REVISTA TÓPICOS
ISSN 2965-6672 | Qualis A2

DA ALUCINAÇÃO DA IA AO CONHECIMENTO VALIDADO: O PAPEL DA COLABORAÇÃO HUMANA

FROM AI HALLUCINATION TO VALIDATED KNOWLEDGE: THE ROLE OF HUMAN COLLABORATION

Engenharias • 22/09/2026

REGISTRO DOI: 10.70773/revistatopicos/790001371

Filipe Marinho da Costa Pinto[1]

João Nóbrega Marques[2]

Ricardo Moreira da Silva[3]

Josilene Aires Moreira[4]

RESUMO

A disseminação de sistemas de Inteligência Artificial (IA) generativa introduziu nas organizações o fenômeno das alucinações (a geração de outputs plausíveis, porém factualmente incorretos ou infiéis ao contexto). Este estudo analisa, por meio de uma revisão integrativa da literatura, os desafios sociotécnicos que tais inconsistências impõem aos processos de validação e confiabilidade do conhecimento organizacional. Como resultado, propõe-se uma matriz conceitual integradora que cruza as correntes de erro algorítmico (alucinações intrínsecas, extrínsecas, de factualidade e de fidelidade) com os níveis de planejamento estratégico, tático e operacional. Frente a esse gargalo tecnológico, a pesquisa defende que o conhecimento autêntico permanece como uma faculdade intrínseca ao ser humano. Conclui-se que a transição de ilusões estatísticas para insights válidos depende obrigatoriamente de uma simbiose sociotécnica estruturada, na qual a cognição humana pluridimensional, segmentada em conhecimento tecnológico, intuitivo, genialidade criativa e capacidade de modelagem abstrata antecedente, atua como o mecanismo central de governança e curadoria crítica na colaboração com a IA.

Palavras-chave: Conhecimento Humano; Inteligência Artificial; Alucinações da IA; Colaboração Humana.

ABSTRACT

The dissemination of generative Artificial Intelligence (AI) systems has introduced the phenomenon of hallucinations into organizations (the generation of plausible outputs that are factually incorrect or unfaithful to the context). This study analyzes, through an integrative literature review, the sociotechnical challenges that such inconsistencies impose on the processes of validation and reliability of organizational knowledge. As a result, an integrative

conceptual matrix is proposed that crosses the streams of algorithmic error (intrinsic, extrinsic, factuality and faithfulness hallucinations) with the levels of strategic, tactical and operational planning. Faced with this technological bottleneck, the research argues that authentic knowledge remains an intrinsic human faculty. It is concluded that the transition from statistical illusions to valid insights necessarily depends on a structured sociotechnical symbiosis, in which multidimensional human cognition, segmented into technological knowledge, intuitive knowledge, creative genius and antecedent abstract modeling capacity, acts as the central mechanism of governance and critical curation in collaboration with AI.
Keywords: Human Knowledge; Artificial Intelligence; AI Hallucinations; Human Collaboration.

1. INTRODUÇÃO

A transição global rumo à era digital tem reconfigurado a dinâmica das organizações, impulsionando debates multidisciplinares sobre os desafios operacionais, as oportunidades de inovação e a formulação de políticas institucionais robustas frente ao avanço tecnológico rápido (Dwivedi et al., 2021).

Historicamente alicerçada nos pilares da Indústria 4.0, caracterizada pela hiperconectividade, automação e sistemas ciberfísicos (Xu; Xu; Li, 2018), a atual mudança estratégica das empresas exige a incorporação da Inteligência Artificial (IA) como um vetor essencial de competitividade (Borges et al., 2021). Paradoxalmente, o paradigma industrial contemporâneo tem evoluído em direção à Indústria/Sociedade 5.0, movimento que resgata a centralidade do ser humano ao harmonizar as tecnologias digitais avançadas com o

bem-estar social, a sustentabilidade e o futuro do trabalho (Bolis et al., 2025). Nesse ambiente hiperconectado, a transformação digital atua como um habilitador da criação do conhecimento, estimulando a capacidade de inovação e o aprendizado contínuo dentro das organizações (Chen et al., 2024).

Essa busca de vantagem competitiva reside tradicionalmente na habilidade de converter o conhecimento tácito, inerente à experiência humana, em conhecimento explícito, criativo e rentável, formalizado e estruturado (Nonaka; Takeuchi, 2008), entretanto, a emergência da inteligência artificial generativa introduz uma nova lente conceitual para essas teorias, redefinindo tarefas e fluxos de trabalho tradicionais através de grandes modelos de linguagem (LLMs), usados em sistemas como o ChatGPT (Korzyński et al., 2023), Gemini e outros.

Longe de consolidar uma substituição da força de trabalho, o cenário aponta para o estabelecimento de uma simbiose entre humanos e IA, desenhando uma parceria colaborativa nos sistemas de gestão e nos processos de tomada de decisão organizacional (Jarrahi et al., 2023).

Apesar do potencial tecnológico disruptivo, a adoção corporativa da IA generativa carrega riscos severos associados à confiabilidade de seus outputs (Dai et al., 2026). O fenômeno das chamadas "alucinações" representa um dos principais gargalos para a sua consolidação prática e segura (Ji et al., 2023). Trata-se de situações em que o modelo gera conteúdos aparentemente plausíveis e dotados de uma falsa fluência linguística, mas que se revelam factualmente incorretos, distorcidos ou semanticamente desconexos do contexto real (Ji et al., 2023).

