Filosofické problémy umělé inteligence

Úvod

Na problematiku umělé inteligence je možné nahlížet z pozice minimálně dvou důležitých motivačních proudů:

- motivace aplikační, inženýrská (v současnosti převládající)
- motivace badatelská – snaha o poznání povahy mentálních procesů, zejména lidských

Samo zakotvení významu „umělé inteligence“ je značně obtížné, a to nejen z důvodu, že neumíme bezproblémově definovat samotný pojem „inteligence“, ba dokonce při bližším zkoumání zjišťujeme, že nejsme ani schopni jasně stanovit, co je přirozené a co umělé. Na jedné straně umělá inteligence jako obor často využívá exaktních algoritmických a matematických postupů, na straně druhé je předmět jejího zkoumání vymezen v mnohém intuitivně.

Z pohledu aplikačního zřetelně vychází Minského definice umělé inteligence: „Umělá inteligence je věda o vytváření strojů nebo systémů, které budou při řešení určitého úkolu užívat takového postupu, který – kdyby ho dělal člověk – bychom považovali za projev jeho inteligence.“

Stejné východisko zastává i definice Richové: „Umělá inteligence se zabývá tím, jak počítačově řešit úlohy, které dnes zatím zvládají lidé lépe.“

Motivačně více univerzální je Kotkova definice: „Umělá inteligence je vlastnost člověkem uměle vytvořených systémů vyznačujících se schopností rozpoznávat předměty, jevy a situace, analyzovat vztahy mezi nimi, a tak vytvářet vnitřní modely světa, ve kterých tyto systémy existují, a na tomto základě pak přijímat účelná rozhodnutí za pomoci schopností předvídat důsledky těchto rozhodnutí, a objevovat nové zákonitosti mezi různými modely nebo jejich skupinami.“ Problém zde spočívá zejména v „…člověkem uměle vytvořených …“ (co když mohou být výše zmíněné systémy vytvářeny jinými umělými systémy – stroji; co když systémy umělé inteligence vznikají „spontánně“ následkem evolučních principů) a „… vytvářet vnitřní modely světa …“ (jak bude ukázáno dále, podle nových přístupů k umělé inteligenci nemusí systém vytvářet vnitřní modely světa).

Alternativní formulace základního problému umělé inteligence může být založena na sémiotickém přístupu. Vychází z triadického pojetí symbolu podle C. S. Peirce: symbol je vykládán jako trojúhelník s vrcholy 1) *znak*, neboli *representamen*; 2) *interpretant*; 3) *bezprostřední objekt*, resp. idea, kterou znak vyvolal v interpretantovi. Na svět lze tak nahlížet jako na nepřetržitý výkladový proces (sémiosis), kde o světě nelze tvrdit nic určitého vyjma výše popsané triády.

V současné době lze klasický vztah mezi člověkem a počítačem (resp. strojem obecně) popsat tak, že v roli representamen mohou být fyzické stopy v hardwaru počítače (např. světelné body na stínítku monitoru), interpretant je uživatel počítače a objekt je například číslo na monitoru, které si interpretant nějakým způsobem vykládá. Snahou umělé inteligence je, aby v roli interpretanta stál sám počítač.

Velmi podstatný je však podíl umělé inteligence v takzvané *kognitivní vědě* a s tím spojená snaha o pochopení a případné modelování fenomenální složky mentálních procesů. Tato problematika je velmi bohatým zdrojem filosofických otázek v oblasti umělé inteligence, mezi něž patří například velmi aktuální tázání „zda a případně jak může mít stroj vědomí“, „jak se liší stroj a člověk, resp. neživý objekt a živá bytost“, „jaké jsou důsledky vzniku systémů umělé inteligence pro člověka a jeho svět“, a podobně.

Filosofické otázky umělé inteligence lze orientačně rozdělit do následujících skupin: ontologické, epistemologické, metodologické, futurologické, etické.

Paradigmata umělé inteligence

„Tradiční“ umělá inteligence

Shodneme-li se na intuitivním chápání pojmu „přirozený“ (ve smyslu způsobu bytí věci ve světě), pak můžeme o nějaké věci (vlastnosti, činnosti) tvrdit, že je *umělá*, pakliže: a) existuje *přirozená* věc, logicky připouštějící duplikaci; b) existuje *záměr* vytvořit duplikát oné věci; c) došlo k *provedení záměru*.

