Vzťah medzi symbolizmom a konekcionizmom v kognitívnej vede

Vladimír Kvasnička
Ústav aplikovanej informatiky FIIT STU
Bratislava
email: kvasnicka@fiit.stuba.sk

Abstrakt. V rámci kognitívnej vedy existuje už od čias jej vzniku neutíchajúca diskusia o tom, aký je vzťah medzi symbolickým a konekcionistickým prístupom k štúdiu kognitívnych procesov prebiehajúcich v ľudskej mysle/mozgu. V *symbolickom prístupe* centrálnou ideou je predstava o symboloch, ktoré sú transformované na iné symboly pomocou hierarchicky usporiadaných pravidiel. *Konekcionistický prístup* k štúdiu kognitívnych procesov spočíva na predstave, že tieto procesy sa odohrávajú v neurónovej sieti obsahujúcej množstvo navzájom poprepájaných elementárnych procesorových jednotiek nazývaných neuróny. Dokážeme dve vety, podľa ktorých, každý proces prebiehajúci v neurónovej sieti môže byť ekvivalentne vyjadrený pomocou konečno-stavového stroja transformujúceho vstupné symboly na výstupne symboly, a naopak.
Kľúčové slová: umelá inteligencia - kognitívna veda – myseľ - kognitívna architektúra - konekcionistická reprezentácia - symbolická reprezentácia.

A relation between symbolic and connectionist representations in cognitive science

Vladimír Kvasnička
Institute of Applied Informatics FIIT STU
Bratislava
email: kvasnicka@fiit.stuba.sk

Since the beginning of cognitive science, hot debates focus on interrelationship between symbolic and connectionist approaches to the study of cognitive processes in human mind/brain. In symbolic approaches basic elementary units are symbols that are transformed into other symbols by hierarchically organized rules. A connectionist approach to cognitive processes is based on an idea that they are running in neural network composed of huge number of mutually interconnected elementary processor units – neurons. We will prove two theorems, which show that any process running in a neural network may be equivalently represented by a finite-state machine that transforms input symbols onto output symbols, and vice versa.
Key words: artificial intelligence – cognitive science – mind – cognitive architecture – connectionist representation – symbolic representation.

1. Úvod

Cieľom tohto článku je formulovať základné myšlienky dvoch paradigiem kognitívnej vedy, a to konekcionistickej paradigmy a symbolickej paradigmy, pre vysvetlenie kognitívnej architektúry ľudskej mysle. S*ymbolická paradigma* je založená na predstave, že základnou elementárnou entitou kognitívnych procesov prebiehajúcich v ľudskej mysli sú symboly, ktoré sú transformované na iné symboly pomocou hierarchicky usporiadaných pravidiel. Táto predstava vyústila v reprezentáciu ľudského mozgu ako procesora, ktorý je schopný spracovávať symbolickú informáciu. *Konekcionistická paradigma* interpretácie kognitívnych procesov spočíva v predstave, že tieto procesy sa odohrávajú v neurónovej sieti obsahujúcej množstvo navzájom poprepájaných elementárnych procesorových jednotiek nazývaných neuróny. V prednáške ukážeme, že tieto dva zdanlivo diametrálne odlišné prístupy sú navzájom ekvivalentné. V práci použijeme jednoduchú metaforu neurónových sietí založenú na logických neurónoch, ktorých základná idea bola formulovaná v r. 1943 McCullochom a Pittsom [12] (poznamenajme, že táto publikácia je považovaná za jednu z prvých "zakladateľských" prác nielen umelej inteligencie ale aj kognitívnej vedy)[1]. Neurónové siete (ktoré sa v kognitívnej vede nazývajú „*konekcionizmus*") v súčasnosti patria medzi významnú časť počítačovo orientovanej umelej inteligencie, kde zaujali postavenie univerzálneho matematicko-informatického prístupu k štúdiu a modelovaniu procesov učenia, adaptácie umelých kognitívnych systémov založených na metafore ľudského mozgu. Okrem umelej inteligencie neurónové siete majú nezastupiteľné uplatnenie aj v kognitívnej vede, lingvistike, neurovede, riadení procesov, prírodných a spoločenských vedách, kde sa pomocou nich modelujú nielen procesy učenie a adaptácie, ale aj široké spektrum rôznych problémov klasifikácie objektov a taktiež problémov riadenia zložitých priemyselných systémov. V tejto súvislosti musíme upozorniť, že najväčší a principiálny význam majú neurónové siete v neurovede a v kognitívnej vede, kde patria medzi základné teoretické metódy pre interpretáciu rôznych aktivít nášho mozgu. V týchto dvoch oblastiach vznikli základné konekcionistické teoretické prístupy (neurónové siete) a bola preukázaná ich vhodnosť a efektívnosť

[1] V r. 1940 americký matematik a kybernetik C. E. Shannon obhájil na MIT MSc dizertáciu „*A Symbolic Analysis of Relay and Switching Circuits*", ktorá je dostupná na web adrese ftp://math.chtf.stuba.sk/pub/vlado/thesis_Shannon/. V tejto dizertácii ukázal, že ľubovoľná Boolova funkcia môže byť „fyzicky" simulovaná pomocou vhodného spínacieho obvodu, čím sa dostal neobyčajne blízko k výsledkom McCullocha a Pittsa založených na logických neurónoch.

