Oponentský posudek na disertační práci Filipa Tvrdého
Turingův test: Filosofické aspekty umělé inteligence

Mohou stroje myslet? Otázka, kterou si kladli filosofové minimálně od dob Reného Descarta uvažujícího o strojovosti lidského těla, otázka, která se ve dvacátém století dostala do popředí zájmu v souvislosti s vývojem informatiky a kybernetiky. Touto otázkou je fascinován i Filip Tvrdý, který se ve své disertační práci rozhodl stopovat bouřlivou debatu, jíž rozpoutal návrh testu Alanem Turingem (T-testu) majícího prokázat inteligenci (nejen) člověkem vyrobených strojů.

Nutno říci, že Filip Tvrdý se svého úkolu zhostil s přehledem a bravurou. Jeho text je psán velmi svěžím a přístupným stylem, je plný břitkého humoru i zaujatých stanovisek a čtenář se ani chvíli nenudí. Autorovi se navíc podařilo oprostit od akademické manýry libující si v přebujelém technicistním pojmovém aparátu, nepředpokládá jako mnozí jiní, že v nejasnosti se skrývá hloubka a nezdráhá se ani vysvětlovat pojmy či myšlenky specialistovi možná notoricky známé. Tvrdého práce by si tak určitě našla čtenáře i mimo úzce specializovanou akademickou obec a po mírném přepracování by mohla vyjít též jako kniha popularizační. Škoda jen, že diplomant vedle extenzivní anglosaské literatury nereflektoval i domácích autory. Tématu se například explicitně věnují v několika publikacích brněnští kolegové Marek Picha či Miloš Dokulil.

I když práci Filipa Tvrdého oceňuji velmi vysoko, přece jen se mi zdá, že se mu nepodařilo zcela vybřednout z pojmové konfůze často doprovázející otázku, zda stroje mohou myslet. Je totiž třeba dobře vyjasnit, co je míněno výrazem „stroj" a výrazem „myslet". Za stroj může být považováno i lidské tělo a myslet může v optice některých filosofů třeba i hora. Tvrdý je si problému dobře vědom a na mnoha místech se mu věnuje a o definice se pokouší. Strojem Tvrdý rozumí Turingův stroj (T-stroj) (s.19), tedy digitální číslicový počítač, stroj, který dokáže mechanicky vypočítat jakoukoli funkci, je-li tato algoritmizovatelná, a to tak, že veškeré matematické operace převádí pomocí principu konečných diferencí na sčítání (logický součin a součet), které vykonávají dva jednoduché obvody. Horší je to s termínem „myslet". Turing definoval myšlení jako schopnost uspět v imitační. Nic víc a nic méně. Takto definované myšlení má ovšem s lidskou „myslí" společné jen málo. Používat stejný termín „myslet" či „mít mentální stavy" pro výkony T-strojů a výkony vědomých bytostí se mi zdá zmatečné. Lidská mysl má mnohem více rozměrů, než které je schopen zachytit T-test a pro lidskou mysl není „inteligence" ve smyslu schopnosti projít T-testem esenciální vlastností. Lidé, kteří v T-testu neuspějí, nejsou o nic méně lidmi. Nicméně původní Turingova otázka, zda úzce definované T-stroje mohou projít úzce definovaným T-testem[1] bývá u mnoha autorů (a zdá se mi, že i u Tvrdého) nepozorovatelně transformována do otázky, zda stroje mohou myslet po způsobu lidském či dokonce zda mohou mít vědomí. Jedná se zde tedy o důkaz kruhem. T-test je předkládán jako kritérium inteligence, přičemž inteligence je definována jako schopnost projít T-testem.

Otázka kterou by si měl Tvrdý položit, nezní, jestli je entita, která projde T-testem, inteligentní (to dozajista je), ale zda má být zařazena do kategorie morálních bytostí vyžadujících naše morální zohlednění. Inteligence (ve smyslu T-testu), podle mého tímto kritériem být nemůže a to rozhodně nikoli jako nutná podmínka a pravděpodobně ne ani jako

[1] Turing (a pravděpodobně i Tvrdý) byl přesvědčen, že T-stroj principálně T-testem projít může, nicméně v současnosti to vypadá, že nikoli. Už jen proto, že přirozený jazyk není formalizovatelný, tedy z defince nemůže být implementován T-strojem. To pochopitelně neznamená, že by nemohl být implementován strojem jiného typu, třeba nějakým konekcionistickým systémem, který nepracuje na principu formálního kalkulu.

podmínka postačující. Nebo snad diplomant trpí výčitkami svědomí, když vypíná svůj počítač, ve kterém má nahraný sofistikovaný konverzační program?

