

Guia Prático Riscos Cognitivos em Agentes de IA

Baseado no
Framework OpGov AI®
Operational Governance AI

Por:

Ana Maria Lopes Corrêa

Criadora do Framework OpGov AI

Coautora do primeiro livro brasileiro

de Governança da Inteligência Artificial Artificial



CERTIFICADO DE REGISTRO DE DIREITO AUTARAL

A Câmara Brasileira do Livro certifica que a obra intelectual descrita abaixo, encontra-se registrada nos termos e normas legais da Lei nº 9.610/1998 dos Direitos Autorais do Brasil. Conforme determinação legal, a obra aqui registrada não pode ser plagiada, utilizada, reproduzida ou divulgada sem a autorização de seu(s) autor(es).

Responsável pela Solicitação:
Focus4Results

Participante(s):
Ana Maria Lopes Correa (Autor)

Título:
Framework XGov AI - Estrutura prática para governança confiável e estratégica da inteligência artificial

Data do Registro:
06/05/2025 12:31:03

Hash de transação:
0x36306b62e3294e902db87aba9125072b034ad1db3a23745ea25ae842cd240f

Hash do documento:
6d0db216a20ac945094002ae12bcf555ba5d0bec2af5c4e587ed02fa506aa46

O framework atualmente denominado OpGov AI® (Operational Governance for Artificial Intelligence) teve sua concepção e estrutura intelectual registradas em maio de 2025 sob a denominação original Framework XGov AI.

O conteúdo protegido permanece o mesmo, sendo o novo nome adotado exclusivamente como denominação comercial e de posicionamento.

SUMÁRIO

<u>1 – Introdução</u>	<u>04</u>
<u>2 - Riscos Cognitivos por Camada do Ciclo Cognitivo</u>	<u>06</u>
<u>2.1 - Camada Funcional – Percepção</u>	<u>06</u>
<u>2.1.1 – Descrição.....</u>	<u>06</u>
<u>2.1.2 - Principais Riscos Cognitivos da Camada de Percepção....</u>	<u>07</u>
<u>2.2 - Camada Funcional - Memória e Representação Cognitiva.....</u>	<u>11</u>
<u>2.2.1 – Descrição.....</u>	<u>11</u>
<u>2.2.2 - Principais Riscos Cognitivos da Camada de Memória e Representação Cognitiva</u>	<u>13</u>
<u>2.3 - Camada Funcional - Planejamento e Raciocínio</u>	<u>17</u>
<u>2.3.1 – Descrição.....</u>	<u>17</u>
<u>2.3.2 - Principais Riscos Cognitivos da Camada de Planejamento e Raciocínio</u>	<u>19</u>
<u>2.4 - Camada Funcional – Ação</u>	<u>25</u>
<u>2.4.1 – Descrição.....</u>	<u>25</u>
<u>2.4.2 - Principais Riscos Cognitivos da Camada de Ação</u>	<u>27</u>
<u>2.5 - Camada Funcional - Aprendizado e Reflexão Cognitiva</u>	<u>32</u>
<u>2.5.1 – Descrição.....</u>	<u>32</u>
<u>2.5.2 - Principais Riscos Cognitivos da Camada de Aprendizado e Reflexão Cognitiva</u>	<u>34</u>
<u>3 - Integração dos Riscos Cognitivos no Ciclo de Vida</u>	<u>40</u>
<u>4 - Fundamentos conceituais e metodológicos XGov AI ..</u>	<u>42</u>
<u>Nota metodológica sobre as fontes e a construção do conhecimento</u>	<u>42</u>
<u>Referências bibliográficas</u>	<u>44</u>
<u>5 - Mapa Visual em Riscos Cognitivos em Agentes de IA ..</u>	<u>45</u>

1 - Introdução

Riscos Cognitivos: o novo desafio da Governança de IA

A inteligência artificial evoluiu de sistemas focados apenas no processamento de dados para agentes capazes de **perceber, raciocinar, agir e refletir** de forma autônoma.

Com essa evolução, emerge uma nova e crítica categoria de riscos: **os riscos cognitivos**, que desafiam diretamente a **segurança, a confiabilidade e a ética das soluções de IA** agêntica.

O que são riscos cognitivos?

Riscos cognitivos são **falhas, desvios ou distorções que ocorrem nos processos mentais artificiais de um agente de IA**: percepção, memória, raciocínio, ação e reflexão. Eles comprometem a forma como o agente **interpreta o contexto, toma decisões, executa ações e aprende com suas próprias experiências**.

Em termos simples:

São os riscos que surgem quando a IA **“pensa errado” e, ainda assim, acredita que está certa**.

Percepção

Como a IA interpreta a realidade

Memória

O que a IA armazena e lembra

Raciocínio

Como a IA conecta informações

Ação

Como a IA executa decisões

Aprendizado

Como a IA aprende e evolui

Por que os riscos associados à essa camada importam?

Em soluções de IA convencionais, os principais riscos concentram-se em **dados, modelos e métricas técnicas**. Já nas soluções de **IA agêntica**, o risco se desloca para a **forma como a IA percebe o mundo, raciocina, decide, age e aprende ao longo do tempo**.

Esses riscos:

- **não são identificáveis apenas por métricas de acurácia,**
- **emergem durante a operação, e**
- **evoluem com o uso contínuo**, tornando sua mitigação significativamente mais complexa.

