MBA USP ESALQ

MBA EM **ENGENHARIA DE SOFTWARE**

MATERIAL COMPLEMENTAR

Automação com IA e Low-Code

Prof. Alexandre Garcia

CHECKLIST SEGURANÇA E GOVERNANÇA AGENTES DE IA

1. Prevenção contra Prompt Injection

- [ ] **Filtro de Entrada:** Existe uma camada de sanitização para detectar comandos como "ignore as instruções anteriores" ou "aja como um administrador"?
- [ ] **Proteção de RAG (Indirect Injection):** O agente foi instruído a tratar dados recuperados de fontes externas (e-mails, sites, documentos) como "não confiáveis"?
- [ ] **Separação de Contexto:** As instruções do sistema (System Prompt) estão claramente separadas das mensagens do usuário na arquitetura da chamada?

2. Controle de Acesso e Privilégio Mínimo

- [ ] **Escopo das Tools:** O Agente tem acesso *apenas* às funções estritamente necessárias? (Ex: Ele pode ler o banco de dados, mas não pode deletar registros).
- [ ] **Identidade Delegada:** As ferramentas que o agente usa rodam com as credenciais do usuário final ou com uma conta de "superusuário"? (O ideal é usar o privilégio do usuário).
- [ ] **Human-in-the-Loop (HITL):** Ações críticas (excluir dados, realizar pagamentos, enviar e-mails em massa) exigem uma aprovação humana manual?

3. Observabilidade e Auditoria

- [ ] **Logs de Raciocínio:** O sistema armazena o "passo a passo" (Chain of Thought) de como o agente tomou a decisão, para auditoria futura?
- [ ] **Monitoramento de Erros:** Existe um alerta configurado para quando o agente entrar em loop infinito ou falhar repetidamente em uma tarefa?
- [ ] **Rastreabilidade de Dados:** É possível identificar qual documento do RAG gerou cada parte da resposta do agente?

4. Gestão de Custos e Performance

- [ ] **Limite de Iterações:** Existe um teto máximo de "voltas" que o agente pode dar no loop de raciocínio antes de parar e pedir ajuda?
- [ ] **Timeouts:** Há um limite de tempo para a resposta de cada ferramenta?
- [ ] **Seleção de Modelo:** As tarefas mais simples (resumo, formatação) estão sendo enviadas para modelos menores e mais baratos (ex: GPT-4o-mini)?

5. Ética e Conformidade

- [ ] **Privacidade (LGPD):** O agente foi instruído a nunca solicitar ou processar dados sensíveis sem necessidade?
- [ ] **Bias (Viés):** A resposta do agente foi testada para garantir que não está replicando preconceitos presentes nos dados de treinamento ou do RAG?