Podmínkou možnosti zařazení do třídy morálních bytostí je dle mého schopnost mít zájem (tedy mít vědomí). S detekcí vědomí jakožto privátní kvality je pochopitelně problém. Netušíme, jak vědomí vzniká, čím je generováno, takže konec konců nelze vyloučit, že jej má i digitální počítač. Nicméně podobně jako Searlovi se mi zdá, že u něj nevzniká na základě toho, že manipuluje se symboly podle formálních pravidel. Tvrdý proti Searlově argumentaci artikuluje tři protiargumenty:

1. že experiment s čínským pokojem je prakticky těžko realizovatelný (s. 70). To nepochybně. Vždyť právě o jeho realizaci se pokoušejí generace programátorů. Ovšem kdyby uspěli, kdyby vyvinuli program schopný obstojně konverzovat v čínštině, pak by šlo Searlův test snadno provézt. Tvrdému by se na obrazovce objevila sada čínských znaků jakožto otázka, zmáčknutím tlačítka by Tvrdý pomocí slovníku-počítače vygeneroval sadu znaků jakožto odpověď a mohl by pak posoudit, zdali rozumí čínsky. (Nechť mi autor promine tento myšlenkový experiment. Jeho nedůvěru k myšlenkovým experimentům sdílím.)

2. že čínsky rozumí celý systém, i když ne jeho část (s.71-2). Tvrdý argumentuje tak, že symboly nabývají sémantický obsah díky interakci s prostředím. Stačí prý proměnit počítač na robota přidáním smyslových receptorů a vše se změní. Syntaktickým symbolům dávají jejich sémantické vlastnosti právě kauzální vztahy, které mají k nesymbolickým věcem ve světě.[2] Vnitřní symboly systému nabývají podle funkcionalistů svého intencionálního obsahu právě na základě kauzálních funkcionálních rolí, které se uplatňují při úspěšném zprostředkování adaptivní interakce s prostředím.[3]

I tato námitka je však dle Searla neadekvátní. Rozlišení mezi sémantikou a syntaxí podle něj neúprosně zvrací i tento manévr. Předpokládáme-li, že robot pobíhající po světě má místo mozku digitální počítač, pak nedokáže přeskočit od čínské syntaxe k sémantice, i kdyby se choval, jakoby čínsky rozuměl. Opět se to dá ilustrovat obměnou myšlenkového experimentu „Čínského pokoje“.[4] Stačí si představit sebe sama sedícího v hlavě robota a pořádajícího na základě formálních pravidel abstraktní symboly přicházející z robotových senzorů. Mám-li pouze symbol a nevím-li, co znamená, nemohu vědět, odkud se vzal. Kauzální interakce mezi robotem a okolním světem je podle Searla irelevantní, pokud není reprezentována v mysli. To je ale zcela vyloučeno, pokud takzvaná mysl sestává pouze z čistě formálních syntaktických operací. Systém pracující výlučně na principu digitálního počítače, T-stroje, nikdy nemůže nabýt subjektivní zkušenosti.

Proti tomu by se sice dalo namítnout, že podobná situace nastává v případě vizuálního vědomí, kdy oko transformuje světelné vzruchy do symbolů a odešle je ke zpracování zrakovým analyzátorem. Odkud tedy berou zrakové vjemy svůj sémantický obsah, když podráždění jednotlivých receptorů sítnice jej nemají?[5] Nevíme, ale vše

---

[2] Vždyť i děti se asi učí sémantice ostenzivně. Jednou z prvních verbálních komunikačních dovedností, kterou zvládla má dcera, bylo na otázku, „Jak dělá kravička?“, odpovědět: „Bůů!!“. V té době ještě žádnou krávu nikdy neviděla. Její odpověď byla tedy podmiňováním dobře nacvičená behaviorální reakce na zvukový stimul bez sémantického obsahu - odpověď dobře naprogramovaného robota. Teprve posléze, když jsme jí opakovaně krávu ukázali (na obrázku i ve skutečnosti), když poprvé uslyšela její zabučení, se v její mysli sémantický obsah výrazu „kravička“ a „bů“ vytvořil a zpřesnil. Jaké sémantické obsahy, jaké představy, si asi spojovala se slovy jako oblaka, měsíc, bučení krávy, atd., Helena Kellerová, hluchoslepá dívka, jejíž inteligence se díky obětavé ošetřovatelce rozvinula natolik, že byla schopna napsat autobiografii? Srv.: Keller, H.: Sourde, muette, aveugle; Payot 1991.

[3] Srv. např.: Gulick, R.: Vedomie, vlastní intencionalita a stroje, ktoré rozumejú samy sebe; in.: Gál, E., Kelemen, J. (Ed.); Myseĺ, telo, stroj; Bradlo 1992.