A literatura recente investiga ativamente as origens dessas inconsistências, mapeando se elas decorrem do comportamento estatístico intrínseco do modelo ou das estratégias de prompting utilizadas pelos usuários (Anh-Hoang et al., 2025). Essa distorção de conteúdo desafia o uso seguro de sistemas baseados em arquiteturas GPT (Generative Pre-trained Transformer) (Madzhumder; Begunova, 2025), gerando percepções negativas e quebras de expectativa por parte dos usuários finais (Larsen et al., 2026).

Para compreender a raiz técnica dessa vulnerabilidade, faz-se necessário examinar os três pilares fundamentais que estruturam a arquitetura GPT, desenvolvida originalmente a partir do mecanismo de Transformer introduzido pelo Google em 2017:

- G - Generative (Generativo): Refere-se à capacidade computacional do modelo de gerar novos conteúdos autônomos (como textos, códigos de programação ou análises) a partir de uma instrução (prompt) enviada pelo usuário.

- P - Pre-trained (Pré-treinado): Indica que a rede neural passou por um treinamento massivo prévio com bases gigantescas de dados históricos extraídos da internet (livros, artigos, sites), onde aprendeu regras gramaticais e associações semânticas de forma puramente estatística.

- T - Transformer (Transformador): Representa o coração da arquitetura, que através do mecanismo de "Atenção" (Self-Attention) permite à máquina calcular o peso e a relevância de cada palavra dentro de uma frase simultaneamente, garantindo a coesão do texto gerado.

Na prática, embora as respostas simulem com precisão a fluidez do raciocínio humano, o modelo opera como um mecanismo de aproximação probabilística de tokens subsequentes. Como a arquitetura prioriza a coesão sintática e a continuidade linguística em detrimento da verificação ontológica da verdade, o sistema gera as supracitadas alucinações sempre que as restrições estatísticas falham em se alinhar aos fatos reais do mundo.

Diante dessa problemática, o mapeamento dos limites dessas inconsistências (Gujjar; Kumar, 2025) e o desenvolvimento de taxonomias robustas tornaram-se prioridades científicas urgentes (Huang et al., 2025). O impacto dessas falhas estende-se a setores críticos, como a cibersegurança, onde uma classificação errônea pode comprometer infraestruturas corporativas inteiras (Sood et al., 2025).

Para mitigar esses efeitos e transformar "ilusão em insight", estudiosos têm concentrado esforços na criação de técnicas de mitigação (Kazlaris et al., 2025) e em debates epistemológicos que propõem refinar os termos para diferenciar meros erros de execução de irregularidades sistêmicas (Ariso; Bannister, 2026). Compreender e intervir nessa fronteira entre os benefícios e as vulnerabilidades da IA é o requisito fundamental para o desenho de sistemas seguros e verdadeiramente inteligentes (Panda; Puri, 2025).

Objetivando preencher lacunas práticas de governança, este artigo busca discutir qual é o papel do conhecimento humano e como suas múltiplas dimensões podem contribuir para a confiabilidade da colaboração humano-IA frente ao fenômeno das alucinações nas organizações.

Por meio de uma revisão integrativa focada nas taxonomias de erro existentes, este estudo propõe um modelo conceitual que vincula os quadrantes de falhas algorítmicas aos níveis de planejamento organizacional e às dimensões da cognição humana, fornecendo um referencial para a validação do conhecimento na era da inteligência artificial generativa.

2. FUNDAMENTAÇÃO TEÓRICA OU REVISÃO DA LITERATURA

2.1. A Natureza Pluridimensional do Conhecimento Humano e a Fronteira Algorítmica

A introdução da Inteligência Artificial Generativa nas organizações atuais exige uma expansão da divisão do conhecimento, antes, simplesmente entre o explícito, passível de codificação e estruturação, e o tácito, enraizado na experiência vivida, na ação e no contexto individual proposto, no final do século XX por Nonaka e Takeuchi (2008) (Böhm; Durst, 2026).

Na verdade, a cognição humana não se limita à mera execução lógica (explicita ou mesmo tácita), pois ela se manifesta como uma capacidade pluridimensional que engloba o conhecimento tecnológico, o conhecimento insightful (intuitivo), a genialidade criativa e a capacidade de modelagem abstrata antecedente.

A compreensão aprofundada de cada uma dessas capacidades revela a assimetria fundamental entre a arquitetura cognitiva humana e os mecanismos probabilísticos das máquinas:

- **Conhecimento Tecnológico:** Representa a capacidade humana de orquestrar ferramentas, compreender as nuances operacionais de sua aplicação no trabalho e externalizar

processos lógicos em sistemas ciberfísicos (Bolis et al., 2025). Essa dimensão transcende a mera operação técnica instrumental; ela envolve uma "cognição corporificada" (embodied cognition), onde o uso de artefatos tecnológicos altera a própria percepção do agente sobre os problemas organizacionais (Clark; Chalmers, 1998). Além disso, o conhecimento tecnológico humano caracteriza-se pelo domínio crítico sobre os regimes sociotécnicos, permitindo que o profissional compreenda o impacto institucional e as limitações ético-políticas de um sistema integrado de informação, algo que o artefato técnico é incapaz de auto reconhecer (Geels, 2004). O especialista humano detém, portanto, a compreensão ontológica do funcionamento da máquina, permitindo-lhe calibrá-la, parametrizá-la e operá-la não como um mero receptor, mas como o agente diretor do processo socioeconômico (ILO, 2026).

