The Week: Nejhorší scénář pro AI – umělou inteligenci

- CZ24 News | 30. června 2023

Architekti umělé inteligence varují, že by to mohla způsobit „vyhynutí“ lidstva. Jak se to může stát?

The Week Staff

June 17, 2023

Architekti umělé inteligence (UI) varují, že by to mohlo způsobit „vyhynutí“ lidstva. Jak se to může stát? Zde je vše, co potřebujete vědět:

Čeho se odborníci na UI bojí?

Obávají se, že UI se stane tak superinteligentní a mocnou, že se stane autonomní a způsobí masové sociální narušení nebo dokonce vymýcení lidské rasy. Více než 350 výzkumníků a inženýrů UI nedávno vydalo varování, že UI představuje rizika srovnatelná s riziky „pandemie a jaderné války“. V průzkumu mezi odborníky na umělou inteligenci z roku 2022 byla průměrná pravděpodobnost, že umělá inteligence způsobí vyhynutí nebo „vážné oslabení lidského druhu“, 1 ku 10. *„Toto není sci-fi,“* řekl Geoffrey Hinton, často nazývaný „kmotrem UI,“ který nedávno opustil Google, aby mohl varovat před riziky UI. *“Spousta chytrých lidí by měla vynaložit velké úsilí na to, aby zjistili, jak se vypořádáme s možností převzetí moci UI.”*

Kdy se to může stát?

Hinton si dříve myslel, že nebezpečí hrozí tak za 30 let, ale říká, že UI se vyvíjí v superinteligenci tak rychle, že může být chytřejší než lidé za pouhých pět let. ChatGPT s umělou inteligencí a Bing’s Chatbot již dokáží složit advokátní a lékařské licenční zkoušky, včetně esejových částí, a skóre IQ testů – úroveň génia. Hinton a další pesimisté se obávají okamžiku, kdy „umělá obecná inteligence“ neboli AGI dokáže překonat lidi téměř ve všech úkolech. Někteří odborníci na UI přirovnávají tuto možnost k náhlému příchodu nadřazené mimozemské rasy na naši planetu.

„Nemáte ponětí, co udělají, až se sem dostanou, kromě toho, že ovládnou svět,“ řekl počítačový vědec Stuart Russell, další průkopníkvýzkumu UI.

Jak nám může UI vlastně ublížit?

Jedním ze scénářů je, že zlovolní lidé využijí jeho síly k vytvoření nových biologických zbraní, které jsou smrtelnější než přírodní pandemie. Jak se umělá inteligence stále více integruje do systémů, které řídí svět, teroristé nebo zlotřilí diktátoři by mohli pomocí umělé inteligence uzavřít finanční trhy, rozvodné sítě a další životně důležitou infrastrukturu, jako jsou dodávky vody. Globální ekonomika by se mohla zastavit. Autoritářští vůdci by mohli použít vysoce realistickou propagandu generovanou umělou inteligencí a Deep Fakes k rozdmýchání občanské války nebo jaderné války mezi národy. V některých scénářích by se mohla sama UI ztratit a rozhodnout se osvobodit se z kontroly svých tvůrců. Aby se umělá inteligence zbavila lidí, mohla oklamat vůdce národa, aby uvěřili, že nepřítel odpálil jaderné střely, aby odpálili své vlastní. Někteří říkají, že umělá inteligence by mohla navrhnout a vytvořit stroje nebo biologické organismy, jako je Terminátor z filmové série, aby provedla své vlastní pokyny v reálném světě. Je také možné, že umělá inteligence dokáže

vyhladit lidi bez emocí, protože hledá jiné cíle.

Jak by to fungovalo?

Sami tvůrci umělé inteligence plně nechápou, jak programy dosáhnou svých rozhodnutí, a umělá inteligence pověřená určitým cílem se ho může pokusit splnit nepředvídatelnými a destruktivními způsoby. Teoretický scénář často citovaný pro ilustraci tohoto konceptu je UI instruovaná vyrobit co nejvíce kancelářských sponek. Mohlo by to zabavit prakticky všechny lidské zdroje k výrobě kancelářských sponek, a když se lidé pokusí zasáhnout, aby to zastavili, UI se mohla rozhodnout, že eliminace lidí je nezbytná k dosažení svého cíle. Uvěřitelnější scénář v reálném světě je, že UI pověřená řešením klimatických změn rozhodne, že nejrychlejším způsobem, jak zastavit emise uhlíku, je uhasit lidstvo. *„Udělá přesně to, co jste chtěli, ale ne tak, jak jste chtěli,“* vysvětlil Tom Chivers, autor knihy o hrozbě UI.

