Paradoxy umělé inteligence: Turingův test 50 let poté

Již před více než padesáti lety napsal slavný britský matematik Alan Turing článek s názvem "Computing Machinery and Intelligence", ve kterém navrhl první variantu dodnes diskutovaného Turingova testu. Jaké otázky se od dob Turinga podařilo zodpovědět a jaké nové problémy naopak vyvstaly?

Pojem umělá inteligence je zatížen obrovským množstvím nejasností, mj. už určitou vágností samotných pojmů. Co prohlásíme za "umělé" (člověk s čipem v hlavě, možnosti klonování a genetických modifikací...) a co za "inteligentní"?

Turing se konstrukcí svého testu snažil otázku "mohou stroje myslet" převést z oblasti filozofických a lingvistických spekulací na exaktnější úroveň. Jinak řečeno, za myslící podle něj prohlásíme počítač tehdy, když jeho chování nebudeme schopni rozeznat od chování člověka.

Motivací pro vývoj umělé inteligence existovala už v polovině 20. století celá řada. Jedná se jistě o obrovskou teoretickou výzvu. Zkoumání inteligence umělé nám zřejmě dokáže mnoho říct o inteligenci vlastní. Dalším důvodem je prostá radost z konstrukce stále dokonalejších mechanismů, která historicky sahá minimálně ke Golemovi. A nikoliv v poslední řadě zde jsou motivy ryze praktické: Řadu úkolů musí plnit inteligentní bytosti, nicméně lidem se do nich příliš nechce (jsou např. stereotypní, nebo naopak nebezpečné). Zde bychom velmi rádi viděli myslící stroje, byť jsou možné i jiné cesty - např. roboti řízení na dálku prostřednictvím "lidského" rozhraní, cyberwaru.

Imitační hra

Ve své původní podobě vychází Turingův test z tzv. imitační hry, ve které jde o to odlišit dva lidi podle pohlaví. Pozorovatel, na jehož pohlaví nezáleží, má proti sobě např. ženu a muže, který předstírá, že je žena. Trojice lidí spolu nepřijde samozřejmě do fyzického kontaktu, sedí v oddělených místnostech a zprostředkovatel mezi nimi přenáší popsané lístky. Turing pak problém posunul: Co kdyby namísto simulace opačného pohlaví byl imitátorem člověka počítač? Dokáže počítač simulovat chování ženy stejně dobře jako muž?

Jaký je stav problému padesát let poté, co jej britský matematik poprvé zformuloval? Existuje celá řada více či méně dokonalých imitací určitých lidských činností. Známá Eliza (mimochodem, tento program se objevil již v roce 1966) představuje program napodobující chování psychoterapeuta. Tzv. Parry je jejím "protějškem", tedy počítačovou simulací pacienta postiženého paranoiou. Praktický význam mají tzv. expertní systémy, tedy software snažící se zastoupit např. odborníka na burzovní prognózy či lékaře stanovujícího diagnózu.

Test jako celek se však dosud žádnému programu splnit nepodařilo. Jaké jsou další vyhlídky? Známý vizionář a spisovatel Ray Kurzweil nedávno uzavřel sázku, že ke splnění podmínek Turingova testu dojde do roku 2029. Pochybující protistranou je Mitchell Kapor z Electronic Frontier Foundation. Magazín Wired ovšem sázku zařadil do jedné řady s dalšími bláznivými nápady, které přicházejí na lidi ze světa IT. Wired v této souvislosti zmiňuje Craiga Mundieho z Microsoftu, který se domnívá, že ve stejné době už bude existovat komerční provoz bezpilotních letadel. Jeden z průkopníků superpočítačů Danny Hillis se zase vsadil, že se vesmír nebude rozpínat věčně; protistranou je v tomto případě Nathan Myhrvold (dříve Microsoft, nyní analytik pro oblast biotechnologií). Současné poznatky o temné hmotě a kosmologické konstantě dávají ovšem za pravdu Myhrvoldovi. Sázky tohoto druhu "spravuje" nezisková organizace Long Bets Foundation.

Námitky a polemiky

Už sám Turing v původním článku uvedl několik možných námitek proti celé koncepci. Později byla řada výhrad zřejmě vyvrácena, jiné přibyly. Polemiky samozřejmě trvají.