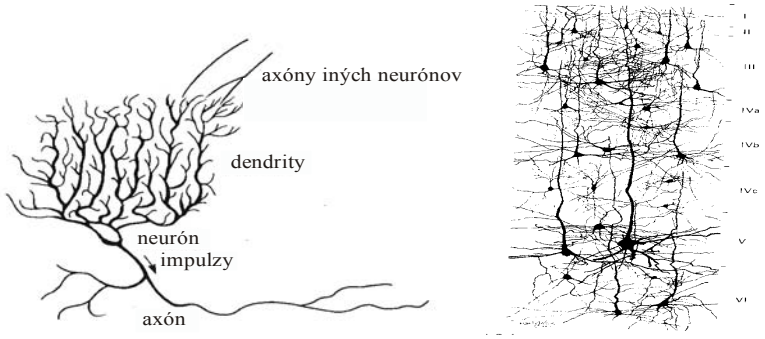
pre štúdium a modelovanie najrozličnejších aktivít a aspektov ľudského mozgu. Konekcionizmus reprezentuje dôležitý pojmový a argumentačný aparát, ktorý umožňuje interpretovať a vysvetľovať kognitívne aktivity ľudského mozgu spôsobom, ktorý je v súlade s našimi predstavami o štruktúre a fyziológii mozgu. O význame a postavení konekcionizmu v systéme kognitívnych a informatických vied píše Ivan M. Havel v jeho článku venovanom filozofickým problémom myslenia [7]. Ďalší teoretický prostriedok, ktorý využijeme v tejto publikácii budú konečnostavové stroje (automaty), ktoré sú schopné transformovať vstupné symboly na požadované výstupné symboly, pričom vlastnosti tohto výpočtového zariadenia sú špecifikované pomocou prechodovej funkcie a výstupnej funkcie. Tieto zariadenia budeme pokladať za typické symbolické (podobne, ako považujeme neurónovú sieť, za zariadenie typické konekcionistické) výpočtové zariadenie, pre ktoré americký matematici a informatici Kleene [8] a Minsky [14] ukázali ekvivalentnosť medzi týmito dvoma zariadeniami. Pomocou tejto teoretickej vlastnosti dokážeme hlavný výsledok tohto článku, a to, že symbolický a konekcionistický prístup sú navzájom ekvivalentné a komplementárne duálne[2].

2. Základné princípy neurónových sietí – inšpirácie z neurobiológie

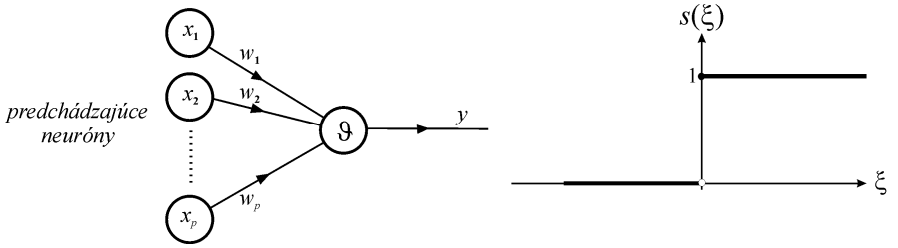
[2] Pojem „komplementárnosti“ bol do vedy zavedený dánskym fyzikom Nielsom Bohrom v 30. rokoch minulého storočia [1] ako integrálna súčasť tzv. Kodanskej interpretácie kvantovej mechaniky. Táto interpretácia sa zriekla snahy získania jednotného, na pozorovateľovi nezávislého, popisu javov prebiehajúcich na atomárnej alebo subatomárnej úrovni. Bohrov princíp komplementárnosti požaduje, aby kvantové javy boli popisované dvojicou parciálnych, vzájomne „komplementárnych“ prístupov (časticového a vlnového), ktoré, aj keď sú súčasne neaplikovateľné, sú potrebné pre plný popis mikrojavov. Bohr sa snažil zovšeobecniť princíp komplementárnosti do všetkých oblastí ľudského poznania, ako nový epistemologický princíp, ktorý je nápomocný k spájaniu protikladných, zdanlivo nekompatibilných pohľadov, na tú istú skutočnosť. Mimo fyziku je najznámejší Bohrov pokus rozšírenia princípu komplementárnosti do biológie, kde existujú dva vzájomne sa vylučujúce pohľady na život. Prvý je *holistický* pohľad, podľa ktorého živý organizmus je neredukovateľný na menšie celky, musí byť študovaný vždy ako celok. Druhý je pohľad *biochemický*, ktorý sa snaží redukovať biologické javy na procesy prebiehajúce na chemickej úrovni. Podľa Bohrovho princípu komplementárnosti, tieto dva pohľady sa navzájom nevylučujú, predstavujú dva navzájom komplementárne pohľady na tú istú skutočnosť. Aplikácia princípu komplementárnosti mimo kvantovú mechaniku môže byť chápaná ako všeobecný epistemologický princíp podporujúci tzv. redukcionizmus v rôznych vedných oblastiach (napr. v chémii, sociológii alebo psychológii) a ktorý je integrálnou súčasťou zjednoteného pohľadu na vedu.

Konekcionizmus v umelej inteligencii a v kognitívnej vede sa chápe ako spôsob paralelného spracovania informácie. Na rozdiel od klasického - *symbolického prístupu*, kde sa sériovo pracuje so symbolmi pomocou hierarchicky usporiadaných logických pravidiel, v konekcionistickom prístupe sa uplatňuje paralelné spracovanie informácie pomocou jednoduchých výpočtov realizovaných neurónmi. V konekcionistickom prístupe je informácia reprezentovaná "z pohľadu" jednotlivých neurónov v sieti jednoduchým sledom impulzov, kde je dôležité, ktoré neuróny v rámci štruktúry siete sú aktívne. Konekcionistické modely sú založené na metafore ľudského mozgu, interpretujú a modelujú kognitívne vlastnosti mozgu pomocou teoretických predstáv, ktoré majú svoj pôvod v neurovede. V konekcionizme sa vychádza zo základného postulátu neurovedy, že základným stavebným kameňom ľudského mozgu je neurón, ktorý má tieto základné vlastnosti [4,10,13,16,17]:

(1) neurón *prijíma signály* z okolia od ostatných neurónov,
(2) neurón *spracováva* (*integruje*) prijaté signály,
(3) neurón *posiela spracované signály iným neurónom zo svojho okolia*.



Obrázok 1. Ľavý obrázok znázorňuje typickú neurónovú bunku, ktorá obsahuje rozsiahly dendritický systém a dlhý vetviaci sa axón. Prostredníctvom dendritického systému do neurónu vstupujú signály z iných neurónov a prostredníctvom axónu z neurónu vystupuje signál charakterizujúci stav neurónu. Pravý obrázok znázorňuje vzájomné prepojenie neurónov pomocou spojov medzi dendritmi a axónmi.

