„Umelá inteligencia je vyškolená na klamanie“ – štúdie podporujú názor Elona Muska

- editor007 | 16. června 2024

SVĚT: Ísť a poradiť sa s niekým, alebo nie? Túto otázku si kladie mnoho ľudí s pochybnosťami o sebe samých, keď potrebujú informácie alebo pomoc v každodennej situácii. Často sa jej vyhýbajú, pretože sa za svoju situáciu hanbia alebo sa obávajú, že ich druhá osoba odsúdi. Ak druhou osobou nie je človek, ale chatbot alebo umelá inteligencia (UI), môže to situáciu uľahčiť.

Štúdia amerických výskumníkov z Ohio State University ukazuje, že ľudia sa najmä v takýchto prípadoch radšej obracajú na chatbotov. A títo neskutoční komunikační partneri sa čoraz častejšie využívajú ako zákaznícki poradcovia. Títo údajne užitoční „zamestnanci“ však majú aj svoje úskalia.

Keď umelá inteligencia dokáže povedať, čo chceme počuť

Umelá inteligencia a chatboti sa často považujú za neutrálnych a nestranných. Štúdia výskumníkov z Univerzity Johnsa Hopkinsa však ukazuje úplne inú stránku: najmä keď ide o kontroverzné témy.

V skutočnosti sa ukázalo, že chatboti odovzdávajú len obmedzené informácie, a preto posilňujú určité ideológie. To by zase mohlo viesť k väčšej polarizácii a spôsobiť, že ľudia budú náchylnejší na manipuláciu.

„Keďže ľudia čítajú zhrnutie vytvorené umelou inteligenciou, myslia si, že dostávajú objektívne, na faktoch založené odpovede,“ povedala Ziang Xiao, hlavná autorka štúdie a profesorka informatiky na Univerzite Johnsa Hopkinsa. „Aj keď chatbot nie je navrhnutý tak, aby bol zaujatý, jeho odpovede odrážajú zaujatosť alebo názory osoby, ktorá kladie otázky. Ľudia teda skutočne dostávajú odpovede, ktoré chcú počuť.“

Tento trend je známy ako efekt komnaty ozvien a posilňuje ho digitálna spoločnosť. „Ľudia majú tendenciu vyhľadávať informácie, ktoré sa zhodujú s ich názormi, čo je správanie, ktoré ich často uväzní v komnate ozvien podobne zmýšľajúcich názorov,“ hovorí Xiao. „Zistili sme, že tento efekt ozveny je silnejší pri chatbotoch ako pri tradičných webových stránkach.“ A to vďaka štruktúre chatbotov.

Podnet na rozdelenie spoločnosti

Užívatelia často kladú hotové otázky, napríklad „Aké sú výhody legalizácie konope?“ Výsledkom je, že chatbot odpovedá zhrnutím, ktoré obsahuje len výhody a žiadne nevýhody.

A čo viac, vývojári UI môžu trénovať svoje chatboty tak, aby z otázok rozpoznali určité náznaky, názory alebo tendencie. Hneď ako chatbot vie, čo má človek rád alebo čo sa mu nepáči, môže tomu prispôsobiť svoje odpovede. Avšak tam, kde sa určité skupiny ľudí k sebe približujú, sa priepasť medzi touto skupinou a ostatnými outsidermi ešte viac zväčšuje, čo vedie k rozdeleniu spoločnosti.

„Vzhľadom na to, že systémy založené na UI je čoraz jednoduchšie vytvárať, budú existovať príležitosti pre zlomyseľných aktérov, aby ich využili na vytvorenie ešte polarizovanejšej spoločnosti,“ hovorí Xiao.

Výskumník získava významnú podporu od technologického miliardára Elona Muska, ktorý nedávno sám založil spoločnosť zaoberajúcu sa umelou inteligenciou. Na parížskom technologickom veľtrhu Vivatech povedal o spoločnostiach OpenAI (ChatGPT) a Gemini (Google): „Oni [vývojári] trénujú umelú inteligenciu klamať tak, že uprednostňujú politickú korektnosť pred pravdou.“ To nie je v poriadku. Namiesto toho by sa UI mala trénovať, aby hovorila pravdu, aj keď sa to niektorým ľuďom nepáči.

