Picha, Marek

Myšlenkové experimenty ve filozofii : Turingův test

Studia philosophica. 2022, vol. 69, iss. 2, pp. 55-62

ISSN 1803-7445 (print); ISSN 2336-453X (online)

Stable URL (DOI): https://doi.org/10.5817/SPh2022-2-5
Stable URL (handle): https://hdl.handle.net/11222.digilib/digilib.77295
License: CC BY-NC-ND 4.0 International
Access Date: 28. 11. 2024
Version: 20230122

Terms of use: Digital Library of the Faculty of Arts, Masaryk University provides access to digitized documents strictly for personal use, unless otherwise specified.

MUNI ARTS Masarykova univerzita Filozofická fakulta

Digital Library of the Faculty of Arts, Masaryk University
digilib.phil.muni.cz

Studia **philosophica** 69 / 2022 / **2**
https://doi.org/10.5817/SPh2022-2-5

Myšlenkové experimenty ve filozofii: Turingův test

Thought Experiments in Philosophy: The Turing Test

Marek Picha

Úvod

Technologický pokrok znejišťuje. Nemine rok, abychom se nepodivili nad další věcí, kterou stroje dokáží lépe než my. Stroje řídí jiné stroje, rozpoznávají obličeje, překládají, diagnostikují nemoci, skládají hudbu a malují obrazy. Stále se zmenšuje okruh činností, které jsme dříve považovali za výlučně lidské. Je pak zcela pochopitelné, že s každým takovým technickým řešením znovu vyvstává základní otázka po povaze člověka: Čím jsme výjimeční?

Odpověď na tuto otázku je pro nás důležitá z mnoha důvodů. Jako zvídaví lidé hledáme klid ve znalostech; jako nejistí lidé hledáme útěchu v pocitu vlastní nenahraditelnosti; jako lidé usilující o spravedlnost hledáme ty, kteří si spravedlivé zacházení zaslouží. Metafyzická otázka po povaze člověka má hluboké praktické důsledky. Kdykoli dojde na práva, člověk si v našich očích zasluhuje zvláštní přístup, a připisujeme-li lidem privilegia, potřebujeme vědět, koho všeho mezi lidi počítat.

Tradičně se výjimečnost člověka spatřovala v jeho rozumnosti.[1] Vyšší poznávací schopnosti jako reflexivní uvažování, abstrahování a zobecňování byly tím, co nás odlišovalo od zvířat. Rozum byl vlastní jen člověku, rozdíl mezi osobou a tělesem spočíval výhradně ve schopnosti myslet. S rozvojem strojů schopných řešit překvapivě složité intelektuální úkoly ale pochopitelně

1 ARISTOTELÉS. *Etika Níkomachova* I 13. Přeložil Antonín KŘÍŽ. Praha: P. Rezek 1996, s. 46–48.

vyvstaly pochybnosti. Je hranice mezi osobou a tělesem vytyčena správně? Chceme-li upřít lidská práva pokročilým strojům, které vlastnosti jsou exkluzivně lidské? Je něco, co lidé dokáží, avšak stroje principiálně ne?

Výslovně si tyto otázky pokládá René Descartes v *Rozpravě o metodě*. Údajně inspirován zdařilými hydraulickými loutkami za zahrad zámku Saint-Germain uvažuje o rozdílu mezi lidmi, zvířaty a stroji. Píše, že lidské tělo se může hýbat samo a že:

> [...] [t]o se nebude zdát nikterak podivné těm, kteří [...] budou považovat lidské tělo za stroj, jenž, byv vytvořen rukama Božíma, je nesrovnatelně lépe uspořádán a schopen pohybů podivuhodnějších, než kterýkoli stroj jiný, jenž může být vynalezen lidmi.[2]

Najdeme však dva způsoby, píše Descartes, jak takový stroj odhalit. Prvním důvodem je neschopnost stroje dobře používat jazyk. Stroj nebude nikdy schopen dlouhodobě inteligentně jazykově reagovat na podněty, neboť jazykové projevy stroje jsou omezené a jsou výsledkem jiných vnitřních postupů než jazykové projevy člověka. Druhým důvodem je praktická nepružnost stroje. Descartes se domnívá, že každý stroj je specializovaný a dokáže řešit jen dílčí skupinu problémů. Jinak řečeno, vnitřní i vnější konstrukce předurčuje stroj ke zvládnutí jen omezené třídy úkolů. Člověk je naproti tomu vybaven rozumem, který je podle Descarta všestranným nástrojem řešení problémů, dokáže se přizpůsobovat nečekaným situacím a je na rozdíl od stroje flexibilní.

