



O custo invisível da IA não está no modelo — está em como você o opera

Por que apostar em uma IA previsível

Junho de 2026 | Don Schuerman, CTO e Head de Marketing



A próxima fase da IA não é sobre possibilidade. É sobre responsabilidade.

Nos últimos dois anos, a IA corporativa deixou de ser curiosidade e virou capacidade real. O que antes exigia equipes especializadas e longos ciclos de desenvolvimento hoje se resolve em poucos dias. Gerar código, colocar agentes no ar e automatizar fluxos de trabalho nunca foi tão rápido para as equipes.

Essa virada mudou o teor da conversa entre executivos.

Ano passado, os líderes perguntavam: *O que a IA pode fazer por nós?*

Hoje a pergunta é mais direta: *Quanto ela custa e que valor estamos de fato vendo?*

Conforme as organizações deixam os projetos-piloto para trás e começam a escalar a IA em operações reais, as expectativas sobem. A IA saiu do orçamento de inovação. Ela passou a fazer parte do funcionamento do negócio. E isso significa que ela precisa resistir ao escrutínio em termos de confiabilidade, governança e custo.

Em todos os setores, cresce uma percepção: a promessa da IA é clara. O custo dela não é. Uma pesquisa recente da Gartner¹ mostra que a maioria das iniciativas de IA ainda tem dificuldade em gerar retorno relevante, e muitas organizações enxergam pouco ou nenhum ROI mensurável no investimento inicial.

Isso não é uma falha da IA. É o sinal de uma nova fase: aquela em que o sucesso dependerá menos da capacidade de construir e mais da eficácia em operar em escala.

¹ Pesquisa da Gartner sobre o valor da IA (gartner.com/en/articles/ai-value)

Criar com IA está cada vez mais fácil, mas operá-la com eficiência está cada vez mais difícil

A primeira onda de adoção da IA foi impulsionada pelo acesso. Com modelos poderosos amplamente disponíveis, as organizações correram para experimentar.

A inovação acelerou. Mas, conforme esses experimentos chegam à produção, outra realidade aparece.

Custos que na escala do piloto pareciam sob controle começam a disparar. A qualidade dos resultados oscila. E, em muitos casos, fica difícil medir a ligação entre o investimento em IA e o impacto no negócio. No centro desse desafio há uma questão estrutural: A maioria dos sistemas de IA hoje é projetada para **experimentação, não para operação contínua.**

Muitos deles dependem bastante do *raciocínio em tempo de execução*, no qual o modelo avalia cada interação do zero. Essa flexibilidade pode ser útil, mas tem um preço. Cada interação consome processamento. Cada salto de complexidade aumenta esse consumo. E, conforme o uso cresce, esses custos se acumulam.

Toda vez que a IA precisa reavaliar uma tarefa, interpretar o contexto ou decidir o próximo passo, gera custo adicional. Em pequena escala, isso é administrável. Em escala corporativa, o efeito se acumula, porque o sistema não está apenas executando o trabalho: está refazendo o mesmo raciocínio sem parar. Na prática, muitas organizações enfrentam hoje algo próximo de um "imposto dos tokens": cada etapa adicional de raciocínio traz um custo marginal que se acumula em escala.

E, em escala, isso muda por completo a equação de custo. Em muitas implantações corporativas, o custo da inferência – manter os modelos rodando continuamente em produção – logo supera o custo inicial de construí-los, transformando a IA de investimento pontual em despesa operacional contínua.

O custo real da IA não está no que você cria. Está no que você precisa manter rodando sem parar. As organizações pagam de forma contínua para "descobrir" um trabalho que poderia ter sido estruturado de antemão. É por isso que soluções com bom desempenho em um projeto-piloto controlado costumam ter dificuldade para escalar com eficácia no mundo real. **A limitação não está no potencial. Está na eficiência.**



² MIT Technology Review, "AI's energy usage and climate footprint" (2025)

O custo real da IA está na forma como o trabalho é feito

Quando os líderes tentam gerenciar o gasto com IA, o instinto costuma ser olhar para os modelos – comparando fornecedores, recursos e estruturas de preço.

Mas existe uma pergunta mais importante: **Onde, na sua operação, a IA está fazendo o trabalho? E como esse trabalho está estruturado?**

Em muitas organizações, a IA cresceu de forma orgânica: equipes distintas implantaram modelos, agentes e fluxos de trabalho diferentes, com pouca coordenação. O resultado não é só complexidade. É duplicação. As mesmas decisões são tomadas várias vezes, em lugares diferentes e, muitas vezes, com lógicas diferentes.