[4] Searle, J.: Mysl, mozek a věda; Mladá fronta, Praha 1994, s. 36.

[5] Na s. 72 se Tvrdý ptá, čím se mozek liší od běžného počítače. Jeden od druhého se dozajista liší v mnoha ohledech, ale relevantní je v tomto kontextu asi otázka, zda-li mozek jako celek pracuje na základě digitálního či analogového principu přenosu informace. Pro digitální (binární) kód by mluvil neuronální zákon „vše, nebo nic“.

nasvědčuje tomu, že vědomí nepovstává z komputace. Koneckonců i chudákovi Ottovi z myšlenkového experimentu na s. 72, kterému zmizela explicitní paměť, zůstalo aspoň bazální vědomí zajišťující schopnost rozumět.

3. že není rozdílu mezi syntaxí a sémantikou. V tom má Tvrdý možná pravdu, ale Searlův argument to nijak nezvrátí. V knize *Záhada vědomí* Searle svůj argument ještě více zobecnil a prohloubil. Myšlenkový experiment „Čínský pokoj" se snažil ukázat, že sémantika vnitřně nevyplývá ze syntaxe. Searle ale časem došel k závěru, že dokonce ani syntax vnitřně nevyplývá z fyziky. Zeptáme-li se totiž, co proměňuje elektrické impulzy v nitru počítače na symboly, musíme si odpovědět, že se jedná o to stejné, co proměňuje v symboly skvrny tuše na stránce knihy: tj. na subjektu závislé relativní interpretační stanovisko. Jednoduše řečeno: ani syntax není v přírodě, nýbrž pouze ve vědomé mysli. Elektrické impulzy jsou na pozorovateli pravděpodobně nezávislé, jejich komputační interpretace však nikoli: ta je odvislá právě od pozorovatele, uživatele či programátora nadaného sémantickým vědomím. V komputačně funkcionalistickém pojetí mysli jakoby se skrývalo reziduum ontologického dualismu. Jakkoli se komputacionismus pyšní svým materialismem, ve skutečnosti je materialismem příliš málo. Připusťme, že mozek se svými elektrochemickými procesy, jež mysl nějak způsobují, je určitý druh organického stroje. Avšak kalkul není elektrochemický proces. Je to proces abstraktně matematický, který existuje jen v myslích vědomých interpretů. A svévolně připisovat tuto abstrakci digitálním strojům je pochopitelně velmi ošidné. Na konkrétní realizaci, alespoň v případě vědomí, totiž zatraceně záleží.

Lze tedy shrnout: na otázku, zda mohou digitální stroje myslet, můžeme bez váhání odpovědět kladně. Ano, pokud myšlení definujeme jako kalkul, tak mohou - svým strojovým způsobem. Mohou dokonce mít i určitou intencionalitu, i když intencionalitu odvozenou, intencionalitu nepravou – ve stejném smyslu, jako má intencionalitu třeba termostat.

Ovšem specifikum přirozených (biologických) myslí netkví ve schopnosti takto definovaného myšlení ani intencionality prvního řádu, nýbrž v čemsi, co intuitivně nazýváme vědomí - tj. schopnost mít zájem[6], schopnost pociťovat subjektivní stavy, schopnost pociťovat utrpení. A na otázku, zda jsou digitální počítače tohoto schopny, musíme odpovědět rozhodné nikoli.

Na každou nervovou buňku nasedá průměrně několik tisíc (až 100 000) synapsí (z nichž některé jsou navíc tlumivé – inhibiční) a záleží pak na počtu podrážděných synapsí a na časovém sledu jejich dráždění (časové sumaci) i na jejich poloze (synapse na vzdáleném dendritu ovlivňují výsledek méně než synapse blíže položené), zda-li výsledná depolarizace dosáhne prahové hodnoty a rozšíří se tak i na axon. Buď nevznikne žádné podráždění, nebo vznikne akční potenciál charakteru „vše, nebo nic" a rozšíří se na všechna zakončení neuritu (axonu), kde způsobí výlev mediátoru. Ovšem velikost výlevu závisí na amplitudě akčního potenciálu. Jedná se zde tedy zjevně o přenos analogový. Neurony nejsou pouhá relé, která by jednoduše předala nebo nepředala přijatý signál. Pokud máme užít počítačovou metaforu, tak jsou to spíše procesory. Typický neuron získává signál z mnoha zdrojů, tuto informaci integruje, transformuje a kodifikuje do komplexních signálů a dále je distribuuje mnoha dalším buňkám. Navíc se zdá, že jednotlivý elektrický impulz není sám o sobě nositelem informace, nýbrž tuto funkci plní spíše změna frekvence. Když není neuron drážděn, vysílá samovolně impulzy s poměrně nízkou četností (tzv. pozadí) často mezi 1 - 5 hertzy. Poté, co je neuron vybuzen množstvím excitačních signálů, počet vydávaných impulzů vysoce vzroste, frekvence činí obvykle 50-100 hertzů nebo více. Na krátkou dobu může dosáhnout četnost vydávaných impulzů až 500 hertzů. Dostane-li však neuron nadbytek tlumivých signálů, je s to vyslat impulzů daleko méně, než odpovídá „pozadí". Činnost nervové soustavy je tedy spíše analogová a šla by snad šla přirovnat k neustále hrajícímu dobře vyladěnému hudebnímu tělesu a změny stavů mysli pak ke změnám rytmu a melodie. Špatně sladěný orchestr by potom vyluzoval zvuky, jež by šlo přirovnat k mysli patologické. A hudba je pochopitelně signál analogový.