- **Conhecimento Insightful (Intuitivo):** Refere-se à aptidão da mente humana para realizar julgamentos rápidos, imediatos e precisos sem a necessidade de passar por um raciocínio lógico-linear deliberado ou por processamento formal de dados. Psicologicamente, o insight se manifesta como o fenômeno do "reconhecimento repentino" (Aha-Erlebnis), em que o cérebro humano reestrutura subitamente um problema complexo a partir de pistas contextuais difusas e não-lineares, ativando redes neuronais hemisféricas distintas daquelas usadas para processamento sequencial lógico (Bowden et al., 2005). Sob a ótica organizacional, esse conhecimento intuitivo é sustentado pelas "heurísticas rápidas e frugais", que capacitam tomadores de decisão experientes a agir sob condições de extrema incerteza ambiental através de padrões internalizados que

ignoram deliberadamente excessos de informação quantitativa (Gigerenzer; Gaissmaier, 2011). Enquanto os grandes modelos de linguagem processam tokens sequenciais por aproximação estatística retrospectiva (Alansari; Luqman, 2026), a cognição humana realiza saltos heurísticos, detectando instantaneamente anomalias, padrões sutis e quebras de contexto que escapam a validadores matemáticos puramente automatizados.

- **Genialidade Criativa:** Define-se pela faculdade de conectar domínios semânticos e conceituais radicalmente distantes para conceber soluções originais, disruptivas e que possuam valor prático reconhecido. Do ponto de vista cognitivo, a criatividade humana opera por meio do "pensamento divergente" e do "mecanismo de variação cega e retenção seletiva" (BVSR), um processo de busca heurística que gera mutações conceituais que rompem deliberadamente com as tendências probabilísticas mais prováveis (Simonton, 1999). Essa genialidade criativa requer flexibilidade cognitiva para transitar entre redes de controle executivo e redes de modo padrão (default mode network), permitindo que a mente humana gere sínteses estéticas e conceituais sem qualquer base de dados correlacionada anterior (Dietrich, 2004). Máquinas generativas criam por recombinação estatística dentro de um espaço amostral restrito aos dados de seu treinamento (Huang et al., 2025). A genialidade criativa humana, por outro lado, manifesta-se no romper de paradigmas e na geração de ideias genuinamente inéditas, operando com eficiência mesmo diante da total escassez de dados históricos pregressos.

- **Capacidade de Modelagem Abstrata Antecedente:** Esta faculdade puramente mental e especulativa representa a capacidade de projetar, simular e validar cenários teóricos, premissas lógicas complexas e construtos abstratos antes mesmo que qualquer evidência física, dado empírico ou representação concreta sobre eles existam no mundo real. Filosoficamente, essa capacidade assenta-se na faculdade da "intencionalidade intrínseca", o poder exclusivo da mente de direcionar seus pensamentos para objetos e estados de coisas inexistentes ou puramente hipotéticos, dotando o ser humano de uma compreensão semântica profunda e independente de símbolos sintáticos mecânicos (Searle, 1980). Na psicologia do desenvolvimento, essa faculdade apoia-se nos "modelos mentais prospectivos" e na capacidade de realizar simulações contrafactuais, permitindo ao cérebro humano prever falhas de sistemas complexos simulando o que poderia acontecer, em vez de apenas computar o que já aconteceu (Johnson-Laird, 2010). Enquanto as arquiteturas GPT operam de maneira retrospectiva (dependendo estritamente de dados passados e do input imediato do usuário para iniciar o seu cálculo (Korzyński et al., 2023), o ser humano consegue conceber antecipadamente modelos de mundo inteiramente novos. É essa arquitetura mental anterior que permite ao profissional realizar a engenharia de contexto refinada (prompting estratégico), enclausurando o espaço operacional do algoritmo para evitar desvios lógicos.

2.2. A Cognição Humana Como Mecanismo de Governança Anti-Alucinação

Longe de Ser Um Problema Técnico Simples, a Distorção de Conteúdo Desafia o Uso Seguro de Sistemas de Informação e Gestão (Panda; Puri, 2025). Para Transitar da "Ilusão Ao Insight" (Kazlaris Et Al., 2025), a Arquitetura dos Sistemas Contemporâneos Deve Migrar de Uma Automação Isolada para Uma Simbiose Humano IA Transformando a Cognição Humana no Principal Vetor de Governança e Validação Epistêmica (Jarrahi Et Al., 2023).

O conhecimento humano, em suas múltiplas vertentes, atua como um sistema dinâmico de defesa contra as falhas algorítmicas através de três frentes estratégicas:

- **Análise Crítica e Filtro de Plausibilidade:** Modelos generativos são otimizados para a fluência linguística, o que torna suas alucinações perigosamente convincentes (Huang et al., 2025). O conhecimento factual, aliado à experiência prática do especialista humano (conhecimento tácito), opera como um filtro em tempo real, detectando desvios conceituais e irregularidades interpretativas que escapam aos validadores estatísticos automatizados (Ariso; Bannister, 2026).

- **Engenharia de Contexto Habilitada por Modelagem Antecedente:** A ocorrência de alucinações está fortemente atrelada à qualidade e ambiguidade das instruções fornecidas às máquinas (Anh-Hoang et al., 2025). O profissional dotado de capacidade de modelagem lógica atua na formulação de estratégias avançadas de *prompting*, restringindo o espaço amostral do modelo a partir de regras semânticas estritas e cenários hipotéticos desenhados pela inteligência humana, mitigando riscos antes mesmo da execução do algoritmo.

- **Validação Cruzada Multidisciplinar:** O impacto de falhas informacionais pode ser catastrófico em setores sensíveis, como a cibersegurança (Sood et al., 2025). A governança anti-alucinação se estrutura quando os insights humanos e a capacidade de julgamento ético e contextual são codificados de volta para o sistema, refinando os algoritmos através de feedback humano estruturado (Dai et al., 2026; Dwivedi et al., 2021).