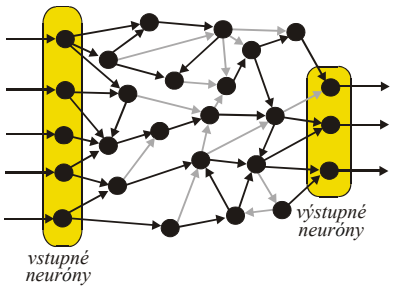


Obrázok 2. Ľavý obrázok znázorňuje prácu neurónu, ktorý obsahuje spoje s p inými neurónmi. Neurón je buď v stave reprezentovanom 0, potom nevysiela signály do okolia pomocou axónu, alebo je v stave 1, potom pomocou axónu vysiela signály do okolia. Ak váhovaná suma vstupných aktivít je väčšia ako prah $-\vartheta$, tak stav neurónu je y=1, v opačnom prípade, ak váhovaná suma je menšia ako prah $-\vartheta$, tak neurón je v stave y=0. Pravý obrázok je graf priebehu krokovej funkcie (1.1b).

Podľa týchto požiadaviek (pozri vlastnosť 4) konekcionistický model obsahuje tzv. *aktivačnú funkciu*, pomocou ktorej sa transformujú vstupné signály na výstupný signál. Jednoduchý model tejto aktivačnej funkcie má tvar krokovej funkcie (pozri obr. 2)

$$y = s(\xi) = s\left(\sum_{i=1}^{p} w_i x_i + \vartheta\right) \tag{1a}$$

$$s(\xi) = \begin{cases} 1 & (\text{if } \xi \geq 0) \\ 0 & (\text{otherwise}) \end{cases} \tag{1b}$$



Obrázok 3. Schematické znázornenie neurónovej siete, kde neuróny reprezentované veľkými bodmi sú pospájané orientovanými spojmi. Čierne šípky sú priradené existujúcim spojom, zatiaľ čo svetlé (sivé) šípky sú priradené virtuálnym spojom, ktoré aktuálne neexistujú, ale potenciálne môžu existovať. Toto rozlišovanie spojov na reálne a virtuálne je umožnené plasticitou neurónovej siete, kde v priebehu učenia siete môžu spoje tak zanikať, ako aj vznikať. Neurónová sieť, ktorá neobsahuje (obsahuje) orientované cykly sa nazýva ***dopredná*** (***rekuretná***) neurónová sieť.

Niekoľko poznámok o význame vyššie uvedených základných princípov konekcionizmu pre umelú inteligenciu a o všeobecných dôsledkoch z nich bezprostredne vyplývajúcich (pozri obr. 3). Neuróny a ich spoje sú *extrémne jednoduché výpočtové zariadenia*, ktoré sú schopné spracovávať resp. prenášať len sekvencie jednoduchých signálov - impulzov. Predstava o tom, že už na úrovni neurónov je spracovávaná symbolická informácia (t.j. štruktúrovaná inak, ako do jednoduchej sekvencie impulzov) je *a priori* chybná. Preto konekcionizmus v umelej inteligencii a v kognitívnej vede stojí pred určitým paradoxom, ako

pomocou jednoduchých "subsymbolických" výpočtových jednotiek je možné vysvetliť a interpretovať vlastnosť ľudského mozgu ako celku, ktorý evidentne je schopný nielen manipulovať so symbolickou informáciou, ale je schopný ju aj ukladať a aj vyberať z pamäti. Vzťah medzi symbolizmom a konekcionizmom v umelej inteligencii a v kognitívnej vede nie je jednoduchý, vo všeobecnosti možno konštatovať, že sa jedná o dve rôzne hierarchické úrovne nazerania na spracovanie informácie v neurónových sieťach. Prvý pohľad je mikroskopický – subsymbolický, ktorý používa pojmy a koncepcie na úrovni neurónov a ich spojov. Druhý alternatívny pohľad je makroskopický – symbolický, ktorý operuje so symbolmi a s ich transformáciou na iné symboly pomocou pravidiel. To znamená, že konekcionizmus nám poskytuje pojmový a argumentačný aparát na interpretáciu symbolického prístupu pomocou takých elementárnych pojmov, akými sú napr. aktivity jednotlivých neurónov, charakter spojov medzi neurónmi (amplifikačný alebo inhibičný) a pod. Podrobnosti o vzťahu medzi subsymbolickým a symbolickým prístupom v umelej inteligencii a v kognitívnej vede možno nájsť v literatúre [2,3,4,5,6,7,9].

Logické neuróny sú schopné simulovať logické spojky, ktoré sú charakterizované ako lineárne separovateľné (napr. disjunkciu, konjunkciu, implikáciu a negáciu). Logické spojky, ktoré nie sú lineárne separovateľné (napr. ekvivalencia a exkluzívna disjunkcia XOR) nemôžu byť simulované logickým neurónom. Táto skutočnosť naznačuje, že logický samotný neurón nie je univerzálne výpočtové zariadenie, existujú úlohy (napr. logická spojka XOR), ktoré nie sú riešiteľné pomocou logického neurónu.

Bolo dokázané už Mc Cullochom a Pittsom, že ***ľubovolná Boolova funkcia je simulovateľná pomocou doprednej neurónovej siete, ktorej vstupné neuróny špecifikujú premenné funkcie a výstupný logický neurón špecifikuje funkčnú hodnotu simulovanej Boolovej funkcie. Táto neurónová sieť už obsahuje tzv. skryté neuróny, ktoré spracovávajú vstupné aktivity (pravdivostné hodnoty premenný Boolovej funkcie) tak, aby boli lineárne separovateľné pre výstupný neurón.*** Táto vlastnosť môže byť zosilnená tak, že neurónová sieť má univerzálny charakter trojvrstvovej neurónovej siete.