[6] V Singerově smyslu. Srv. např.: Singer, P.: Questions d´éthique pratique; Bayard Éditions 1997.

Rád bych se ještě na závěr vyjádřil k otázce, zda-li je úspěch v T-testu „postačující podmínkou pro detekci inteligence u námi vytvořených strojů." (s. 120). Otázku zde chápu navzdory předchozím užším definicím v širším smyslu. Pod inteligenci zde zahrnuji (podobně jako Tvrdý) i schopnost vědomí a pod strojem chápu i jiné stroje než T-stroje. Neboť má dozajista pravdu Tvrdý, když tvrdí, že: „je naprosto bezdůvodné strojům apriorně upírat něco, o čem vůbec nevíme, čím ve skutečnosti je." (s. 103). Jestliže v budoucnu nějaký stroj projde T-testem, přiznáme mu morální práva?

William James, byť poněkud v jiné souvislosti, uvažuje o „automatické milence". Má na mysli robotickou milenku, která by nejen prošla Turingovým testem, ale která by prošla i testem v posteli: „Má bezduché tělo, absolutně nerozeznatelné od duchem stvořené panny, která se směje, mluví, červená se, stará se o nás a vykovává všechny ženské role tak něžně a líbezně, jako by měla duši."[7] Považoval by ji James za plnohodnotný ekvivalent? Nikoli. A proč? „Protože náš egoismus, jelikož jsme takoví, jací jsme, se dožaduje především vnitřního spolucítění a uznání, lásky a obdivu. Vnější starostlivost hodnotíme především jako výraz, manifestaci souběžného vědomí, o kterém jsme přesvědčeni. Přesvědčení o automatické milence by potom pragmaticky nefungovalo a ve skutečnosti ji nikdo nepovažuje za seriózní hypotézu."[8]

Otázkou však je, jak by James mechanickou milenku poznal, když není zřejmé, co vědomí způsobuje. I když je behaviorální evidence v principu nedostatečná, jiné než behaviorální kritérium bohužel nemáme. Co je pak v situaci morální nerozhodnosti přijatelnější: připsat stroji vědomí s tím, že se možná, hrůza hrůz, budeme milovat s bezduchou milenkou jako s vědomou osobou, nebo upřít stroji vědomí, a možná jednat s cítící osobou jako s bezduchým strojem a tak přinejmenším ranit její city? A toho, kdo nyní váhá, se ptám: Jsi si jistý, že tvá žena není automatická milenka?[9] Raději připíšeš jinému člověku vědomí s tím, že možná budeš jednat s bezduchým strojem jako s vědomou bytostí, nebo mu vědomí upřeš a budeš riskovat jednání s vědomou bytostí jako s bezduchým strojem?

Anebo: můžeme si být jisti, že Tvrdého práci nenapsal stroj? Jak ji hodnotit? Já za sebe hodnotím doktorandův text jako promyšlený a inteligentní, plně vyhovující požadavkům kladeným na práci disertační a **doporučuji** jej nejen k obhajobě, ale i k případné publikaci.

V Kunčicích pod Ondřejníkem 3.10. 2011 Doc., Mgr. Marek Petrů, Ph.D.

[7] James, W.: Pragmatistický výklad pravdy a jeho nesprávne pochopeniea. In.: Višňovský, E., Mihina, F. (Eds.): Pragmatizmus; Iris, Bratislava 1998, s. 257.

[8] Tamtéž. Dodejme, že vesmír bez Boha by byl Jamesovi přesně takový. James pomyšlení na něj odmítá ze stejných pragmatických důvodů jako automatickou milenku.

[9] Srv.: Capgrasův syndrom, v rámci nějž jsou postižení přesvědčeni, že blízké osoby, které podle vzhledu poznají, jsou ve skutečnosti podvodníci či mimozemšťané, kteří na sebe vzali jejich podobu, a pouze předstírají, že jsou osobami známými a blízkými.