MINISTERSTVO INVESTÍCIÍ, REGIONÁLNEHO ROZVOJA A INFORMATIZÁCIE SLOVENSKEJ REPUBLIKY

CSIRT.SK

Bezpečnostné zásady pri používaní systémov s umelou inteligenciou

V dnešnom dynamickom svete sme vystavení potrebe využívať svoj čas čo najefektívnejšie a snažíme sa hľadať spôsoby, ako to dosiahnuť. Každý deň vznikajú nové technológie, ktoré nám môžu pomôcť s pracovnými úlohami a majú pozitívny vplyv na zvýšenie efektivity práce. S každou novou technológiou sa spájajú aj bezpečnostné riziká, ktoré nesmieme podceňovať. Je nevyhnutné pristupovať zodpovedne k systémom, ktoré využívajú prvky umelej inteligencie (ďalej len „AI“), akým sú napríklad populárne modely ChatGPT, Gemini, Llama, DeepSeek, Grok či Claude. Zodpovednosť za zdieľanie informácií a interpretovanie výsledkov vždy preberá jednotlivec a nie je možné preniesť zodpovednosť na stroj. Účelom modelov AI je inšpirovať a vytvárať nové myšlienky, nie bezchybne odpovedať na otázky.
Preto by sme radi upozornili na bezpečnostné opatrenia pre správne a bezpečné využívanie systémov, ktoré obsahujú prvky umelej inteligencie.

Bezpečnostné zásady a opatrenia:

- Neposielajte a nevkladajte do modelov AI žiadne osobné údaje.
- Neposielajte a nevkladajte interné, citlivé a dôverné informácie. Všetky vložené údaje si AI uchováva a používa na učenie. Myslite na to, že zadávané údaje môžu ľahko získať externí používatelia.
- Nepoužívajte prídavné moduly pre prehliadače, ktoré využívajú AI, pretože budú zbierať a posielať údaje na jej učenie.
- Nepoužívajte modely AI ako vyhľadávač na internete. Modely AI generujú aj neexistujúce a falošné fakty, tzv. halucinujú. Poskytujú inšpiráciu pre ďalšie vyhľadávanie. Pre vyhľadávanie informácií ako sú aktuálne športové výsledky, či vedecké fakty a zdroje, používajte vyhľadávače.
- Kontrolujte a overujte všetky odpovede modelov AI. Je potrebné overiť si fakty a posúdiť, či odpoveď dáva zmysel. AI dokáže zložiť zrozumiteľný text, no nerozumie súvislostiam a jeho významu. Odpoveď nie je možné považovať automaticky za pravdivú.
- Kontrolujte a overujte zdroje, ktoré po vyžiadaní uvedie model AI. Stáva sa, že si ich AI vymyslí.
- Pristupujte opatrne pri používaní AI pre generovanie inovatívnych myšlienok.
- Správajte sa zodpovedne a opatrne pri používaní výstupu. Vygenerovaný kód môže spôsobiť neočakávané správanie, grafický výstup môže obsahovať obrázky chránené autorskými právami, a pod..
- Zvážte trvalé vypnutie histórie konverzácií pre prípad, že by sa v otázkach nevedome použili citlivé informácie. Konverzácie vedené pri vypnutej histórii nie sú použité pri trénovaní modelov AI.

Správajte sa k AI ako k dieťaťu, ktoré dokážu používatelia ľahko zmiasť.
V učiacej fáze získa AI rovnaké odchýlky od správneho úsudku ako majú materiály, na ktorých sa učila.

TLP: White

Odporúčania pre manažment:

- Namiesto úplného zákazu používania sa zamerajte na dôkladné vzdelávanie používateľov čo je povolené robiť a čo nerobiť pri používaní generatívnych modelov AI.
- Urobte analýzu rizík.
- Vytvorte usmernenie pre zamestnancov, ako môžu používať modely AI v súvislosti so svojou pracovnou činnosťou.
- Zakážte prídavné moduly pre prehliadače, ktoré využívajú AI, pretože budú zbierať a posielať údaje na jej učenie.
- Pokiaľ máte implementovaný model AI v infraštruktúre organizácie, zamedzte mu prístup k citlivým údajom, resp. prístup na internet, aby neunikli.

Návrhy na použitie:

- Generovanie reportov a textov, ktoré sú pre zamestnanca inak časovo náročné, no nevyžadujú analytický prístup (napr. formulovanie do súvislého textu už spracovaných podkladov, kontrola gramatiky). Zadávaný text nesmie obsahovať citlivé údaje. Môže ísť napríklad o správy určené na zverejnenie.
- Získanie inšpirácie, akým spôsobom riešiť problém alebo úlohu. Problém či úloha nesmie mať interný charakter, ani obsahovať osobné údaje.
- AI môže slúžiť ako vhodná a efektívna alternatíva internetových vyhľadávačov pri získavaní prvotného prehľadu o danej téme. Výsledok je potrebné skontrolovať a overiť v nezávislom zdroji.
- AI môže pomôcť pri tvorbe a odlaďovaní skriptov a kódu, pokiaľ sa nejedná o citlivé projekty.

Zdroje:

- *https://www.trendmicro.com/en_us/research/23/b/review-what-gpt-3-taught-chatgpt-in-a-year.html*
- https://gizmodo.com/chatgpt-ai-samsung-employees-leak-data-1850307376
- https://openai.com/blog/new-ways-to-manage-your-data-in-chatgpt
- Komunikácia so zahraničnými tímami združenými v organizácii FIRST
- Meusers von Wissmann, Richard. Umělá inteligence se může stát zabijákem Googlu. Chip, ISSN 1210-0684, 2023, vol.33, no.04, pp 16-18