Veľkú zásluhu na pochopení a správnej interpretácii práce McCullocha a Pittsa má Minsky, ktorý vo svojej knihe "*Computation: Finite and Infinite Machines*" z r. 1967 [14] dôkladne analyzoval ich výsledky a vykonal niekoľko ďalších zovšeobecnení ich prístupu (najmä na konečno-stavové automaty). Navyše, štýl, akým bola písaná pôvodná práca McCullocha a Pittsa, nemá ďaleko od úplnej nezrozumiteľnosti. Minsky

preformuloval ich výsledky do zrozumiteľnej formy a uviedol ich aj do kontextu vtedajšej počítačovej vedy.

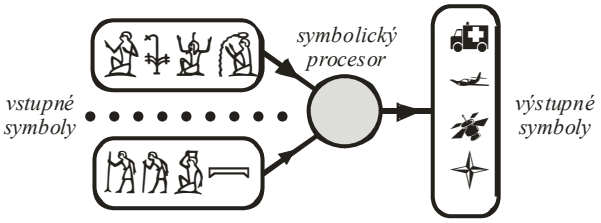
3. Klasická (symbolická) umelá inteligencia a kognitívna veda

V dobe vrcholného rozkvetu európskeho racionalizmu, filozof Leibniz sa pokúšal zostrojiť formálny systém, ktorý by nahradil verbálne metódy usudzovania formálnymi (t. j. hlavne matematickými) manipuláciami s formulami. Postuloval formálny systém s dvoma časťami: (1) jazyk logiky – *lingua characteristica*, pomocou ktorého je možné reprezentovať každý výrok a (2) *calculus ratiocinator* – pomocou ktorého je možné uskutočňovať usudzovanie systematickým a matematicky presným spôsobom. Žiaľ, trvalo ešte ďalších 200 rokov než sa naplnila táto Leibnizova idea, keď v polovici 19. st. anglický matematici A. de Morgan a G. Boole zostrojili „kalkulus“ výrokovej logiky.

Pre Leibnizovho súčasník anglického filozofa T. Hobbesa myslenie už bolo len špeciálny druh výpočtu. Táto hypotéza, ktorá na prelome 16. a 17. storočia znela veľmi neobvykle ba až exoticky, až v súčasnosti bola plne akceptovaná a realizovaná pomocou umelej inteligencie a kognitívnej vedy, kde má postavenie centrálnej paradigmy. Základné problémy klasickej umelej inteligencie sú

- reprezentácia poznatkov,
- procesy usudzovania,
- riešenie problémov,
- komunikácia v prirodzenom jazyku,
- robotika,
- ….

Všetky tieto (a iné) problémy klasickej umelej inteligencie a kognitívnej vedy sú riešenie v rámci ***komputačnej paradigmy***, kde symbolická reprezentácia procesov z vyššie uvedeného zoznamu je transformovaná pomocou symbolického procesoru (pozri obr. 4).



Obrázok 4. Symbolický procesor, ktorý je schopný transformovať vstupné symboly na výstupné symboly.

Symbolická paradigma sa v súčasnosti obvykle považuje za správny prístup pre riešenie vyšších kognitívnych aktivít (komunkácia v prirodzenom jazyku, riešenie problémov, logické usudzovanie,...). Táto paradigma je veľmi efektívnou tým, že vedie k používaní matematického formalizmu, ktorý je ľahko zrozumiteľný a aplikovateľný k riešeniu daných kognitívnych problémov. Tento prístup má aj principiálne ohraničenia, menovite problém učenia je v ňom ťažko implementovateľný, pretože sa jedná obvykle o proces inkrementálneho charakteru, čo je v priamom protiklade so symbolickou reprezentáciou, ktorá má diskrétny charakter. Podobne, vysvetlenie pamäti (menovite rozdiel medzi krátkodobou a dlhodobou) predstavuje pre symbolickú paradigmu neriešiteľný problém.

3.1 Vzťah medzi subsymbolickým a symbolickým prístupom ku štúdiu kognitívnych aktivít ľudského mozgu

Moderný pohľad na vzťah medzi mozgom a mysľou vychádza z neurovednej paradigmy (pozri ref. [10,17]), podľa ktorej, architektúra mozgu je špecifikovaná spojmi medzi neurónmi, ich inhibičným, alebo excitačným charakterom a taktiež aj ich intenzitou. Ľudský mozog vykazuje neobyčajnú plasticitu, v priebehu učenia neustále vznikajú (ale taktiež aj zanikajú) synaptické spoje. *Schopnosť mozgu vykonávať nielen kognitívne aktivity, ale byť aj pamäťou a aj riadiacim centrom pre našu motoriku, je plne zakódovaná jeho architektúrou*. Predstava o ľudskom mozgu ako o počítači, sa musí formulovať tak, že mozog je *paralelný distribuovaný počítač* (obsahujúci mnoho miliárd neurónov, elementárnych procesorov, ktoré sú medzi sebou poprepájané do zložitej neurónovej siete). Program v tomto paralelnom počítači je priamo zabudovaný do architektúry neurónovej siete, t. j. ľudský mozog je jednoúčelový paralelný analógový počítač reprezentovaný neurónovou sieťou, ktorý nie je možné preprogramovať bez zmeny jeho architektúry. V symbolickej paradigme [5] počítač sa obvykle chápe ako výpočtové zariadenie von Neumannovského typu, kde pamäť je striktne separovaná od procesoru transformujúceho vstupné symboly na výstupné symboly. Z pohľadu konekcionistickej paradigmy táto predstava je a-priori nesprávna, mozog chápaný ako neurónová sieť môže byť interpretovaný len ako analógový počítač, kde vykonávaný program je zabudovaný priamo v architektúre neurónovej siete. Z tejto skutočnosti vyplýva aj taká prozaická pravda, že predstavy niektorých autorov sci-fi literatúry o tom, že v ľudskom mozgu sa môže usadiť nejaká cudzia bytosť, ktorá pomocou tohto mozgu vykonáva svoje vyššie kognitívne aktivity je úplne scestná a nereálna.