Následně můžeme říci, že umělá inteligence (jako obor) se pokouší o popis a realizaci *umělého myšlení*. Budeme-li předpokládat, že myšlení (lidská mysl) sestává z *mentálních procesů*, pak cílem umělé inteligence je *modelování* těchto procesů. Z teoretického hlediska můžeme rozeznávat dvě komponenty mentálních procesů: *performační* a *fenomenální*. Performační komponenta je ta část procesu, která je objektivně („zvnějšku“) popsatelná, zatímco fenomenální komponenta je to, co subjekt „vnitřně“ prožívá. Některé přístupy (zejména funkcionalistické a behavioristické) popisující mysl redukují mentální procesy toliko na jejich performační složku. Přesněji lze tedy tvrdit, že „tradiční“

umělá inteligence se zabývá modelováním performační složky mentálních procesů.

Modely v umělé inteligenci mívají podobu tzv. *metaforických* modelů, u nichž (na rozdíl od *věrných modelů*) i odlišnost od modelované skutečnosti pomáhá něco vysvětlit. Přístupy k modelování v umělé inteligenci lze v zásadě rozlišit podle toho, kterou úroveň „reality“ chceme modelovat pokud možno věrně: přístup „shora“ a přístup „zdola“. V tuto chvíli chápejme „nad“ a „pod“ víceméně intuitivně – např. úroveň buněk je „nad“ úrovní molekul, ta je sama nad úrovní atomů, úroveň mentálních procesů je nad úrovní neuronů, úroveň neuronů je nad úrovní dynamiky přenosu signálu na synapsích, apod. *Základní úroveň analogie* je u daného modelu taková zvolená úroveň popisu, „pod“ níž již není důležité, zda si prvky zachovávají podobnost s modelovanou skutečností.

Přístup „shora“ klade v umělé inteligenci základní úroveň analogie na rovinu *logicko-symbolických* reprezentací a jejich sériového zpracování, přičemž jej můžeme nazvat *tradičním paradigmatem* umělé inteligence. Právě k tomuto paradigmatu se vážou výše citované definice umělé inteligence a obrovské množství velmi dobře rozpracovaných metod, postupů a algoritmů pro řešení rozličných specifických problémů, úloh a performačních složek některých typů mentálních procesů (řešení úloh prohledáváním stavových prostorů, hraní her, dokazování a odvozování logických vět a hypotéz, produkční systémy, sémantické sítě, rámce, strukturální metody rozpoznávání, plánování pomocí pravidel, atd.).

V souvislosti s tradičním paradigmatem se objevuje i názor *počítačového funkcionalismu*, neboli *„silné“ umělé inteligence*, který je možno vyjádřit tezí: „Povaha mysli je algoritmická, přičemž není podstatné, v jakém médiu jsou algoritmy (programy) uloženy.“ Podle této teze přítomnost myšlení nezáleží na fyzickém médiu (hardwaru), ale na vhodném algoritmu (softwaru) – mysl tedy může mít jak člověk, tak i počítač, dokonce i „Čínská komora“ (viz dále). Naproti tomu tzv. *„slabá“ umělá inteligence* uvažuje pouze o *modelování* lidské mysli (případně jejích částí) na počítači, nikoli o její *replikaci*.

Emergentismus

Upustíme-li od požadavku, že *umělá* vlastnost musí být před svým vytvořením popsatelná (konstrukčním) projektem, můžeme se pokusit vytvořit ji nepřímo, a to připravením nebo sestrojením (umělého) aktivního média, v němž ona vlastnost vznikne „samovolně“. Tento přístup odpovídá postupu „zdola“, kde je základní úroveň analogie položena na nižší, mnohdy fyzikálně popsatelné složky reality. Samozřejmě se zde zásadně mění výsledný vztah konstruktéra k jeho výtvoru.

Je-li nějaký systém konstrukčně vytvořen na určité „nízké“ základní úrovni analogie, pak při vhodném uspořádání a množství prvků této úrovně mohou na některé z „vyšších“ úrovní *vyvstávat* (neboli *emergovat*) jevy, které nejen že nemusely být na základě analýzy fungování jednotlivých prvků „nízké“ úrovně očekávány, ale ani pro ně neexistuje v rámci „nízké“ úrovně popis. Tak například jednotlivé buňky květiny nemají ani ponětí o tvaru květu či vůni, přesto z jejich komplexní organizace může vyvstat růže.

Právě zde nabývá velkého významu otázka, zda fenomenální komponenty mentálních procesů mohou být emergentními jevy (umělého) média, čili zda například radost či strach (a zejména jejich *vědomé* prožívání) vyvstávají z obrovské komplexity lidského (nebo i zvířecího) mozku. Epistemologickým problémem samozřejmě zůstává, zda jsme toto schopni vůbec kdy zjistit. Nicméně v případě, že by tomu tak bylo, mohly by být stroje alespoň principiálně nadány vědomím (resp. myslí).