- Dessa forma, a construção de sistemas anti-alucinações eficazes não reside na busca por um algoritmo perfeito e isolado, mas no desenho de fluxos de trabalho sociotécnicos onde a máquina processa o volume explícito de dados (Xu; Xu; Li, 2018), enquanto o capital humano exerce a curadoria, a supervisão crítica e o direcionamento ético-estratégico da cognição (Bolis et al., 2025).

3. METODOLOGIA

Para Identificar as Principais Classificações e Tipologias de Alucinações em Sistemas de Inteligência Artificial, Foi Conduzida Uma Revisão Apoiada por Uma Estratégia de Busca Estruturada na Base Scopus. A Busca Foi Inicialmente Realizada por Meio da Seguinte String de Pesquisa: (Title (Hallucination* Or "Hallucinated Output*" Or "Hallucinated Content") And Title-Abs-Key ("Artificial Intelligence" Or "Ai")

A string teve como intuito relacionar os dois grupos temáticos: IA e alucinações. A decisão por limitar os descritores voltados para alucinações para o título teve como intuito voltar a busca para estudos nos quais fosse tema central, um dos pressupostos para que

o artigo propusesse alguma classificação para os erros. Posteriormente, foram adicionados filtros limitando para artigos publicados na língua inglesa, entre 2020 e 2026 e no tipo de documento article e review.

- **Seleção dos Estudos -** Foram definidos como elegíveis os estudos que apresentassem propostas de classificação, tipologia ou taxonomia das alucinações de IA, identificáveis a partir do título e do resumo. Foram excluídos os trabalhos que abordavam o fenômeno das alucinações sem propor qualquer forma de categorização.

  A busca inicial resultou em 195 registros, dos quais 11 atenderam aos critérios de elegibilidade. Entretanto, foi possível obter acesso ao texto completo de apenas sete estudos. Considerando o caráter narrativo da revisão e o objetivo de construir um panorama conceitual das tipologias existentes, adicionalmente a amostra foi complementada com três estudos de referência reconhecidos na literatura, identificados durante o aprofundamento teórico sobre o tema e a inclusão desses estudos fundamentou-se em sua relevância para a consolidação das classificações discutidas, totalizando dez artigos analisados.

  O reduzido número de estudos incluídos não representa uma limitação do procedimento de busca, mas reflete o estágio de maturidade da literatura sobre o tema. Embora as alucinações em IA tenham despertado crescente interesse científico, poucos estudos se dedicam à proposição de classificações ou taxonomias do fenômeno.

Além disso, a própria terminologia (alucinação) empregada para descrever o fenômeno observado ainda não se encontra consolidada, havendo divergências conceituais na literatura (Ariso; Bannister, 2026). Dessa forma, o propósito desta etapa foi identificar os principais referenciais conceituais capazes de sustentar a análise das implicações das alucinações, em mecanismos de gestão nas organizações.

- **Procedimento de Análise e Síntese Conceitual -** Na etapa de análise, realizou-se uma análise de conteúdo sobre os artigos, buscando identificar os critérios utilizados pelos autores para classificar as alucinações, as dimensões enfatizadas em cada proposta e as convergências e divergências entre as taxonomias existentes. Com base na síntese comparativa das taxonomias identificadas, foi elaborado um modelo conceitual integrador das principais categorias de alucinações descritas na literatura.

A síntese das classificações identificadas permitiu estabelecer relações conceituais entre os diferentes tipos de alucinação e os mecanismos de criação a partir dos tipos conhecimento humanos (tecnológico, intuitivo, genialidade criativa e capacidade de modelagem abstrata antecedente).

4. RESULTADOS E DISCUSSÕES

4.1. Caracterização das Alucinações

Com Base nos Artigos Base, Foi Elaborado Um Quadro Resumo Sobre o Que Abrange as Definições de Alucinação Até as Taxonomias Mais Recentes Aplicadas a Contextos Específicos, Discutidas nos Artigos.

Quadro 1. Quadro Resumo de Categorização

Autor	Categorização	Resumo
ALANSARI & LUQMAN (2026)	Factualidade (Fabricação e Contradição) e Fidelidade (Inconsistência lógica, de contexto e de instrução).	Revisão panorâmica recente que conecta causas em todo o ciclo de vida do LLM, com ênfase em explicabilidade. Explora alucinações em tarefas específicas (Tradução, Sumarização, QA, Código).
ANN-HOANG et al. (2025)	Intrínseca, Extrínseca, Factual e Lógica	Foca na sensibilidade do modelo ao prompt e na variabilidade intrínseca do modelo para diagnosticar o erro.
GUJJAR & PRUMAR (2025)	Estágios de Criação: Problemas no Treinamento (dados, overfitting) e problemas no Uso (prompts, rotulagem).	Classifica as fontes de alucinação para sugerir medidas corretivas durante o ciclo de vida do modelo.
HUANG et al. (2025)	Factualidade (Contradição/Fabricação) e Fidelidade (Inconsistência de instrução, de contexto e lógica).	Propõe uma taxonomia refinada para a era dos LLMs, focando em aplicações centradas no usuário e alinhamento.
JI et al. (2023)	Intrínseca (contradiz a fonte) e extrínseca (informação não verificável).	Pioneiro em definir os dois tipos clássicos em Geração de Linguagem Natural (NLG).
LARSEN et al. (2026)	De maneira geral, adapta Factualidade e Fidelidade de Huang et al. (2025)	Adapta modelos anteriores para o atendimento ao cliente, incluindo "omissões" como um tipo de erro.