As projeções já apontam que as cargas de trabalho de IA corporativa devem multiplicar por mais de 20 o uso de tokens nos próximos anos³. Conforme os agentes ganham escala no negócio, esse crescimento não amplia só o uso – amplia o custo. Com isso, a IA costuma ser sobreposta aos processos existentes sem alterar de fato o modo como eles funcionam.

Isso gera melhoria incremental, mas não transformação, e cria dois problemas sistêmicos:

1. Preserva a ineficiência. Se o processo por baixo é fragmentado ou complexo demais, acrescentar IA não elimina essa complexidade. Apenas a acelera.
2. Joga responsabilidade demais para o tempo de execução. Em vez de definir de antemão como o trabalho deve acontecer, as organizações deixam a IA determinar o próximo passo dinamicamente, interação por interação.

Essa flexibilidade pode parecer poderosa. Mas traz variabilidade, eleva o custo e dificulta o controle dos resultados. No fim das contas, essa é uma questão de arquitetura:

- Sistemas bem projetados reduzem o esforço antes mesmo de o trabalho começar.
- Sistemas mal estruturados geram custo toda vez que são executados.

Quando a IA é aplicada sem repensar como o trabalho circula entre sistemas e equipes, o custo passa a crescer com o uso, em vez de se estabilizar com a escala.

Por isso muitas organizações veem o gasto crescer mais rápido do que os resultados.

³ Goldman Sachs, "AI agents forecast to boost tech cash flow as usage soars"

Adicionar IA a processos quebrados não os conserta, e sim escala seus problemas

Um padrão que aparece com frequência é o que se poderia chamar de abordagem "acoplada" da IA. A empresa identifica um fluxo de trabalho – atendimento ao cliente, onboarding, sinistros – e insere uma camada de IA para melhorá-lo. A experiência fica mais conversacional. O sistema fica mais responsivo.

Mas o processo por baixo continua o mesmo. Vários sistemas continuam envolvidos. As decisões seguem fragmentadas. O trabalho segue passando por uma série de etapas desconectadas. Para compensar, a IA assume mais responsabilidade. Ela interpreta, encaminha e se adapta em tempo real, na prática tentando manter o processo de pé.

Essa abordagem pode melhorar a experiência no curto prazo – mas não altera a estrutura de custos. Na verdade, com frequência a encarece. Isso porque o sistema agora realiza tanto o trabalho original quanto a interpretação contínua necessária para manter tudo funcionando junto.

Toda vez que o sistema precisa reavaliar ou reconstruir o fluxo de trabalho, consome recursos adicionais. Multiplique isso por milhares ou milhões de interações e o impacto fica evidente. Nesse modelo, a IA se torna uma camada de interpretação contínua assentada sobre a complexidade estrutural.

Isso não se sustenta em escala. Se o objetivo é mudar a equação de custo, o ponto de partida precisa ser o próprio processo.

O controle de custos começa no design, não na otimização em tempo de execução.

As organizações que vêm avançando de verdade com a IA encaram o problema de outra forma. Elas estão migrando para um modelo orquestrado – no qual a IA é coordenada, estruturada e incorporada ao modo como o trabalho é projetado e executado.

Em vez de contar com a IA para decidir no calor do momento, recorrem a ela mais cedo – no design de como o trabalho deve ser feito. Nessa etapa, a IA pode ser extremamente poderosa. Ela ajuda a analisar os processos existentes, identificar ineficiências e propor formas melhores de estruturar o trabalho entre sistemas e equipes.

Definida essa estrutura, a execução fica muito mais eficiente. Em tempo de execução, o sistema deixa de precisar "descobrir" cada etapa. Ele percorre um caminho definido e recorre à IA de forma seletiva, onde ela agrega mais valor – como interpretar intenções ou lidar com entradas não estruturadas.

Essa passagem do *raciocínio em tempo de execução* para a *inteligência em tempo de design* muda tanto o perfil de custos quanto a consistência dos resultados. Com isso, as organizações conseguem:

- Reduzir processamento desnecessário
- Eliminar decisões repetidas
- Garantir que o trabalho seja executado de maneira repetível e governada

Nesse modelo, a IA passa a integrar um sistema de execução – e não a substituí-lo.

A previsibilidade é o que viabiliza a IA em escala corporativa

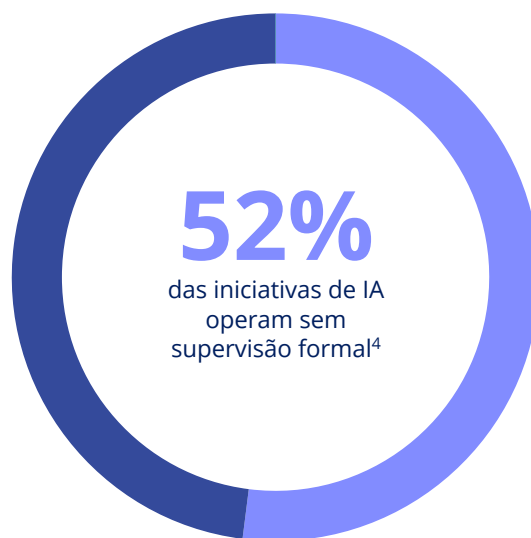
Para que a IA se torne estrutural na empresa, precisa operar com a mesma disciplina dos demais sistemas centrais.

