Turingův test: Filosofické aspekty umělé inteligence

Autoreferát k disertační práci

Filip Tvrdý

Předložená disertační práce se zabývá problematikou připisování myšlení jiným entitám, a to pomocí imitační hry navržené v roce 1950 britským filosofem Alanem Turingem. Zatímco u lidských bytostí se jedná o vcelku triviální úkol, za hranicemi našeho biologického druhu se může situace značně komplikovat. Jedná se především o identifikaci myslí u jiných živočichů, ale i u hypotetických, zatím neexistujících počítačů a robotů. Právě v případě lidmi zkonstruovaných strojů se často hovoří o takzvaném Turingově testu, který je jednoduchým kritériem pro nebo proti připsání inteligence. Test byl Turingem poprvé popsán v eseji "Computing Machinery and Intelligence", jenž se stal jedním z nejvíce citovaných a komentovaných filosofických textů druhé poloviny 20. století. Turingovo řešení je velice originální, i když do určité míry vychází ze starších zdrojů, mezi něž patří Descartes, de Cordemoy nebo Leibniz. Imitační hry se účastní tři hráči: člověk, stroj a lidský tazatel. Tazatel je v jiné místnosti než účastníci hry a jeho úkolem je prostřednictvím série otázek zjistit, kdo je stroj a kdo člověk. Stroj předstírá, že je člověk, zatímco člověk se snaží tazateli pomáhat při rozhodování. Komunikace je prováděna písemně nebo zprostředkovaně, například pomocí dálnopisu nebo terminálu, aby se tazatel nemohl řídit fyzickými znaky účastníků. Zkouška není časově ani tematicky omezená, a proto je možné klást dotazy z naprosto všech oblastí. "Lidskost" účastníka je posuzována podle kvality odpovědí, jejich přirozenosti, smysluplnosti, vtipnosti apod., tedy úplně stejně jako posuzujeme intelekt našeho partnera v běžné mezilidské konverzaci. Turingův test byl interpretován různě, a to buď jako pokus o operacionální definici inteligence, nebo jako induktivní metoda pro formulování hypotézy o přítomnosti myšlení. Nejpřesvědčivější je ale podle mě interpretace, jež test považuje za postačující podmínku myšlení: každá entita, která uspěje v Turingově testu, myslí.

Turing věnuje podstatnou část eseje vyvrácení devíti potenciálních námitek. Lze říct, že vlastně předešel všechny relevantní způsoby kritiky více či méně úspěšně. Jedná se o následující teze: teologická námitka, námitka "hlavy v písku", matematická námitka, argument z vědomí, argument z různých neschopností, námitka lady Lovelaceové, argument ze spojitosti nervové soustavy, argument z neformálnosti chování a argument z mimosmyslového vnímání. Nejcennější je matematická námitka, a proto je jí v disertační práci věnována zvláštní pozornost. Vyplývá z Gödelova teorému o neúplnosti, podle něhož "v každé dostatečně silné logické soustavě lze formulovat tvrzení, která nemohou být v rámci soustavy dokázána ani vyvrácena, pokud sama soustava není nekonzistentní" (Turing 1950: 444). Turing z věty o neúplnosti vyvozuje, že "existují určité věci, které stroj nemůže udělat", ale zároveň je přesvědčen o přítomnosti podobných omezení v lidském rozumu. Domnívám se, že vliv Gödelova teorému na myšlení lidí a počítačů je neproblematický. Všechny varianty námitky u Lucase nebo Penrose je možné překonat prostým použitím Tarského rozlišení mezi objektovým jazykem a metajazykem, díky němuž lze "vystoupit ven" z příslušné soustavy vět. Je pravda, že věta *p* je nedokazatelná v soustavě S_1, ale my můžeme tuto nedokazatelnost konstatovat v nadřazené soustavě S_2. A i když se v soustavě S_2 setkáme opět s neúplností, můžeme postup opakovat povýšením na úroveň S_3 a tak dál *ad infinitum*.

Hlavní část disertační práce se týká pozdější recepce Turingova testu v letech 1950 až 2010. Postupně jsou diskutovány tyto hlavní způsoby kritiky: Searlův argument čínského pokoje poukazující na nutnost sémantiky pro myšlení; Blockův návrh simulovat inteligenci pomocí brutální výpočetní síly; Frenchova hypotéza o nenaučitelnosti subkognitivních vlastností; Michieho konstatování nemožnosti inference vědomí z behaviorálních projevů. Práce má ambici vypořádat se s každou z uvedených námitek, a to za použití postupů, o nichž v češtině neexistuje žádná nebo téměř žádná literatura. Proti Searlovi stavím takzvanou systémovou námitku, která připisuje schopnost komunikovat v čínštině celému pokoji, přičemž člověk v něm zavřený plní funkci řídícího procesoru. Za ještě závažnější považuji Rapaportovu koncepci syntaktické sémantiky, jejímž modelovým příkladem je způsob, jakým si osvojila jazyk hluchoslepá Helen Kellerová. Při analýze Blockova přístupu podrobuji kritice použití myšlenkových experimentů ve filosofii, neboť jejich slabá vazba na empirii a epistemologická

komplikovanost znemožňují spolehlivé vyhodnocení. Opírám se přitom o závěry experimentálních filosofů popírajících udržitelnost "armchair philosophy". Frenchovu a Michieho kritiku odmítám z empirických důvodů - skutečnost, že současný stav našeho vědeckého poznání nedokáže vysvětlit nějaký fenomén, ještě neznamená, že jej nebudeme schopni vysvětlit nikdy. Frenchovi činí potíže představit si stroj vítězící ve vědomostní soutěži, ale takové stroje již dnes existují; Michie si nesvede představit stroje vybavené vědomím, ale neexistuje žádný důvod, proč by počítače s myšlením vyššího řádu nemohly existovat.

Situace ve filosofii umělé inteligence je zatížena intelektuální nepoctivostí, kterou Hauser nazývá "posouvání branek". Pravidla pro připsání inteligence strojům jsou *ad hoc* měněna tak, aby určití adepti nesplnili nějakou zkoušku nebo kritérium. Důvodem je snaha o udržení výsadního postavení člověka v universu, kterého se mnozí antropocentricky zaměření myslitelé nechtějí vzdát. I když počítače zatím fatálně selhávají v soutěžích, jakou je Loebnerova cena, přesto nelze odmítnout možnost budoucích inteligentních strojů kategoricky. Případný neúspěch digitálních technologií totiž nemůže být předpovězen pomocí apriorních úvah, nýbrž konstatován za použití experimentálního testování empirických hypotéz. Projevuje se tak podle mého názoru důležitý princip, kterým je priorita vědy před filosofií. Filosofové přesto nemusí rezignovat na pozitivní roli při rozšiřování poznání, protože mohou úspěšně ovlivňovat vývoj kognitivních věd, i když do něj příliš aktivně nezasáhnou. Podobně jako Turing ve čtyřicátých letech popisoval programování ještě neexistujících počítačů, tak se i my můžeme pouštět do opatrných předpovědí o dalším směřování našeho pochopení mysli. Se skromnou autoritou, kterou filosofie mým slovům propůjčuje, mohu například deklarovat své přesvědčení, že Turingův test je postačující podmínkou pro detekci inteligence. Důrazněji řečeno, behaviorismus je při identifikaci myslí u nehumánních entit obligatorní, nikoli pouze fakultativní.