Stejnému problému, tedy principiální odlišitelnosti stroje od člověka, se věnuje i britský matematik Alan M. Turing v textu *Computing Machinery and Intelligence*.[3] Také pro něj důležitou roli při hledání rozumnosti hraje jazyk a schopnost flexibilně reagovat.

V následujícím textu nejprve naznačím Turingovu proměnu Descartovy úvahy do podoby realistického konverzačního testu, poté představím myšlenkový experiment, který se s tímto konverzačním testem úzce pojí. Krátce se budu se věnovat i některým kritickým výhradám vůči tomuto myšlenkovému experimentu.

2 DESCARTES, René. *Rozprava o metodě*. Praha: Svoboda 1992, s. 40–41.

3 TURING, Alan M. Computing Machinery and Intelligence. *Mind*. 1950, **59**(10), s. 433–460.

Imitační hra

Inspirován starší společenskou kratochvílí navrhuje Turing jednoduchý diagnostický nástroj rozumnosti, slepý konverzační test. Uvažuje dialog se zvláštními pravidly:

(i) cílem protagonisty je určit, zda je antagonista člověk, nebo stroj,
(ii) cílem antagonisty je vzbudit zdání, že je člověkem,
(iii) dialog se vede zprostředkovaně.

Tento dialog, *Imitační hra*, domnívá se Turing, pokrývá vše, co hraje při rozpoznávání rozumnosti roli. Bude-li stroj schopen konverzovat tak, že protagonista nedokáže spolehlivě rozlišovat mezi odpověďmi stroje a odpověďmi člověka, musíme stroji připsat rozumnost úplně stejně, jako ji připisujeme jiným lidem.

Podívejme se blíže na některé dílčí charakteristiky Imitační hry.

(ad i) Turing ve svém návrhu vychází z tradiční otázky, zda mohou stroje myslet. Píše, že odpověď na takovou otázku ovšem předpokládá znalost významů slov „stroj" a „myslet". Protože však jde o slova nepřesná, u kterých se nenabízí žádný přijatelný způsob precizování, formuluje raději otázku jinou: Mohou počítače obstát v Imitační hře? Tato nová otázka je podle Turinga s tou původní, tradiční otázkou „těsně spojena".

Povahu tohoto těsného spojení bohužel Turing blíže nepopisuje, což otevírá prostor různým výkladům. Podle jednoho možného výkladu je vztah mezi myšlením a úspěchem v Imitační hře těsný v tom smyslu, že bez myšlení je dlouhodobý úspěch v Imitační hře vyloučen. Myšlení je tedy nezbytnou prerekvizitou konverzačního úspěchu, konverzační úspěch je *spolehlivým* ukazatelem myšlení.[4] Podle druhého možného výkladu je onen vztah těsný proto, že konverzační úspěch je *nejlepším dostupným* ukazatelem myšlení. Dlouhodobý úspěch v Imitační hře bez myšlení vyloučen sice není, nemáme však k dispozici žádný spolehlivější způsob, jak se o myšlení přesvědčit.