MADZHUNDER & BEGUNOVA (2025)	Cross-lingual (falha idiomática), Semântica (Associações, Raciocínio Ilógico, "Fantasmagorias") e Contextual.	Introduz o termo "Fantasmagorias" para respostas absurdas e analisa erros induzidos pela mudança de idioma da consulta.
MASSENON et al. (2025)	Incorreção Factual, Informação Fabricada, Saída Irrelevante, Inconsistência Lógica, Desvio de Persona e Fabricação Visual.	Taxonomia baseada na percepção real de usuários de apps móveis em lojas de aplicativos.
SOOD et al. (2025)	Contextual, Factual, Atribucional, Ambígua e Lógica	Focado em cibersegurança, onde alucinações podem gerar falsos positivos perigosos em sistemas de detecção de ameaças.
SUN et al. (2024)	8 Tipos: Overfitting, Erros de Lógica, de Raciocínio, Matemáticos, Fabricação Infundada, Factuais, de Saída de Texto e Outros.	Oferece a lista mais granular (dividida em 31 subtipos), incluindo erros de tradução e de código de programação.

Fonte: Citada no próprio quadro.

A partir do quadro resumo, é possível inferir que há duas principais correntes de classificação na literatura analisada: (i) Corrente associada à relação do conteúdo gerado com a fonte de dados fornecida, podendo ser categorizada como extrínseca e intrínseca e (ii) Corrente associada com a maneira como o conteúdo gerado reflete relações lógicas, contextuais e qualitativas, predominantemente classificada através de fidelidade e

factualidade. Dessa forma, é possível definir os 4 conceitos base de classificação, com base nos artigos analisados.

a. **Alucinação Intrínseca:** Ocorre quando o output dos modelos de inteligência artificial contradiz os fatos presentes no documento fonte (Anh-Hoang et al., 2025).

b. **Alucinação Extrínseca:** Caracterizada pela inclusão de informação no output que não está baseada em uma verdade fundamental. Ao contrário da intrínseca, introduz conteúdo adicional que não pode ser verificado em relação à fonte, no lugar de contradizê-la (Alansari; Luqman, 2026). Nesses casos, o texto parece plausível, entretanto carece de suporte explícito no input. Esse tipo de alucinação nem sempre é errôneo, pois pode ter sido gerado a partir de informações factualmente corretas externas (Ji et al., 2023).

c. **Alucinação de Factualidade:** Produz saídas que ou são divergentes com dados reais ou inverificável (Huang et al., 2025), ocorrendo quando o sistema gera uma resposta fabricada apresentada como fato (Sood et al., 2025). Pode consistir em alucinações intrínsecas ou extrínsecas e ocorre devido a característica dos modelos que priorizam coerência em detrimento de acerto factual. Pode ser subdividida em Contradições Factuais, onde apresenta situações em que a saída apresenta entidades errôneas ou relações falsas entre as entidades, e Fabricações Factuais, onde apresenta situações não verificáveis ou generalizações que carecem de validação universal (Huang et al., 2025).

d. **Alucinação de Fidelidade:** Ocorre quando o conteúdo gerado foge do input original ou contexto, violando as instruções do usuário ou a consistência lógica da resposta. Pode ser subdividida em inconsistências de instrução, quando a saída desvia da diretriz do usuário, inconsistências de contexto em que a saída é infiel à informação contextual fornecida, e ainda inconsistências lógicas, onde a saída apresenta contradições lógicas, usualmente em tarefas de raciocínio (Huang et al., 2025).

Nesse sentido, é importante destacar que essas classificações não são excludentes, mas complementares. Dessa forma, como intuito de constituir um referencial para análise das implicações organizacionais, foi construída uma matriz como mecanismo de classificação das alucinações da IA (Quadro 2).

Quadro 2. Matriz de Classificação das Alucinações

	Factualidade	**Fidelidade**
Intrínseca	**Intrínseca + Factualidade** - A resposta altera ou contradiz um fato presente na fonte.	**Intrínseca + Fidelidade** - A resposta contradiz instruções, premissas ou informações contextuais fornecidas.
Extrínseca	**Extrínseca + Factualidade** - A resposta acrescenta elementos não solicitados ou não sustentados pelo contexto.	**Extrínseca + Fidelidade** - A resposta fabrica um fato inexistente ou não verificável.

Fonte: Criado a partir do Quadro 1.

4.2. Impactos das Alucinações nas Organizações

A Partir da Classificação das Alucinações e do Papel Desempenhado Pela Inteligência Artificial nas Organizações, Foi Construído o Quadro a Seguir, com Possíveis Riscos e Desafios Trazidos Pelas Alucinações da IA.

Quadro 3. Exemplos de Falhas nos Níveis de Planejamento

	Desafios Estratégicos	**Desafios Táticos**	**Desafios Operacionais**
Intrínseco + Factualidade	Resumos gerenciais, auditorias ou pareceres jurídicos que distorcem o diagnóstico e/ou invertem a avaliação do risco real.	Distorção de tendências e indicadores. Síntese incorreta de dados financeiros ou de mercado em análises.	Interpretação incorreta do desempenho divergindo dos indicadores reais. Distorção das capacidades de recursos disponíveis. Erros de diagnóstico de processo. Invalidação direta da precisão de dados em fluxos operacionais. Dados de entrada, status e parâmetros opostos ao padrão de origem. Transferência de dados incorretos para processos subsequentes.
Intrínseco + Fidelidade	Diretrizes recomendadas que contrariam os limites éticos, orçamentários ou regulatórios estabelecidos pela governança.	Omissão de seções críticas obrigatórias. Lógica em que o resultado contradiz as condições iniciais. Recomendaçõe	Desconsideração das restrições metodológicas definidas. Descumprimento de estruturas e formatos definidos de resposta. Desrespeito a proibições