Z vyššie uvedených všeobecných úvah vyplýva, že myseľ s mozgom tvoria jeden integrálny celok, ktorý je charakterizovaný ***komplementárnym dualizmon***. Myseľ je v tomto prístupe chápaná ako program vykonávaný mozgom, pričom tento program je špecifikovaný architektúrou distribuovanej neurónovej siete reprezentujúcej mozog. Mozog a myseľ tvoria dva rôzne pohľady na ten istý objektu:

(1) Keď hovoríme o mozgu, myslíme tým „hardwarovú“ štruktúru, biologicky realizovanú neurónmi a ich synaptickými spojmi (formálne reprezentovanú neurónovou sieťou), v opačnom prípade,

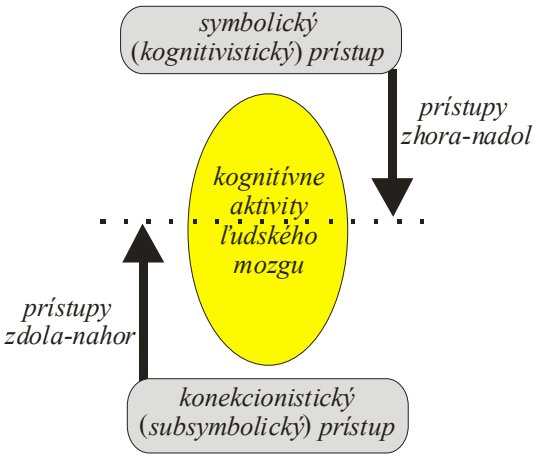
(2) keď hovoríme o mysli, myslíme tým kognitívne a iné podobné aktivity mozgu, ktoré sú vykonávané na symbolickej úrovni, kde prebieha transformácia symbolickej informácie na základe (jednoduchých) pravidiel.

Komplementárny dualizmus mozgu a mysle spôsobuje určité ťažkosti pri interpretácii kognitívnych aktivít mysle. Čisto neurálny prístup k interpretácii kognitívnych aktivít mysle sa upriamuje na hľadanie neurálnych korelátov medzi aktivitami neurónov a kognitívnymi aktivitami (*konekcionizmus* alebo taktiež *subsymbolický prístup*). Použitie neurálnej paradigmy k interpretácii symbolických kognitívnych aktivít má „vedľajší“ efekt v tom, že tieto sa nám „rozpúšťajú“ v ich mikroskopickom popise, symboly sa nám akoby strácajú v detailnom popise aktivít neurónov, intenzít synaptických spojov a pod. V opačnom prípade, absolutizovanie symbolickej paradigmy pri interpretácii kognitívnych aktivít mysle (*kognitivizmus* alebo *symbolizmus*) a ignorovanie skutočnosti, že myseľ je pevne ukotvená v mozgu, vedie k snahe stotožňovať umelú inteligenciu s kognitívnou vedou, t.j. navrhovať alebo preberať priamo z umelej inteligencie rôzne „symbolické“ algoritmy a pomocou nich interpretovať kognitívne procesy ľudskej mysle na fenomenologickej úrovni odvodenej od koncepcie symbolu. Obvykle sa ignoruje požiadavka plauzibility týchto symbolických modelov so súčasnými neurovednými predstavami o ľudskom mozgu.

Prvý pokus o prekonanie konceptuálnych rozdielov medzi kognitivistickým (symbolickým) a konektivistickým (subsymbolickým) pochádza od Smolenskeho [17], ktorý navrhol hierarchický konekcionistický model, pomocou ktorého sa pokúsil zosúladiť kognitivizmus s konekcionizmom. Pri postupe zdola nahor v rámci tohto

modelu nachádzame poznatkové atómy s veľkým stupňom implicitnosti, ktoré sa v určitom priblížení môžu interpretovať už ako symboly.

Realistický názor v kognitívnej vede a v umelej inteligencii na tieto dva prístupy je, že poskytujú dva alternatívne pohľady na ten istý problém, pričom symbolizmus (kognitivizmus) je vhodný na interpretáciu vyšších kognitívnych aktivít ľudského mozgu, zatiaľ čo konekcionizmus (subsymbolizmus) interpretuje nižšie kognitívne aktivity (napr. vnímanie). Názorná formulácia tohto pohľadu je, že symbolizmus sa môže chápať ako prístup "zhora-nadol", ktorý interpretuje vyššie kognitívne aktivity pomocou symbolických prístupov známych z umelej inteligencie, pričom si musíme uvedomovať skutočnosť, že navrhovaný model musí mať aj určitú "konekcionistickú" plauzibilitu, že substrátom ľudského myslenia je mozog, ktorého architektúra je výlučne konekcionistická. Na druhej strane, konekcionistické prístupy k interpretácii kognitívnych aktivít ľudského mozgu sú založené na neurónových sieťach a predstavujú prístup "zdola-nahor", pozri obr. 5



Obrázok 5 Schematické znázornenie vzťahu medzi konekcionistickým a symbolickým prístupom k interpretácii kognitívnych aktivít ľudského mozgu.

Pri aplikovaní konekcionistických metód k interpretácii vyšších kognitívnych aktivít ľudského mozgu je však potrebné zavádzať hypotetické bloky (moduly) vykonávajúce špeciálne aktivity, ktoré už majú veľmi blízko k blokovej štruktúre symbolického prístupu. V ideálnom prípade budeme očakávať, že sa tieto dva prístupy stretnú na polceste (vyznačenej prerušovanou horizontálnou čiarou na obr. 5, tak napr. konekcionistické prístupy budú poskytovať interpretáciu modulov používaných v symbolickom prístupe. Inak povedané, konekcionizmus poskytuje symbolickému prístupu "mikroskopickú" teóriu jeho

fenomenologických pojmov, ktorá je v súlade so súčasnými predstavami o štruktúre a fyziológii ľudského mozgu.