(ad ii) Ať už je výklad „těsného spojení" jakýkoli, podstatou Turingovy úvahy je proměna metafyzické otázky *Mohou stroje myslet?* na epistemologickou

4 Tento výklad vedl později ke vzniku tzv. turingovského funkcionalismu, což je pojetí mentálních stavů, jež schopnost uspět v Imitační hře chápe jako jeden z výskytů myšlení. (PUTNAM, Hilary. Minds and Machines. In HOOK, Sidney (ed.). *Dimensions of Minds*. New York: New York University Press 1960, s. 138–164.)

otázku *Kdy máme strojům připsat myšlení?*, kterou dále mění na otázku konstrukční *Mohou stroje obstát v Imitační hře?* Zatímco první, metafyzická otázka podle Turinga neumožňuje žádnou smysluplnou diskusi, v souvislosti se třetí, konstrukční otázkou předpověděl, že na přelomu tisíciletí již budou existovat stroje, které v pětiminutové Imitační hře uspějí v sedmi případech z deseti.[5]

(ad iii) Má-li protagonista připisovat lidství na základě rozumnosti, musí být konverzace nepřímá. Pro Turinga to znamená, že rozhovor musí probíhat prostřednictvím strojově psaných replik (tj. nikoli verbálně, nikoli pomocí rukou psaných poznámek). Jde o formu jednoduchého zaslepení testu. Turing se obává zkreslení protagonistova hodnocení podružnými okolnostmi; jak on sám píše, chce „vést ostrou hranici mezi fyzickými a intelektuálními schopnostmi člověka".[6]

Expozice: Turingův test

Neexistuje žádné široce přijímané vymezení myšlenkového experimentu. Definiční návrhy se pohybují na škále od velmi volných, podle kterých je myšlenkovým experimentem jakákoli trochu vyprávěcí pasáž textu, po velmi striktní, jež chápou myšlenkové experimenty jako konkrétní logické konstrukce.[7] Budu zde chápat myšlenkový experiment jako sadu pokynů, které adresátovi říkají, co si má představit a co má v dané představě sledovat.[8] Tyto pokyny mohou být vyjádřeny názorným jazykem a mohou být plné detailů, stejně tak ale mohou být popsány velmi stroze a mohou klást velký důraz na adresátovu tvořivost a imaginaci. Podstatné je, aby šlo o pokyn, že se má adresát snažit něco *zjistit pomocí představivosti*.

Jaký myšlenkový experiment je popsán v Turingově textu? Mělo by být patrné, že samotná Imitační hra podle výše přijatého vymezení myšlenkovým experimentem není. Turing navrhuje *reálný test* konverzačních schopností, předkládá sadu realistických pokynů, co se má skutečně odehrát. Myšlenkový

5 Od roku 1990 do roku 2019 byla každoročně vypisována tzv. Loebner Prize pro algoritmy hrající omezenou verzi Imitační hry. Pozoruhodná byla především vysoká úspěšnost relativně jednoduchých algoritmů.

6 TURING, Alan M. Computing Machinery and Intelligence. *Mind*. 1950, **59**(10), s. 434.

7 SORENSEN, Roy A. *Thought Experiments*. Oxford a New York: Oxford University Press 1992.

8 PICHA, Marek. *Kdyby chyby: epistemologie myšlenkových experimentů*. Olomouc: Nakladatelství Olomouc 2011.

experiment je v textu vyjádřen až *v souvislosti* s Imitační hrou, a to následovně: „Jsou představitelné digitální počítače, které by si slušně vedly v Imitační hře?“[9]

Tuto větu můžeme chápat jako pokyn k řešení problému pomocí imaginace. Turing nás vybízí k tomu, abychom si představili stroj, který je schopen dlouhodobé inteligentní konverzace – s tím, že sledovaným parametrem má být samotná vnitřní soudržnost takové představy. Nemáme hledat odpověď na otázku, zda by takový stroj měl svědomí či si zasloužil určitá práva; máme se ptát, zda tato naše představa při pečlivějším promýšlení neodhalí nějaký skrytý nesoulad. Tentýž myšlenkový experiment můžeme formulovat i jinak, možná názorněji: Představte si, že by existovaly stroje, které by při konverzaci dokázaly dobře předstírat, že jsou lidé. Existuje něco, co by takové stroje prozradilo?

