



White paper

O guia do CIO para engenharia agêntica: Criando apps corporativos na velocidade da IA sem comprometer a confiabilidade

Um guia para COOs, CXOs e vice-presidentes de operações



A pressão é real, e o problema também.

Os conselhos de administração já não perguntam se devem adotar a IA. Agora perguntam quando, em quanto tempo e qual será o retorno. Para COOs e líderes de operações, essa pressão se transformou em uma onda de projetos-piloto com agentes de IA autônomos – implantações em que a IA cuida das interações com clientes, encaminha casos, verifica políticas e gerencia fluxos de trabalho com intervenção humana mínima.

Os primeiros resultados pareciam promissores: resolução mais rápida. Redução na demanda por pessoal. Menos escalonamentos manuais. Mas, conforme esses projetos-piloto ganharam escala em produção, um cenário diferente apareceu: os custos não se comportam como os modelos previam, os resultados não são tão consistentes quanto as demonstrações sugeriam, e o controle operacional que os líderes achavam ter foi se perdendo silenciosamente.

56% dos CEOs afirmam hoje não obter nenhum benefício financeiro mensurável com os investimentos em IA ([29ª Pesquisa Global de CEOs da PwC, janeiro de 2026](#)). Isso não é uma falha da tecnologia. É uma falha da arquitetura. A questão não é se a IA é capaz de gerar eficiência operacional. A questão é se o modo como a maioria das empresas a implanta é estruturalmente capaz de entregar isso.

Para a maioria, a resposta é não.



O custo invisível do raciocínio

A IA transformou profundamente a economia da automação corporativa. Os tokens – a unidade que mede quanto um sistema de IA está "pensando" – tornaram-se um dos itens mais decisivos no orçamento de transformação de uma empresa. E, na maioria das organizações, esse item está se comportando de um jeito que ninguém previu.

Quando os líderes de negócios aprovaram as implantações de agentes de IA, o modelo de custo era direto: trocar o trabalho manual pelo raciocínio automatizado, reduzir custos com pessoal e ficar com a margem. O que esse modelo não previu foi o quanto o consumo de tokens se tornaria imprevisível em escala corporativa. Duas forças alimentam essa imprevisibilidade e, juntas, vêm corroendo o ROI prometido aos conselhos de administração.

A primeira é a natureza cumulativa do uso agêntico de tokens. Agentes de IA encarregados de tarefas longas não raciocinam uma só vez e seguem em frente. Eles recalculam o raciocínio a cada etapa, checando o contexto, pesando opções e produzindo resultados sem parar. O consumo de tokens para manter esses sistemas em operação vem por cima do consumo que já foi gasto para criá-los e desenvolvê-los. No volume de uma operação corporativa de médio porte, a conta cresce mais rápido do que a maioria dos modelos financeiros previa.

É o que chamamos de imposto dos tokens.

A segunda força é a oscilação dos preços. No início, os fornecedores de IA praticavam preços subsidiados e ilimitados, o que dava às empresas uma estrutura de custos previsível para se planejar. Esse tempo acabou. Como a economia da IA passou a esbarrar em limites de chips, energia e capacidade de data center, e como os fornecedores de IA caminham para a abertura de capital, as empresas vêm absorvendo sucessivas mudanças de preço que deslocam o modelo da tarifa fixa para a cobrança por consumo. O custo de uma mesma carga de trabalho deixou de ser fixo. Em junho de 2026, Sam Altman admitiu em público que os custos de tokens de IA viraram "um problema enorme" até para os próprios fornecedores ([Axios, junho de 2026](#)).

O resultado é que as empresas vêm consumindo seus orçamentos de transformação sem enxergar com clareza o caminho até o ROI. [Segundo a Gartner, mais de 40% dos projetos de IA agêntica devem ser cancelados até o fim de 2027, com os custos operacionais em alta e imprevisíveis entre as causas principais \(Gartner, junho de 2025\).](#)

Hoje os executivos já não perguntam se devem usar a IA. Perguntam como chegar a um ROI positivo quando a estrutura de custos por baixo dela muda o tempo todo. A resposta passa por inverter a lógica: pagar por resultados, não por tokens. Na prática, é montar uma arquitetura que reserve o gasto de tokens para o trabalho que realmente exige raciocínio e deixe todo o resto rodando sobre uma estrutura de custos previsível, fixa e sem efeito cumulativo.

Quando a IA foge do roteiro

Os custos imprevisíveis são apenas uma dimensão do problema. A outra é o que acontece quando agentes de IA passam a decidir sobre aquilo que nunca deveria estar em suas mãos.

Os modelos de linguagem de grande porte são treinados para ser prestativos e fluentes. Eles não são treinados para seguir as políticas de compliance da sua empresa. Eles não compreendem por conta própria os seus SLAs, as suas obrigações regulatórias, as suas regras de escalonamento ou os compromissos específicos que a sua marca assumiu com os clientes. Sem restrições rígidas, agentes autônomos em fluxos de trabalho voltados ao cliente podem abrir uma distância considerável entre o que a IA considera razoável e o que a sua equipe de compliance aprovaria.

A questão não se resume à tolerância ao risco. É uma questão de arquitetura.

As decisões que trazem maior exposição jurídica e reputacional – elegibilidade, resolução de disputas, exceções a políticas, comunicações reguladas – são exatamente aquelas que têm uma resposta correta definida. Não há ambiguidade nelas, e inferência probabilística não acrescenta nada. Encaminhar um caso pelo status da conta, verificar se uma solicitação cumpre um checklist documentado, aplicar uma exceção de política conforme critérios definidos – tudo isso são operações determinísticas. Delegar essas decisões ao raciocínio de um agente de IA não agrega valor algum. O que se ganha é variabilidade onde se exige consistência e exposição jurídica onde se exige precisão.

