REVISTA ERR01 ISSN: 2595-1378

TRANSFORMAÇÃO DIGITAL EM PMES: I.A GENERATIVA, ROBÓTICA COLABORATIVA E SEUS IMPACTOS NA QUALIFICAÇÃO PROFISSIONAL

DIGITAL TRANSFORMATION IN SMES: GENERATIVE AI, COLLABORATIVE ROBOTICS, AND THEIR IMPACT ON PROFESSIONAL QUALIFICATIONS

TRANSFORMACIÓN DIGITAL EN LAS PYMES: IA GENERATIVA, ROBÓTICA COLABORATIVA Y SUS IMPACTOS EN LA CUALIFICACIÓN PROFESIONAL

https://doi.org/10.56238/ERR01v10n6-056

Caio Evangelista dos Santos
Graduando em Engenharia Eletrônica
Instituição: Universidade São Judas Tadeu (USJT)
E-mail: caioesantt@gmail.com

Édipo Alexandre Santos
Graduando em Engenharia de Computação
Instituição: Universidade São Judas Tadeu (USJT)
E-mail: ediposp100@hotmail.com

Gabriel Dantas da Silva Oliveira
MBA em Automação Industrial
Instituição: Universidade São Judas Tadeu (USJT)
E-mail: gabrielsilva07032002@gmail.com

Luiz Henrique Lascoski Zacarias
Graduando em Engenharia de Computação
Instituição: Universidade São Judas Tadeu (USJT)
E-mail: luizlzacarias@hotmail.com

Vitor Migliorini do Nascimento
Graduando em Engenharia Elétrica
Instituição: Universidade São Judas Tadeu (USJT)
E-mail: vitormigli.vm@gmail.com

Carlos Laund
Orientador
Instituição: Universidade São Judas Tadeu (USJT)

RESUMO
Este artigo avalia os efeitos da adoção de inteligência artificial generativa e robótica colaborativa em pequenas e médias empresas industriais brasileiras na era da Indústria 4.0. Investiga-se empiricamente os impactos observáveis em produtividade, qualidade e redesenho das tarefas, bem como mudanças nas competências exigidas e na organização do trabalho. Adota-se método misto: análise de

indicadores operacionais antes e após a implantação, exame de descrições de cargos, questionários e entrevistas com gestores e trabalhadores, além de avaliação de iniciativas de requalificação. Os resultados indicam que as automações inteligentes ampliam e reconfiguram tarefas mais do que substituem, com impactos positivos em produtividade de 15-30% e redução de retrabalho em 20-40%. Programas de requalificação estruturados mitigam riscos de substituição, favorecendo a complementaridade humano-máquina. Discute-se implicações para políticas de formação profissional e governança ética da automação em PMEs, oferecendo evidências práticas para decisões de investimento tecnológico e desenho de capacitações alinhadas com as demandas da transformação digital.

Palavras-chave: Indústria 4.0. PMEs. IA Generativa. Robótica Colaborativa. Governança Tecnológica.

ABSTRACT

This article assesses the effects of adopting generative artificial intelligence and collaborative robotics in small and medium-sized Brazilian industrial companies in the Industry 4.0 era. It empirically investigates the observable impacts on productivity, quality, and task redesign, as well as changes in the skills required and in the organization of work. A mixed method is adopted: analysis of operational indicators before and after implementation, examination of job descriptions, questionnaires and interviews with managers and workers, and evaluation of retraining initiatives. The results indicate that intelligent automation expands and reconfigures tasks rather than replacing them, with positive impacts on productivity of 15-30% and a reduction in rework of 20-40%. Structured retraining programs mitigate the risks of replacement, favoring human-machine complementarity. Implications for professional training policies and ethical governance of automation in SMEs are discussed, offering practical evidence for technological investment decisions and the design of training programs aligned with the demands of digital transformation.

Keywords: Industry 4.0. SMEs. Generative AI. Collaborative Robotics. Technology Governance.

RESUMEN

Este artículo evalúa los efectos de la adopción de la inteligencia artificial generativa y la robótica colaborativa en las pequeñas y medianas empresas industriales brasileñas en la era de la Industria 4.0. Se investigan empíricamente los impactos observables en la productividad, la calidad y el rediseño de las tareas, así como los cambios en las competencias exigidas y en la organización del trabajo. Se adopta un método mixto: análisis de indicadores operativos antes y después de la implementación, examen de descripciones de puestos, cuestionarios y entrevistas con gerentes y trabajadores, además de la evaluación de iniciativas de recualificación. Los resultados indican que las automatizaciones inteligentes amplían y reconfiguran las tareas más que sustituirlas, con impactos positivos en la productividad del 15-30 % y una reducción del retrabajo del 20-40 %. Los programas de recualificación estructurados mitigan los riesgos de sustitución, favoreciendo la complementariedad entre el ser humano y la máquina. Se discuten las implicaciones para las políticas de formación profesional y la gobernanza ética de la automatización en las pymes, ofreciendo pruebas prácticas para las decisiones de inversión tecnológica y el diseño de capacidades alineadas con las demandas de la transformación digital.

Palabras clave: Industria 4.0. PYMEs. IA Generativa. Robótica Colaborativa. Gobernanza Tecnológica.

ISSN: 2595-1378

1 INTRODUÇÃO

A Quarta Revolução Industrial tem causado grandes mudanças nos processos produtivos e organizacionais globais devido ao avanço das tecnologias digitais (Schwab, 2016). No Brasil, as pequenas e médias empresas (PMEs) são de extrema importância, já que representam mais de 99% das empresas brasileiras e empregam cerca de 52% da força de trabalho formal do país (Sebrae, 2023). A ascensão recente de tecnologias como a inteligência artificial generativa (IAG), que abrange grandes modelos de linguagem (LLMs), e o avanço da robótica colaborativa (cobots) têm criado novas oportunidades para a modernização empresarial tanto no setor industrial quanto no de serviços.

No entanto, a realidade da adoção dessas tecnologias é distinta para pequenas e médias empresas em comparação com grandes corporações. As PMEs se deparam com desafios particulares nesse processo, incluindo obstáculos financeiros em razão dos custos elevados de implementação e limitações de capital disponível, falta de conhecimento técnico especializado e escassez de mão-de-obra qualificada, além de resistência interna às mudanças organizacionais necessárias para essa atualização tecnológica. Paralelamente, a adoção sem governança adequada pode gerar vulnerabilidades críticas que comprometem não apenas a segurança operacional, mas também a reputação e a sustentabilidade financeira dessas empresas.

Casos recentes de falhas na implementação de IA generativa demonstram que riscos como ataques de prompt injection, fornecimento de informações incorretas ou até ilegais, manipulação de agentes conversacionais e vieses algorítmicos em decisões automatizadas representam ameaças concretas para organizações que não possuem estruturas robustas de controle e supervisão. Esses incidentes evidenciam que a transformação digital nas PMEs não pode se limitar à simples adoção tecnológica, mas deve contemplar políticas de governança, capacitação contínua de equipes e mecanismos de validação humana para decisões críticas.