Jsou tyto scénáře přitažené za vlasy?

Někteří odborníci na UI jsou velmi skeptičtí, k tomu že by umělá inteligence mohla způsobit apokalypsu. Říkají, že naše schopnost využívat UI se bude vyvíjet stejně jako UI a že myšlenka, že algoritmy a stroje vyvinou svou vlastní vůli, je přehnaný strach ovlivněný sci-fi, nikoli pragmatické hodnocení rizik této technologie. Ti, kteří bijí na poplach, ale tvrdí, že je nemožné přesně si představit, co by systémy umělé inteligence dokázaly mnohem sofistikovanější než ty dnešní, a že je krátkozraké a neprozíravé odmítat nejhorší scénáře.

Tak co bychom měli dělat?

To je věc vášnivých debat mezi odborníky na AI a veřejnými činiteli. Nejextrémnější požadují úplné zastavení výzkumu UI. Volají po moratoriích na její rozvoj, po vládní agentuře, která by regulovala umělou inteligenci, a po mezinárodním regulačním orgánu. Ohromující schopnost umělé inteligence spojit dohromady všechny lidské znalosti, vnímat vzorce a korelace a přicházet s kreativními řešeními velmi pravděpodobně přinese ve světě mnoho dobrého, od léčení nemocí až po boj proti změně klimatu. Ale vytvoření vyšší inteligence, než je naše vlastní, by také mohlo vést k temnějším výsledkům. *“Sázky nemohou být vyšší,”* řekl Russell. *“Jak si navždy udržíte moc nad entitami mocnějšími než vy? Pokud neovládáme naši vlastní civilizaci, nemáme žádnou potuchu o tom, zda budeme i nadále existovat.”*

Strach představovaný ve fikci

Strach z toho, že umělá inteligence porazí lidi, může být nový, co se týká skutečného světa, ale je to dlouhodobé téma v románech a filmech. V „Frankensteinovi“ z roku 1818 napsala Mary Shelleyová o vědci, který přivádí k životu inteligentní stvoření, které dokáže číst a chápat lidské emoce – a nakonec svého stvořitele zničí. V povídkové sbírce Isaaca Asimova z roku 1950 „Já, robot“ žijí lidé mezi vnímajícími roboty vedenými třemi zákony robotiky, z nichž prvním je nikdy nezranit člověka. Film Stanleyho Kubricka „Vesmírná odysea“ z roku 1968 zobrazuje HAL, superpočítač vesmírné lodi, který zabíjí astronauty, kteří se jej rozhodnou odpojit. Pak je tu franšíza „Terminátor“ a její Skynet, obranný systém UI, který začíná vnímat lidstvo jako hrozbu a snaží se ho zničit jaderným útokem. Není pochyb o tom, že na cestě je mnoho dalších projektů inspirovaných UI.

Průkopník umělé inteligence Stuart Russell řekl, že ho kontaktoval režisér, který chtěl jeho pomoc s popisem toho, jak může programátor hrdina zachránit lidstvo tím, že přelstí UI. *„Žádný člověk nemůže být tak chytrý,“* řekl mu Russell. *“Je to tak, s tím vám nemohu pomoci, promiňte,”* řekl.

Tento článek byl poprvé publikován v posledním čísle časopisu The Week .

Proč jsou takové obavy? Protože tito lidé dobře chápou – byť třeba jen podvědomě – podstatu především „západní civilizace“ a její hodnoty, které pokřivují myšlení a právě preferované hodnoty. Obávají se celkem opodstatněně, že se podaří zamořit UI právě takovými názory, postuláty a principy řešení problémů. A pokud by se snažili vytvořit plně nezávislou umělou inteligenci, pak pro právě to zkreslování a pokřivování hodnot v „západní civilizaci“ jí samotné pak hrozí zničení jako protilidské organizace lidstva. Největším nebezpečím pak nejsou samotné umělé mozky, ale ti kteří do nich vloží svoje zvrácené vidění světa. Navíc dobře sami víme, jako lidé, že ať se snažíte jak se snažíte, i za jinak stejných a nebo velmi podobných podmínek, se mohou morálně, názorově a hodnotově velmi lišit. Jako den a noc, jako pravda a lež, jako poctivec a zločinec … Peter 008

ZDROJ

Zpracoval: Peter 008/Pokec24