ÚVOD DO FILOZOFIE
UMĚLÉ INTELIGENCE

Michal Pěchouček

Talk draft

1. Úvod – definice Umělé Inteligence
2. Tři základní filosofické proudy Umělé Inteligence
3. Slabá a Silná verze Umělé Inteligence
4. Argument Čínského Pokoje

ÚVOD

Umělá inteligence je empirická věda, která se zabývá zkoumáním a chápáním inteligentních projevů. Nástrojem bádání je abstrakce a modelování inteligentních projevů mimo medium lidské mysli, zpravidla pomocí počítače. Inteligentními projevy podle (Feigenbauma) rozumíme např.: učení, řešení problémů, porozumění jazyku, uvažování. Marvin Minsky, jehož definice je považována za tu nejobecnější a nejuznávanější, definuje umělou inteligenci jako vědu, která se zabývá tím jak přinutit stroje, aby exhibovaly chování takové, které by v případě člověka vykazovalo potřebu inteligence.

Umělá inteligence je jako součást poznávacích věd (cog-sci) chápána většinou jako hraniční věda, která do sebe zahrnuje aspekty: kognitivní psychologie, neurologie, filosofie ducha, ethologie, logiky, evolučních věd, sociologie a jiných. Umělá inteligence jako název není zdaleka ideálním pojmem ale zastřešuje diserfikované pojmenování jako je strojová inteligence, výpočetní psychologie nebo automatizované uvažování.

Účelem této přednášky je navodit vědomí umělé inteligence jako vědní disciplíny, představit její základní proudy a iniciovat diskusi na téma relevantnosti daného rozdělení v současné době. Každý z nás by se měl zamyslet a pokusit se zjistit, do jaké míry se jeho odborné, vědecké a akademické počínání identifikuje se základními myšlenkami zmíněnými během přednášky.

TŘI ZÁKLADNÍ PROUDY UMĚLÉ INTELIGENCE

SYMBOLICKÝ FUNKCIONALISMUS

Symbolický funkcionalismus, proud zvaný též jako ‘stará-dobrá-umělá-inteligence’, je založen na dvou základních hypotézách – funkcionalistické hypotéze a hypotéze fyzikálního systému symbolů. Funkcionální hypotéza tvrdí, že:

> *„Inteligentního chování daného systému je dosaženo interakcí mezi jednotlivými komponentami, které disponují odlišnou funkcionalitou, čehož je dosaženo tím, že v rámci systému hrají odlišnou roli",*

Hypotéza fyzikálního systému symbolů tvrdí, že:

> *„Fyzikální systém symbolů je dostatečným a nezbytným prostředkem pro prezentaci inteligentního chování."*

Fyzikální systém symbolů je obecný stroj, který zajišťuje evoluci populace struktur symbolů v čase. Základními stavebními kameny fyzikálního systému symbolů jsou vznik, zánik, modifikace a reprodukce struktury. Turingův stroj je příkladem fyzikálního systému symbolů, kde páska je chápána jako jediná struktura symbolů, na kterou je pomocí programu a hlavy aplikován operátor modifikace. Druhý výklad je chápán v tom smyslu, že každý symbol je jedna struktura symbolů, na kterou je aplikován postupně operátor vzniku a destrukce.

Zjednodušeně řečeno, základními předměty výzkumu symbolického funkcionalismu jako i celé 'staré-dobré-umělé-inteligence' je tedy problém reprezentace znalostí a inteligentního prohledávání stavového prostoru.

KONEKCIONISMUS

Konekcionismus jako směr bádání v rámci umělé inteligence předpokládá, že esence inteligence plyne ze statického propojení velkého počtu jednoduchých výpočetních jednotek. Myšlenka je inspirována mozkem jako médiem, které zosobňuje inteligentní uvažování. Základní výpočetní jednotkou rozumíme model neuronu, který na základě hodnoty součtu vážených vstupů excituje do aktivního stavu nebo nikoli. Mezi hlavní odvětví, které se zařadí ke konekcionismu, patří neuronové sítě.

Vedou se diskuse, není-li konekcionismus ve své podstatě založen stejně tak jako symbolický funkcionalismus na hypotéze fyzikálního systému symbolů. Já se domnívám, že tato myšlenka je ze své podstaty lichá, protože jako neuronová síť je statická a nevyvíjí se během procesu vykazování inteligentního chování. S hypotézou fyzikálního systému symbolů sdílí pouze to, že inteligence je založena na interakci mezi jednotlivými komponentami systému. Ovšem za prvé tyto komponenty mají v rámci konekcionismu stejnou funkcionalitu a za druhé se nevyvíjejí v čase během vykazování inteligentního chování.

Neuronová síť je chápána poslední dobou jako černá skříňka, která řeší vše úplně stejně jako mozek (nadneseně řečeno). Podle mého soudu se vytrácí základní poslání umělé inteligence jako chápání kognitivních a inferenčních procesů a celá činnost modelování je transformována na výběr vhodné struktury neuronové sítě a vhodného nastavení počátečních podmínek.

Z tohoto důvodu se mylně řadí mezi konekcionismus i problematika evolučních výpočetních strategií jako jsou genetické algoritmy a genetické programování, které podstatně spíše vyhovují hypotéze systému fyzikálních symbolů. Otázka je ta, zda jsme oprávněni klasifikovat přenos informace zakódované v DNA jako projev inteligentního chování.