Současná umělá inteligence uplatňuje čistě v rámci emergentismu dvě základní strategie – *konekcionismus* a *evoluční přístup* (v obou se oproti tradičnímu paradigmatu uplatňuje paralelní zpracování). Paradigma konekcionismu je založeno na spojování množství základních relativně jednoduchých funkčních prvků s jednoduše definovaným chováním, přičemž lokální chování systému je kvůli tomu poměrně snadno popsatelné, zatímco globální chování je díky principu emergence kvalitativně značně odlišné. Konekcionistické systémy jsou například *neuronové sítě*. Evoluční přístup (např. *genetické algoritmy*) je inspirován přírodovědeckou evoluční teorií a na základě selekčního tlaku se snaží prosazovat z daného hlediska „nejschopnější jedince“ (algoritmy). Uplatňují se zde principy náhodné mutace, křížení, apod.

Vztah tradiční a konekcionistické umělé inteligence klade tyto otázky:

1. Může u konekcionistických systémů existovat „vyšší“ úroveň, ve které by samovolně probíhaly logicko-symbolické procesy podobné těm, které lze realizovat metodami tradiční umělé inteligence?
2. Lze tyto samovolné procesy *cíleně vyvolat* pomocí projektů definovaných pro „nižší“ úroveň (tj. neuronovou síť)?
3. Lze takto (cíleně) vyvolat navíc i takové procesy, které na vyšší úrovni nedovedeme předem popsat jako projekty, a tedy ani realizovat metodami tradiční umělé inteligence?

Tyto dva přístupy dále konfrontujme následujícími dvěma tezemi:

A. Silná konekcionistická teze: „Mentální stavy a procesy jsou emergentními jevy na některé vyšší úrovni dostatečně složitého konekcionistického systému.

B. Teze symbolického paradigmatu: „K tomu, aby fyzický systém vykazoval obecnou inteligentní činnost, je nutnou a postačující podmínkou, aby to byl fyzický symbolový systém.“

Je zřejmé, že tyto teze jsou vzájemně naprosto neslučitelné. Z teze B vyplývá, že podle tradiční umělé inteligence je fenomenální komponenta myšlení irelevantní nebo je popsatelná formálními symbolickými zákonitostmi, zatímco teze A tvrdí, že myšlení jako takové je emergentní, a tedy prostředky konstrukční úrovně nepopsatelný jev.

V předchozím textu byl často zmiňován pojem „úroveň“, jehož náplň byla doposud chápána intuitivně. Obecnější pojetí předpokládá, že přirozený reálný svět je členěn do množství dílčích „oblastí zájmů“ - *domén*, v rámci nichž se daná část světa jeví srozumitelně pro určitou činnost. Například živé organismy mohou být popisovány třeba z hlediska domény proteinů, nebo mnohobuněčných organizmů, buněk, či nukleových kyselin. Každá z těchto domén představuje nedílnou část fungování živého organismu, přičemž o žádné z nich nemůžeme apriori tvrdit, že je nadřazena či podřízena kterékoli jiné, případně že by mohla existovat odděleně od ostatních. Zároveň je však přechod z jedné domény do druhé *nekanonický* – tj. není jednoznačný a přechodem z jedné domény do jiné se vždy nějaká informace „ztrácí“ (resp. ztrácí svůj „smysl“), zatímco jinou je možné „získat“ (tj. svůj „smysl“ teprve dostává).

O každé doméně lze říci, že není možné přesně a ostře stanovit její hranice – od „středu“ domény směrem k jejím okrajům se náš zájem o ni jakoby vytrácí a rozostřuje a vztahy blízko okrajů nejsou zdaleka tak zřetelné a zřejmé, jako ty obsažené v jádru. Každá doména má tedy svůj *obzor*.

Důležitým hlediskem při zkoumání domén jsou *měřítka veličin*, tedy například prostorová měřítka objektů či časová měřítka procesů obsažených v daných doménách. Takto může dojít k dělení domén na „mikrosvět“ a „makrosvět“, „rychlé“ a „pomalé“, apod. Domény, jež lze takto charakterizovat, nazveme *škálové domény*. Nyní uvažujme vhodnou skupinu škálových domén, jež lze podle daného měřítka lineárně seřadit (např. podle velikosti) – dostáváme tak *hierarchii úrovní*.

