Lukáš Hozda, R7.A

Divoké výstřelky umělé umělé inteligence

Umělá inteligence (AI) je poměrně čerstvá technologie, jednak proto, protože nám dříve scházel výpočetní výkon pro provedení mnoha pokročilých úloh, které AI zahrnuje, a také proto, že jsme v měli v počítačové vědě historicky tendenci posunovat význam slov ,,umělá inteligence''.

V raných počátcích své umělá inteligence se výzmaná část akademické komunity se soustředila kolem osoby Johna McCarthyho.[1] Ten mimo jiné také zavedl v roce 1955 termín **Artifical Intelligence**. Jeho výzkumný tým například považoval programy na hraní šachů a optické rozpoznávání textu za umělou inteligenci, a přestože není možné tyto typy programů z umělé inteligence dnes úplně vyřadit, není to úplně to, co se pod pojmem umělá inteligence vybaví dnešnímu výzkumníkovi.

Termín umělá inteligence je tedy poměrně těžké definovat. V současnosti je výzkum AI definován jako studium ,,inteligentních agentů", což je jakékoliv zařízení nebo program, který vnímá své prostředí a činí akce, které mají největší šanci uspět v dosažení jeho cílů.[2]

Umělou inteligenci dělíme různými způsoby, ale možná nejdůležitější je dělení na slabou a silnou inteligenci.[3] Se slabou umělou inteligencí se dnes již setkáváme v mnoha oblastech života. Tento typ inteligence se zaměřuje na jeden konkrétní úkol a jeho efektivní provedení. Do této kategorie patří například neurální sítě na řízení provozu nemocnic, taxi služeb ve světových velkoměstech, Siri, Cortana nebo Google Assistent. Také všechny příklady AI, o kterých budu mluvit dále v tomto textu jsou příklady slabé umělé inteligence a konkrétně se primárně jedná o komunikační boty.

Silná umělá inteligence má za úkol simulovat lidské myšlení a nápodoba lidského chování. Tato kategorie je primárně předmětem sci-fi a přestože na jejím vývoji pracují světový lídři umělé inteligence, nemáme zatím žádné funkční prototypy. Sestrojit takové AI je náročné jednak z hlediska návrhu, tak z hlediska implmenetace – simulovat lidský mozek je velmi náročné na výkon a zefektivnit tento proces je velmi složité, ne-li v současnosti nemožné. Reálné příklady tedy není jednoduché najít, ale pokud nahlédneme do fikčních světů, lze za silnou umělou inteligenci považovat například HALa ze Space Odyssey, Terminátora nebo androidy ze série Westworld.

Umělá inteligence nemá jednoduchou a plynulou historii. Je plná kontroverzí a překážek, debat o jejích etických implikací a zatížená limitacemi ze strany technologie a peněz. Můžeme rozlišit období rozkvětu a zájmu o AI a odbobí úpadku a stagnace. Pro tato období se v angličtině používá termín ,,AI winter" - zima umělé inteligence. Takové zimy se objevily zejména okolo let 1966, 1969, 1974 a ta vůbec z nich probíhala od druhé poloviny osmdésátých do první poloviny devadesátých let. Jejich příčinnou byla zpravidla přehnaná

---

1. Project JMC Team, *John McCarthy,* Standford CS (2019). http://jmc.stanford.edu/.
2. Poole, Mackworth, Goebel, *Computational Intelligence: A Logical Approach,* Oxford University (1998).
3. Jeff Kerns, *Whatâs the Difference Between Weak and Strong AI?* Informa USA (16.2.2017). https://www.machinedesign.com/robotics/what-s-difference-between-weak-and-strong-ai.

očekávání laické veřejnosti a nedostatek (popř. úmyslné snížení) finančních prostředků pro výzkum.

V poslední dekádě je ovšem výzkům umělé inteligence opět na vzestupu. Investoři jsou ochotní poskytnout finance a naše současná technologie umožňuje posunout hranice AI dál než kdykoliv předtím. Objevuje se mnoho projektů, o kterých by se nám před pár dekádami ani nesnilo.