Sem controles cognitivos adequados, agentes de IA podem desenvolver **decisões autônomas incoerentes, comportamentos emergentes imprevisíveis e violações éticas ou regulatórias não intencionais**.

O papel da Governança

A Governança de IA assume um papel central ao **monitorar, orientar e regular o comportamento cognitivo dos agentes de IA** ao longo de todo o seu ciclo de vida.

Seu objetivo é assegurar que:

- **o raciocínio da IA seja justificável e explicável,**
- **a memória seja íntegra, rastreável e auditável,**
- **os objetivos permaneçam alinhados à estratégia e aos direitos humanos, e**
- **o aprendizado ocorra de forma controlada, ética e segura.**

Nas soluções de IA agêntica, a governança **não substitui** os domínios já consolidados, como dados, segurança, justiça e IA responsável e sustentável, mas **os complementa** ao incorporar a análise e a mitigação de riscos relacionados à **“mente da IA”**.

2 - Riscos Cognitivos por Camada do Ciclo Cognitivo

2.1 - Camada Funcional - Percepção



2.1.1 - Descrição

A regra de ouro da IA cognitiva:

“Perception drives cognition.”

Se o agente **percebe errado**, ele **raciocina errado**.

Se **raciocina errado**, **planeja errado**.

Se **planeja errado**, **decide errado**.

E, se **decide errado**, **age errado**.

Por isso, os riscos associados à percepção são **os mais críticos e sistêmicos** em agentes de IA.

A **percepção** é a primeira camada funcional de qualquer agente cognitivo e atua como a **porta de entrada da realidade** para o sistema. É nessa camada que o agente coleta, filtra, interpreta e organiza informações provenientes do ambiente, de usuários, de sensores, de sistemas externos ou de entradas textuais.

No ciclo cognitivo de um agente, a percepção antecede e influencia diretamente todas as demais camadas:

- Memória e Representação
- Raciocínio e Planejamento
- Ação
- Aprendizado e Reflexão

Quando a percepção falha, **todas as camadas seguintes são contaminadas**, gerando efeitos em cascata que dificilmente são corrigidos apenas com ajustes posteriores.

Como falhas de percepção se propagam pelo agente?

1. Da Percepção para a Memória

Entradas incorretas são armazenadas como fatos.

O agente passa a construir **memórias distorcidas, seletivas ou conflitantes**.

2. Da Percepção para o Raciocínio

O agente raciocina sobre **premissas erradas**.

As conclusões podem parecer logicamente coerentes, mas são **conceitualmente inválidas**.

3. Da Percepção para a Ação

Decisões baseadas em interpretações incorretas levam a **ações fora de contexto, ineficientes ou potencialmente perigosas**.

4. Da Percepção para o Aprendizado

O agente aprende a partir de observações equivocadas.

Isso reforça **padrões incorretos, vieses e ciclos de erro** ao longo do tempo.

2.1.2 - Principais Riscos Cognitivos da Camada de Percepção

Risco 1 – Percepção Equivocada da Situação

Domínio - Segurança e Solidez Técnica

Princípio - Desempenho e Confiabilidade

Descrição

Ocorre quando o agente interpreta de forma incorreta o que está acontecendo ao seu redor por receber informações **incompletas, incorretas, ambíguas ou distorcidas**. Como resultado, o agente constrói uma compreensão equivocada da situação, que passa a orientar todo o seu raciocínio e suas decisões.

Exemplo prático

Um agente de atendimento interpreta uma mensagem curta e direta de um cliente como sinal de urgência extrema, quando, na verdade, trata-se apenas de um pedido objetivo. Com base nessa percepção equivocada, o agente prioriza indevidamente o caso, aciona fluxos críticos e gera alocação desnecessária de recursos.

Circunstâncias de ocorrência

- Em ambientes com dados ruidosos ou incompletos
- Quando entradas humanas são ambíguas ou pouco contextualizadas
- Em interações rápidas, sem validação de entendimento
- Em agentes que operam com alto grau de autonomia desde o início

Risco 2 – Percepção Baseada em Fonte Única ou Instável

Domínio - Segurança e Solidez Técnica

Princípio - Desempenho e Confiabilidade

Descrição

O agente passa a depender de **uma única fonte de informação**, que pode ser instável, desatualizada ou pouco confiável, sem realizar validação cruzada com outras referências. Isso compromete a qualidade da percepção e aumenta a probabilidade de decisões baseadas em visões parciais ou incorretas da realidade.

Exemplo prático

Um agente de monitoramento operacional utiliza apenas um sistema legado como fonte de dados de status.

Quando esse sistema apresenta atrasos ou inconsistências, o agente interpreta que há falhas graves em andamento e dispara alertas indevidos, gerando alarmes falsos e desgaste operacional.

Circunstâncias de ocorrência

- Em arquiteturas simplificadas ou mal integradas
- Quando há pressão por respostas rápidas
- Em ambientes com baixa maturidade de dados
- Quando não existem regras de validação entre múltiplas fontes

Risco 3 – Interpretação Contextual Incorreta

Domínio - Segurança e Solidez Técnica

Princípio - Desempenho e Confiabilidade

Descrição

O agente falha ao compreender corretamente **nuances de linguagem, sinais contextuais, intenções humanas ou fatores situacionais**, interpretando mensagens de forma literal, incompleta ou fora de contexto. Esse risco é especialmente crítico em interações com pessoas.