Při doménové fragmentaci světa jsou pro naše chápání klíčové *kauzální vztahy*, a to nejen ve smyslu fyzikálním, ale i mentálním, případně jakémkoli dalším přirozeně myslitelném. *Kauzální doména* nechť je jakákoli oblast (výsek, fragment, komponenta) skutečnosti, v jejímž rámci se nám kauzální vztahy jeví jako *zjevné*, *srozumitelné* a *vzájemně koherentní*. Z tohoto pohledu lze nyní formulovat *zobecněnou emergentistickou tezi*:

C. Zobecněná emergentistická teze: „Mentální stavy a procesy lze pojmout jako emergentní jevy nad rozsáhlou množinou vzájemně vázaných kauzálních domén.“

Podle teze C tedy nejde pouze o jevy v jedné konkrétní kauzální doméně, jež vyvstávají za podpory procesů z kauzálních úrovní nižších, ale o jevy povstávající z celé globální sítě vzájemných vazeb napříč doménami. Zatím však neexistuje žádná ucelená teorie zabývající se touto problematikou, nicméně zobecněná emergentistická teze může takové budoucí teorii poskytnout určité základy.

Zjednávací přístup

Obě předchozí paradigmata umělé inteligence v zásadě sdílejí společnou představu, že myšlení je samostatný a svébytný proces, pojmově i funkčně oddělitelný od okolního prostředí. Určitým způsobem počítají u inteligentního stroje s vytvářením vnitřní reprezentace okolního světa (byť v případě konekcionistickém, resp. emergentistickém, nemá tato reprezentace explicitní symbolický charakter), což principiálně umožňuje rozlišovat *správnou* a *chybnou* reprezentaci, nicméně kritérium takového rozlišování je v rukou *vnějšího pozorovatele*.

Odlišný přístup považuje stroj za plně autonomní systém, který si „žije vlastním životem“, čili jeho zkušenostní svět je od lidského diametrálně odlišný a tedy i jeho inteligence se podřizuje jiným kritériím. Z této pozice hovoří jeden z nových směrů v kognitivní vědě (resp. umělé inteligenci) – *zjednávací přístup* – u něhož je důraz na vnitřní reprezentaci nahrazen důrazem na vnímání a jednání stroje *ve světě*, který je tak vlastně spoluvytvářen. Tuto myšlenku vyjádřil R. Brooks: „Svět sám je svou nejlepší reprezentací.“ Příkladem zjednávacího přístupu je *reaktivní princip* v robotice.

Zjednávací přístup přisuzuje strojům (resp. obecně jakýmkoli *agentům*) tzv. *vtělenou kognici*, která nespočívá v reprezentování světa nějakou předem danou myslí, nýbrž v jeho průběžném *zjednávání* – tvarování světa (včetně mysli) v průběhu *historie jednání*.

Oblast umělé inteligence se tak může rozdělit na dva směry: a) konstrukce *kognitivních robotů* (agentů), kteří napodobují nebo rozšiřují lidské performační schopnosti za použití reprezentace lidského světa; b) konstrukce *zjednávacích robotů* (agentů), čili umělých organismů vybavených specifickými prostředky pro zjednávání jejich vlastního světa. Zajímavá pak je kombinace obou – množství zjednávacích agentů vytváří prostředí, v němž mohou

emergovat určité performační schopnosti reprezentující lidský svět u kognitivního agenta, jehož součástmi zjednávací agenti jsou.

Další otázky umělé inteligence

Myšlenkové experimenty

Myšlenkové experimenty tvoří od dob Galileových důležitou součást vědecké (a z dob ještě dřívějších i filosofické) metodologie. Nejznámějším myšlenkovým experimentem v oblasti umělé inteligence je tzv. „Turingův test“ (1950). Alan Turing v něm navrhuje způsob, jak rozhodnout, zda stroj myslí. Představme si dvě komory – v jedné je uzavřen stroj, ve druhé člověk. Vnější experimentátor má na základě dialogu s oběma komorami rozhodnout, ve které z nich je člověk a ve které z nich je stroj (stroj se přitom bude vydávat za člověka). Na první pohled vypadá test zajímavě, nicméně při hlubším zamyšlení dojdeme k závěru, že není příliš průkazný ani inspirativní. Je totiž založen čistě na behavioristickém pojetí myšlení (zcela ignoruje možnou fenomenální komponentu), navíc testuje značně specifickou věc, a to jak dobře dokáže stroj předstírat, že není stroj. Mezi další protiargumenty můžeme zařadit i námitku, že v nejstriktnější verzi testu nemůže uspět žádný stroj, neboť postrádá jakékoli lidské (nikoli však strojové...) sociální a kulturní vazby (stejně jako neví, jaké to je mít (být?) lidské tělo), tudíž na tato témata nedokáže adekvátně reagovat (ačkoli zůstává otevřená otázka, zda by nebylo možné stroji jeho lidské tělesné počitky „simulovat“, stejně tak i vytvořit mu „virtuální“ lidské sociální prostředí).