3. Konečnostavové stroje (automaty)

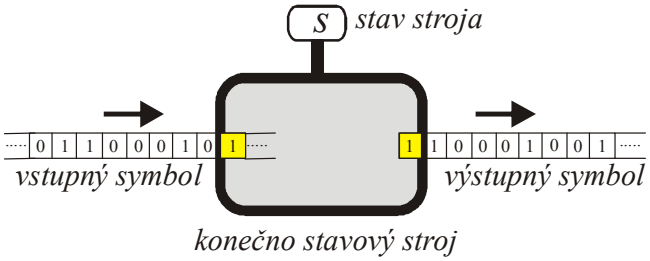
Konečnostavový stroj [8,15,16] pracuje v diskrétnych časových okamžikoch 1, 2,...,t , t+1,... . Obsahuje dve pásky: vstupných symbolov a výstupných symbolov, pričom nový stav je určený pomocou vstupného symbolu a aktuálneho stavu stroja (pozri obr. 6). Stav systému a výstupný symbol sú určené pomocou dvoch funkcií (pozri obr. 7):

(1) ***prechodová funkcia*** f určuje nasledujúci stav stroja na základe aktuálneho stavu a vstupného symbolu

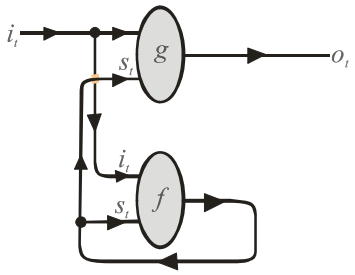
$$stav_{t+1} = f\left(vstupný\ symbol_t, stav_t\right) \tag{2a}$$

(2) ***výstupná funkcia*** g určuje nasledujúci výstupný symbol na základe aktuálneho stavu a vstupného symbolu

$$výstupný\ symbol_{t+1} = g\left(vstupný\ symbol_t, stav_t\right) \tag{2b}$$



Obrázok 6. Konečnostavoý stroj, ktorý transformuje vstupné symboly na výstupné symboly, pričom jeho stav s je určený vstupným symbolom stavom v predchádzajúcom okamžiku. Takto definované zariadenie môžeme chápať ako univerzálne výpočtové zariadenie, ktoré na symbolickej úrovni transformuje vstupné symboly na výstupné symboly.

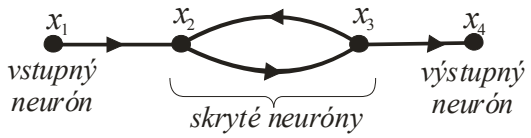


Obrázok 7. Ľavý diagram znázorňuje schému „zapojenia“ konečnostavového stroja, ktorý obsahuje dva oválne bloky reprezentujúce funkcie f a g špecifikované formulami (2a-b). Táto reprezentácie konečnostavového stroja umožňuje pomerne ľahko implementovať toto zariadenie, tak aby vykonávalo požadovanú transformáciu vstupných symbolov na výstupné symboly. Pravý diagram vyjadruje konečnostavový stroj ako jeden blok s vnútorným stavom.

Veta 1. Každá neurónová sieť môže byť reprezentovaná ekvivalentným konečnostavovým strojom.

Táto veta bola dokázaná nezávisle Kleeneom [8] a Minskym [14] (aj keď určité náznaky dôkazu sú už aj v pôvodnej publikácii Mc Cullocha a Pittsa [12], pozri ref. [8]).

Príklad 1. Uvažujeme neurónovú sieť znázornenú na obr. 8. Vytvoríme z tejto siete konečnostavový stroj, tak, že množiny symbolov sú určené pomocou príslušných aktivít neurónov. Pomocou neurónovej siete, znázornenej na obr. 8 zostrojíme tabuľky, ktoré špecifikujú prechodovú funkciu a výstupnú funkciu. Týmto sme vlastne zostrojili konečnostavový stroj, ktorý simuluje danú neurónovú sieť, tento konštruktívny proces môžeme chápať ako konštruktívny dôkaz vety 1.



Obrázok 8. Jednoduchá neurónová sieť obsahujúca 1 vstupný neurón, 2 skryté neuróny a 1 výstupný neurón. Množiny symbolov sú tvorené pomocou aktivít neurónov, $I = \{i = (x_1)\}$, $S = \{s = (x_2, x_3)\}$, $O = \{o = (x_4)\}$.

Obrátená veta k vete 1 sa zaoberá konečnostavovým strojom a možnosťou zostrojiť takú neurónovú sieť, ktorá je ekvivalentná s týmto konečnostavovým strojom.

Veta 2 Každý konečnostavový stroj môže byť reprezentovaný rekurentnou neurónovou sieťou.

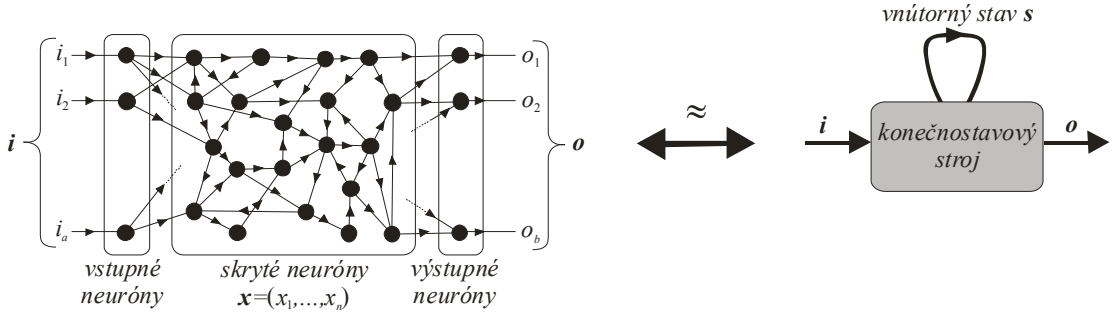
Príklad 2. Nech konečnostavový stroj je špecifikovaný pomocou tabuliek určujúcich prechodovú funkciu a výstupnú funkciu. „Schéma zapojenia“ týchto dvoch funkcií je znázornená na obr. 7, t. j. môžeme potom zostrojiť konečnostavový stroj, ktorý transformuje vstupné symboly na výstupné symboly, pričom tieto transformácie sú určené aj daným aktuálnym stavom stroja. V ďalšej etape navrhneme binárnu reprezentáciu symbolov, každému symbolu použitom v stroji priradíme binárny vektor fixnej dĺžky. Tak napr. ak máme 4 vstupní symboly, $I = \{i_1, i_2, i_3, i_4\}$, potom tieto 4 symboly môžeme reprezentovať pomocou binárnych vektorov {(0,0),