Všimněme si, že uvedený myšlenkový experiment zkoumá něco jiného než schopnost stroje obstát v konverzaci – k tomu přece máme samotnou Imitační hru; zkoumá naši představu o hranicích strojově řešitelných intelektuálních úkolů. Stejně jako dříve Descartes nás Turing vybízí, abychom se pokusili nalézt nějakou mentální činnost, jíž nelze automatizovat. Pokud žádný takový strojově neřešitelný intelektuální úkol nenajdeme, nemáme dobrý důvod upírat strojům rozum.

Reakce

Turingův text svým důrazným trváním na objektivních kritériích připisování rozumnosti nechá málokoho chladným. Během více než sedmi dekád od jeho publikování se objevilo mnoho reakcí, jež Imitační hru chápou jako mezní a spolehlivý test myšlení, či dokonce jako finální konstruktérský cíl; objevilo se pochopitelně i mnoho reakcí, jež s různými pasážemi klíčového textu polemizují.[10] Sám Turing se přímo v *Computing Machinery and Intelligence* věnuje námitkám,[11] které velmi vhodně rámují téměř celou další kritickou

9 TURING, Alan M. Computing Machinery and Intelligence. *Mind*. 1950, **59**(10), s. 442.

10 Některé z těchto reakcí mají dokonce podobu samostatných myšlenkových experimentů, viz např. KIRK, Robert. Sentience and behaviour. *Mind*. 1974, **83**(1), s. 43–60; BLOCK, Ned. Troubles with functionalism. *Minnesota Studies in the Philosophy of Science*. 1978, **9**, s. 261–325; SEARLE, John. Minds, brains, and programs. *Behavioral and Brain Sciences*. 1980, **3**(3), s. 417–457.

11 Stejně jako před ním Descartes při publikování svých *Meditací o první filosofii* (1641).

diskusi nad Turingovým testem. Tyto námitky lze rozdělit do tří kategorií: na námitky, jež kritizují důsledky připsání rozumu strojům, námitky, jež kritizují předpoklady Turingova testu, a námitky, jež poukazují na výjimečné schopnosti lidské mysli.

Námitky z důsledků hledají nepřijatelné dopady, jež by mělo připsání rozumnosti strojům. Turing se věnuje dvěma takovým námitkám. Začíná teologickou, podle které je připsání rozumnosti strojům v rozporu s náboženskou doktrínou přiznávající duši výhradně lidem, poté pokračuje dystopickou námitkou, jež cílí na blíže neurčený neblahý dopad existence myslících strojů. Jakkoli mohou být tyto hlasy za zvláštních okolností relevantní, nehrají žádnou roli při promýšlení Turingova testu jako takového; týkají se výhradně Imitační hry. Teologická námitka tvrdí, že Imitační hra je špatné kritérium, neboť neumožňuje rozlišit mezi člověkem a bezduchým strojem, dystopická námitka tvrdí, že Imitační hra je špatné kritérium, neboť nás nutí připsat rozum robotům, což nechceme.[12]

Námitky z předpokladů odmítají některá východiska širší Turingovy úvahy. Lidský mozek, tvrdí jedna z takových námitek, je svojí výpočetní architekturou odlišný od stroje popsaného v Turingově textu. Lidský mozek realizuje vysoce distribuovanou neuronovou síť, zatímco Turing ale bere do úvahy jen konečné automaty. Tato námitka, jakkoli věcně správná, nemá žádný dopad nejen na Turingův test, ale dokonce ani na Imitační hru. Rozdíl ve výpočetní architektuře se totiž nijak v konverzaci neprojeví, reakce systému s distribuovanou výpočetní architekturou budou stejné jako reakce systému s architekturou klasickou.

Druhá námitka, které Turing věnuje překvapivě značnou pozornost, je námitka vycházející z telepatie. Kdyby protagonista Imitační hry dokázal na dálku odlišit mentální monolog člověka od vnitřního monologu stroje, mělo by to vliv na výsledek Turingova testu. Nebyli bychom odkázáni na behaviorální projevy systému, měli bychom přímý vhled dovnitř. V posledku by to ale znamenalo jen tolik, dodává Turing, že Imitační hra by se musela hrát v „telepatiitěsných" místnostech.[13]

Třetí námitka z předpokladů se týká vědomé zkušenosti: člověk podle ní vnímá skutečné kvality a prožívá opravdové emoce, stroj nikoli. Jazykové cho-

12 Turing obě námitky hodnotí jako nerelevantní a zaslepeně antropocentrické. Podstatou teologické námitky je apriorní odmítnutí Imitační hry, podstatou dystopické námitky je poznávací chyba tzv. toužebného přání.