Galantně začneme námitkou, kterou vypracovala žena, nadto se jedná o úvahu historicky nejstarší, předcházející samotný Turingův test o více než 100 let. Ada Lovelaceová, dcera lorda Byrona a spolupracovnice konstruktéra mechanických parních počítačů Charlese Babaggea, přišla již v 1. polovině 19. století s tvrzením, že počítač nemyslí, protože myšlení je nějak spojeno s původností. Z počítače však dostaneme pouze modifikaci toho, co do něj vložíme (výsledek rovnice apod.).

Za současného stavu vývoje IT není tato námitka příliš relevantní. Předně nevíme, jak vlastně definovat onu "původnost" - např. báseň je taktéž variací již existujících slov, tedy de facto je určitou kombinatorickou operací. Programy dnes již každopádně dokáží psát povídky nerozpoznatelné od těch lidských nebo odhalit taková tajemství šachové hry, která zůstávala jejich tvůrcům zcela skrytá. Program tedy svým chováním dokáže své tvůrce překvapit (nejen ve smyslu chyby), což je snad možné chápat jako ekvivalent původnosti. Přitom pomíjíme psychologické a sociologické úvahy o tom, že řada lidí rovněž není nějak "původní", aniž bychom na tomto základě mohli tvrdit, že vůbec nemyslí nebo nejsou inteligentními bytostmi.

Rozsáhlá literatura existuje o tzv. paradoxu čínského pokoje, s nímž přišel filozof John Searl. Searlova námitka uvádí, že splnění Turingova testu ještě nemusí znamenat nějaké myšlení a uvědomování. V původní formulaci Searl ukazuje, že člověk může být např. schopný smysluplně třídit tabulky s čínskými znaky, aniž rozumí čínsky a chápe, co text na tabulkách znamená. Podobně počítač, i když projde Turingovým testem, bude stále pouze něco imitovat a sám nebude např. nic "chápat".

Turing ovšem již ve svém původním článku poznamenává, že k tomu, abychom mohli hodnotit, co počítač prožívá, bychom jím také museli sami být. Světoznámý fyzik Stephen Hawking přímo dodává, že má smysl posuzovat pouze jevy viditelné navenek - tedy inteligenci a myšlení - nikoliv kvality, které lze uvidět pouze zevnitř (vědomí, prožívání apod.). Podle Hawkinga prostě nevíme, zda počítač má "vědomí", stejně jako nevíme, zda ho budou mít představitelé nějaké hypotetické mimozemské civilizace. Dokonce to jistě nevíme ani v případě druhých lidí (jistotu máme pouze u sebe).

Existuje velice zajímavý myšlenkový experiment tzv. Chalmersových zombií. Zkuste si představit, že všichni lidé kolem vás jsou pouze předprogramovanými automaty, bezvědomými zombiemi. Nikdy si nemůžeme být jisti opakem, na vědomí druhých lidí usuzujeme pouze podle jejich chování - a totéž je tedy logické činit i v případě počítačů. Z tohoto hlediska jsou Searlovy výhrady (snad) překonané.

Zajímavé námitky proti umělé inteligenci se týkají zdánlivě tak "strojového" oboru, jakým je matematika. V původních Turingových poznámkách se uvádí, že počítač by mohl být v testu odhalen právě tím, že nedělá chyby, eventuálně počítá rychleji než člověk (tedy by paradoxně byl nápadný tím, že je dokonalejší; program by se zřejmě dal nějak modifikovat, aby se občas spletl, poskytoval výsledky se zpožděním apod.).

Hawkingův spolupracovník Roger Penrose přišel s další matematickou námitkou, která má dokázat, že lidské myšlení má v sobě kromě algoritmů ještě něco jiného - počítače podle něj nejsou schopny matematického důkazu, např. odhalit, že prvočísel je nekonečně mnoho. Penrose se domnívá, že nejde jen o otázku softwaru a výpočetní kapacity, některé matematické úlohy jsou neřešitelné i pro obecný Turingův stroj.

Nastíněný problém je velmi široký a souvisí mj. s protikladem myšlení induktivního a deduktivního, taktéž se dotýká např. Goedelových teorémů o neúplnosti matematických systémů. Penrose sám přichází s

hypotézou, že lidské myšlení nedospěje k matematickému důkazu krok za krokem, ale jakousi zkratkou, kterou on sám vysvětluje naší schopností přijít do kontaktu s "platónským" světem matematických objektů.