Dalším důležitým myšlenkovým experimentem je „Čínská komora“ (Searle, 1980), který má vyvracet paradigma počítačového funkcionalismu (viz výše). Představme si tuto situaci: jsem v uzavřené komoře, do níž mi vnější experimentátor vkládá rozličné nápisy psané čínskými znaky. Já v rozsáhlém manuálu (psaném česky) vyhledám toliko na základě tvaru mně předložených znaků sestavu jiných čínských znaků, jež mám překreslit na papír a předat ven z komory. Po určité době dozajista nabude rodilý Číňan (který je oním vnějším experimentátorem) přesvědčení, že komora (či alespoň něco v ní) umí čínsky, zatímco já *vím*, že čínsky neumím. Budu-li v místnosti nahrazen počítačem provádějícím totéž, není tedy žádný důvod si myslet, že tento počítač čínsky *umí*.

Searlův experiment s sebou přináší značné množství problémů i dalších úvah, které budou hlouběji probrány na přednáškách. Každopádně se při hlubším rozboru ukazuje, že nelze přímočaře klást kauzální vztahy mezi úrovní (sub) symbolové manipulace a úrovní rozumění jazyku (z hlediska fenomenálního).

Eliza

Všechny pokusy o stanovení podmínek poznání, zda stroj vědomě prožívá, narážejí na jakousi epistemologickou „bariéru“ - dospíváme k názoru, že takový poznatek nemůžeme s jistotou učinit ani o jiném člověku. To v případě, že výchozím bodem pro zjednávání světa je *člověk*. Uznáme-li za výchozí bod *stroj*, situace je jiná – stačí pouze měnit pojetí a chápání člověka jako bytosti tak dlouho, dokud bezvýhradně nesplní kritéria kladená na stroj. Pak máme problém vědomí vyřešen. Otázka je, v jakém stavu by se tou dobou nacházela lidská společnost. Mám za to, že právě napětí mezi „strojovostí“ a „člověkovostí“ (ve smyslu v zásadě metafysickém) lidské bytosti může být jedním z konstituujících prvků lidského života, jak jej prožíváme.

Módní otázka zní například: „Může stroj cítit a prožívat lásku, jako ji cítí a prožívá člověk?“. Odpovědět lze: „Ano.“ – stačí pojem „láska“ dostatečně fyzikalisticky redukovat, degradovat a vyprázdnit. Na cestě od člověka ke stroji tím urazíme značný kus a lidstvo bude mít pocit, že zlepšilo své stroje a rozšířilo znalosti o podstatě a fungování světa. Nepoznané však zůstane, že zásadní změna nastala v lidech, nikoli ve strojích.

Jeden směr umělé inteligence se zabývá navrhováním různých dílčích systémů, které ve specifických problémech dokáží nahradit člověka. Tato oblast je (jako většina vědeckého bádání) vystavena možnosti zneužití svých poznatků proti jednotlivcům či skupinám lidí. Zároveň však dokáže výrazně přispět ke kvalitě lidského života (podobně jako například jaderný výzkum). Výzkum jako takový nelze zastavit (asi by to ani nebylo moudré), neboť emerguje z mnohem nižších úrovní lidské činnosti, avšak je třeba, aby každý jednotlivec odpovědně postavil svůj vztah k produktům umělé inteligence (a nejen jí) na stabilním udržení rovnováhy mezi výše zmíněnou „strojovostí“ a „člověkovostí“.

Druhý směr – snaha o vytvoření „umělého“ člověka – je (dle mého názoru) pro umělou inteligenci odsouzen k nezdaru. Proč tomu tak je, bude diskutováno na přednáškách, kde bude též probrán pohled a názory Josepha Weizenbauma, tvůrce programu Eliza, který ukazuje, jak mocná je touha lidí připisovat strojům vědomí a zrcadlit v nich své vlastní myšlenky.

Pozn.: V textu je mnohde odkazováno na a citováno z publikací Ivana Havla (Přirozené a umělé myšlení jako filozofický problém,; Causal Domains and Emergent Rationality), Jozefa Kelemena (Budoucí Altamira; Kybergolem; Berušky, andělé a stroje – spoluautor Anton Markoš) a Josepha Weizenbauma (soubor Mýtus počítače).