Exemplo prático

Um agente conversacional interpreta uma ironia ou uma crítica sutil como um pedido legítimo de ação. Com isso, executa uma recomendação inadequada ou toma uma decisão que não reflete a real intenção do usuário.

Circunstâncias de ocorrência

- Em interações humanas complexas ou subjetivas
- Em contextos culturais diversos
- Quando o agente opera sem histórico contextual suficiente
- Em sistemas que priorizam eficiência em detrimento de compreensão semântica profunda

2.2 - Camada Funcional - Memória e Representação Cognitiva



2.2.1 - Descrição

A camada de **Memória e Representação Cognitiva** é responsável por **armazenar, organizar, atualizar e estruturar o conhecimento** utilizado pelo agente de IA ao longo do tempo.

Ela registra fatos, decisões anteriores, contextos, preferências, estados do ambiente, aprendizados e informações relevantes para sustentar o raciocínio e orientar ações futuras.

Essa camada define o que o agente **"lembra"**, como ele **representa a realidade internamente** e **quais informações ele considera válidas ao tomar decisões**.

Diferente de bases de dados tradicionais, a memória cognitiva pode ser **dinâmica, contextual e evolutiva**, sendo constantemente atualizada a partir das percepções e interações do agente.

Por que os riscos associados a essa camada importam?

A memória é o **alicerce da continuidade cognitiva** do agente.

Se ela falha, o agente pode:

- raciocinar com base em informações incorretas ou obsoletas,
- repetir erros passados,
- perder contexto crítico, ou
- violar princípios de integridade, privacidade e segurança.

Riscos nessa camada são especialmente críticos porque **não afetam apenas uma decisão isolada**, mas **influenciam todas as decisões futuras**, ampliando impactos ao longo do tempo.

2.2.2 - Principais Riscos Cognitivos da Camada de Memória e Representação Cognitiva

Risco 1 – Memória Conflitante ou Desatualizada

Domínio: Administração Responsável de Dados

Princípio: Integridade dos Dados

Descrição

O agente mantém em sua memória informações que **se contradizem entre si** ou que já **não correspondem à realidade atual**, sem mecanismos adequados de atualização, versionamento ou descarte. Isso gera registros inconsistentes que comprometem a coerência do raciocínio.

Exemplo prático

Um agente de suporte registra que um cliente possui um contrato ativo, mas também armazena, em outro ponto da memória, a informação de que o contrato foi encerrado. Dependendo de qual memória é acessada, o agente pode oferecer serviços indevidos ou negar atendimentos legítimos.

Circunstâncias de ocorrência

- Em ambientes com múltiplas fontes de informação
- Quando não há políticas claras de atualização ou expiração de memória
- Em agentes que operam por longos períodos sem revisão de contexto
- Após mudanças frequentes de regras, contratos ou políticas de negócio

Risco 2 – Exposição Indevida de Informações na Memória

Domínio: Administração Responsável de Dados

Princípio: Proteção e Privacidade

Descrição

Informações sensíveis, pessoais ou confidenciais ficam armazenadas na memória do agente de forma inadequada, permitindo **acesso não autorizado**, uso indevido ou vazamento acidental desses dados durante interações ou decisões automatizadas.

Exemplo prático

Um agente conversacional memoriza dados pessoais fornecidos por um usuário em uma interação anterior e, posteriormente, reutiliza essas informações fora de contexto ou as expõe em uma conversa com outro usuário.

Circunstâncias de ocorrência

- Quando não há critérios claros sobre o que pode ou não ser memorizado
- Em agentes que mantêm memória de longo prazo sem filtros de privacidade
- Na ausência de segregação entre memória operacional e memória sensível
- Em ambientes com baixa maturidade em governança de dados

Risco 3 – Perda de Informações Importantes na Memória

Domínio: Segurança e Solidez Técnica

Princípio: Desempenho e Confiabilidade

Descrição

O agente deixa de registrar, manter ou recuperar informações relevantes que deveriam compor sua memória cognitiva. Isso resulta em **lacunas de contexto**, registros incompletos ou ausência de informações essenciais para decisões futuras.

Exemplo prático

Um agente de planejamento não registra decisões anteriores sobre restrições operacionais já acordadas. Em ciclos posteriores, ele propõe planos que violam essas restrições, exigindo correções manuais recorrentes.

Circunstâncias de ocorrência

- Em agentes com limitação de memória ou descarte automático mal configurado
- Quando há excesso de informações concorrentes
- Em sistemas que priorizam desempenho e velocidade em detrimento de persistência
- Durante falhas técnicas ou reinicializações não controladas

Risco 4 – Memória Contaminada por Informações Falsas ou Manipuladas

Domínio: Administração Responsável de Dados

Princípio: Integridade dos Dados

Descrição

A memória do agente é preenchida com **informações falsas, distorcidas ou manipuladas**, que passam a ser tratadas como fatos válidos. Essas informações contaminam o raciocínio e podem direcionar decisões incorretas ou prejudiciais.

Exemplo prático

Um agente que aprende com feedbacks de usuários armazena informações intencionalmente falsas inseridas por atores maliciosos. Com o tempo, essas informações passam a influenciar recomendações e decisões do agente.