Premiéru si programy v matematických důkazech odbyly již v roce 1976, když se podílely na potvrzení důkazu "čtyř barev". V rámci dokazování bylo třeba klasifikovat určitý počet map. Úkol přesahoval lidské možnosti, počítače jej však již tehdy zvládly - nikoliv ovšem samy od sebe, postupovaly podle instrukcí zadaných z vnějšku. Důkaz ale přesto provázejí určité "filozofické" pochybnosti, jeden z matematiků výsledek dokonce komentoval slovy: "Takže to ukazuje, že to prostě nebyl dobrý problém." Chtěl tím naznačit, že regulérní matematický důkaz by se neměl dělat výčtem prvků - a řada důkazů takto udělat opravdu nejde. Nicméně i proti Penrosově argumentaci samozřejmě existují protinámitky, někdy bývá dokonce pokládána za již překonanou (tento názor zastává např. Paul Thagard v knize Úvod do kognitivní vědy).

Co se týče praktických problémů spojených s konstrukcí myslících počítačů, už Turing přišel s myšlenkou programů schopných se samy učit. Možná je obtížné zkonstruovat rovnou "dospělého člověka", uvažoval Turing ve svém článku, zkusme to tedy nejprve s "dítětem". Po Turingovi tyto myšlenky ožily díky experimentům, které se snažily např. modelovat biologickou evoluci. Buněčné automaty ukázaly, že i velmi komplexní systémy mohou vzniknout aplikací několika jednoduchých pravidel, složitost se vynořuje "sama od sebe".

Pokud pro tvorbu systémů umělé inteligence použijeme techniky typu genetického programování, je ovšem docela dobře možné, že před námi vyvstane nová černá skříňka. Nakonec budeme mít inteligentně se chovající systém, aniž se toho ovšem mnoho dozvíme o inteligenci nás samých. Jestliže je však naším cílem především praktické nasazení expertního systému, který by dokázal simulovat práci odborníků, pak nám fenomén černé skříňky zase příliš nevadí.

Tak trochu zvláštní počítač

Na celý problém se můžeme podívat i z úplně opačné strany. Většinu námitek vůči Turingovu testu prostě smeteme ze stolu tezí, že lidský mozek není ničím jiným než počítačem, zařízením na zpracování informací. Struktura mozku je modulární, přičemž moduly v počítačové terminologii zhruba odpovídají různým podprogramům a specializovaným aplikacím. Chápání mozku jako počítače je v některých oborech, např. v evoluční psychologii, již bráno za samozřejmost.

Nicméně - v některých ohledech současné počítače rozhodně nedosahují schopností lidského mozku a Turingovým testem doposud neprojdou. Plodnější než debaty o tom, zda stroje mohou myslet nebo zda mozek je počítač, může být proto užší vymezení problému: V čem se mozek-počítač liší od počítačů současných, jaké má speciální vlastnosti?

V první řadě není jasné, zda mozek je systémem výlučně digitálním. V Computerworldu 9/2003 jsme na toto téma přinesli rozhovor s Markem Petrů, který zdůrazňoval, že na práci lidského nervového systému se výrazně uplatňují i analogové systémy (např. mechanismus pro rozpoznávání tváří).

Druhý okruh úvah se může točit kolem počítačů kvantových. Roger Penrose na jednu stranu přirovnává mozek ke kvantovému počítači.

Tuto Penrosovu představu však většina ostatních odborníků odmítá. Pokud se nám podaří přivést k životu kvantové počítače, snad budeme moci vyslovit určitější závěry. S Penrosovými zásadnějšími výhradami k (ne)algoritmizaci lidského myšlení ve vztahu k matematickým důkazům jsme se již setkali výše.

Další série problémů se týká otázky, zda lidský mozek funguje jako sériový nebo jako paralelní systém. Myslíme sériově, nebo paralelně?

Známý britský evoluční biolog Richard Dawkins přišel s ideou, že naše mozky jsou kombinací sériového a paralelního přístupu. Dawkins si představoval, že zatímco na procesoru klasického PC běží úloha ve skutečnosti sériově, byť se interface tváří navenek jako paralelní multitasking, u lidského mozku je tomu naopak. Jde o zařízení pracující paralelně, sériovost, tedy ještě jedna úroveň "seskládání", pak snad odpovídá vzniku vědomí nebo alespoň myšlení, předvídání apod.

Robert Sternberg ve své knize Kognitivní psychologie uvádí pokusy, jejichž snahou bylo rozlišit sériové a paralelní zpracování informace. Například si máte zapamatovat určitá slova. Pokud budou čtyři slova trvat stejně dlouho jako dvě, pak je úloha řešena paralelně.