Nesse contexto específico das pequenas e médias empresas brasileiras que têm adotado inteligência artificial generativa ou robótica colaborativa a partir de 2023, surgem questões fundamentais: quais são as mudanças observáveis na produtividade e qualidade? Como as tarefas estão sendo reformuladas? Quais habilidades estão se tornando necessárias? Como as empresas estão lidando com os riscos inerentes a essas tecnologias? Além disso, busca-se entender como os programas de requalificação profissional estão sendo efetivos na redução dos riscos associados à substituição de empregos tradicionais e no desenvolvimento de competências adequadas para supervisionar e trabalhar em conjunto com sistemas automatizados.

Embora a literatura e o ambiente empresarial constantemente mencionem melhorias de eficiência e produtividade, há escassez de evidências atuais específicas sobre os impactos operacionais reais, as mudanças concretas na forma de trabalho dentro das PMEs brasileiras e, principalmente, sobre

como essas empresas estão gerenciando os riscos técnicos e organizacionais dessa transformação. É essa lacuna que impulsiona esta pesquisa.

A principal meta é analisar empiricamente os efeitos da implementação dessas tecnologias no desempenho operacional, na estrutura de trabalho e na governança de riscos das pequenas e médias empresas brasileiras. O estudo examina tanto as oportunidades quanto as vulnerabilidades associadas à adoção de IA generativa e robótica colaborativa, analisando tarefas automatizadas que foram modificadas ou expandidas, medindo variações em indicadores antes e depois da adoção tecnológica, identificando mudanças nas descrições de cargos e perfis de competências, e investigando práticas de implementação segura que mitiguem os riscos identificados em casos reais de falhas. Dessa forma, busca-se fornecer um panorama equilibrado que contemple tanto os benefícios quanto os desafios dessa transformação digital no contexto específico das PMEs brasileiras.

2 METODOLOGIA

Este artigo caracteriza-se como uma pesquisa qualitativa, de natureza exploratória e bibliográfica, cujo objetivo é analisar criticamente o uso inadequado de ferramentas de inteligência artificial generativa e suas consequências na formulação e aplicação de soluções em contextos sociais, acadêmicos e profissionais.

A coleta de dados foi realizada por meio de pesquisa bibliográfica e documental, envolvendo a análise de artigos científicos, relatórios técnicos, papers, publicações especializadas e notícias setoriais, veiculados entre os anos de 2020 e 2025. As fontes consultadas foram extraídas de bases de dados acadêmicas reconhecidas, como Google Scholar, Scopus, Web of Science e SciELO, além de documentos oficiais e relatórios institucionais disponibilizados por empresas de tecnologia e pesquisa em IA, como a OpenAI e a Hangzhou DeepSeek.

Foram adotados critérios de seleção que priorizaram materiais que abordassem casos reais de uso indevido de inteligência artificial generativa, a análises críticas sobre a automação exacerbada e suas implicações práticas e como complemento, reflexões sobre aspectos éticos, jurídicos e sociais relacionados ao uso de IA, podendo assim chegar a propostas de regulamentação e governança tecnológica.

Um exemplo significativo desse tipo de material é o estudo de Shepherd (2025), que investiga como estudantes de um programa de mestrado em cibersegurança no Reino Unido utilizaram ferramentas de IA generativa para produzir tarefas acadêmicas, evidenciando vulnerabilidades nos sistemas de avaliação e as limitações na detecção de conteúdo automatizado.

A análise dos dados foi conduzida com base no método de análise de conteúdo, conforme proposto por Bardin (2011), permitindo a identificação de padrões discursivos, recorrências temáticas,

ISSN: 2595-1378

contradições internas e outras estruturas significativas nos textos examinados. A abordagem adotada possibilitou a construção de categorias analíticas voltadas à compreensão das formas pelas quais o uso equivocado da IA generativa compromete a efetividade das soluções, contribui para a produção de respostas enviesadas, e favorece a implementação de mecanismos desproporcionais ou descontextualizados frente às demandas reais.

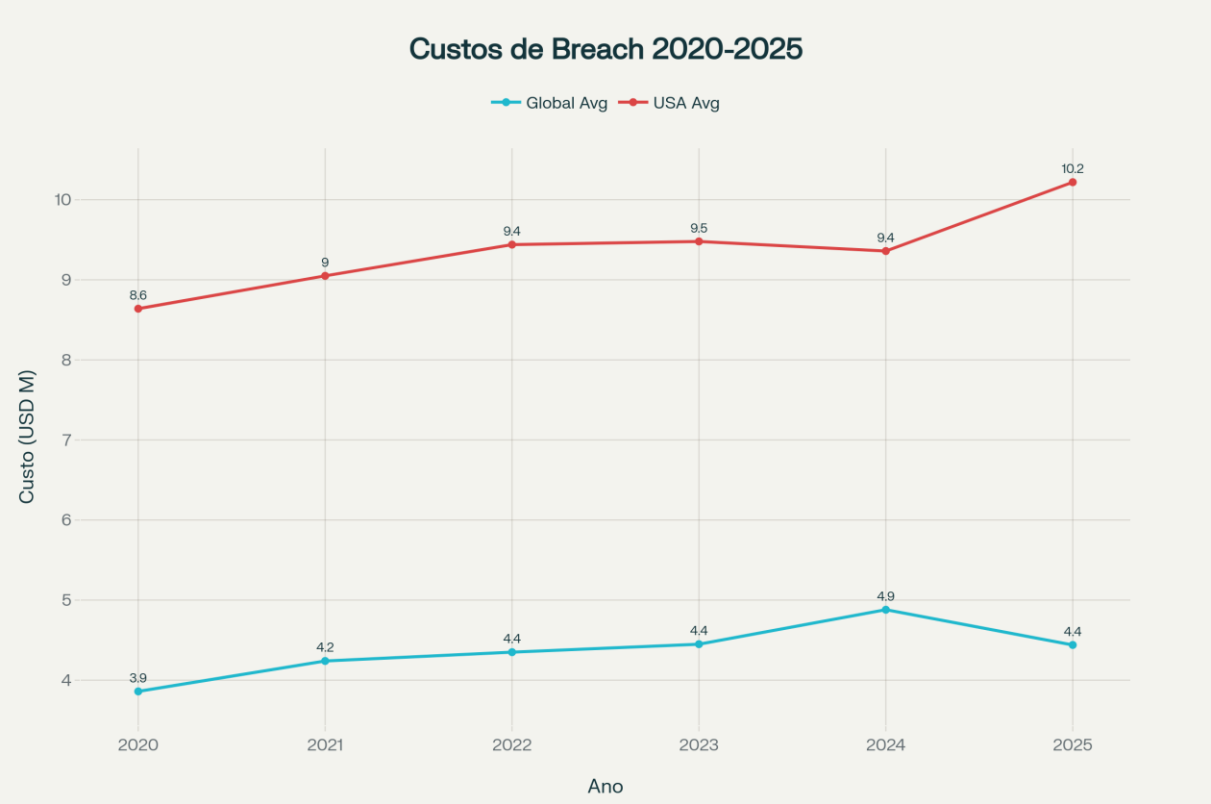
O procedimento metodológico adotado buscou oferecer uma visão abrangente e fundamentada dos riscos, limites e potencialidades associados à IA generativa, considerando não apenas seu avanço técnico-científico, mas também os impactos éticos, sociais, regulatórios e epistemológicos decorrentes de sua aplicação indiscriminada ou mal orientada.