Em produção e em escala, alucinações de IA deixam de ser exceção. São o resultado esperado de qualquer sistema que produz respostas por inferência probabilística, e não por regras determinísticas. Onde há pouco em jogo, uma alucinação não passa de um transtorno. Em uma disputa com o cliente, em um fluxo de sinistros, em uma interação de serviços financeiros ou em uma comunicação de setor regulado, ela vira um passivo.

[Entre as principais previsões da Gartner para 2026 aparece uma estimativa de mais de 2.000 ações judiciais por "morte causada por IA" no mundo, resultado de governança e controles de risco frágeis sobre o que a IA gera \(Gartner, outubro de 2025\).](#) Nada disso é risco abstrato para um futuro distante. É exposição concreta e presente em qualquer organização que já entregou a agentes de IA autônomos autoridade relevante sobre fluxos voltados ao cliente ou regulados.

A resposta mais produtiva está na arquitetura: são as regras de negócio que determinam o que os agentes de IA podem e não podem fazer. A IA nunca deveria ser o sistema de registro da sua lógica de compliance. Ela deve operar dentro desse sistema. Em decisões que não deixam margem a dúvida, quem define a ação correta é a lógica de negócio, aplicada na camada de orquestração – e não um modelo capaz de raciocinar até chegar a outra resposta amanhã. Com essa separação em vigor, o compliance deixa de ser vigilância e passa a ser estrutura. A IA não consegue fugir do roteiro, porque o roteiro não se deixa contornar por raciocínio.

A arquitetura que garante os dois lados

O problema operacional da maioria das empresas não é excesso de IA. É que a IA está operando nas partes erradas da arquitetura de processos.

Há um argumento econômico legítimo para a IA em operações: custo de atendimento menor, mais volume processado e experiências de cliente mais consistentes em escala. Esses são resultados alcançáveis. Mas exigem uma visão nítida de onde a IA agrega valor e de onde ela acrescenta custo e risco sem retorno equivalente.

A arquitetura que entrega esses resultados de forma consistente tem duas camadas atuando em conjunto. A primeira é uma camada de orquestração determinística: um mecanismo

baseado em regras que gerencia o fluxo do processo, faz valer a lógica de negócio, acompanha o estado e encaminha o trabalho conforme critérios definidos. Essa camada opera com custo marginal por transação próximo de zero. Ela não raciocina. Ela executa. É completamente previsível, totalmente auditável e não alucina.

A segunda camada é a inferência de IA, acionada exatamente nos pontos do processo em que a lógica determinística não basta. Análise de sentimento do cliente. Interpretação de documentos não estruturados. Geração de respostas em linguagem natural. Resumo de casos complexos. Aqui, o caráter probabilístico da IA joga a favor em vez de virar passivo, justamente porque a tarefa não tem uma resposta única que regras conseguiriam calcular.

Separadas como devem ser, essas duas camadas reduzem drasticamente o imposto dos tokens. O gasto de tokens passa a se concentrar apenas no trabalho que realmente exige inferência. Todo o resto é executado de forma determinística. Ao mesmo tempo, o risco de compliance cai, porque a camada de orquestração faz valer as regras de negócio independentemente do que a camada de IA produza. A IA atua dentro de limites que raciocínio nenhum consegue driblar.

Apenas 12% dos CEOs dizem ver a IA gerar ganhos mensuráveis em custo e receita ao mesmo tempo ([PwC, janeiro de 2026](#)). As empresas desses 12% não usam mais IA que as demais. Elas a implantam com muito mais disciplina estrutural.



A vantagem competitiva está no controle

Quem vai extrair ROI duradouro da IA não são os líderes de operações que implantaram mais rápido nem os que soltaram mais as rédeas dos seus agentes. São aqueles que seguraram o controle da arquitetura de processos e colocaram a IA exatamente onde ela gera valor.

Isso significa tratar a orquestração como ativo estratégico, e não como uma amarra do legado. Significa separar com clareza o que é decisão determinística do que exige inferência e montar uma arquitetura que faça essa fronteira valer. Significa acompanhar o gasto com IA transação a transação, não apenas no nível da plataforma, para que o imposto dos tokens fique visível antes de virar um rombo no orçamento.

O conselho de administração acerta ao dizer que a IA pode baixar o custo de atendimento e melhorar a experiência do cliente. O caminho até lá não passa por agentes autônomos tentando decifrar os seus fluxos de trabalho. É uma arquitetura de processos governada e orquestrada, na qual a IA entra com precisão.

Quer assumir o controle? Use a [calculadora de TCO](#) e meça a diferença entre o seu gasto atual com tokens e uma orquestração estruturada nos fluxos de trabalho de maior volume.





Obrigado

Sobre a Pega

A Pega fornece a principal plataforma impulsionada por AI para transformação empresarial. As organizações mais influentes do mundo confiam na nossa tecnologia para repensar como o trabalho é realizado, automatizando fluxos de trabalho, personalizando a experiência do cliente e modernizando sistemas legados. Desde 1983, nossa arquitetura escalável e flexível vem alimentando uma inovação constante, ajudando clientes a acelerar a jornada rumo à Empresa Autônoma.

[PEGA.COM/PT-BR](https://www.pega.com/pt-br)