RIZIKA A MITIGACE AI AGENTŮ

Tři rizika, která musíte umět ošetřit už před pilotem

1. Halucinace

Co to je: agent vygeneruje informaci, která ve zdrojích není, nebo je nesprávná. Působí přesvědčivě, ale opírá se o domněnku.

Typické projevy v úřadu:

- Agent doplní chybějící datum jednání nebo číslo jednací, protože „to tak bývá".
- Do shrnutí vloží výrok, který v textu není.
- Vytvoří odůvodnění, které zní odborně, ale stojí na domněnce.

Mitigace:

- Lidská kontrola u všech výstupů s právním nebo reputačním dopadem.
- U klíčových tvrzení vyžadujte odkaz na konkrétní část zdroje.
- Do role agenta vložte větu: „Pokud údaj není ve zdroji, napiš nenalezeno ve zdroji a navrhni krok k doplnění."

2. Zkreslení (bias)

Co to je: systematická odchylka ve výstupech – agent „má tendenci" hodnotit některé situace jinak, než by měl. Projeví se až na větším množství případů.

Typické projevy v úřadu:

- Shrnutí dává některým typům podání menší váhu.
- Návrhy odpovědí se liší tónem podle toho, kdo podání poslal.
- Při třídění preferuje určitou kategorii, protože tu „viděl v datech".

Mitigace:

- U rizikových kategorií ruční kontrola každého výstupu.
- Sledujte, kde se agent pravidelně mýlí, a upravujte štítky/příklady.
- U nových modelů nebo změn pravidel krátké testování na vzorku.

3. Citlivá data

Co to je: únik osobních údajů, identifikátorů nebo neveřejných informací. U agentů se citlivá data objevují na více místech než u chatu.

Čtyři místa, kde mohou citlivá data utéct:

- Vstupní dokumenty (osobní údaje, údaje o dávkách, zdravotní informace).
- Průběžné záznamy (logy) zachycující identifikátory během běhu agenta.
- Výstupy, které opakují detaily, jež neměly opustit zdrojový dokument.
- Testovací data, když si „pomůžete" reálnými případy.

Mitigace:

- Anonymizace (nevratně) tam, kde agent nepotřebuje znát osoby.
- Pseudonymizace (kódování s odděleným klíčem), když musíte zachovat vazby.
- Auditní stopa: kdo, kdy, co dostal na vstupu, co vrátil, kdo schválil.