Vyhladí nás AI? Človek, ktorý ju vyvíjal, varuje pre vážnym rizikom

11.9.2026 - Lenka Ivančíková | Zajtra studio spol.

Umelá inteligencia napreduje rýchlejšie, než mnohí očakávali. Bývalý výskumník spoločností OpenAI a Anthropic Jacob Coxon teraz varuje, že vývoj najpokročilejších modelov môže priniesť riziká, na ktoré zatiaľ nemáme riešenie.

Po odchode z Anthropic **vystúpil v televízii CNN**, kde vysvetlil, prečo sa rozhodol o svojich obavách hovoriť verejne.

Coxon patrí medzi ľudí, ktorí pracovali priamo na vývoji veľkých AI modelov. Posledné tri roky sa venoval najmä ich počiatočnému trénovaniu, najskôr v OpenAI a neskôr v Anthropic. Zo spoločnosti Anthropic odišiel 8. septembra 2026. Následne zverejnil sériu príspevkov, v ktorých obvinil popredné AI firmy z príliš rýchleho vývoja systémov bez dostatočného riešenia bezpečnosti.

Zdroj: YouTube/CNN

Obavy sa netýkajú dnešného ChatGPT ani Claude

Coxon v rozhovore s Andersonom Cooperom zdôraznil jednu podstatnú vec. Podľa neho dnešné AI modely nepredstavujú existenčné riziko pre ľudstvo.

Jeho obavy sa týkajú toho, čo by mohlo prísť v ďalších rokoch. Konkrétne systémov, ktoré by dokázali samostatne pracovať na ďalšom vývoji umelej inteligencie a postupne zlepšovať samy seba.

Takémuto procesu sa hovorí **rekurzívne sebazlepšovanie**. Teoreticky by mohol viesť k situácii, keď by AI zvyšovala svoje schopnosti rýchlejšie, než by ľudia dokázali jej vývoj kontrolovať. Coxon tento **scenár označuje za jednu z najväčších hrozieb** ďalšieho vývoja.

Podľa jeho slov môže byť nebezpečná najmä kombinácia vysokej inteligencie, autonómneho rozhodovania a prístupu k digitálnym systémom. Ako príklady rizík spomenul útoky na kritickú infraštruktúru či zneužitie AI pri vývoji nebezpečných biologických technológií. Ide však o možné (hypotetické) budúce scenáre, nie o schopnosti, ktorými by súčasné bežne dostupné chatboty disponovali.

> I resigned from Anthropic today. I spent the last three years doing pretraining research at both OpenAI and Anthropic. Neither company is acting responsibly. They are racing straight to self-improving superintelligence and gambling with our lives. More thoughts below.
>
> — Jacob Coxon (@hilbertspaess) September 9, 2026

Viac než 10 % riziko

Coxonove vyjadrenia získali pozornosť aj preto, že s časťou jeho obáv verejne súhlasí **Evan Hubinger**, ktorý v Anthropic vedie výskum zameraný na zosúladenie správania AI s ľudskými

záujmami.

Hubinger uviedol, že podľa jeho osobného odhadu existuje viac než 10 % pravdepodobnosť, že by veľmi pokročilá AI mohla počas nasledujúcej dekády, tedy do roku 2030, viesť až ku katastrofickému scenáru pre ľudstvo.

Zároveň priznal, že Anthropic zatiaľ nemá vyriešený spôsob bezpečného riadenia prípadnej superinteligencie. Zdôraznil však, že riziko súčasných modelov považuje za nízke.

Prečo firmy jednoducho nespomalia?

Coxon tvrdí, že ľudia vo veľkých AI spoločnostiach si podľa neho riziká uvedomujú. Problémom je konkurenčný boj. Každá firma sa obáva, že ak vývoj spomalí, prvenstvo získa iná spoločnosť alebo zahraničný konkurent.

Na otázku, či spoločnosti ako Anthropic skutočne chcú reguláciu, Coxon odpovedal, že ich požiadavky považuje za úprimné. Podľa neho by firmy privítali systém, ktorý by umožnil všetkým najväčším hráčom spomaliť súčasne bez toho, aby tým niektorý z nich získal konkurenčnú nevýhodu.

Anthropic tvrdí, že bezpečnosť berie vážne

Spoločnosť Anthropic sa proti tvrdeniu, že by bezpečnostné riziká ignorovala, ohradila. Pre CNN uviedla, že od začiatku otvorene upozorňuje nielen na možné výhody AI, ale aj na jej bezprecedentné riziká.

Anthropic zároveň testuje schopnosti svojich systémov v oblastiach, ako je kybernetická bezpečnosť alebo biológia, a výsledky časti týchto testov verejne zverejňuje.

Coxon napriek tomu tvrdí, že súčasný spôsob vývoja nepovažuje za dostatočne bezpečný. Podobne kritizoval aj OpenAI, kde pred príchodom do Anthropic pracoval.

Dôležitá bude kontrola ďalšej generácie AI

Súčasné AI nástroje už dnes dokážu programovať, analyzovať dokumenty, pracovať s obrazom či samostatne vykonávať niekoľko krokov za sebou. Stále však potrebujú rozsiahlu infraštruktúru, prístupové oprávnenia a v mnohých prípadoch aj dohľad človeka.

Práve hranica medzi nástrojom ovládaným človekom a systémom schopným dlhodobo konať samostatne bude v nasledujúcich rokoch jednou z najdôležitejších tém vývoja umelej inteligencie.

Či sa naplnia extrémne scenáre opisované Coxonom, dnes nikto nevie. Jeho vystúpenie však ukazuje, že diskusia o bezpečnosti AI už nie je iba témou ľudí stojacich mimo technologického priemyslu. O rovnakých problémoch verejne hovoria aj výskumníci, ktorí najpokročilejšie modely sami vyvíjajú.

https://www.mojandroid.sk/vyhladi-ai-ludstvo-vyskumnik-z-anthropic-varuje