(0,1), (1,0), (1,1)}. Na záver použijeme vlastnosť, že každú Boolovu funkciu môžeme vyjadriť pomocou neurónovej siete, ktorá obsahuje logické neuróny Mc Cullocha a Pittsa. Použitím všeobecnej vlastnosti neurónových sietí s logickými neurónmi (pozri zvýraznený text v 2. kapitole tohto príspevku), ľubovolná Boolova funkcia môže byť vyjadrená pomocou neurónovej siete s logickými neurónmi, potom aj bloky funkcií *f* a *g* znázornené na obr. 7 môžeme vyjadriť pomocou neurónových sietí, použitím „schémy zapojenia“ z tohto obrázku, zostrojíme rekurentnú neurónovú sieť, ktorá sa správa podobným spôsobom ako daný konečnostavový stroj, čím sme vlastne dokázali vetu 2 konšteruktívnym spôsobom.

Vety 1 a 2 umožňujú študovať klasický problém konekcionizmu (neurónových sietí), vzájomný vzťah medzi ***konekcionistickou reprezentáciou*** (neurónovými sieťami) a ***symbolickou reprezentáciou*** (ktorá sa využíva v klasickej umelej inteligencii) (pozri obr. 9):

(1) Podľa vety 1, každá neurónová sieť môže byť transformovaná na ekvivalentný symbolický konečnostavový stroj, kde symboly sú vytvárané pomocou binárnych vektorov aktivit neurónov.

(2) Podľa vety 2, každý symbolický konečnostavový stroj môže byť pretransformovaný na ekvivalentnú neurónovú sieť. Symboly z konečnostavového stroja sú reprezentované pomocou binárnych vektorov.



Obrázok 9. Ekvivalentnosť medzi konekcionistickou neurónovou sieťou s logickými neurónmi a symbolickým konečnostavovým strojom. Neurónová sieť z ľavej strany znázorňuje neurónovú sieť s logickými neurónmi, podľa vety 1 a 2 táto sieť je ekvivalentná so symbolickým konečnostavoým strojom, ktorý pre daný stav ***s*** transformuje vstupný symbol ***i*** na výstupný symbol ***o***.

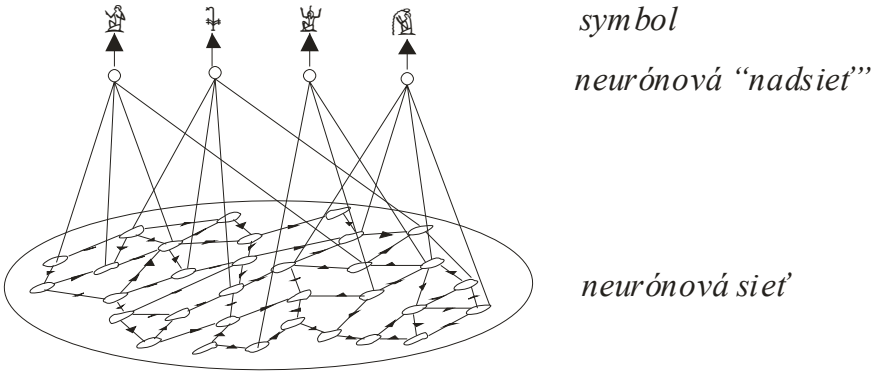
4. Záverečné poznámky

Na záver zosumarizujeme hlavné výsledkov tejto práce, ktoré sú venované diskusii vzťahu medzi konekcionistickou a symbolickou paradigmou pri

ich použití na interpretáciu kognitívnych aktivít ľudského mozgu, t. j. vo všeobecnej rovine, vzťahu medzi mozgom a mysľou. Ak akceptujeme súčasné neurovedné predstavy a skutočnosti o ľudskom mozgu, potom musíme taktiež akceptovať hlavnú ideu konekcionizmu, podľa ktorej „hardvérový substrát" na ktorom prebiehajú kognitívne aktivity ľudského mozgu sú neurónové siete, ktoré vo všeobecnosti môžeme chápať ako vysoko prepojené jednoduché elementárne procesory – neuróny, pričom spoje môžu jednosmerne prenášať len jednoduchú neštruktúrovanú informáciu v podobe elektrochemických impulzov. V tomto konekcionistickom pohľade nemôžeme od seba separovať hardverovú a softverovú charakteristiku študovaného systému mozgu, ktoré sú neoddeliteľné a sú charakterizované komplementárnym dualizmom. Z tejto skutočnosti vyplýva, že aj keď študujeme vlastnosti ľudskej mysle na fenomenologickej symbolickej úrovni, nemôžeme plne odhliadnuť od skutočnosti, aby navrhované algoritmy boli neurálne plauzibilné. V kognitívnej vede práve táto požiadavka plauzibilnosti navrhovaných modelov spolu s ich experimentálnou verifikáciou kognitívnou psychológiou je primárnym momentom ich akceptovateľnosti. Touto požiadavkou môžeme ***operatívne odlíšiť kognitívnu vedu od umelej inteligencie***. Tieto dve príbuzne disciplíny, aj keď majú rovnaký predmet výskumu – algoritmické modelovanie „inteligentných" aktivít ľudského mozgu – môžeme operatívne od seba odlíšiť dôležitou podmienkou, aby modely kognitívnej vedy boli okrem iného aj neurálne plauzibilné. Podmienka neurálnej plauzibility pre umelú inteligenciu nie je dôležitá, hlavným cieľom metód umelej inteligencie je modelovať „inteligentné aktivity", pričom sa ignoruje nielen ich neurálna plauzibilita ale (žiaľ) aj ich výpočtová zložitosť.