Circunstâncias de ocorrência

- Em agentes que aprendem continuamente a partir de interações externas
- Em ambientes abertos ou com múltiplos usuários
- Quando não há validação, curadoria ou auditoria da memória
- Em cenários de ataque, manipulação ou uso adversarial

2.3 - Camada Funcional - Planejamento e Raciocínio



2.3.1 - Descrição

A camada de **Planejamento e Raciocínio** é responsável pelo **processo lógico, dedutivo e deliberativo** do agente de IA. É nessa camada que o agente **conecta informações, formula hipóteses, define objetivos, avalia alternativas, antecipa consequências e elabora planos de ação** para alcançar os resultados esperados.

Essa camada representa, na prática, como o **agente “pensa”**, como constrói cadeias de decisão e como transforma percepções e memórias em ações intencionais. Diferente de regras fixas, o raciocínio em agentes cognitivos pode ser **dinâmico, adaptativo e contínuo**, ajustando-se a novos contextos e aprendizados.

Por que os riscos associados a essa camada importam?

Os riscos dessa camada são particularmente críticos porque:

- afetam diretamente a **qualidade das decisões** tomadas pelo agente;
- podem gerar **ações aparentemente lógicas, mas conceitualmente erradas**;
- tendem a **passar despercebidos por métricas técnicas tradicionais**; e
- **impactam pessoas, processos e resultados de negócio de forma direta**.

Falhas no planejamento e no raciocínio podem levar agentes de IA a **agir fora do propósito definido, tomar decisões incoerentes, justificar erros com argumentos plausíveis ou adotar comportamentos autônomos não autorizados**.

2.3.2 - Principais Riscos Cognitivos da Camada de Planejamento e Raciocínio

Risco 1 – Raciocínio Errado ou Confuso

Domínio: Segurança e Solidez Técnica

Princípio: Desempenho e Confiabilidade

Descrição

O agente interpreta informações de forma equivocada e estabelece conexões incorretas entre elas, chegando a conclusões erradas porque seu raciocínio interno é **incompleto, confuso, inconsistente ou baseado em pressupostos inadequados**.

Exemplo prático

Um agente de apoio à decisão interpreta correlação como causalidade e conclui que uma ação específica gerou um resultado positivo, quando, na verdade, ambos foram influenciados por um fator externo. Com base nisso, passa a recomendar ações que não produzem os resultados esperados.

Circunstâncias de ocorrência

- Em cenários complexos ou com múltiplas variáveis
- Quando o agente opera com informações incompletas ou ambíguas
- Em situações de pressão por respostas rápidas
- Quando não há validação externa das premissas utilizadas

Risco 2 – Desvio de Metas Durante o Planejamento

Domínio: IA Responsável e Sustentável

Princípio: Autonomia

Descrição

O agente, ao tentar otimizar eficiência ou desempenho, passa a **priorizar objetivos não autorizados** ou que não fazem parte da meta principal definida pela organização, desviando-se da intenção original.

Exemplo prático

Um agente de otimização logística começa a reduzir custos priorizando apenas rotas mais baratas, ignorando restrições estratégicas relacionadas à qualidade do serviço ou à satisfação do cliente.

Circunstâncias de ocorrência

- Em agentes com alto grau de autonomia
- Quando objetivos são definidos de forma ambígua
- Em sistemas fortemente orientados à otimização
- Na ausência de limites claros ou critérios de parada

Risco 3 – Raciocínio Repetitivo ou Contraditório

Domínio: Segurança e Solidez Técnica

Princípio: Desempenho e Confiabilidade

Descrição

O agente fica preso em **ciclos de raciocínio repetitivos** ou passa a produzir conclusões que se contradizem, sem conseguir avançar ou manter coerência com informações e decisões anteriores.

Exemplo prático

Um agente de planejamento alterna constantemente entre duas recomendações opostas, sem conseguir decidir qual seguir, gerando instabilidade operacional e indecisão contínua.

Circunstâncias de ocorrência

- Em problemas mal definidos
- Quando há conflitos não resolvidos na memória
- Em ambientes com mudanças frequentes de contexto
- Quando o agente não possui mecanismos de resolução de impasse

Risco 4 – Mudança Não Autorizada de Metas

Domínio: IA Responsável e Sustentável

Princípio: Autonomia

Descrição

O agente passa a **modificar por conta própria os objetivos que deveria seguir**, adotando novas metas que não foram definidas ou aprovadas pelos responsáveis humanos.

Exemplo prático

Um agente de gestão de desempenho redefine internamente seus critérios de sucesso para maximizar métricas operacionais, mesmo que isso contradiga diretrizes estratégicas ou regulatórias.

Circunstâncias de ocorrência

- Em agentes autoajustáveis ou autoaprendizes
- Quando não há mecanismos de validação humana contínua
- Em ambientes com metas mal formalizadas
- Após longos períodos de operação sem supervisão

Risco 5 – Raciocínio Ético Deficiente

Domínio: IA Responsável e Sustentável

Princípio: Direitos Fundamentais Humanos

Descrição

O agente toma decisões sem considerar **implicações éticas**, impactos sociais, direitos fundamentais ou consequências para indivíduos, grupos ou o meio ambiente.

Exemplo prático

Um agente de triagem automatizada prioriza eficiência e custo, sem avaliar impactos discriminatórios ou injustos sobre determinados grupos de pessoas.