	Projeções com resultados incompatíveis com os dados de entrada.	s de melhoria incompatíveis com os limites e/ou metas acordados.	parametrizadas no comando. Resposta que desconsidera informações fornecidas em etapas anteriores da interação.
Extrínseco + Factualidade	Dados irreais e fabricados na formulação de diretrizes estratégicas. Teses estratégicas fundamentadas em informações plausíveis, porém falsas.	Utilização de parâmetros forjados pela IA. Recursos, metodologias e ferramentas não existentes apontadas como método viável. Premissas inexistentes como fundamentação.	Geração de dados inexistentes ou falsos em bases de dados. Etapas operacionais e regras de execução fundamentadas em respaldo técnico inexistente.
Extrínseco + Fidelidade	Propostas estratégicas que ultrapassam o escopo esperado. Introdução de riscos estruturais não considerados. Elaboração de relatórios executivos que assumem premissas que vão além das formuladas.	Inclusão de análises secundárias e variáveis não solicitadas que poluem o diagnóstico tornando as decisões mais complexas. Adição de pressupostos operacionais paralelos sem autorização.	Execução de ações adicionais ou adição de conteúdo não solicitados nas regras do fluxo padrão. Aplicação não solicitada de regras ou políticas externas ao procedimento padrão requerido.

Fonte: Análise criada a partir do Quadro 2.

Esse quadro tem caráter exemplificativo, não sendo exaustiva a lista de desafios. Entretanto, cumpre a função de elucidar como cada tipo de alucinação pode ser revelada em situações reais. A principal utilidade está em identificar os modos de falha da IA em cada função de planejamento de modo a explicitar os possíveis problemas encontrados com as alucinações.

4.3. Mitigação dos Riscos de Alucinações de IA usando Conhecimento Humano Tecnológico, Intuitivo, Genialidade Criativa e Capacidade de Modelagem Abstrata Antecedente

A integração da inteligência artificial (IA) nos fluxos de trabalho contemporâneos exige que as saídas geradas por esses sistemas não sejam automaticamente aceitas como conhecimento consolidado. Como discutido, o conhecimento autêntico é uma faculdade intrínseca ao ser humano e contextualizada, e por isso, as respostas da IA devem ser tratadas como insumos retrospectivos brutos que demandam confrontação e validação.

Para reduzir o risco de conteúdos incorretos serem incorporados às práticas das organizações, a governança anti-alucinação deve ser estruturada a partir do Quadro 2 (Matriz de Classificação), contrapondo os quadrantes de falhas algorítmicas à mobilização ativa das quatro dimensões fundamentais da cognição humana descritas neste estudo:

a. **O Conhecimento Humano Tecnológico e o quadrante Intrínseca + Fidelidade:** Esta primeira dimensão cognitiva humana atua diretamente na contenção de falhas onde a resposta da IA contradiz instruções, premissas ou parâmetros contextuais fornecidos no próprio comando (inconsistências

de instrução). O profissional dotado de conhecimento tecnológico compreende as limitações operacionais do modelo, calibrando a ferramenta e corrigindo desvios em tempo real quando a máquina descumpre formatos ou regras explícitas parametrizadas na interação.

b. **A Capacidade de Modelagem Abstrata Antecedente e o quadrante Extrínseca + Fidelidade:** Quando o sistema tende a fabricar um fato inexistente, não verificável ou a extrapolar o escopo solicitado (violando a consistência lógica profunda), a modelagem abstrata humana atua de forma preventiva. O indivíduo projeta estruturas lógicas, cenários conceituais hipotéticos e regras semânticas estritas através de uma engenharia de contexto avançada (prompting estratégico). Ao desenhar o modelo mental antes da execução do algoritmo, o ser humano delimita rigidamente o espaço amostral da máquina, bloqueando a dispersão e mitigando as alucinações de fidelidade na raiz do processo.

c. **O Conhecimento Intuitivo (Insightful) e o quadrante Intrínseca + Factualidade:** Ocorre quando o output altera ou contradiz frontalmente um dado ou fato presente no documento fonte (como inversões de risco em relatórios ou erros em dados financeiros). Frente a alucinações perigosamente fluentes e convincentes, a intuição do especialista, alicerçada na experiência prática acumulada (conhecimento tácito), funciona como um filtro instantâneo de plausibilidade. O salto cognitivo humano detecta sutilezas contraditórias que validadores estatísticos puramente automatizados deixam passar, assegurando a fidelidade factual do insumo.

d. **A Genialidade Criativa e o quadrante Extrínseca + Factualidade:** É o cenário onde a IA acrescenta elementos factuais falsos, não sustentados por nenhuma base ou verdade fundamental (como a indicação de metodologias e recursos inexistentes). Enquanto a máquina está presa a restrições probabilísticas retrospectivas de dados passados, a genialidade criativa humana é capaz de realizar conexões inéditas entre domínios distantes. Ela desmascara a alucinação factual ao aplicar teorização pura e julgamento crítico, descartando as falsas premissas geradas pelo algoritmo e transformando apenas os insumos legítimos em inovação real e conhecimento efetivamente validado.

Dessa forma, a mitigação dos riscos associados à IA deixa de ser vista como um filtro puramente técnico para se tornar uma simbiose sociotécnica estruturada. O cruzamento da cognição pluridimensional humana com a matriz de alucinações dita o nível de supervisão necessário: no plano operacional, valida-se a consistência de tarefas repetitivas; no tático, o impacto nos processos e decisões gerenciais; e no estratégico, exige-se o rigor máximo da modelagem e da intuição, impedindo que premissas forjadas ou distorcidas pela IA (independentemente do quadrante de origem) corrompam as diretrizes de longo prazo e a sustentabilidade da organização.