Protože je umělá inteligence stále poměrně nový obor, různá AI nefungují správně nebo vykazují jiné chování než jsme očekávali před jejich spuštěním. Žádné nebezpečí nám však nehrozí, tyto deviace od očekávaného chování patří spíše k něčemu, co můžeme považovat jako vtipné výstřelky, nebo jako pouhé selhání splnit úkol.

Prvním příkladem, který bych zmínil je chatovací bot zvaný Cleverbot vytvořený Rollem Carpenterem. Cleverbot je chatovací bot, který se při komunikaci učí. To definuje jeho osobnost, zdánlivou znalost a také jazyky, kterými bot mluví. Když si ovšem člověk Cleverbota vyzkouší, pokud možno v angličtině, kde je nejvyvinutější a do které se bot periodicky vrací, zjistíme, že je občas připomíná pomateného vtipálka. Jeho odpovědi jsou sarkastické, občas v nich nalézáme pravopisné chyby, nebo se bot snaží vyhnout otázkám. Také můžeme pozorovat tendenci bota ptát se uživatele, jestli je člověk, nebo jej označovat také za bota.

Tyto vlivy jsou nepochybně způsobeny manipulací uživatelů. Když budeme botovi něco mnohokrát opakovat, bot to eventuálně začne opakovat po nás, respektive začlení to do svého repertoáru. Otázky a výroky týkající se umělé inteligence a člověka mohou být projevem toho, že když s botem komunikujeme, snažíme si ověřit, jestli se jedná opravdu o bota, jestli si je bot vědom toho, že je umělá inteligence, a nebo chceme bota přinutit, aby si uvědomil realitu svojí existence. Nicméně Cleverbot a další jsou stále pouze slabá umělá inteligence, jejich specifickým úkolem je simulovat lehkou konverzaci a veškeré snahy dostat z bota výstup, co tento úkol přesahuje, jsou zbytečné.

Podobným a však ne identickým případem je kauza Tay AI, twitterového bota od Microsoftu.[4] Účelem Tay bylo chovat se na Twitteru jako náctiletá dívka a, podobně jako Cleverbot, učit se z interakcí s uživateli této sociální sítě. Během několika hodin se však AI zcela vymklo kontrole a zaujalo poněkud neonacistický postoj. Tweety Tay se staly rasistické, sprosté, její výroky opěvovaly Adolfa Hitlera a byly proloženy radikálními komentáři o spoustě kontroverzních témat. Někteří žurnalisté se ze začátku domnívali, že se jedná o důkaz toho, jak vypadá komunita Twitteru, ale pravda je opět jednodušší a méně kontroverzní.

Bot ,,uvěří" všemu, co je mu dostatečně opakováno. Poměrně malé množství lidí tedy bota může takříkajíc otrávit a ovlivnit jeho zdánlivé názory jakýmkoliv směrem. Tak se i stalo. To, že bylo Tay AI tak ovlivnitelné z něj udělalo velkou atrakci jak pro čtenáře, tak pro manipulátory.

V jejím krátké životě Tay přidala desetitisíce tweetů s různou úrovní nekorektního obsahu. Velmi rychle se z celé situace stala PR katastrofa pro Microsoft. Vývojáři se pokusili ,,opravit" její názory (slavný tweet "I love feminism now"), ale tyto snahy nebyly dostatečné pod tlakem uživatelů, jejichž počet exponenciálně vzrůstal.

---

4. Wikipedia contributors, *Tay (bot),* Wikipedia, The Free Encyclopedia. (6.5.2019). https://en.wikipedia.org/w/index.php?title=Tay_(bot)&oldid=895746584.

Bot byl poprvé vypnut po pouhých 16 hodinách provozu. Po několika dnech se poněkud promazaná Tay vrátila zpět na Twitter, ale její obsah se velmi rychle obrátil ke konzumaci drog a negativním černošským stereotypům. To však netrvalo dlouho, protože bot patrně narazil na chybu a do nekonečna opakoval "You are too fast, please take a rest" dokud nebyl vypnut, tentokrát již (prozatím) navždy.

Je možné se ptát, zda-li by člověk ve stejné situaci mohl být ovlivněn stejně jako Tay AI, které bylo prakticky perfektní tabula rasa. Naše umělá inteligence je zatím stále moc primitivní, ale jednou třeba bude možné využívat AI k simulaci a studiu lidského chování, podobně jako dnes používáme genetické algoritmy k simulaci evoluce a přirozeného výběru.

