BEZPEČNOST AI VE FIRMĚ · CHECKLIST

12 otázek, než AI agenta pustíte k firemním datům

Pro AI agenty, automatizace s AI a asistenty napojené na poštu, dokumenty nebo systémy. Stejné otázky lze poslat dodavateli nástroje.

Jazykový model spolehlivě nerozliší váš pokyn od textu, který zpracovává (OWASP LLM01:2025). Podle OWASP není jasné, zda proti prompt injection existuje spolehlivá prevence. Proto se neřeší vyloučení útoku, ale jeho dopad: co agent čte, co smí a jestli nevratné akce schvaluje člověk. Zaškrtněte, co platí.

A. CO AGENT ČTE A CO MŮŽE UDĚLAT

1. ☐ **Zdroje:** je sepsáno, z jakých zdrojů agent čte (pošta, web, úložiště, CRM, nahrané soubory) a které obsahují obsah od lidí mimo firmu?
2. ☐ **Akce:** je sepsáno, jaké akce agent může provést (odeslat, změnit, smazat, sdílet, zaplatit, spustit) a ke kterým systémům má napojení?
3. ☐ **Nejhorší případ:** je jasné, co nejhoršího se stane, když agent poslechne cizí pokyn? Pokud „pošle data ven“ nebo „zaplatí“, platí otázky 6 a 7.
4. ☐ **Potřebuje to být agent?** Byla zvážena varianta, kdy AI jen připraví návrh a člověk ho provede? (OWASP: zásada nejmenší samostatnosti, Least-Agency)

B. OPRÁVNĚNÍ A SCHVALOVÁNÍ

5. ☐ **Nejmenší oprávnění:** má agent jen nezbytná práva, nejlépe jen pro čtení, a je to nastaveno v cílovém systému, ne jen v instrukcích pro AI?
6. ☐ **Lidské schválení:** nevratné akce (odeslání ven z firmy, platba, mazání, změna práv, publikování) schvaluje člověk dřív, než proběhnou?
7. ☐ **Náhled před potvrzením:** vidí schvalující člověk, co přesně agent udělá (komu, co, kolik), ne jen „agent chce pokračovat“?
8. ☐ **Vlastní identita:** má agent vlastní účet nebo token s omezeným rozsahem, ne sdílené heslo správce? Jde zrušit jedním krokem?

C. PROVOZ A LIDÉ

9. ☐ **Logy:** je zapnuté logování toho, co agent udělal, kdy a na základě čeho? Je určeno, kdo logy kontroluje?
10. ☐ **Hlášení a vypnutí:** vědí lidé, komu hlásit, když agent udělá něco nečekaného, a jak ho rychle vypnout?
11. ☐ **Registr a smlouva:** je agent zapsaný v registru AI nástrojů (dodavatel, umístění dat, rozsah oprávnění) a je s dodavatelem zpracovatelská smlouva, pokud pracuje s osobními údaji?
12. ☐ **Doporučení není ověření:** vědí lidé, že návrh agenta neplatí jako kontrola, a platí u plateb, změn účtů a přístupů stejná kontrola jako u požadavku od člověka?

Stop pravidlo: pokud u otázek 5, 6 nebo 10 odpovíte „ne“, agenta zatím nenasazujte. Zbytek lze dotáhnout za provozu.

Zdroj: aiscale.cz/bezpecnost-ai/prompt-injection-ai-agenti/ · Vychází z OWASP Top 10 for LLM Applications 2025 (LLM01, LLM06) a OWASP Top 10 for Agentic Applications 2026 (ASI01, ASI02, ASI09)

Verze říjen 2026 · informativní materiál, nejde o právní poradenství