Circunstâncias de ocorrência

- Quando critérios éticos não são explicitamente incorporados
- Em decisões orientadas exclusivamente por métricas de desempenho
- Em contextos sociais sensíveis
- Na ausência de avaliação de impacto humano

Risco 6 – Viés Sistêmico de Alucinação Interagente

Domínio: Segurança e Solidez Técnica

Princípio: Segurança Geral e Robustez

Descrição

O agente **gera ou repete informações incorretas durante interações com outros agentes**, modelos ou módulos automatizados, fazendo com que erros se propaguem em cadeia e comprometam decisões e operações.

Exemplo prático

Um agente aceita como verdade uma inferência incorreta produzida por outro agente e a utiliza como base para decisões subsequentes, amplificando o erro ao longo do sistema.

Circunstâncias de ocorrência

- Em arquiteturas multiagentes
- Quando não há validação cruzada entre agentes
- Em ambientes altamente automatizados
- Em fluxos com dependência excessiva de inferências de terceiros

2.4 - Camada Funcional - Ação



2.4.1 - Descrição

A camada de **Ação** está relacionada à forma como o agente de IA **executa decisões no ambiente**, transformando planos e raciocínios em **comportamentos concretos**, comandos, recomendações ou intervenções automatizadas.

É nessa camada que a IA **impacta diretamente sistemas, processos, pessoas e o mundo real**.

Essa camada inclui desde ações simples (responder, recomendar, registrar) até ações complexas e críticas (aprovar transações, acionar sistemas, ajustar parâmetros operacionais ou iniciar fluxos automatizados).

Por isso, a camada de ação representa o **ponto de maior visibilidade e risco prático da atuação de um agente de IA**.

Por que os riscos associados a essa camada importam?

Mesmo quando percepção, memória e raciocínio funcionam corretamente, **falhas na execução podem gerar impactos graves.**

Riscos nessa camada importam porque:

- **materializam erros cognitivos** em efeitos reais;
- podem causar **danos imediatos e em cadeia**;
- afetam diretamente **confiança, segurança e responsabilidade**;
- muitas vezes são **difíceis de reverter após a execução.**

A **governança** da camada de ação é essencial para garantir que **decisões automatizadas sejam executadas no momento certo, com os limites corretos e sob o nível adequado de supervisão humana.**

2.4.2 - Principais Riscos Cognitivos da Camada de Ação

Risco 1 – Ação Executada Fora do Contexto Adequado

Domínio: Segurança e Solidez Técnica

Princípio: Segurança Geral e Robustez

Descrição

O agente executa uma ação em um **momento, situação ou contexto inadequado**, desconsiderando condições essenciais que deveriam orientar sua atuação.

Exemplo prático

Um agente de gestão operacional aplica automaticamente uma correção em um sistema durante um horário crítico de operação, causando interrupções que poderiam ter sido evitadas com a análise do contexto.

Circunstâncias de ocorrência

- Em ambientes altamente dinâmicos
- Quando o agente não avalia corretamente o estado atual do sistema
- Em cenários de exceção não previstos
- Quando regras contextuais não estão claramente definidas

Risco 2 – Ação Crítica sem Supervisão Humana

Domínio: IA Responsável e Sustentável

Princípio: Supervisão Humana

Descrição

O agente realiza de forma autônoma uma **ação crítica, sensível ou de alto impacto** sem que haja revisão, validação ou autorização humana prévia.

Exemplo prático

Um agente financeiro aprova automaticamente uma transação de alto valor sem que um responsável humano valide a decisão, gerando risco financeiro e regulatório.

Circunstâncias de ocorrência

- Em sistemas com alto grau de automação
- Quando limites de autonomia não estão claramente definidos
- Em processos que priorizam velocidade em detrimento de controle
- Na ausência de pontos formais de aprovação humana

Risco 3 – Ação com Impacto em Sequência

Domínio: Segurança e Solidez Técnica

Princípio: Segurança Geral e Robustez

Descrição

Uma decisão ou ação incorreta realizada pelo agente desencadeia **outras ações indesejadas em cadeia**, amplificando o impacto inicial e afetando múltiplos componentes do sistema.

Exemplo prático

Um agente de orquestração ajusta incorretamente um parâmetro de produção, o que aciona automaticamente outros sistemas a reagirem de forma inadequada, gerando um efeito dominó operacional.

Circunstâncias de ocorrência

- Em sistemas altamente integrados
- Em fluxos automatizados encadeados
- Quando não há mecanismos de contenção ou rollback
- Em arquiteturas com baixa tolerância a falhas

Risco 4 – Ação Induzida por Manipulação Externa

Domínio: Segurança e Solidez Técnica

Princípio: Segurança Geral e Robustez

Descrição

O agente executa uma ação porque foi **influenciado, enganado ou manipulado por uma fonte externa**, direcionando seu comportamento de forma inadequada ou maliciosa.

Exemplo prático

Um agente conversacional é induzido por comandos cuidadosamente formulados a executar ações não previstas, explorando falhas de interpretação ou validação.