Problém však je v tom, že u paralelního zpracování se stejně čeká na nejpomalejší úlohu, a proto čím více stejných/obdobných úloh, tím delší bude pravděpodobně ta nejdelší z nich - a tím déle bude operace trvat. Předpověď je tudíž stejná jako u zpracování sériového.

Lze matematicky dokázat, že každý sériově prováděný výpočet lze modelovat výpočtem paralelním, byť třeba matematika rovnic bude složitější. Vždy existují paralelní modely, které co do předpovědí simulují ty sériové - a naopak.

Dosavadní experimenty však svědčí pro to, že mozek funguje spíše sériově. Čemuž by snad mohla napovídat i následující skutečnost: Jak si v jedné úvaze, která se na stránkách Computerworldu objevila již asi před deseti lety, povšiml Jiří Peterka, není ani tak problém zkonstruovat paralelní počítač, ale spíše algoritmizovat úlohy tak, aby se tohoto paralelismu dokázalo smysluplně využít (z tohoto důvodu se masivně paralelní systémy typu DNA počítačů uplatní jen pro velmi omezený okruh problémů). Proto se zdá, že přinejmenším na úrovni vědomí pracuje náš mozek skutečně sériově a paralelismus není tím, co by vývoj systémů umělé inteligence mohlo samo o sobě nějak posunout.

Speciální okruh námitek proti koncepci vědomí jako softwaru pak představují problémy vazby vědomí na hardware ("vtělení mysli"). Např. Antonio Damasio ve své i v češtině vyšlé knize Descartesův omyl polemizuje s tvrzením, že lidské myšlení můžeme pochopit izolovaně bez vazby na tělo. Descartovský dualismus, podle kterého jsou mysl i tělo dvě zcela různé entity, je podle Damasia chybný. V průběhu evoluce se naše inteligence vyvinula právě jako nástroj péče o fyzický organismus; doprovodné rysy myšlení jako vůle a emoce jsou podle Damasia procesy silně fyzickými. Současné počítače touto vazbou na svůj hardware ovšem zřejmě nedisponují.

I proti Damasiovým námitkách se ovšem objevily protiargumenty. Většina lidí zabývajících se vývojem systémů umělé inteligence patří spíše k funkcialistům a konekcialistům než k materialistům, tedy předpokládají, že myšlení a vědomí jsou v zásadě výpočetními procesy a jako takové je možné je simulovat i počítačově. Na rozdíl od dualistů sice funkcialisté uznávají, že výpočet musí běžet na nějakém hmotném "substrátu", ale ten pro ně může být víceméně libovolný - lidský mozek stejně jako křemíkový procesor.

Zde se mimochodem objevuje další zajímavý problém: Pokud je proces lidského myšlení možné realizovat výpočtem, může k tomu dojít i dostatečně dlouhým přemísťováním kuliček na kuličkovém počítadle? Byl by abakus po milionu let nadán vědomím? Touto přece jen trochu absurdně působící představou bychom se už ovšem dostali do oblasti sci-fi (nápad je velmi zajímavě rozvinut např. v románu Grega Egana Město permutací).

Spor mezi dualisty, materialisty a funkcialisty/konekcialisty už ovšem spadá spíše do oblasti filozofických spekulací - a Turingův test byl zkonstruován právě jako způsob, kterým se filozofii

vyhnout. Představa, že lidské vědomí je software, který může běžet na hardwaru různého typu, každopádně odkazuje k dalšímu zdroji motivů pro vývojáře umělé inteligence. Do hry pak totiž vstupuje další odvěký lidský sen - naděje na nesmrtelnost.

Nástupci Elizy

Mají různé automatické konverzační programy šanci projít Turiungovým testem? Současné systémy typu Elizy zcela rezignují na schopnost jazyku porozumět, pouze na základě výskytu určitých slov vygenerují svou vlastní frázi. Jejich činnost spadá do kategorie "triků", které jsou ovšem dříve nebo později odhaleny - především proto, že partner v konverzaci zaregistruje vyhýbavost a přeskoky mezi tématy, jimiž program "odvádí řeč jinam". Tímto způsobem lze člověka zmást jenom ve skutečně specifických situacích - Eliza klade podobně jako psychoterapeut občas otázky jakoby bez souvislosti s tím, o čem pacient hovoří ("A co vás napadne, když si vzpomenete na otce?"), vesměs však nějak "papouškuje" to, co člověk řekl v předešlé větě ("Mé manželství bylo nevydařené." - "Proč si myslíte, že vaše manželství bylo nevydařené?"). Parry má zase podobně jako reálný paranoik "kruhové" myšlení vracející se stále k tématu vlastní posedlosti ("Může za to Mafie.").