E isso começa pela previsibilidade. Os líderes precisam de segurança em três frentes:

- 1. Quanto vai custar**
- 2. Que resultado vai produzir**
- 3. Como se comportará em condições reais**

Sem isso, escalar vira risco.

Uma pesquisa recente da EY mostra que mais da metade das iniciativas de IA funciona sem supervisão formal – o que expõe as organizações a estouros de orçamento, lacunas de compliance e resultados inconsistentes conforme a adoção ganha escala.



É aqui que uma abordagem mais estruturada da IA começa a ganhar terreno. Quando a IA opera dentro de fluxos de trabalho bem definidos, com o apoio de orquestração e governança, algo importante acontece:

- Os resultados ficam mais consistentes.
- Os custos ficam mais fáceis de antecipar e cada vez mais atrelados a decisões, transações e resultados concretos de negócio.
- As operações ficam mais fáceis de gerenciar.

E, o que é mais importante, essa previsibilidade não vale apenas para o uso de tokens. Ela alcança todo o custo total de propriedade – operações, governança, compliance e o esforço contínuo de evoluir os sistemas ao longo do tempo. É aqui que ele fica mais nítido, indo além dos tokens para incluir o custo de operar, governar, manter e escalar a IA.

Em vez de tratar a IA como um recurso aberto e sem contornos, as organizações passam a encará-la como parte controlada e mensurável do modo como o trabalho é feito. Essa mudança permite que elas deixem de medir consumo e passem a medir resultados. E esse é um passo decisivo para fazer as contas da IA fecharem.

⁴ EY, "Autonomous AI adoption surges as oversight falls behind" (2026)

As organizações que tiverem sucesso vão mudar o jeito de pensar a IA

Muitas organizações ainda trabalham sob uma falsa escolha: avançar rápido com a IA ou manter o controle. Na prática, a limitação não é a velocidade – é a arquitetura.

Quem extrair todo o valor da IA não será apenas quem chegou primeiro.

As empresas que acertam nesse ponto repensam o modo como o negócio opera. Na prática, isso significa concentrar-se em outro conjunto de prioridades:

Deslocar o foco das ferramentas para os sistemas

A pergunta central não é o que a IA consegue fazer isoladamente, e sim como ela se integra ao fluxo de trabalho de toda a empresa.

Redesenhar antes de automatizar

Aplicar IA a um processo quebrado raramente resolve o problema. Repensar o processo antes é o que cria a base para um impacto real.

Medir resultados, não atividade

O sucesso precisa estar atrelado a resultados de negócio – trabalho concluído, experiências melhores, menos atrito – e não ao volume de uso.

Tratar a IA como infraestrutura operacional

Conforme a IA se incorpora aos processos centrais, precisa atender aos mesmos padrões de confiabilidade e governança de qualquer outro sistema corporativo.

O ponto é menos adotar novas tecnologias e mais aplicar disciplina ao modo como essa tecnologia é usada.

Da experimentação à execução

Estamos diante de um ponto de virada na IA corporativa. A primeira fase mostrou o que era possível. A próxima fase vai definir o que é prático e o que é escalável.

Conforme as organizações ampliam o uso da IA, o desafio deixa de ser gerar insights ou automatizar tarefas isoladas. Está em operar esses recursos de forma estável, eficiente e alinhada aos resultados de negócio.

É essa a diferença entre um sucesso isolado e um impacto que atravessa a empresa inteira.

As organizações que fecharem essa lacuna vão avançar mais rápido, operar com mais eficiência e se adaptar melhor às mudanças.

Mas não chegarão lá apenas implantando mais IA.

Chegarão lá construindo a base certa para o modo como a IA opera dentro do negócio.

Em resumo

A IA já vem transformando o que as organizações conseguem fazer.

Agora o foco passa para como ela se comporta em condições reais, em escala, em meio a sistemas complexos e dentro dos limites de custo e controle.

O maior custo oculto da IA não está no modelo em si. Ele vem da forma como o trabalho é estruturado, executado e repetido ao longo do tempo.

As organizações que resolverem esse ponto não vão apenas reduzir custos. Elas vão construir um sistema capaz de sustentar de fato a operação do negócio.

E, na próxima era da IA corporativa, são essas as organizações que saem na frente – porque construíram uma IA que conseguem controlar, medir e operar como parte do negócio.



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)