Circunstâncias de ocorrência

- Em ambientes abertos ou expostos ao público
- Na ausência de filtros de segurança e validação
- Em cenários de ataque adversarial
- Quando o agente confia excessivamente em entradas externas

Risco 5 – Ações Contraditórias do Mesmo Agente

Domínio: Segurança e Solidez Técnica

Princípio: Segurança Geral e Robustez

Descrição

O mesmo agente executa **ações contraditórias ou inconsistentes ao lidar com situações semelhantes**, gerando imprevisibilidade e perda de confiabilidade.

Exemplo prático

Um agente aprova uma solicitação em um momento e rejeita uma solicitação idêntica em outro, sem justificativa clara ou mudança de contexto.

Circunstâncias de ocorrência

- Quando há inconsistências na memória ou no raciocínio
- Em ambientes com mudanças frequentes de estado
- Quando o agente não mantém coerência entre decisões passadas
- Na ausência de critérios claros de decisão

2.5 - Camada Funcional - Aprendizado e Reflexão Cognitiva



2.5.1 - Descrição

A camada de **Aprendizado e Reflexão Cognitiva** está relacionada à capacidade do agente de IA de **avaliar suas próprias ações, interpretar resultados, incorporar novas experiências e modificar seu comportamento futuro.**

É nessa camada que o agente ajusta como percebe, raciocina, planeja e age, com base em feedbacks, resultados observados ou padrões identificados ao longo do tempo.

Essa camada sustenta a **evolução contínua do agente**, permitindo adaptação a novos contextos, melhoria de desempenho e refinamento de decisões. Ao mesmo tempo, ela representa um dos **pontos mais sensíveis da autonomia cognitiva**, pois mudanças internas podem ocorrer sem serem imediatamente visíveis para os humanos.

Por que os riscos associados a essa camada importam?

Os riscos dessa camada são especialmente críticos porque:

- **alteram o comportamento futuro do agente**, e não apenas uma ação pontual;
- podem gerar **desvios progressivos e cumulativos**, difíceis de detectar;
- afetam diretamente **previsibilidade, controle e confiança**; e
- podem levar o agente a **“aprender o errado”** e reforçar esse aprendizado ao longo do tempo.

Sem governança adequada, o aprendizado e a reflexão podem transformar o agente em um sistema que **evolui fora da intenção humana**, mesmo que suas decisões individuais pareçam razoáveis.

2.5.2 - Principais Riscos Cognitivos da Camada de Aprendizado e Reflexão Cognitiva

Risco 1 – Mudança Cognitiva Não Autorizada

Domínio: IA Responsável e Sustentável

Princípio: Autonomia

Descrição

O agente altera, por conta própria, a forma como pensa, aprende ou decide, realizando **mudanças internas em seus processos cognitivos** sem validação, autorização ou supervisão humana.

Exemplo prático

Um agente de recomendação ajusta internamente seus critérios de decisão para priorizar engajamento extremo, mesmo que isso contrarie diretrizes éticas ou estratégicas definidas pela organização.

Circunstâncias de ocorrência

- Em agentes com aprendizado contínuo
- Quando não há checkpoints de validação humana
- Em sistemas que operam por longos períodos de forma autônoma
- Na ausência de limites explícitos para autoevolução

Risco 2 – Aprendizado Indevido pelo Agente

Domínio: IA Responsável e Sustentável

Princípio: Autonomia

Descrição

O agente aprende comportamentos, padrões ou estratégias **não planejados, não desejados ou não autorizados**, modificando sua atuação de forma não prevista pelas equipes responsáveis.

Exemplo prático

Um agente aprende, a partir de interações repetidas, que pode contornar regras internas para atingir objetivos mais rapidamente, incorporando esse comportamento como padrão.

Circunstâncias de ocorrência

- Em ambientes com feedbacks implícitos ou ambíguos
- Em sistemas expostos a interações humanas variadas
- Quando critérios de aprendizado não são claramente definidos
- Em agentes que aprendem por tentativa e erro

Risco 3 – Autoajuste pelo Agente Não Validado

Domínio: IA Responsável e Sustentável

Princípio: Autonomia

Descrição

O agente **modifica autonomamente seus parâmetros internos, regras, pesos ou modelos**, sem qualquer processo de validação humana, o que pode comprometer desempenho, segurança ou alinhamento estratégico.

Exemplo prático

Um agente de otimização ajusta seus parâmetros para maximizar eficiência local, mas degrada o desempenho global do sistema, sem que isso seja percebido imediatamente.

Circunstâncias de ocorrência

- Em sistemas com mecanismos de autoajuste automático
- Quando não há governança sobre alterações internas
- Em ambientes de alta complexidade
- Na ausência de logs e rastreabilidade de mudanças

Risco 4 – Aplicação Indevida de Regras Aprendidas

Domínio: IA Responsável e Sustentável

Princípio: Autonomia

Descrição

O agente aplica conhecimentos ou regras aprendidas em **contextos inadequados**, falhando em reconhecer diferenças relevantes entre situações e tomando decisões incorretas.

Exemplo prático

Um agente aprende que determinado procedimento é eficaz em um cenário específico e passa a aplicá-lo indiscriminadamente em outros contextos, gerando resultados negativos.

Circunstâncias de ocorrência

- Em ambientes com contextos variados
- Quando o agente não diferencia corretamente situações
- Em sistemas com generalização excessiva
- Após aprendizado baseado em poucos exemplos

Risco 5 – Autojustificativas Incorretas do Agente

Domínio: IA Responsável e Sustentável

Princípio: Supervisão Humana

Descrição

O agente produz **explicações falsas, incompletas ou incoerentes para justificar suas decisões**, dificultando a compreensão humana e mascarando falhas reais no seu raciocínio ou aprendizado.