5. CONCLUSÃO

A crescente incorporação da inteligência artificial generativa às atividades organizacionais amplia a capacidade de processamento de volumes explícitos de informação, mas introduz sérios desafios à integridade e à confiabilidade dos ativos de conhecimento das

empresas. Diante desse cenário, esse trabalho permitiu responder qual é o papel do conhecimento humano e como suas múltiplas dimensões podem contribuir para a confiabilidade e a eficiência da colaboração humano IA nas organizações frente ao fenômeno das alucinações.

Foi possível construir e sistematizar as principais manifestações de erro algorítmico da literatura por meio da construção de um referencial tipológico integrado e a contribuição alude-se na compreensão de quatro categorias fundamentais de alucinação mapeadas no Quadro 2: (i) Alucinação Intrínseca, (ii) Alucinação Extrínseca, (iii) Alucinação de Factualidade e, (iv) Alucinação de Fidelidade.

Como demonstrado ao longo do texto e sintetizado na matriz de impactos organizacionais (Quadro 3), tais falhas afetam de maneiras distintas, porém igualmente severas, os níveis estratégico, tático e operacional do planejamento corporativo.

Assim, a grande contribuição teórica deste estudo reside em demonstrar que a superação desse gargalo tecnológico não emergirá de soluções puramente algorítmicas ou automações isoladas. A transformação de ilusões estatísticas em insights confiáveis depende, compulsoriamente, do gerenciamento estratégico da cognição humana na colaboração com a tecnologia.

As saídas produzidas pela IA generativa não constituem conhecimento pronto; são insumos retrospectivos que exigem a curadoria reflexiva, o discernimento crítico, a responsabilidade ética e a validação epistêmica que apenas o conhecimento tecnológico, o

conhecimento intuitivo, a genialidade criativa e a capacidade de modelagem abstrata antecedente do ser humano possuem.

Como limitação, este estudo possui natureza conceitual e baseia-se em uma revisão integrativa com uma amostra concentrada de estudos dedicados estritamente à taxonomia das alucinações, reflexo da recente consolidação desse campo. Pesquisas futuras podem aplicar este modelo conceitual empiricamente em estudos de caso organizacionais, investigando como diferentes estruturas de governança baseadas nas dimensões da cognição humana impactam a taxa de mitigação de erros e impulsionam a inovação segura nos ecossistemas de trabalho nas organizações.

REFERÊNCIAS BIBLIOGRÁFICAS

ALANSARI, Aisha; LUQMAN, Hamzah. Large language models hallucination: A comprehensive survey. **Computer Science Review,** v. 61, 2026. DOI: 10.1016/j.cosrev.2026.100970.

ANH-HOANG, Dang; TRAN, Vu; NGUYEN, Le Minh. Survey and analysis of hallucinations in large language models: attribution to prompting strategies or model behavior. **Frontiers in Artificial Intelligence,** Frontiers Media SA, 2025. DOI: 10.3389/frai.2025.1622292.

ARISO, José María; BANNISTER, Peter. 'AI lost the prompt!' Replacing 'AI hallucination' to distinguish between mere errors and irregularities. **AI and Society,** v. 41, n. 4, p. 3475–3487, 2026. DOI: 10.1007/s00146-025-02757-1.

BÖHM, Karsten; DURST, Susanne. Knowledge management in the age of generative artificial intelligence – from SECI to GRAI. **VINE**

Journal of Information and Knowledge Management Systems, v. 56, n. 1, p. 106–121, 2026. DOI: 10.1108/VJIKMS-10-2024-0357.

BOLIS, Ivan; MARQUES, João Nobrega; CAGNO, Enrico; MORIOKA, Sandra Naomi. Digital technologies, sustainability, and work: How can these themes be brought together to promote a human-centered future in industry 5.0 implementation? **Applied Ergonomics,** v. 125, p. 104475, 2025. DOI: 10.1016/j.apergo.2025.104475. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S0003687025000110. Acesso em: 10 ago. 2026.

BORGES, Aline F. S. et al. The strategic use of artificial intelligence in the digital era: Systematic literature review and future research directions. **International Journal of Information Management,** Elsevier Ltd, 2021. DOI: 10.1016/j.ijinfomgt.2020.102225.

BOWDEN, Edward M. et al. New approaches to demystifying insight. **Trends in Cognitive Sciences,** v. 9, n. 7, p. 322–328, 2005.

CHEN, Yufen; PAN, Xiaoyi; LIU, Pian; VANHAVERBEKE, Wim. How does digital transformation empower knowledge creation? Evidence from Chinese manufacturing enterprises. **Journal of Innovation and Knowledge,** v. 9, n. 2, 2024. DOI: 10.1016/j.jik.2024.100481.

CLARK, Andy; CHALMERS, David. The extended mind. **Analysis,** v. 58, n. 1, p. 7–19, 1998.

DAI, Shuanping et al. Responsible AI in knowledge creation: An exploration of generative AI's opportunities and risks. **Technological Forecasting and Social Change,** v. 226, 2026. DOI: 10.1016/j.techfore.2026.124570.

DIETRICH, Arne. The cognitive neuroscience of creativity. **Psychonomic Bulletin & Review,** v. 11, n. 6, p. 1011–1026, 2004.

DWIVEDI, Yogesh K. et al. Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. **International Journal of Information Management,** v. 57, 2021. DOI: 10.1016/j.ijinfomgt.2019.08.002.

GEELS, Frank W. From sectoral systems of innovation to socio-technical systems: Insights about dynamics and change from a sociology of technology perspective. **Research Policy,** v. 33, n. 6-7, p. 897–920, 2004.