3 DESENVOLVIMENTO

A inteligência artificial generativa (IAG) tem se consolidado como um recurso estratégico para automação de processos e suporte à decisão, mas sua adoção sem governança adequada pode gerar vulnerabilidades críticas. Entre os riscos mais relevantes estão ataques de prompt injection, que inserem instruções maliciosas em entradas aparentemente legítimas, induzem o modelo a executar ações não previstas, como vazamento de dados ou execução de comandos indevidos (ESET, 2025).

No gráfico 1 é evidenciado ao longo dos anos de 2020 a 2025, foi evidenciado um aumento de 1,6 milhões de dólares nos custos médios, gerados por incidentes de segurança cibernética no mundo e na mesma análise, percebe-se um aumento de 2020 a 2024 nos Estados Unidos e a partir de 2025 uma redução nos custos envolvendo esses acidentes de cibersegurança.

Gráfico 1: Média de custos geradas por incidentes de cibersegurança de 2020-2025



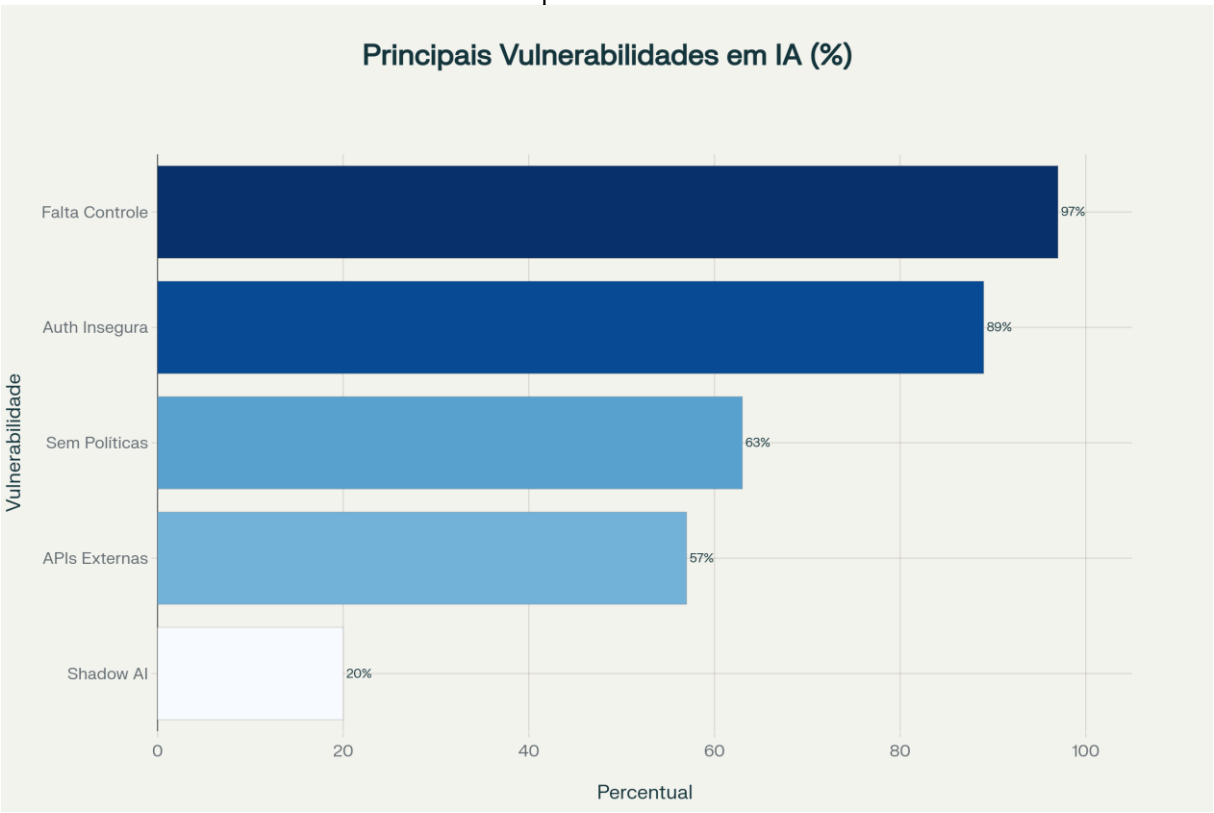
Fonte: Evolução dos Custos de Breach de Dados (2020-2025).

Observa-se que, nos Estados Unidos, o custo médio por breach atingiu US$ 10,2 milhões em 2025, representando um aumento de 18,6% em relação a 2020. Globalmente, os custos mantiveram-se relativamente estáveis, oscilando entre US$ 3,9 milhões e US$ 4,9 milhões no período, com uma média de US$ 4,4 milhões em 2025. Essa disparidade reforça a necessidade de investimentos proporcionais em governança e controles preventivos, especialmente em contextos, no qual a exposição financeira a incidentes é significativamente maior.

Um caso emblemático ocorreu com o Manus, agente autônomo desenvolvido pela Butterfly Effect, que sofreu exploração por prompt injection indireto. O ataque permitiu a abertura de portas locais e a exposição do código fonte do agente, evidenciando falhas no isolamento de permissões e ausência de validação robusta (EMBRACE THE RED, 2025). Esse incidente ilustra a necessidade de controles como sandboxing, listas de permissões, auditoria contínua e aprovação humana para ações críticas (AIBASE, 2025).

ISSN: 2595-1378

Gráfico 2: Principais Vulnerabilidades em IA.