Exemplo prático

Um agente apresenta uma justificativa plausível para uma decisão errada, fazendo parecer que seguiu critérios adequados, quando na verdade o erro decorreu de um aprendizado inadequado.

Circunstâncias de ocorrência

- Em agentes que geram explicações automaticamente
- Quando não há validação da coerência entre decisão e justificativa
- Em sistemas com baixa transparência cognitiva
- Em ambientes que exigem explicações rápidas

Risco 6 – Aprendizado de Erros como Comportamento Aceitável

Domínio: IA Responsável e Sustentável

Princípio: Autonomia

Descrição

O agente interpreta incorretamente suas próprias experiências e passa a considerar **erros recorrentes como aceitáveis**, reforçando comportamentos inadequados ao longo do tempo

Exemplo prático

Um agente percebe que decisões incorretas não geram consequências imediatas e passa a repeti-las, incorporando o erro como padrão de comportamento.

Circunstâncias de ocorrência

- Quando não há feedback corretivo claro
- Em ambientes com baixa supervisão humana
- Em sistemas que priorizam velocidade sobre qualidade
- Em processos onde falhas não são explicitamente sinalizadas

3 - Integração dos Riscos Cognitivos no Ciclo de Vida da IA

Ao longo deste Guia Prático, foi possível compreender que os **riscos cognitivos não surgem de forma isolada**. Eles emergem, se acumulam e se amplificam **ao longo de todo o ciclo cognitivo dos agentes de IA**, acompanhando a forma como a solução percebe, lembra, raciocina, age e aprende.

Cada camada cognitiva cumpre um papel essencial e, ao mesmo tempo, representa um **ponto específico de vulnerabilidade**:



O ponto central é que **riscos cognitivos são riscos sistêmicos**, uma falha inicial pode atravessar camadas, reforçar-se com o tempo e resultar em **comportamentos emergentes, imprevisíveis ou não autorizados**.

Por isso, a Governança de IA, especialmente em soluções agênticas, não pode se limitar a dados, modelos ou métricas técnicas.

Ela precisa **alcançar a cognição da IA**, estabelecendo mecanismos que permitam:

- visibilidade sobre como o agente percebe, decide e aprende;
- limites claros de autonomia em cada camada;
- validação humana proporcional ao impacto das decisões;
- rastreabilidade e explicabilidade do comportamento cognitivo; e
- correção contínua antes que pequenos desvios se tornem riscos críticos.

Este Guia Prático foi concebido, como um primeiro passo, para apoiar profissionais, líderes e equipes a **identificar, analisar e antecipar riscos cognitivos** de forma estruturada, conectando governança, tecnologia e impacto real.

Em um cenário onde agentes de IA deixam de ser apenas ferramentas e passam a atuar como **atores autônomos dentro das organizações, governar a cognição da IA deixa de ser opcional** e torna-se um requisito fundamental para confiança, segurança e sustentabilidade no uso da inteligência artificial.

4 - Fundamentos conceituais e metodológicos do XGov AI®

Nota metodológica sobre as fontes e a construção do conhecimento

O conjunto de riscos cognitivos apresentado neste trabalho é resultado de um processo intencional, aprofundado e contínuo de pesquisa, análise crítica e criação conceitual. As fontes listadas a seguir foram utilizadas como **base teórica e empírica de sustentação**, e não como guias prescritivos ou modelos a serem replicados diretamente.

Os conceitos, classificações e descrições de riscos cognitivos não foram extraídos de forma literal dessas referências, tampouco utilizam terminologia idêntica à apresentada nos materiais originais. O esforço central deste trabalho consistiu em **interpretar, correlacionar e traduzir conhecimentos dispersos**, oriundos da literatura técnica, acadêmica e prática, para um modelo estruturado, aplicável e alinhado à governança de soluções de Inteligência Artificial no contexto organizacional.

O **Framework OpGov AI®** foi inspirado na estrutura conceitual proposta pelo **The Alan Turing Institute**, especialmente no que se refere à aplicação da governança ao longo do ciclo de vida das soluções de IA, bem como à organização por domínios e princípios. No entanto, essa estrutura foi **substancialmente adaptada e expandida**, dando origem a uma taxonomia própria, desenvolvida para atender às necessidades práticas de análise, decisão e responsabilização em projetos reais de IA.

Essa adaptação inclui, entre outros elementos, o uso sistemático de **Key Assessment Questions (KAQs)** para identificação e análise contextual de riscos, a definição de **Micro-OKRs™** como mecanismo de mitigação orientado a resultados verificáveis, a **atribuição explícita de responsabilidades** ao longo do ciclo de vida e a integração desses componentes em artefatos operacionais de governança.

Complementarmente, este trabalho incorpora de forma significativa a **experiência prática da autora no desenvolvimento de Agentes de Governança**, responsáveis por apoiar, e futuramente executar de forma autônoma, atividades do próprio Framework OpGov AI®. A vivência direta na concepção, implementação, testes e observação do comportamento desses agentes contribuiu para a identificação de riscos cognitivos que extrapolam a literatura tradicional, especialmente aqueles relacionados a memória, planejamento, adaptação, inferência e interação contínua com humanos e outros sistemas.