13 Případně bychom byli nuceni uvažovat stroje, které dokáží imitovat nejen konverzaci, ale i údajnou telepatií spolehlivě zjišťovanou vnitřní aktivitu.

vání zapříčiněné kvalitami či emocemi je potom odlišné od chování zapříčiněného automatickými procesy. Tato námitka nedává v kontextu Turingova testu dobrý smysl bez toho, aniž by bylo blíže určeno, v čem se odlišně zapříčiněné projevy budou nutně rozcházet.

Námitky z neschopností jsou sadou výhrad, které se přímo týkají Turingova testu. Poukazují na různé intelektuální úkoly, jež je údajně nemožné automatizovat. Turing zmiňuje úkoly týkající se charakterových rysů (laskavost, nápaditost, iniciativnost, přátelskost, kreativita, originalita) či schopností (smysl pro humor, cit pro dobro, schopnost vychutnat si jahody, schopnost učit se, schopnost mýlit se, schopnost parafrázovat, schopnost zamilovat se, vkus, schopnost improvizovat a jednat neformálně).

Jeho reakce na tento typ výhrad je taková, že každý z uvedených příkladů představuje problém konstrukční, nikoli však problém principiální. Stroj může být podle Turinga sestaven tak, aby generoval náhodné chyby, referoval o svých vnitřních stavech, učil se a měl vždy dostatek paměti; na základě těchto bazálních schopností pak stroj zvládne vše ostatní.

Mezi námitky z neschopnosti lze řadit i tzv. matematickou námitku, která říká, že stroje nedokáží překonat formální výpočetní omezení daná algoritmickou nerozhodnutelností některých problémů. Podle této námitky existuje třída úkolů, jež jsou řešitelné lidským intelektem, nikoli však strojovými výpočty, typickým příkladem je řešení tzv. problému zastavení.

Turingova reakce je zde jiná než na předchozí sadu námitek z neschopností. Nezpochybňuje skutečnost, že výpočetní schopnosti strojů jsou omezené, zpochybňuje předpoklad, že se daná omezení netýkají lidí. To, že jsou lidé schopni rozpoznat určité strojově neřešitelné problémy, ještě neznamená, že tytéž problémy dokáží i vyřešit. Není důvod předpokládat, že chování stroje bude tváří v tvář těmto matematickým úkolům jiné než chování člověka.

Závěr

Otázka po výjimečnosti člověka je stará snad jako člověk sám, nová není ani myšlenka konverzačního testu lidství. To, co z Turingova testu učinilo možná vůbec nejznámější myšlenkový experiment současnosti, je fakt, že jsme svědky přerodu filozofické fikce ve skutečnost. Na základě reálného konverzačního testu vznikl hypotetický příklad vybízející ke spekulaci, ten se ale s rozvojem umělé inteligence nezvykle brzy vrátil zpět k reálným kořenům.

Turing se domníval, že zhruba v této době již budeme bez nadsázky hovořit o strojovém myšlení.[14] Možná tam ještě úplně nejsme, ale jsme blízko.

doc. PhDr. Marek Picha, Ph.D.

Katedra filozofie, Filozofická fakulta, Masarykova univerzita
Arna Nováka 1, 602 00 Brno, Česká republika
istvan@mail.muni.cz

14 TURING, Alan M. Computing Machinery and Intelligence. *Mind*. 1950, **59**(10), s. 442.

Toto dílo lze užít v souladu s licenčními podmínkami Creative Commons BY-NC-ND 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode). Uvedené se nevztahuje na díla či prvky (např. obrazovou či fotografickou dokumentaci), které jsou v díle užity na základě smluvní licence nebo výjimky či omezení příslušných práv.