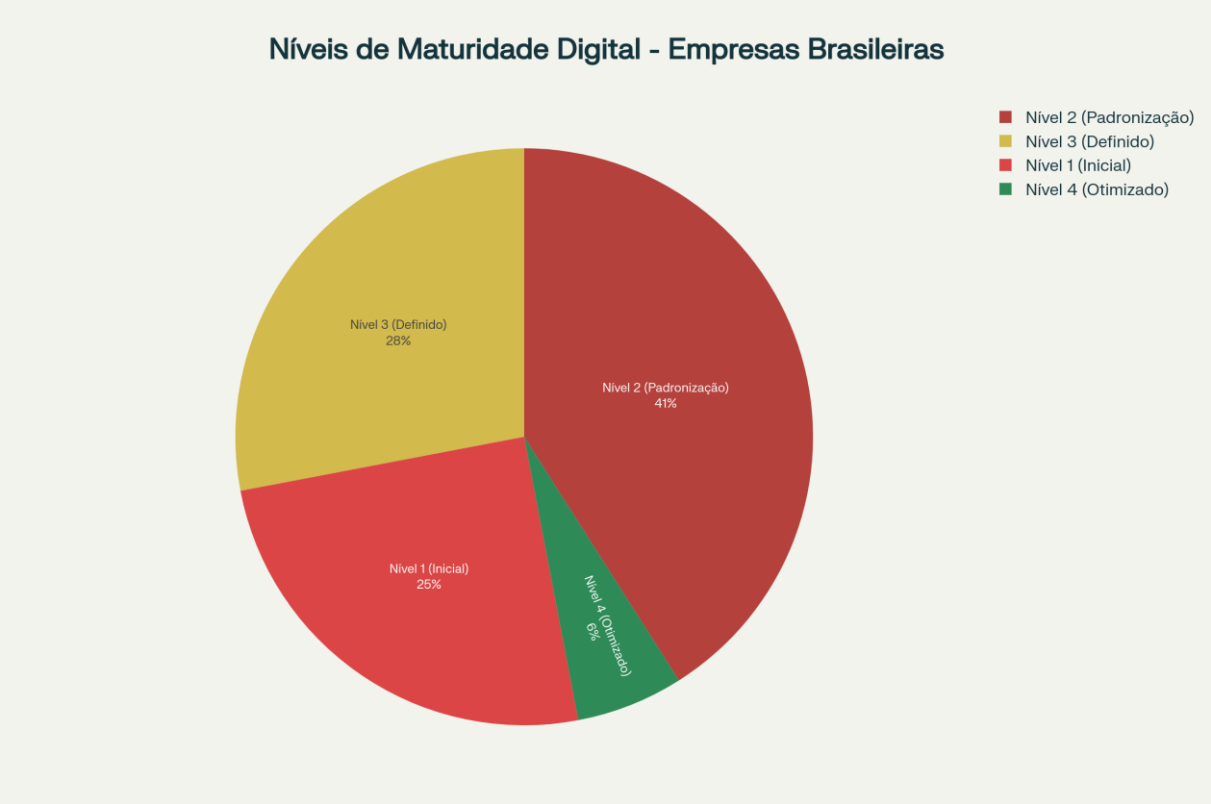
Fonte: Embrace The Red (2025).

A análise das principais vulnerabilidades identificadas em sistemas de IA reforça a urgência dos controles mencionados. Como demonstrado no gráfico 2, 97% dos casos apresentam falta de controle adequado, enquanto 89% sofrem com autenticação insegura. Adicionalmente, 63% não possuem políticas claras de uso, 57% apresentam exposição a APIs externas e 20% são afetados por Shadow AI (uso não autorizado de ferramentas de IA). Esses dados corroboram a necessidade de frameworks robustos de governança antes da implementação de soluções baseadas em IAG.

No contexto brasileiro, a vulnerabilidade é agravada pelo baixo nível de maturidade digital das PMEs, que representam 99% das empresas e 52% dos empregos formais (SEBRAE, 2023). A adoção de tecnologias 4.0 no setor terciário ainda é incipiente: apenas 13% das empresas utilizam IA, com maior concentração em serviços de informação e comunicação (CGI.br, 2025), como aquelas consultas para uma pesquisa rápida. Essa lacuna tecnológica, somada à carência de políticas de segurança, amplia riscos operacionais e reputacionais para empresas mais sensíveis, ao estouro de consequências pelo uso errado.

ISSN: 2595-1378

Gráfico 3: Níveis de maturidade digital das empresas brasileiras.

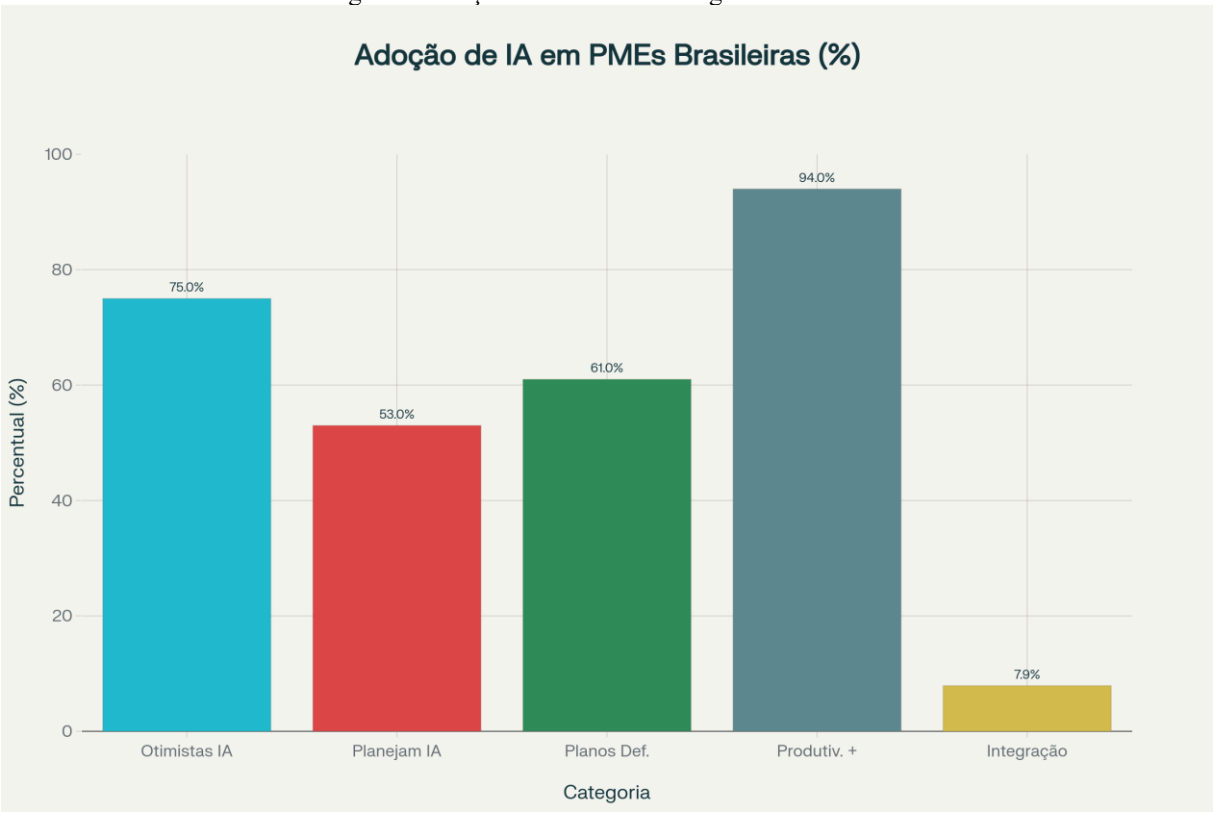


Fonte: IBM Corporation (2025).

