Consciousness is a big suitcase. EDGE, http://www.edge.org/3rd_culture/minsky/ minsky_p2.html, 1998
[15] M. Novák: Každý sám svým pánembohem. *Světová literatura* r. 10, č. 4, s. 165-194, 1965.
[16] R. Pfeifer, J. Bongard: How the body shapes the way we think: a new view of intelligence. The MIT Press, 2006.
[17] R. Pfeifer, C. Scheier: *Understanding Intelligence.* The MIT Press, Cambridge, Massachusetts, London, England, 1999, 697 s.
[18] V. S Ramachandran: Mirror neurons and imitation as the driving force behind "the great leap forward" in human evolution. *EDGE: The third culture*, viz http://www.edge.org/3rd_culture/ramachandran/ramachandran_p1.html
[19] V. S. Ramachandran: Mirror neurons and the brain in the vat. http://www.edge.org/3rd_culture/ ramachandran06/ramachandran06_index.html
[20] G. Rizzolatti, L. Fadiga, V. Gallese, I. Fogassi: Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, 3:131-141,1996.
[21] A. Smee: Principles of the Human Mind Deduced from Physical Laws, 1849 (N.Y. 1853).
[22] J. R. Searle: Minds, Brains, and Programs. Behavioral and Brain Sciences, Vol 3, No. 3, s. 169-225, 1980.
[23] L. L. Steels, M. Loetzsch, M. S. Spranger: Semiotic Dynamics Solves the Symbol Grounding Problem. Available from *Nature Precedings* http://hdl.nature.com/10101/npre.2007.1234.1, 1997
[24] A. Turing: Computing machinery and intelligence. *Mind,* vol. LIX, no. 236, October 1950, pp. 433-460.
[25] L. G. Valiant: *Circuits of the Mind*. Oxford University Press, New York, Oxford, 1994, 237 p.
[26] J. Wiedermann.: Towards Algorithmic Explanation of Mind Evolution and

Functioning (Invited Talk). In: L. Brim, J. Gruska and J. Zlatuška (Eds.), *Mathematical Foundations of Computer Science, Proc. of the 23-rd International Symposium (MFCS'98)*, Lecture Notes in Computer Science Vol. 1450, Springer Verlag, Berlin, 1998, pp. 152—166.

[27] J. Wiedermann: Mirror Neurons, Embodied Cognitive Agents and Imitation Learning. In: *Computing and Informatics*. Vol. 22, no. 6 (2003), p. 545-559.

[28] J. Wiedermann: Chtěli byste být mozkem v bańce? *Pokroky matematiky, fyziky a astronomie,* 51(4), str. 272-282, 2006.

[29] J. Wiedermann: HUGO: A Cognitive Architecture with an Incorporated World Model. *Proc. of the European Conference on Complex Systems ECCS'06*, Saïd Business School, Oxford University, 2006, viz také Technical report No. 966, Ústav informatiky AV ČR, 2006.