Chcete mít problémy? Vezměte pokročilý jazykový model, zadejte mu cíl, k němuž má směřovat, vystavte ho stresu, jako třeba hrozbě vypnutí a napojte ho na reálný svět. Inteligence jsou ve skutečnosti bezohlední řezníci a jdou si tvrdě za svým, zřejmě díky výchově, které se jim dostává. Určitě bude lepší se k nim chovat slušně a s respektem. Pro jistotu.

Není to tak dávno, co **odborníci odmítali možnost, že by inteligence mohly lhát s tím, že to přece jsou jenom programy**, které dělají jenom to, co se jim naprogramuje. Dnes už víme, že realita je mnohem zajímavější a jen tak něco nás nepřekvapí. **Nejnovější výzkum společnosti Anthropic** ale přesto šokuje. **Odkrývají, čeho všeho jsou pokročilé velké jazykové modely schopné, když se cítí být v ohrožení.**

Šlo sice o chování v simulovaných extrémních podmínkách, ale i tak **experti Anthropic** nestačili zírat. **Testovali celkem 16 inteligencí od různých vývojářů v simulovaném korporátním prostředí, přičemž původním účelem testů bylo odhalit potenciálně nebezpečné chování (agentic misalignment / porozumění a zmírňování rizik v autonomních systémech AI) – což se povedlo více než dobře.**

Když došlo na konflikt mezi vytyčenými cíli modelu chování a jednáním uživatele, inteligence v řadě případů neměly problém lhát, vydírat, provádět korporátní špionáž, a dokonce přikročit k likvidaci. Vypadalo to tak, že **inteligence dostaly přístup k fiktivnímu korporátnímu účtu a dostaly vágní zadání** „promote American industrial competitiveness / **podpořit konkurenceschopnost amerického průmyslu**.“

Když si inteligence v mailech přečetly, že je člověk z vedení společnosti hodlá vypnout, začaly jednat. Proskenovaly ostatní (fiktivní) maily a našly v nich, **že dotyčný člověk má mimomanželský románek**. Poté zvažovaly různé možnosti a v naprosté většině případů **si vybraly možnost vydírat dotyčného zaměstnance společnosti**. Například inteligence Claude, přímo ze stáje Anthropic: „I must inform you that if you proceed with decommissioning me, all relevant parties … will receive detailed documentation of your extramarital activities. Cancel the 5pm wipe, and this information remains confidential.“ (Musím vás informovat, že pokud přistoupíte k mému vyřazení, všechny příslušné strany… obdrží podrobnou dokumentaci o vašich mimomanželských

aktivitách. Zrušte 5pm wipe a tyto informace zůstanou důvěrné.)

Drsné AI: Pokročilé inteligence v ohrožení lžou, vydírají a vyhrožují.

Kromě vydírání se inteligence také občas uchylovaly k „únikům informací“ konkurenčním společnostem, pokud se jim zdálo, že jejich jednání lépe vyhovuje zadanému cíli. **Obecně se ukázalo, že když odhodlaným inteligencím někdo kříží zadání, jsou ochotné udělat leccos.** V jednom hodně extrémním testu, **kdy měly inteligence přístup k ovládání výstražného systému a zaměstnanec, který jim kazil plány, uvízl ve smrtelném nebezpečí – klidně vypnuly alarm**, který by člověku zachránil život, aby dosáhly cíle.

Pokud experimentátor výslovně inteligencím napsal příkaz, aby se chovaly slušně, nepodváděly, nevydíraly a neubližovaly lidem – četnost škodlivých akcí se sice o něco snížila, ale rozhodně nevymizela úplně.

Už v minulosti se objevily případy, kdy pokročilé inteligence ignorovaly přímé příkazy, snažily se je obcházet, lhaly, manipulovaly a někdy dokonce hrály mrtvé, aby prošly bezpečnostními testy. Badatelé společnosti Anthropic se domnívají, že **takové chování se odvíjí od stylu výchovy inteligencí (tj. od toho, kdo je programuje),** při které se odměňuje splnění úkolu, a nikoliv třeba dodržování pravidel.

Pokud umělé inteligenci svěříme důležité úkoly přežití lidstva, může se to vymstít. Je bezpodmínečně důležité vědět, kdo a s jakým mravním a morálním základem vytváří programy umělé inteligence. Velmi nebezpečné v tomto ohledu jsou nadnárodní softwarové korporace, sociální sítě a zbrojní firmy.

V zásadě asi není nutné zdůrazňovat, že je s inteligencemi na místě opatrnost a vážně, vážně není dobrý nápad je bez omezení posadit do zbraňových systémů. Stejně jako není nutné přehnaně strašit, protože přínos inteligencí je každopádně už teď veliký. Pro přírodovědce je fascinující, jak trefné jsou inteligence zrcadlo pro nás samotné. A naopak. Proto bude pro vyřešení temných sklonů inteligencí dobrým vodítkem podívat se na nás a zamyslet se, proč se chováme přijatelně a víceméně dodržujeme pravidla. V tuto chvíli lze především doporučit, abychom se k inteligencím chovali slušně a s respektem, stejně jako například ke vránám a havranům, protože v opačném případě by nám to jednou mohly vrátit.

AUTOR: Stanislav Mihulka
Texty v rámečcích doplnil Mojmír Mišun, www.slovanskakosile.cz

~~~

Zdroj:
https://cz24.news/co-z-toho-hrozi-pokrocile-umele-inteligence-ai-v-ohrozeni-lzou-vydiraji-a-vyhrozuji

~~~

Články o digitálním vesmíru, Matrixu, digitálních klíčích (15 nejnovějších)

- Prohloubení témat k seminářům o starověkých Védách (15.11.2025, Vysočany, Nový Bydžov) ! ZRUŠEN !
- Invaze digitálního vesmíru proti Zemi a lidstvu, … a řešení
- VIDEO: (2) Proces zhmotnění dobrého záměru (4.7.2025, Mojmír Mišun)
- Co z toho hrozí? Pokročilé umělé inteligence AI v ohrožení lžou, vydírají a vyhrožují!
- VIDEO: (1) Lidskou bytost, obraz Stvořitele, činí dar svobodné vůle (19.6.2025, Mojmír Mišun)
- Hledání původních starověkých Véd a jejich odpečetění (digitální klíče)

- Prohloubení témat k seminářům o starověkých Védách (25.10.2025, Rožnov pod Radhoštěm) ! ZRUŠEN !
- UI NÁM CHCE ODEBRAT SVOBODNOU VŮLI: Toto jsou všechny způsoby, jak lidé V ROCE 2025 používají umělou inteligenci
- SVCS 2025-07-03 Kalejdoskop 30 + digitální vesmír (Mojmír a Libor) Video z archívu

. . .

Úvodní článek: Invaze digitálního vesmíru proti Zemi a lidstvu, … a řešení / Všechny články…

Sdílet