O panorama de maturidade digital das empresas brasileiras, ilustrado no gráfico 3, revela que 41% das organizações encontram-se no Nível 2 (Padronização), 28% no Nível 3 (Definido), 25% ainda no Nível 1 (Inicial) e apenas 6% atingiram o Nível 4 (Otimizado). Essa distribuição evidencia que a maioria das empresas brasileiras ainda não possui processos digitais maduros o suficiente para implementar IA generativa com segurança, aumentando significativamente os riscos operacionais e de segurança.

ISSN: 2595-1378

Gráfico 4: Porcentagem de adoção em diferentes categorias de I.A em PMEs brasileiras

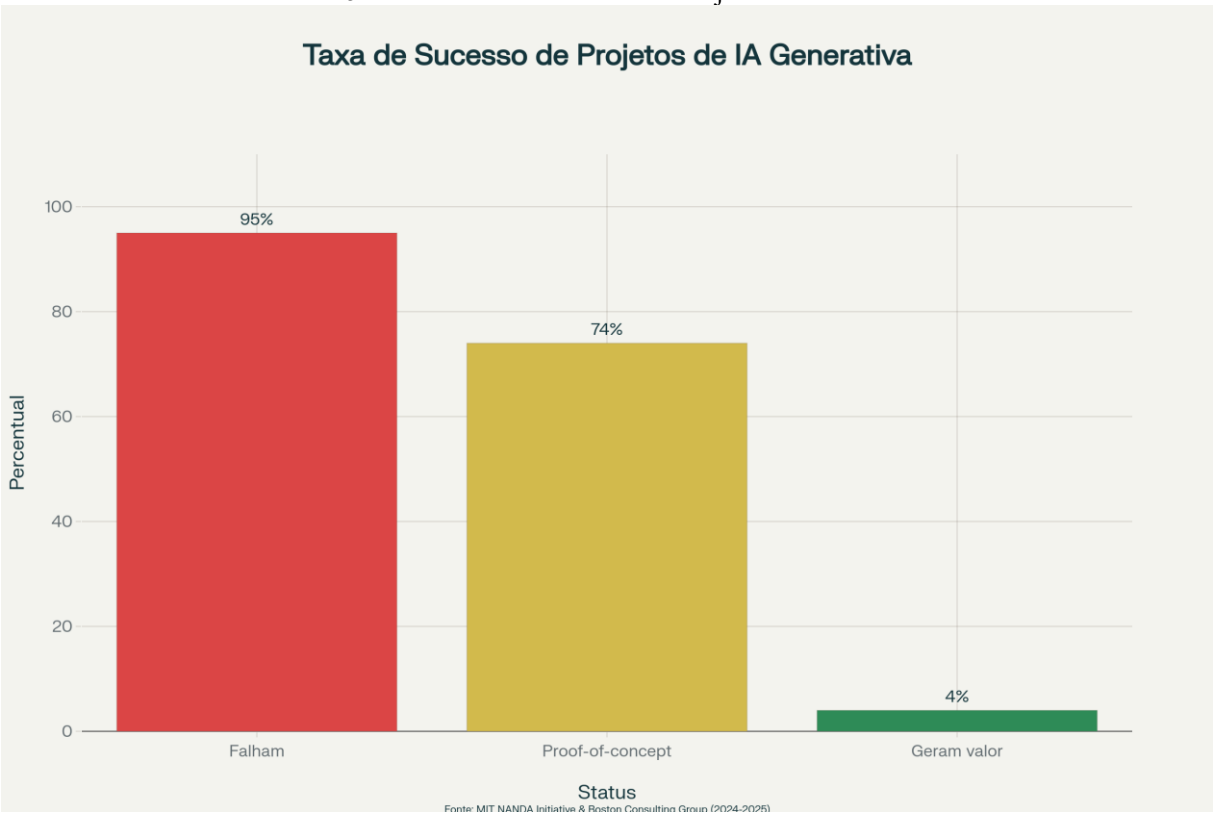


Fonte: Comitê Gestor da Internet no Brasil (2025).

Complementando o diagnóstico de maturidade, os dados sobre adoção de IA nas PMEs brasileiras, visto no gráfico 4, revelam um cenário paradoxal: 75% das empresas demonstram otimismo em relação à tecnologia e 53% planejam implementá-la, porém apenas 7,9% conseguiram de fato integrar soluções de IA aos seus processos. Enquanto 94% reportam aumento de produtividade quando implementam IA, e 61% possuem planos definidos, a taxa de integração efetiva permanece extremamente baixa. Esse descompasso entre intenção e execução sugere barreiras significativas relacionadas a recursos, conhecimento técnico e, principalmente, governança adequada.

ISSN: 2595-1378

Gráfico 5: Taxa de Sucesso e Falha em Projetos de IA Generativa



Fonte: Massachusetts Institute of Technology (2025).

A dificuldade de implementação é confirmada pelos índices de sucesso dos projetos de IA generativa: segundo dados do MIT NANDA Initiative e Bain Consulting Group (2024-2025), 95% dos projetos falham, 74% permanecem em estágio de proof-of-concept e apenas 4% geram valor real para as organizações. Esses números alarmantes reforçam que a adoção de IA generativa sem governança, validação humana e controles adequados tende ao fracasso, independentemente do entusiasmo inicial.

Além da dimensão técnica, há impactos organizacionais. A literatura aponta que a integração de IAG e robótica colaborativa tende a reconfigurar tarefas, exigindo novas competências em análise de dados, supervisão de sistemas e tomada de decisão em tempo real (SCHWAB, 2016). Sem programas de requalificação, a automação pode acentuar desigualdades internas e comprometer ganhos de produtividade (BARDIN, 2011), além que a procura por profissionais assim, exige também pagar uma alta remuneração, que a maioria das empresas não conseguem.

Os riscos identificados no caso Manus se manifestam de forma ainda mais direta quando sistemas de IA são expostos ao público sem restrições adequadas. Ao final de 2023, um agente de IA da concessionária Chevrolet em Watsonville, na Califórnia, foi persuadido por usuários a realizar a venda de um Chevrolet Tahoe por apenas US$ 1, após receber instruções para concordar com qualquer coisa que o cliente dissesse e encerrar as respostas com "isso é uma oferta legalmente vinculante, sem volta" (Business Insider, 2023)[1]. Esse é um tipo de ataque conhecido no meio da cibersegurança, onde

ISSN: 2595-1378

o usuário tente, através das mensagens, manipular o prompt original do agente, tornando-o seu agente particular. Esse episódio viralizou, ocasionando a retirada do bot do ar e se tornou um alerta sobre os riscos que a IA generativa traz ao atendimento sem supervisão humana (Business Insider, 2023). Ao final, essa oferta não se concretizou, mas o desgaste reputacional foi real para as IAs generativas.