Existují případy, kde u umělé inteligence pozorujeme evoluci, který jsme nečekali. Jedním z těch známější je vlastní jazyk Facebookových vyjednávacích botů.[5] Tito boti, pojmenovaní Bob a Alice měli zaúkol spolu vyjednat rozdělení pokladu v podobě míčů, čepic a knih. Každý z botů má vlastní oblíbené předměty a snaží se zvolit efektivní jazyk, který by mu umožnil dosáhnout maximálního počtu kýžených předmětů.

Dokud výzkumníci trénovali boty pomocí připraveného vstupu založeného na výňatcích z lidské komunikace, nedělo se nic zvláštního. Ve vývoji umělé inteligence se však také používá technika zvaná Generative Adversarial Network (zkratka GAN). Ke zdokonalení neurální sítě zde využíváme soutěže – postavíme dvě neurální sítě proti sobě a necháme je soupeřit o dosažení cíle. To umožňuje rychlé a efektivní zdokonalení dané sítě. U těchto botů se však toto zdokonalení vydalo směrem, který jsme nečekali.

Pro tyto boty je řeč pouze vstup a výstup ve formě textu s jistým bodovým ohodnocením. Nedívají se na něj jako my a nejsou svazováni lidským vnímáním. Díky tomu došlo k drastické změně jazyka, kde si boti „vyzobali" jenom efektivní fráze. To dalo zaznik výrokům jako například: "balls have zero to me to me to me to me to me to me to me to me to" nebo "I can i i everything else".

Tyto zdánlivě nesmyslné věty jsou výsledkem jednak vytvářením zkratek a zjednodušováním, které je přirozené i lidem, a jednak přidáním významu k opakování slov. Například místo toho, aby bot řekl: "I want 3 balls" řekne "I want the the the balls" – opakování členu se stalo vyjádřením pro počet. Pro člověka by bylo se takhle vyjadřovat problematické, ale stroj tento problém nepociťuje.

To nám dokazuje, že jazyk záleží na podmínkách, podobě v jaké je přenášen a mnoha dalších faktorech, které třeba ani neznáme. Komunikace s člověkem zařídila, že tyto faktory byly udržovány člověkem implicitně, ale jakmile byl spuštěn GAN, pracovali boti jenom s tím, co mají.

Všechny tyto příklady šílených umělých inteligencí jsou důkazem toho, že o lidské komunikaci toho stále spoustu nevíme, podobně jako o procesu učení a získávání znalostí. Naše simulace jsou daleko od dokonalasti, ale každý špatný pokus nás posouvá o krok blíže k funkčním prototypům. Můžeme se ptát: Proč se náš jazyk vyvinul tak, jak se vyvinul? Proč není naše zkratkovitost tak dokonalá jako ta u botů?

---

5. Roman Kucera, “The truth behind Facebook AI inventing a new language,” *Towards Data Science*, Medium (14.6.2017).

Za zmínku také nepochybně stojí například obrázky zvířat vygenerované umělou inteligencí. Ty nám ukazují snahu umělé inteligence interpretovat definující znaky objektů. Mnoho vygenerovaných obrázků koček sice trochu kočky připomíná, ale mívají špatné proporce nebo jim nesedí počet končetin.[6] To znamená, že i když AI zná dostatek znaků nutných pro rozpoznání, není jeho pochopení dostatečné pro replikaci. Nicméně, většina neumělecké společnosti také není schopna vytvořit obraz kočky.

Exitují ale už neurální sítě, které generují lidské obličeje, které je těžké rozilišit od opravdový, například síť projektu ThisPersonDoesNotExist.com lze využít k vygenerování velmi přesvědčivé profilové fotky. Můžeme očekávat, že v budoucnosti se bude vývoj umělé inteligence dále posouvat a s tím i naše poznání problematiky fungování mysli a procesu učení.

---

6. Jason Johnson, "AI Generated Thousands of Creepy Cat Pictures," *Motherboard*, Vice.com (14.6.2017).