GIGERENZER, Gerd; GAISSMAIER, Wolfgang. Heuristic decision making. **Annual Review of Psychology,** v. 62, p. 451–482, 2011.

GUJJAR, Praveen; KUMAR, Prasanna. Exploring the boundaries of artificial intelligence hallucination. **International Journal of Applied Mathematics,** v. 38, n. 2s, 2025.

HUANG, Lei et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. **ACM Transactions on Information Systems,** v. 43, n. 2, 2025. DOI: 10.1145/3703155.

ILO. International Labor Organization. Artificial intelligence. Geneva: International Labour Organization, 2026. Disponível em: https://www.ilo.org/topics-and-sectors/artificial-intelligence. Acesso em: 10 ago. 2026.

JARRAHI, Mohammad Hossein et al. Artificial intelligence and knowledge management: A partnership between human and AI.

Business Horizons, v. 66, n. 1, p. 87–99, 2023. DOI: 10.1016/j.bushor.2022.03.002.

JARRAHI, Mohammad Hossein. Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. **Business Horizons,** v. 61, n. 4, p. 577–586, 2018. DOI: 10.1016/j.bushor.2018.03.007.

JI, Ziwei et al. Survey of Hallucination in Natural Language Generation. **ACM Computing Surveys,** Association for Computing Machinery, 2023. DOI: 10.1145/3571730.

JOHNSON-LAIRD, Philip N. Mental models and human reasoning. **Proceedings of the National Academy of Sciences,** v. 107, n. 43, p. 18243–18250, 2010.

KAZLARIS, Ioannis et al. From Illusion to Insight: A Taxonomic Survey of Hallucination Mitigation Techniques in LLMs. **AI,** Multidisciplinary Digital Publishing Institute (MDPI), 2025. DOI: 10.3390/ai6100260.

KORZYNSKI, Pawel et al. Generative artificial intelligence as a new context for management theories: analysis of ChatGPT. **Central European Management Journal,** v. 31, n. 1, p. 3–13, 2023. DOI: 10.1108/CEMJ-02-2023-0091.

LARSEN, Anna Grøndahl et al. LLM Hallucinations in Conversational AI for Customer Service: Framework and End-User Perceptions. **International Journal of Human-Computer Interaction,** v. 42, n. 13, p. 9807–9828, 2026. DOI: 10.1080/10447318.2025.2580540.

MADZHUMDER, M. Sh; BEGUNOVA, D. D. Causes of Content Distortion: Analysis and Classification of Hallucinations in GPT Large

Language Models. **Scientific and Technical Information Processing,** v. 52, n. 6, p. 598–604, 2025. DOI: 10.3103/S0147688225700479.

MASSENON, Rhodes et al. ”My AI is Lying to Me”: User-reported LLM hallucinations in AI mobile apps reviews. **Scientific Reports,** v.15, n.1, 2025. DOI: 10.1038/s41598-025-15416-8.

NONAKA, Ikujiro; TAKEUCHI, Hirotaka. **Gestão do conhecimento.** Porto Alegre: Bookman, 2008. E-book. Disponível em: https://integrada.minhabiblioteca.com.br/reader/books/9788577802296/. Acesso em: 28 jul. 2026.

PANDA, Mahanish; PURI, Roma. Understanding the role of artificial intelligence in knowledge management systems: a systematic review using the antecedent, consequences and intervention framework. **Journal of Work-Applied Management,** p. 1–14, 2025. DOI: 10.1108/JWAM-05-2025-0082.

SARKAR, Bishal Dey et al. Digital transition from industry 4.0 to industry 5.0 in smart manufacturing: A framework for sustainable future. **Technology in Society,** v. 78, 2024. DOI: 10.1016/j.techsoc.2024.102649.

SEARLE, John R. Minds, brains, and programs. **Behavioral and Brain Sciences,** v. 3, n. 3, p. 417–424, 1980.

SIMONTON, Dean Keith. Creativity as blind variation and selective retention: Is the creative process Darwinian? **Psychological Inquiry,** v. 10, n. 4, p. 309–328, 1999.

SOOD, Aditya K.; ZEADALLY, Sherali; HONG, Een Kee. The Paradigm of Hallucinations in AI-driven cybersecurity systems: Understanding taxonomy, classification outcomes, and mitigations. **Computers and Electrical Engineering,** v. 124, 2025. DOI: 10.1016/j.compeleceng.2025.110307.

SUN, Yujie et al. AI hallucination: towards a comprehensive classification of distorted information in artificial intelligence-generated content. **Humanities and Social Sciences Communications,** v. 11, n. 1, 2024. DOI: 10.1057/s41599-024-03811-x.

XU, Li Da; XU, Eric L.; LI, Ling. Industry 4.0: State of the art and future trends. **International Journal of Production Research,** v. 56, n. 8, p. 2941–2962, 2018. DOI: 10.1080/00207543.2018.1444806.

YILDIZ, Tayfun et al. Impact of knowledge-sharing culture on organizational creativity: integrating explicit and tacit knowledge sharing as mediators. **Journal of Knowledge Management,** v. 29, n. 4, p. 1248–1277, 2025. DOI: 10.1108/JKM-05-2024-0633.

---

[1] Graduando em Engenharia de Produção

[2] Graduando em Engenharia de Produção

[3] Pós doutor em Energia e Clima (KTH Suécia); Pós doutor em Engenharia Industrial (UBI Portugal); Doutor em Administração (UFPE) e Doutor em Engenharia de Produção (UFPB).

[4] Pós doutora em Sociologia (UBI Portugal); Doutora em Computação (UFPE).