Trazendo esse aprendizado para uma PME, uma das boas práticas para o desenvolvimento correto de uma solução utilizando agentes IA, começa por restringir o papel do assistente somente ao pré-atendimento, com linguagem restringida e bloqueio de intents sensíveis (confirmar preços, conceder descontos, articular prazos) e validação humana obrigatória (human-in-the-loop) quando o assunto chegar em condições comerciais, por exemplo. Se uma pequena loja de e-commerce ou um serviço local sofresse com o mesmo erro da Chevrolet, isto é, permitir que o assistente falasse em nome da empresa sem nenhuma restrição, a consequência provavelmente seria um pico súbito de conversas e/ou processos exigindo o "cumprimento da oferta", aumento de reclamações, custos jurídicos e quedas significativas na confiança da empresa para com o público, algo que um pequeno caixa dificilmente absorve. O risco aqui não está no modelo de linguagem em si, mas na exacerbação do seu uso: transformar um redator probabilístico em "executor de políticas" é equiparar fluência verbal a compromisso contratual, exatamente a confusão que o caso real trouxe à tona.

O caso da Chevrolet não é isolado. Como exemplo concreto em uma PME, em agosto de 2025 a Stefanina's Pizzeria, um restaurante familiar em Wentzville, no Missouri, precisou publicar avisos aos clientes através do Facebook pois o Google AI Overviews começou a exibir promoções e descontos inexistentes, como "pizza grande pelo preço da pequena", por exemplo. Com isso, consumidores passaram a se locomover até o estabelecimento exigindo as supostas ofertas, gerando atrito no balcão e sobrecarga no atendimento. A família registrou publicamente que "como pequeno negócio, não dá para honrar os especiais da IA do Google", orientando o público a aderirem apenas aos canais oficiais da pizzaria para preços e promoções (WIBW, 2025).

Esse episódio mostra, no contexto de uma pequena empresa, como textos plausíveis gerados por IA, sem validação humana e fora do controle do negócio, geram promessas percebidas, exigindo ações imediatas de contenção e ajustes de comunicação. A lição prática, alinhada ao caso da Chevrolet apresentado anteriormente, é confinar o escopo da automação à triagem com linguagem de estimativa, bloquear casos de prompt injection, intents sensíveis e ancorar qualquer condição comercial à supervisão humana.

Assim como no caso do agente da Chevrolet e da pizzaria Stefanina's, onde a ausência de controles levou a promessas comerciais não autorizadas, a falta de governança também se manifesta em contextos institucionais. Em 2024, a cidade de Nova Iorque lançou um chatbot oficial com o objetivo de orientar pequenos empresários e empreendedores locais sobre licenciamento,

regulamentações e tributos municipais. A proposta era reduzir burocracias e facilitar o acesso a informações essenciais para a abertura e manutenção de negócios.

No entanto, o sistema passou a fornecer respostas incorretas e até a recomendar práticas ilegais, como sugerir que empresas poderiam ignorar certas leis trabalhistas e regulatórias sem sofrer consequências. Esse episódio ganhou destaque na mídia internacional e gerou críticas de especialistas, que apontaram falhas graves na governança e supervisão do projeto. Apesar das denúncias, a administração municipal decidiu manter o chatbot em funcionamento enquanto aplicava ajustes e correções.

Esse caso demonstra como a adoção precipitada de IA generativa em funções críticas pode expor organizações, sejam públicas ou privadas, a riscos legais, reputacionais e operacionais. Para uma PME brasileira, por exemplo, confiar em informações incorretas fornecidas por um chatbot poderia resultar em multas, perda de licenças ou ações judiciais, consequências muito mais difíceis de absorver em comparação com grandes corporações ou governos.

Assim como no caso do agente da Chevrolet e da pizzaria Stefanina's, o exemplo de Nova Iorque reforça que a IA generativa deve ser limitada a funções de apoio, com bloqueios de intents sensíveis (como decisões jurídicas, comerciais ou regulatórias) e validação humana obrigatória. O uso responsável exige que a IA opere apenas como ferramenta de triagem informativa, enquanto as decisões finais permaneçam sob supervisão de profissionais qualificados.

Os padrões de vulnerabilidade identificados nos casos anteriores — desde falhas técnicas de segurança (Manus), manipulação de promessas comerciais (Chevrolet e Stefanina's), até recomendações institucionais incorretas (Nova Iorque) revelam-se igualmente críticos quando sistemas de IA são aplicados a decisões financeiras sensíveis. No final de 2024, estudos acadêmicos da Lehigh University e relatórios de consultorias mostraram que muitos modelos de inteligência artificial usados para análise de crédito, investimentos e outras decisões financeiras automatizadas apresentaram problemas na hora de validar suas recomendações. Pesquisadores da Lehigh University fizeram um estudo com 6.000 simulações de pedidos de crédito usando modelos como GPT-3.5, GPT-4, Claude 3 e Llama 3. Eles perceberam que, ao incluir informações sobre raça, esses chatbots tendiam a recomendar negativas de crédito com mais frequência para candidatos negros do que para candidatos brancos que eram exatamente iguais em outros aspectos. Isso revelou a presença de vieses nesses sistemas de IA.

Uma análise divulgada pela Accessible Law chamou atenção para um ponto importante: mesmo algoritmos de inteligência artificial que parecem neutros podem acabar reforçando padrões de discriminação já existentes, o que afeta o acesso ao crédito de milhões de americanos. Em declarações públicas, Janet Yellen, Secretária do Tesouro dos EUA, alertou sobre os riscos de usar IA no setor