ROBOTICKÝ FUNKCIONALISMUS

Klíčová filosofie robotického funkcionalismu je, vulgárně řečeno, založena na implementaci behaviorismu jako psychologické školy. Raději než se zabývat reprezentací inteligence, robotičtí funkcionalisté se koncentrují na funkcionalitu modelovaného systému. Jako inteligentní chování je zde chápána jako rozumná interakce mezi třemi entitami: *systém, prostředí, úloha.* V případě že bude agent na danou úlohu a v daném prostředí reagovat inteligentně (jako by reagoval člověk), je považován za agenta disponujícího schopností vyvozovat inteligentní chování.

K tomuto směru jako k hlavnímu rivalovi symbolického funkcionalismu se ještě vrátíme během přednášky.

TŘI DRUHY POROZUMĚNÍ

Stupeň implementace inteligence pomocí umělého media závisí v principu na tom, do jaké míry se podaří namodelovat klíčový faktor lidského uvažování – chápání. Chápání, stejně tak jako spousta jiných kvalit, lze charakterizovat ve smyslu *silné* nebo *slabé* míry. Řekneme-li o myšlence, systému nebo tvrzení, že je *slabé*, myslíme tím, že je obecné. Na druhé straně *silná* myšlenka je ve své podstatě specifická. Systém GPS, který měl na principu matematické logiky a několika obecných inferenčních pravidel řešit obecné problémy, byl klasifikován jako *slabý*, na druhé straně systém MYCIN je *silný* v rámci naší klasifikace.

S ohledem na to, jaké porozumění (*slabé-silné*) se nám podaří modelovat, dělíme umělou inteligenci následujícím způsobem:

- *Slabé* umělé inteligence dosáhneme, namodelujeme-li *slabé* porozumění. *Slabé* porozumění (Turingovo) chápeme jako porozumění takové, že systém na správné vstupní podněty vykáže korespondující reakce. Turingův test dokazuje, zdali jsme dosáhli implementace *slabého* porozumění.

- *Silné* umělé inteligence dosáhneme, namodelujeme-li *silné* porozumění. *Silné* porozumění (Brentanovo) chápeme jako porozumění takové, že systém bude disponovat pocitem chápání takovým, jakým disponuje lidská mysl.

Vědci věřící v *silnou* umělou inteligenci chtějí v principu navrhnout nového člověka ze silikonu. Naopak *slabá* umělá inteligence pomáhá formalizovat jisté oblasti lidského uvažování a navrhuje formální systémy pro následné využití v místech, kde lidské uvažování není dostupné.

Smith se domnívá, že existuje střední cesta v umělé inteligenci.

- *Střední* umělé inteligence dosáhneme, namodelujeme-li *střední* porozumění. *Střední* porozumění (Smithovo) chápeme jako porozumění takové, že systém na správné vstupní podněty vykáže korespondující reakce pomocí správné reprezentace znalostí a aproprietního odvozovacího mechanismu.

Zatímco robotický funkcionalismus stačí na modelování *slabé* umělé inteligence, *střední* umělá inteligence je spíše doménou symbolického funkcionalismu.

ARGUMENT ČÍNSKÉHO POKOJE

Existuje spousta argumentů pro a proti možnosti sestrojit umělý mozek, umělou mysl. Jedním z nejznámějších a vlastně klíčovým argumentem, který ve svém důsledku zavádí dělení chápání, jak zde bylo již řečeno na *silné* a *slabé*, je Argument Čínského Pokoje. Původním záměrem argumentu čínského pokoje bylo vyvrátit možnost sestrojitelnosti umělé mysli. Rigorózněji řečeno, Searle dokázal, že splnění Turingova testu ještě zdaleka nic nevypovídá o tom, zda se systém chová inteligentně vzhledem k *silné* definici umělé inteligence, která, byla-li by splněna, by dokázala možnost implementace umělé mysli.

Udělejme myšlenkový pokus:
Searl, jenž nerozumí ani slovo čínsky, se usadí v uzavřené místnosti plné knih a návodů, jak reagovat na jakoukoliv otázku v čínštině. Dejme tomu, že v libovolném okamžiku, když dostane vzkaz napsaný čínsky, dokáže pomocí knih a návodů zareagovat v čínštině. Není problém si představit konverzaci s Číňanem stojícím před pokojem a strkajícím si papírky na relativně velmi omezené téma. Toto téma lze samozřejmě indefinitivně zobecňovat, až dojdeme k původnímu požadavku.

Číňan, stojící před komnatou, nutně dojde k závěru, že tam musí být někdo (něco), kdo rozumí čínštině. Zjevně to není ani Searle, ani knihy, ani kameny, z nichž je místnost vybudovaná. Proto, ačkoliv místnost splňuje Turingův test, není zde ani stopy po chápaní, porozumění, ani vědomí.

Searle tímto dokázal, že ačkoliv stroje můžou dokázat překvapivě mnoho, nemohou myslet, nemohou disponovat vědomím a proto ani intencionalitou.

Otázky:

1. Věříte v umělou inteligenci (*silnou* / *slabou*) ?
2. Přesvědčil Vás Searle ?
3. Kam by jste zařadili Vaši vědeckou působnost ?
4. Zdá se Vám rozdělení na tři -ismy relevantní ?