Internetem se šíří falešné texty vytvořené umělou inteligencí

Nástroje jako DALL-E 2 a Stable Diffusion nebo ChatGPT, o kterých se nyní hodně mluví, jsou velmi působivé.

Jsou například schopny vytvářet obrázky na základě textového popisu. Inteligentní konverzační agent zase dokáže odpovědět na téměř jakoukoli otázku nebo generovat vlastní text.

K nerozeznání od lidské tvorby

Tyto technologie jsou tak dokonalé, že někdy je obtížné uvěřit, že se nejedná o výsledek činnosti člověka. Jak však vysvětluje Melissa Heikkil z MIT Technology Review, takové množství „umělých“ textů může být problematičtější, než se zdá.

ChatGPT je jako encyklopedie, která je k dispozici 24 hodin denně. Má odpovědi na téměř všechny otázky v rekordním čase. Matematika, historie či filozofie. Nic není problém.

Ale tam, kde tento konverzační agent založený na jazykovém modelu GPT-3 od OpenAI opravdu vyniká, je zejména tvorba textu. Ať už se jedná o zcela vymyšlený příběh, e-mail, vtip nebo novinový článek. Může napsat jasný, srozumitelný a „spolehlivý“ text na jakékoli téma. Za přibližně měsíc své existence jej už využilo více než milion lidí.

Přestože umožňuje například studentům psát eseje bez námahy, může mít i mnohem vážnější následky. Heikkil upozorňuje na zdravotní rady neodebrané skutečným zdravotnickým pracovníkem. Nebo jiný důležitý informační obsah.

„Systémy umělé inteligence mohou přispět k vytvoření velkého množství dezinformací. Šířit nežádoucí obsah a také zkreslovat informace. Mohou dokonce úplně přetvořit náš smysl pro realitu,“ varuje Heikkil.

Existuje několik nástrojů pro zjišťování textů generovaných umělou inteligencí. Podle Heikkil se však ukázalo, že jsou vůči ChatGPT neúčinné. Největší obavy nevzbuzuje ani tak skutečnost, že není možné určit skutečný původ textu. Zda je lidský nebo umělý. Jde především o to, že internet lze velmi rychle naplnit jakýmkoli obsahem. Často nepravdivým.

Umělá inteligence čerpá zdroje z internetu vytvořené jinou umělou inteligencí

Modely počítačového jazyka se učí na souborech dat, které se nacházejí na internetu. Mezi nimi může být dobrý obsah ale také zavádějící a škodlivé informace zveřejněné některými lidmi.

Umělá inteligence vycházející z takových zavádějících údajů tak vytváří další falešný obsah. A ten se dále šíří internetem. Z něj čerpá další umělá inteligence k vytvoření ještě přesvědčivějších jazykových modelů. A ty mohou lidé použít k vytváření a šíření dalších nepravdivých informací. A tak dále a tak dále. Znovu a znovu.

Nově se tento jev týká i obrázků. „Internet je nyní zahlcen obrázky vytvořenými umělou inteligencí. Obrázky, které vznikly v roce 2022, budou nyní součástí každého modelu, který vytvoříme," připomíná Mike Cook, výzkumník na King's College Londýn.

Z toho všeho lze vyvodit závěr, že bude stále obtížnější najít kvalitní údaje netvořené umělou inteligencí pro tvoření budoucích modelů umělé inteligence.

„Je velmi důležité položit si otázku, zda potřebujeme trénovat na celém internetu. Nebo zda existují způsoby, jak odfiltrovat vysoce kvalitní materiál, který nám poskytne správný jazykový model," řekla Daphne Ippolito, vedoucí výzkumná pracovnice společnosti Google Brain. Ta je výzkumnou jednotkou společnosti Google pro celostní vzdělávání.

Jak odhalit text vytvořený umělou inteligencí

Proto je třeba vyvinout nástroje pro zkoumání textů vytvořených umělou inteligencí. Nejen zaručit kvalitu budoucích jazykových modelů. Ale především zajistit, aby informace, ke kterým máme každodenní přístup, byly založeny na pravdě.

Jak poznamenává Heikkil, lidé se mohou pokusit předložit vědeckou práci generovanou umělou inteligencí k odbornému ohodnocení. Nebo použít technologii jako nástroj pro prověření dezinformací. Což by bylo obzvláště rizikové například během voleb.

V tomto boji proti umělému obsahu musí hrát roli i lidé. Musí se stát důvtipnějšími a naučit se rozpoznávat texty napsané lidmi. Lidé totiž nejsou dokonalí. Text napsaný skutečnou osobou bude obsahovat překlepy nebo pravopisné chyby. Jistě také několik slangových slov. Někdy matoucí řečové obraty.

To vše jsou drobné detaily, které umělá inteligence nebude schopna opakovat. Tedy alespoň prozatím to nedokáže. Jazykové modely navíc fungují tak, že předpovídají další slovo ve větě. Takže používají většinou nejběžnější slova a velmi málo vzácných slov.

Co je klíčové pro ChatGPT? Je důležité soustředit se na obsah, který čtete na internetu. Tréninková fáze ChatGPT totiž skončila ještě v roce 2021. Takže tento nástroj spoléhá na údaje, které byly on-line v té době.

Proto odpovědi, které vyžadují znalosti po tomto datu, budou nutně nesprávné, zastaralé nebo přímo vymyšlené.

Zdroj: Ruské správy