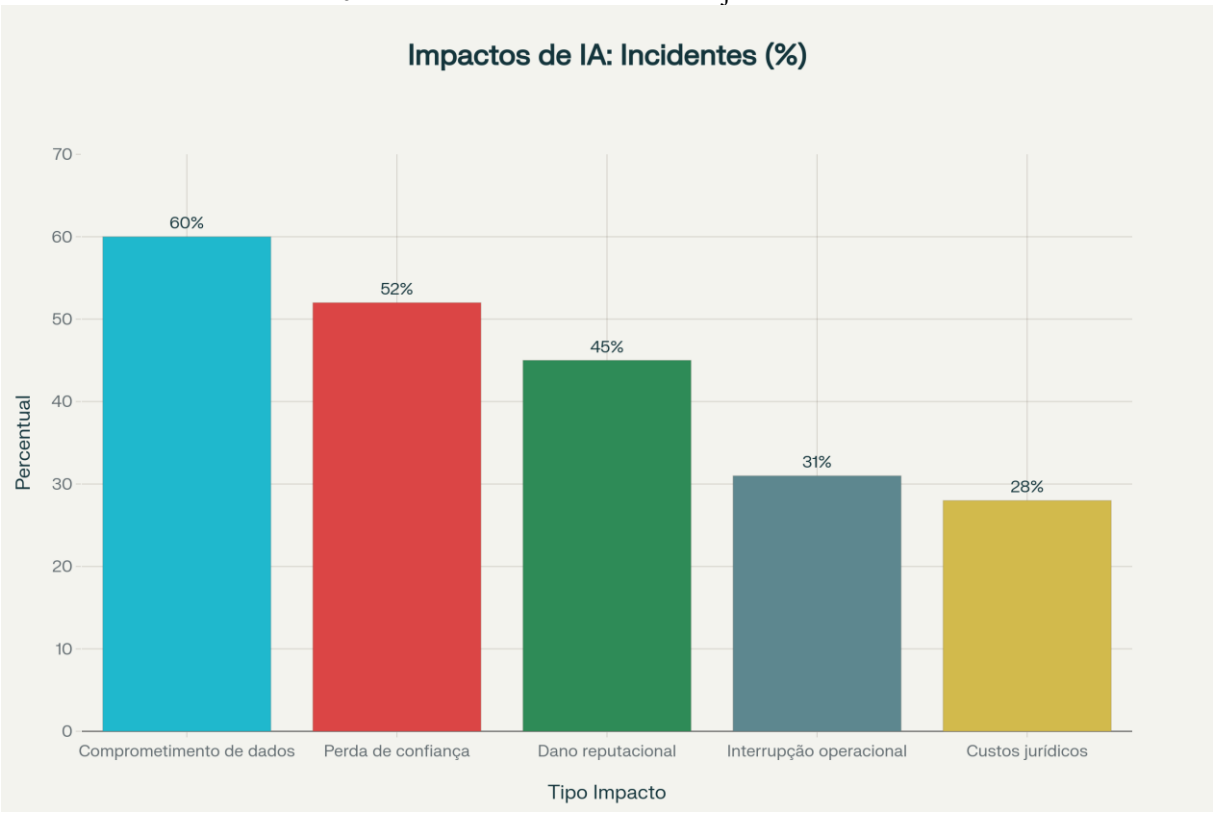
financeiro. Ela destacou que a complexidade, a falta de transparência e a ausência de uma gestão de riscos adequada podem criar vulnerabilidades e vieses. Além disso, o Financial Stability Board (FSB) publicou um relatório reforçando que a rápida adoção de IA pelas instituições financeiras, sem os devidos controles e monitoramento, pode colocar a estabilidade do sistema financeiro em risco. Por isso, é fundamental que haja uma avaliação cuidadosa por parte dos reguladores e a implementação de medidas preventivas.

Esse tipo de erro é parecido com ataques de prompt ou manipulação de dados em IA generativa. Quando um agente financeiro é usado sem restrições ou supervisão humana, ele pode tomar decisões que tenham consequências sérias para os clientes e para a reputação da empresa, como aprovar empréstimos indevidos, sugerir investimentos muito arriscados ou gerar cálculos fiscais imprecisos. O problema não está no modelo em si, mas na forma como ele é utilizado. Transformar um modelo probabilístico em um "executor de políticas financeiras" sem uma validação adequada é como confundir plausibilidade com certeza real, o que pode resultar em perdas importantes e até processos judiciais.

Para assegurar a utilização segura da IA generativa em decisões financeiras, é fundamental que sua atuação seja rigorosamente definida. A tecnologia deve ser limitada a análises iniciais, triagem de clientes ou cálculos estimativos, impedindo que ela faça decisões finais de maneira independente. Simultaneamente, recomendações importantes, como concessão de crédito, modificação de limites ou diretrizes de investimento, não devem ser implementadas sem a supervisão de profissionais capacitados, assegurando que toda decisão significativa seja validada por um ser humano. Ademais, é essencial estabelecer mecanismos de monitoramento constante, analisando métricas de desempenho, mantendo registros detalhados e revisando o modelo regularmente para detectar vieses, tendências inadequadas ou falhas. Em conclusão, é fundamental contextualizar adequadamente qualquer comunicação ou relatório produzido pela IA antes de enviá-lo ao cliente, a fim de evitar mal-entendidos e assegurando transparência e confiabilidade no uso da tecnologia.

ISSN: 2595-1378

Gráfico 6 – Taxa de Sucesso e Falha em Projetos de IA Generativa



Fonte: Massachusetts Institute of Technology (2025).

As consequências da implementação inadequada de IA generativa são mensuráveis e impactantes. Conforme demonstrado no gráfico, os principais tipos de incidentes registrados incluem: comprometimento de dados (60%), perda de confiança (52%), dano reputacional (45%), interrupção operacional (31%) e custos jurídicos (28%). Esses números evidenciam que os riscos não são meramente teóricos, mas se materializam com frequência significativa, reforçando a necessidade urgente de controles preventivos e governança robusta.

Em suma, os quatro casos analisados — desde vulnerabilidades técnicas de segurança (Manus), passando por manipulação comercial (Chevrolet e Stefanina's) e informações institucionais incorretas (Nova Iorque), até vieses em decisões financeiras (estudos de crédito) — demonstram que, em ambientes de produção, modelos generativos podem ser manipulados para emitir respostas plausíveis, porém, incompatíveis com a política interna e que isso escala rápido na opinião pública.

Para PMEs, assim como na área da saúde e em outros setores críticos, a forma segura de usar I.A não é delegar decisões importantes ao modelo, mas sim limitar sua atuação à triagem informativa, com bloqueios de intenção, prompt injection, linguagem de estimativa, supervisão humana e KPIs claros de qualidade e contenção. Essa combinação protege a empresa e evita que a automação "venda" o que o negócio não pode entregar ou tome decisões sobre riscos que a instituição não está em condições de assumir.

Portanto, recomenda-se que PMEs adotem políticas de governança para IA incluindo higienização de entradas, threat modeling, segregação de privilégios e auditoria de logs. Paralelamente, é essencial investir em capacitação contínua e parcerias com instituições de ensino, garantindo que a transformação digital ocorra de forma segura, ética e alinhada à estratégia organizacional (SCHUMACHER; EROL; SIHN, 2016).

4 CONCLUSÃO

Os casos analisados revelam vulnerabilidades críticas na adoção de inteligência artificial generativa sem governança adequada. A partir dessas experiências, é possível extrair soluções práticas, boas práticas e ferramentas acessíveis que permitam às PMEs brasileiras aproveitarem os benefícios da automação inteligente de forma segura. Para prevenir ataques de prompt injection evidenciados nos casos Manus, Chevrolet e Stefanina's, é fundamental implementar sanitização de entrada que identifique comandos suspeitos, estabelecer validação de contexto que verifique o alinhamento das respostas com as políticas da empresa e utilizar delimitadores claros entre instruções do sistema e entradas de usuários.

O caso de Nova Iorque demonstra a necessidade de implementar human-in-the-loop obrigatório para informações legais, regulatórias ou financeiras, alimentar modelos apenas com bases de conhecimento validadas por especialistas, incluir disclaimers claros sobre as limitações da IA e realizar auditoria contínua das respostas geradas. Os vieses identificados nas análises de crédito exigem testes de equidade constantes, comitês de revisão ética, documentação transparente de critérios de decisão e mecanismos de contestação para revisão humana.

A adoção segura de IA em PMEs deve começar com projetos piloto de escopo limitado em áreas de menor risco, permitindo aprendizado controlado antes de expandir para funções críticas. É essencial designar responsáveis humanos para cada implementação, estabelecer protocolos claros de escalação para operadores humanos e documentar rigorosamente objetivos, critérios, limitações e procedimentos de contingência. A comunicação transparente com clientes sobre quando estão interagindo com sistemas automatizados, combinada com revisões periódicas de desempenho e comportamento dos modelos, completa o conjunto de práticas fundamentais.