Niekoľkokrát sa v tejto práci spomenulo, že konekcionistické a symbolické prístupy sú vzájomne komplementárne duálne (pozri poznámku pod čiarou označená indexom [2]). To znamená, že ak označíme „stupeň konekcionizmu" číslom $0 \le \alpha \le 1$, potom „stupeň symbolizmu" je špecifikovaný číslom $\beta = 1 - \alpha$, t. j. komplementárnou hodnotou stupňa α. Obrazne môžeme povedať, že zvyšovanie stupňa symbolizmu napr. pri štúdiu takých kognitívnych aktivít, akými sú napr. riešenie zložitých problémov, vedie k znižovaniu stupňa konekcionizmu, a naopak. V pravom symbolickom prístupe (kde $\beta \to 1$), symboly majú prevažne fenomenologický *ad-hoc* charakter, neskúmame ich podstatu a vznik, ale len relácie medzi nimi, ktoré sú taktiež špecifikované na fenomenologickej *ad-hoc* úrovni realizovanej pomocou systému pravidiel. Určité formálne problémy začínajú vznikať vtedy, ak si kladieme na tejto „vysokej

symbolickej úrovni“, aká je podstata používaných symbolov, kde sa berú v našej mysli, aké je ich ukotvenie na substráte poprepájaných neurónov a ich aktivít (akceptujeme základné dogma súčasnej neurovedy, že procesy prebiehajúce v ľudskom mozgu sú realizované pomocou neurónových sietí). Prvotný pohľad na riešenie tohto problému nám už poskytuje použitý formálny prístup k štúdiu vzťahu medzi neurónovou sieťou s logickými neurónmi a konečnostavovými strojmi. Bolo ukázané, že neurónová sieť je ekvivalentná konečnostavovému stroju, pričom jeho vstupné, výstupné a vnútornostavové symboly sú špecifikované pomocou binárnych vektorov aktivít príslušných neurónov (t. j. symboly sú interpretované ako obrazce aktivít neurónov). Smolensky [17] navrhuje pri interpretácii symbolov použiť prístup hierarchicky usporiadaných neurónov, pričom aktivity neurónov na najvyššej úrovni špecifikujú daný symbol (pozri obr. 10). Táto idea ukotvenia symbolu prostredníctvom aktivít vybraných neurónov z najvyššej vrstvy siete dobre koreluje so známym poznatkom, že symbol buď poznáme alebo nepoznáme, neexistuje „medzistav“ v ktorom by sme daný symbol poznali len čiastočne. Táto skutočnosť vyplýva z predstavy, že symbol je ukotvený prostredníctvom aktivít niekoľkých neurónov. Ak z týchto neurónov zanikne čo len jeden neurónov, silne sa poruší obrazec neurónových aktivít, čo vedie k „zániku“ daného symbolu. Obvykle sa v neurovede postuluje, že dochádza k postupnému zániku našich mentálnych aktivít (angl. *graceful degradation of mental activites*), avšak tento fenomén je hlavne používaný pre nižšie kognitívne aktivity, ktoré sú veľmi robustné vzhľadom k zániku jednotlivých neurónov alebo spojov medzi nimi. Každý pozná z vlastnej skúsenosti, že toto neplatí pre vyššie kognitívne aktivity, ktoré sú skoro výlučné založené na symbolickej reprezentácii, buď si spomíname alebo si nespomíname na daný symbol (meno, obraz, a pod.).



Obrázok 10. Smolenského idea hierarchického usporiadania neurónovej siete, pričom horná „nadsieť“ prostredníctvom aktivít svojich neurónov reprezentuje daný symbol.

Literatúra

[1] Bohr, N.: *Atomic Theory and the Description of Nature*. Cambridge University Press, Cambridge, 1934.

[2] Farkaš I.: Hľadanie kauzálnych vzťahov v probléme mysle a tela z pohľadu reredukcionistického fyzikalizmu. In: *Myseľ, inteligencia a život*. Vydavateľstvo STU, Bratislava, 2007, pp. 3-16.

[3] Farkaš, I.: Konceptuálne východiská pre model stelesnenej mysle. In Kvasnička, V., Kelemen, J., Pospíchal, J.: *Modely mysle*. Europa, Bratislava, 2008, pp. 35-64.

[4] Fodor, J.A. and Pylyshyn, Z.W.: Connectionism and cognitive architecture: A critical analysis. *Cognition*, **28** (1988) 3-71.

[5] Gärdenfors, P.: Symbolic, Conceptual and Subconceptual Representations. In Cantoni. V. et al.(editors): *Human and Machine Perception*, Plenum Press, New York, 1997, pp. 255—270.

[6] Harnad, S.: The Symbol Grounding Problem. *Physica* **D42** (1990), 335-346.

[7] Havel I. M..: Přirozené a umělé myšlení jako filozofický problém. In Mařík V. a kol.: *Umělá inteligence* (3. díl). Academia, Praha, 2001, pp. 17-75.

[8] Kleene, S. C.: Representation of events in nerve nets and finite automata. In C. E. Shannon and J. McCarthy, editors, *Automata Studies*. Princeton University Press, Princeton, 1956, pp. 3-41.

[9] Kvasnička, V., Pospíchal, J.: Deductive rules in holographic reduced representation. Neurocomputing **69**(2006), 2127-2139.

[10] Kvasnička, V., Beňušková, Ľ., Farkaš, I., Kráľ, A., Pospíchal, J. a Tiňo, P.: *Úvod do teórie neurónových sietí*. IRIS, Bratislava, 1997.

[11] Kvasnička, V., Pospíchal, J.: *Matematická logika*, Vydavateľstvo STU, Bratislava, 2006.

[12] McCulloch, W. S., Pitts, W. H.: A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, **5** (1943), 115-133.

[13] Minsky, M. and Papert, S.: *Perceptrons. An Introduction to Computational Geometry*. MIT Press, Cambridge, MA, 1969.

[14] Minsky, M. L.: *Computation. Finite and Infinite Machines*. Prentice-Hall, Englewood Cliffs, NJ, 1967.

[15] Molnár Ľ., Češka, M., Melichar, B.: *Gramatiky a jazyky*. Bratislava, Alfa, 1987.

[16] Rumelhart, D. E. and McClelland, J. L.: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 1-2. MIT Press, Cambridge, MA, 1986.

[17] Smolensky, P.: On the proper treatment of connectionism. *The Behavioral and Brain Sciences,* **11** (1988), 1-74.