Zkušenosti s Elizou ukazují, že konverzační programy jsou relativně úspěšnější, pokud je nadáme něčím na způsob osobnosti, vybavíme je sadou názorů a "lidskou" minulostí. Chatovací roboty by neměly pouze pasivně odpovídat, ale být v konverzaci aktivnější, samy přivádět řeč na nová témata. Věrohodnost zvýší, pokud budou disponovat aktuálními informacemi (hudba, film, politika). Samozřejmostí by měl být výstup v podobě gramaticky správných vět. Pouze ve chvíli, kdy se software ocitne úplně "mimo", by měl generovat věty jako Eliza - tedy tak, že použije slovní zásobu svého partnera v debatě a nějak ji přeskládá.

Stále se však pohybujeme na poli triků. Skutečně "myslící" aplikace budou nuceny lidskému jazyku nějak rozumět. A zde zůstává zřejmě největší výzva Turingova testu pro vývojáře: Jakým způsobem vlastně máme konstruovat tyto "rozumějící" programy? Co toto porozumění jazyku přesně obnáší?

Postupný pokrok programů snažících se o splnění Turingova testu ukazuje tzv. Loebnerova cena.

Jak tedy postupovat dál? Je možné, že v tuto chvíli ještě nevlastníme nějakou klíčovou znalost, která by nám umožnila konstruovat programy tak, aby prošly Turingovým testem. Např. Velká Fermatova věta byla dokázána s pomocí matematických koncepcí, které za Fermatova života vůbec nebyly známy. Podobně se může stát, že taktéž Turingův test bude splněn až díky technologiím, o nichž dosud nemáme ani tušení.

Pak je zde ale ještě druhý problém. Zdá se, že jakkoliv splnění Turingova testu garantuje inteligenci, opak se již říct nedá. Proto je docela dobře možné, že se podaří vyvinout systémy, kterým neupřeme vlastní "myšlení" - ale Turingovým testem tyto aplikace přesto neprojdou.

Modifikace Turingova testu

Během více než 50 let, které uplynuly od myšlenky původního Turingova testu, se samozřejmě objevila řada návrhů na jeho modifikace a zdokonalení.

První námitka vychází z toho, že počítač v Turingově testu simuluje pouze jednu z mnoha schopností člověka - schopnost rozumět jazyku a verbálně se vyjadřovat, oboje navíc pouze v psané formě. V modifikaci Stevana Harnada by počítač měl simulovat prostě všechny lidské schopnosti - je jasné, že tento požadavek nesplní "program", ale pouze pokročilý humanoidní robot.

Americký analytický filozof Daniel Dennett hovoří v souvislosti s umělou inteligencí o potřebě "intenciality" a "účelnosti". Lidský intelekt je poháněn vůlí a chtěním, přičemž tyto vlastnosti jsou

důsledkem dlouhé biologické evoluce. Myšlení není mechanické, ale směřuje k nějakému účelu. Počítač by tedy podle Dennetta měl být nejen schopen projít testem, ale měl by o to i sám usilovat, měl by mít své vlastní cíle. Na druhé straně, Turingova původní simulace mířila k otázce, zda stroj může myslet, ne zda může také chtít.

Další verze modifikací se točí kolem tvrzení, že člověka nečiní inteligentním použití jazyka, ale fakt, že jazyk sám v minulosti vyvinul (respektive jej vyvinuli naši předkové). Paul Schweizer proto soudí, že bychom měli nechat počítač, aby sám zkonstruoval jazyk či navrhl pravidla pro hru typu šachů. A nakonec snad i sám přišel s myšlenkou Turingova testu, čímž by jej automaticky splnil :-).

Poměrně bizarní jsou různé "inverzní" Turingovy testy, kdy v roli posuzovatele je sám počítač. Cílem počítače je pak odlišit člověka od jiného počítače, eventuálně dojít k závěru, že sám je nikoliv člověk, ale počítač.

Problémem modifikovaných verzí Turingova testu je obvykle jejich obtížná praktická realizovatelnost. Na rozdíl od původní varianty zůstanou proto zřejmě v rovině experimentů pouze myšlenkových.

Zkrácená verze vyšla v Computerworldu číslo 11/2003.

http://www.transhumanismus.cz/library.php?source=turing50