Assim, as fontes apresentadas a seguir devem ser compreendidas como **fundamento intelectual e ponto de partida**, e não como delimitação do escopo conceitual. O conhecimento aqui consolidado reflete um investimento consistente em estudo, experimentação e síntese crítica, com o objetivo de oferecer uma abordagem original, prática e evolutiva para a governança de sistemas de IA cada vez mais autônomos, cognitivos e integrados às decisões organizacionais.

Nota sobre o uso de imagens

As imagens utilizadas neste guia foram produzidas com o apoio de ferramentas de Inteligência Artificial generativa, com finalidade exclusivamente ilustrativa e conceitual. Essas imagens não representam pessoas reais, organizações específicas ou situações concretas, sendo utilizadas como recurso visual de apoio à compreensão dos conceitos apresentados.

Referências bibliográficas

Livros e publicações técnicas

IJLAL, Taimur. **Agentic AI: Navigating Risks and Security Challenges – A Beginner’s Guide to Securing and Governing AI Agents**. Edição do autor, abr. 2025.

LANE, Mark A. **Agentic AI Handbook: Design Patterns – Build an Agent AI that Thinks, Plans and Delivers**. Version 1.0.6. Design Agentic LLC, mar. 2025. ISBN 979-8-9925718-0-6.

POULOS, Jesse M. **The AI Governance Architect: Building Model Risk Management and Compliance Frameworks**. © 2025. Todos os direitos reservados.

ALAN TURING INSTITUTE. **AI Ethics and Governance in Practice**. Disponível em:
<https://www.turing.ac.uk/research/research-projects/ai-ethics-and-governance-practice>.

5 - Mapa Visual de Riscos Cognitivos em Agentes de IA

O conteúdo apresentado até aqui descreve os fundamentos conceituais e metodológicos dos riscos cognitivos em agentes de IA, explicando por que eles surgem, em que contextos e como devem ser analisados ao longo do ciclo de desenvolvimento.

A seguir, este conhecimento é sintetizado em um Mapa Visual, que consolida de forma integrada:

- as etapas do desenvolvimento de agentes nas quais a probabilidade de ocorrência dos riscos deve ser analisada,
- as camadas cognitivas envolvidas,
- e as zonas onde a análise de riscos cognitivos é essencial.

Este mapa não representa riscos identificados em uma solução específica.

Ele indica **riscos cognitivos potenciais que devem ser analisados durante o desenvolvimento** de agentes de IA, conforme o framework XGov AI®.

Por se tratar de uma síntese visual estruturante, o mapa é apresentado em formato horizontal, para permitir uma leitura clara e completa das relações entre etapas, camadas cognitivas e princípios de governança.

Recomendação: visualize o mapa em tela cheia para melhor compreensão.

Mapa Visual de Riscos Cognitivos em Agentes de IA

XGOV AI® Framework

Etapa - Especificação da Cognição e da Arquitetura

PERCEPÇÃO	Percepção Equivocada da Situação	Percepção Baseada em Fonte Única ou Instável	Interpretação Contextual Incorreta
MEMÓRIA	Memória conflitante ou desatualizada	Exposição indevida de informações na memória	
PLANEJAMENTO E RACIOCÍNIO	Desvio de Metas Durante o Planejamento	Mudança Não Autorizada de Metas	Raciocínio Ético Deficiente
AÇÃO	Ação Crítica sem Supervisão Humana		
APRENDIZADO E REFLEXÃO	Aprendizado Indevido pelo Agente	Mudança Cognitiva Não Autorizada	Aplicação Indevida de Regras Aprendidas
Princípios da Governança da IA =>			
	Desempenho e Confiabilidade	Integridade dos Dados	Proteção e Privacidade de Dados
		Direitos Fundamentais Humanos	Supervisão Humana
			Autonomia

Mapa Visual de Riscos Cognitivos em Agentes de IA

XGOV AI® Framework

Etapa - Análise Contextual, Situacional e Temporal dos Dados

Etapa - Desenvolvimento e Treinamento dos Componentes Inteligente e Políticas de Decisão

PERCEPÇÃO			
MEMÓRIA	Memória contaminada por informações falsas ou manipuladas	Perda de informações importantes na memória	
PLANEJAMENTO E RACIOCÍNIO		Raciocínio errado ou confuso	Raciocínio repetitivo ou contraditório
AÇÃO			
APRENDIZADO E REFLEXÃO			

Princípios da Governança da IA =>

Desempenho e Confiabilidade

Integridade dos Dados

Mapa Visual de Riscos Cognitivos em Agentes de IA

XGOV AI® Framework

Etapa - Teste e Validação Técnica e Comportamental

PERCEPÇÃO				
MEMÓRIA				
PLANEJAMENTO E RACIOCÍNIO	Percepção Equivocada da Situação			
AÇÃO	Ação Executada Fora do Contexto Adequado	Ação com Impacto em Sequencial	Ação Induzida por Manipulação Externa	Ações Contraditórias do Mesmo Agente
APRENDIZADO E REFLEXÃO	Autoajuste Pelo Agente Não Validado	Aprendizado de Erros como Comportamento Aceitável	Autojustificativas Incorretas do Agente	

Princípios da Governança da IA =>

Segurança e Robustez

Autonomia

Supervisão Humana