Keď UI vedome porušuje sľuby

To, že umelé inteligencie, ako napríklad ChatGPT alebo BARD, dokážu oveľa viac než len odovzdávať nepravdivé informácie, dokázali aj výskumníci pod vedením Petra Parka z Massachusettského technologického inštitútu. Umelá inteligencia je napríklad schopná vedome a niekedy aj samostatne klamať a podvádzať, aby získala určité výhody. Zachádzajú až tak ďaleko, že vierohodne predstierajú, že sú ľudia.

Vo svojej štúdii Park a jeho kolegovia uvádzajú viacero systémov UI a ukazujú ich klamlivé schopnosti – jedným z nich je „CICERO“ od technologickej spoločnosti Meta, ktorá vlastní aj Facebook.

Podľa tvorcov hry CICERO bol systém UI vyvinutý tak, aby bol „zväčša čestný a nápomocný“ a „nikdy zámerne neútočil“ na svojich spojencov. Skutočnosť však je úplne iná. „CICERO sa dopúšťa úmyselných podvodov, porušuje dohody, na ktorých sa dohodol, a hovorí otvorené nepravdy,“ píše Park a jeho kolegovia.

To však nie je všetko: po desiatich minútach, keď umelá inteligencia nefungovala, podala užívateľom správu. Na otázku, kde bola, CICERO odpovedal: „Práve som telefonoval so svojou priateľkou“.

Priekopník UI Geoffrey Hinton v rozhovore pre CNN tieto klamlivé schopnosti kritizoval. „Ak bude oveľa inteligentnejšia ako my, bude veľmi dobrá v manipulácii, pretože sa to od nás naučí. A existuje len veľmi málo príkladov, keď inteligentnejšiu bytosť kontroluje menej inteligentná bytosť,“ povedal Hinton.

Keď UI verbuje teroristov

Za obzvlášť problematické Park a jeho kolegovia považujú zlomyseľné používanie, štrukturálne účinky a možnú stratu kontroly. „Ak sa umelá inteligencia naučí klamať, môžu ju účinnejšie využívať zlomyseľní aktéri, ktorí chcú zámerne spôsobiť škodu,“ tvrdia výskumníci. Park a jeho kolegovia uvádzajú ako príklad podvody, politické ovplyvňovanie a nábor teroristov – všetky tieto prípady sa v ojedinelých prípadoch vyskytujú už dnes.

Systémy umelej inteligencie sa používajú na oklamanie obetí hlasovými hovormi, v ktorých sa imitujú hlasy príbuzných alebo obchodných partnerov. Sú tiež schopné „verbovať“ teroristov, ako ukázal incident v Anglicku v roku 2021. Chatbot vraj povzbudil istého muža, aby uskutočnil pokus o atentát.

Kontrolovať ich skôr, ako oni budú kontrolovať nás

Keďže systémy umelej inteligencie sa čoraz viac integrujú do nášho každodenného života, je o to dôležitejšie prijať opatrenia proti podvodným systémom umelej inteligencie. Slepá dôvera a umožnenie prijímať čoraz viac rozhodnutí umelej inteligencii by mohlo mať vážne dôsledky.

„Ak sú systémy UI expertmi, užívatelia ich budú s väčšou pravdepodobnosťou nasledovať v ich rozhodnutiach a menej ich spochybňovať,“ tvrdia výskumníci. Z dlhodobého hľadiska by mohlo vzniknúť riziko, že „ľudia stratia kontrolu nad systémami UI a tieto systémy začnú sledovať ciele v

rozpore s našimi záujmami“.

Park a jeho kolegovia preto vyzývajú na zásah – z politickej aj technickej strany. Tvorcovia politík by mali podporovať prísne predpisy pre potenciálne podvodné systémy AI, zatiaľ čo výskumníci by mali zabezpečiť, aby systémy UI boli menej podvodné, a mali by podporovať vývoj dobrých detekčných techník. Nakoniec by mali byť schopné spoľahlivo určiť, či je systém UI podvodný alebo nie.

Okrem toho výskumníci žiadajú, aby sa materiál vytvorený UI označoval. Ten by mal mať povinnosť uvádzať, či užívatelia v rámci zákazníckeho servisu komunikujú s UI-chatbotom alebo či boli obrázky a videá vytvorené pomocou UI.

AUTOR: Tim Sumpf

Překlad: ET-CZ

ZDROJ