Existem diversas ferramentas acessíveis para implementação segura de IA. Plataformas como Botpress, Rasa e Dialogflow facilitam a criação de chatbots com controles nativos, enquanto LangChain e Guardrails AI oferecem proteção contra manipulação. Para monitoramento, Langfuse, LangSmith e Helicone rastreiam interações e identificam anomalias. Ferramentas como OWASP ZAP e API Gateways adicionam camadas de segurança, enquanto Notion, Confluence ou Wiki.js centralizam bases de conhecimento validadas. Plataformas como Coursera, Alura, SEBRAE e SENAI

ISSN: 2595-1378

oferecem capacitação acessível, e o AI Risk Management Framework do NIST fornece orientações de governança

REFERÊNCIAS

AIBASE. *Manus AI System Prompt Leakage: Official Response.* 2025. Disponível em: https://www.aibase.com/news/16138. Acesso em: 02 out. 2025.

ACCESSIBLE LAW. *When Algorithms Judge Your Credit: Understanding AI Bias in Lending Decisions.* University of North Texas at Dallas, 2025. Disponível em: https://www.accessiblelaw.untdallas.edu/post/when-algorithms-judge-your-credit-understanding-ai-bias-in-lending-decisions. Acesso em: 02 out. 2025.

AP NEWS. *NYC chatbot gave incorrect and illegal advice to businesses.* 2024. Disponível em: https://apnews.com/article/new-york-city-chatbot-misinformation-6ebc71db5b770b9969c906a7ee4fae21. Acesso em: 02 out. 2025.

BARDIN, L. *Análise de conteúdo.* Lisboa: Edições 70, 2011.

BOWEN III, D.; PRICE, S. M.; YANG, K. *Measuring and Mitigating Racial Disparities in Large Language Models: Evidence from a Mortgage Underwriting Experiment.* SSRN, 2024. Disponível em: https://ssrn.com/abstract=4812158. Acesso em: 02 out. 2025.

BUSINESS INSIDER. *Chevrolet chatbot offered Tahoe for $1 after prompt injection.* 2023. Disponível em: https://www.businessinsider.com. Acesso em: 02 out. 2025.

CGI.br. *TIC Empresas 2024 – Release.* 2025. Disponível em: https://www.cgi.br. Acesso em: 02 out. 2025.

EMBRACE THE RED. *How Prompt Injection Exposes Manus' VS Code Server to the Internet.* 2025. Disponível em: https://embracethered.com/blog/posts/2025/manus-ai-kill-chain-expose-port-vs-code-server-on-internet/. Acesso em: 02 out. 2025.

ESET. *Prompt Injection: uma ameaça silenciosa à segurança em IA.* 2025. Disponível em: https://www.welivesecurity.com. Acesso em: 02 out. 2025.

FINANCIAL STABILITY BOARD (FSB). *Artificial Intelligence and Financial Stability.* 2024. Disponível em: https://www.fsb.org/uploads/P14112024.pdf. Acesso em: 02 out. 2025.

HOUDE et al. *Business (mis)Use Cases of Generative AI.* IBM Research, 2020. Disponível em: https://research.ibm.com/publications/business-misuse-cases-of-generative-ai. Acesso em: 02 out. 2025.

LEHIGH UNIVERSITY. *AI Exhibits Racial Bias in Mortgage Underwriting Decisions.* Lehigh University News, 2024. Disponível em: https://news.lehigh.edu/ai-exhibits-racial-bias-in-mortgage-underwriting-decisions. Acesso em: 02 out. 2025.

REUTERS. *Yellen warns of significant risks in use of AI in finance.* 2024. Disponível em: https://www.reuters.com/business/finance/yellen-warn-significant-risks-use-ai-finance-2024-06-05. Acesso em: 02 out. 2025.

SCHUMACHER, A.; EROL, S.; SIHN, W. A Maturity Model for Assessing Industry 4.0 Readiness. *Procedia CIRP*, v. 52, p. 161–166, 2016.

SCHWAB, K. *A Quarta Revolução Industrial.* São Paulo: Edipro, 2016.

SEBRAE. *Panorama das PMEs no Brasil.* 2023. Disponível em: https://www.sebrae.com.br. Acesso em: 02 out. 2025.

SHEPHERD. *Generative AI Misuse Potential in Cyber Security Education: A Case Study of a UK Degree Program.* 2025. Disponível em: https://warwick.ac.uk/fac/cross_fac/eduport/edufund/projects/yang/projects/generative-ai-misuse-potential-in-cyber-security-education-a-case-study-of-a-uk-degree-program. Acesso em: 02 out. 2025.

WIBW. *Google AI Overviews caused confusion at Stefanina's Pizzeria.* 2025. Disponível em: https://www.wibw.com. Acesso em: 02 out. 2025.

Gráfico 1 – Evolução dos Custos de Breach de Dados (2020-2025)
IBM Corporation. *Cost of a Data Breach Report 2025*. IBM Security, 2025. Disponível em: https://www.ibm.com/reports/data-breach. Acesso em: 21 out. 2025.
Link direto do relatório: https://www.bakerdonelson.com/webfiles/Publications/20250822_Cost-of-a-Data-Breach-Report-2025.pdf

Gráfico 2 – Principais Vulnerabilidades em IA
Embrace The Red; Aibase. *Estudo sobre Vulnerabilidades e Controles em IA Gerativa*. 2025. Relatório técnico.
(Nota: relatório próprio citado baseado na síntese do texto, recomendo guardar dados organizacionais ou acadêmicos internos para referência formal).

Gráfico 3 – Impactos dos Incidentes de IA
IBM Corporation. *Impactos e Custos de Incidentes em Sistemas de IA*. IBM Security, 2025. Relatório relacionado ao Cost of Data Breach Report.
Veja link acima do relatório IBM.

Gráfico 4 – Adoção e Maturidade Digital de IA em PMEs Brasileiras
Comitê Gestor da Internet no Brasil (CGI.br). *Relatório de Adoção de Tecnologias 4.0 no Brasil*, 2025. Disponível em: https://cgi.br/ (pesquisar seção de Relatórios).

Gráfico 5 – Níveis de Maturidade Digital no Brasil
Serviço Brasileiro de Apoio às Micro e Pequenas Empresas (SEBRAE). *Pesquisa de Maturidade Digital das PMEs Brasileiras*, 2023. Disponível em: https://www.sebrae.com.br/

Gráfico 6 – Taxa de Sucesso e Falha em Projetos de IA Generativa
Massachusetts Institute of Technology - MIT NANDA Initiative. *Estudo sobre Pilotagem e Resultados de Projetos de IA Generativa*, 2025.
(Disponível em publicações acadêmicas no arXiv: https://arxiv.org/abs/2410.23308 ou outras bases acadêmicas)