



INFORME

# ***La guía de ingeniería agéntica para el CIO:***

## **Cómo crear aplicaciones empresariales a la velocidad de la IA sin comprometer la fiabilidad**

Una guía para directores de operaciones (COO), directores ejecutivos (CXO) y vicepresidentes de operaciones



# La presión es real, y el problema también.

Los juntas directivas ya no se preguntan si deben implementar la IA. Se preguntan cuándo, con qué rapidez y cómo será el retorno. Para los directores de operaciones y los responsables de operaciones, esa presión se ha traducido en una oleada de proyectos piloto con agentes de IA autónomos: implementaciones en las que la IA gestiona las interacciones con los clientes, deriva los casos, verifica las políticas y administra los flujos de trabajo con una mínima intervención humana.

Los primeros resultados parecían prometedores: tiempos de resolución más rápidos. Reducción de las necesidades de personal. Menos escalamientos manuales.

Pero a medida que estos proyectos piloto se han ampliado a producción, está surgiendo un escenario diferente: los costos no se comportan como predijeron los modelos, los resultados no son tan coherentes como mostraban las demostraciones, y el control operativo que los responsables asumieron que mantenían, se ha ido erosionando silenciosamente.

El 56% de los directores ejecutivos afirman que actualmente no obtiene ningún beneficio financiero cuantificable de las inversiones en IA ([PwC 29th Global CEO Survey, enero de 2026](#)). Eso no es un fallo de la tecnología. Es un fallo de la arquitectura. La cuestión no es si la IA puede ofrecer eficiencia operativa. La cuestión es si la forma en que la mayoría de las empresas lo están implementando es estructuralmente capaz de hacerlo.

**La respuesta, para la mayoría, es no.**



# El costo oculto del razonamiento

La IA ha transformado radicalmente la economía de la automatización empresarial. Los tokens, la unidad de medida de cuanto "piensa" un sistema de IA, se han convertido en una de las partidas de mayor impacto en el presupuesto de transformación de una empresa. Y para la mayoría de las organizaciones, esa partida presupuestaria se está comportando de maneras que nadie había previsto.

Cuando los responsables de las empresas aprobaron la implementación de agentes de IA, el modelo de costos era sencillo: sustituir el trabajo manual por razonamiento automatizado, reducir los costos de personal y conseguir un margen de beneficio. Lo que ese modelo no tuvo en cuenta fue lo impredecible que se volvería el consumo de tokens a escala empresarial. Dos fuerzas provocan esa imprevisibilidad y, juntas, están erosionando el retorno de la inversión que se les prometió a los consejos de administración.

La primera fuerza es la naturaleza acumulativa del uso de tokens agénticos. Los agentes de IA que realizan tareas complejas y prolongadas no razonan una sola vez y pasan a otra cosa. Vuelven a realizar sus razonamientos en cada etapa del proceso, verificando el contexto, evaluando opciones y generando resultados de forma continua. El gasto en tokens utilizado para operar estos sistemas se suma al gasto en tokens utilizado para construirlos y desarrollarlos inicialmente. Con el volumen de operaciones de una empresa mediana, esto se acumula más rápido de lo que la mayoría de los modelos financieros anticipaban.

## **A esto lo llamamos el impuesto de los tokens.**

La segunda fuerza es la inestabilidad de los precios. Los primeros proveedores de IA ofrecían modelos de precios subvencionados y de tarifa plana, que proporcionaban a las empresas una estructura de costos predecible sobre la que planificar. Esa era está llegando a su fin. A medida que la economía de la IA se ha enfrentado a limitaciones de recursos como chips, energía y capacidad de centros de datos, y a medida que los proveedores de IA se preparan para salir a bolsa, las empresas han tenido que asumir repetidos cambios de precios que transforman el modelo de tarifa plana a facturación basada en el consumo. El gasto para la misma carga de trabajo ya no es fijo. Sam Altman reconoció públicamente en junio de 2026 que los costos de los tokens de IA se han convertido en "un gran problema" incluso para los propios proveedores ( [Axios, junio de 2026](#) ).

El resultado es que las empresas están gastando sus presupuestos de transformación sin tener una visión clara del retorno de la inversión. Gartner predice que más del 40% de los proyectos de IA con agentes serán cancelados para finales de 2027, y los costos operativos crecientes e impredecibles se citan como un factor determinante ( [Gartner, junio de 2025](#) ).

La pregunta que se hacen ahora los ejecutivos no es si deben usar la IA. Se trata de cómo lograr un retorno de la inversión positivo cuando la estructura de costos subyacente está en constante cambio. La solución requiere un cambio de enfoque: pasar de pagar por tokens a pagar por resultados. Eso significa construir una arquitectura donde el gasto en tokens de IA se concentre en el trabajo que realmente requiere razonamiento, mientras que todo lo que no requiere razonamiento funciona con una estructura de costos predecible, fija y que no se acumula.

# Cuando la IA se sale del guion

Los costos impredecibles son una de las dimensiones del problema. El otro problema surge cuando se otorga a los agentes de IA autoridad sobre decisiones que, en primer lugar, nunca deberían tomar.

Los modelos de lenguaje a gran escala se entrenan para ser útiles y fluidos. No se entrenan para seguir las políticas de cumplimiento normativo de su empresa. No tienen un conocimiento intrínseco de sus acuerdos de nivel de servicio (SLA), sus obligaciones regulatorias, sus reglas de escalamiento ni los compromisos específicos que su marca ha asumido con los clientes. Cuando los agentes autónomos operan en flujos de trabajo orientados al cliente sin restricciones estrictas, la diferencia entre lo que la IA considera una respuesta razonable y lo que aprobaría su equipo de cumplimiento normativo puede ser significativa.

## **No es una cuestión de tolerancia al riesgo. Es cuestión de la arquitectura.**

Las decisiones que conllevan mayor riesgo legal y reputacional, determinaciones de elegibilidad, resolución de disputas, excepciones a las políticas o comunicaciones reguladas, son también las decisiones que tienen una respuesta correcta definida. No son ambiguas; no se benefician de la inferencia probabilística. Asignar un caso en función del estado de la cuenta, comprobar si un reclamo cumple una lista de verificación documentada o aplicar una excepción de una política de acuerdo con unos criterios definidos: estas son operaciones deterministas. Otorgar a un agente de IA la autoridad para razonar a través de ellas no añade valor. Introduce variabilidad en procesos que requieren consistencia y riesgos en interacciones que requieren precisión.

Las alucinaciones de la IA no son casos excepcionales en entornos de producción a gran escala. Son un resultado predecible de cualquier sistema que genere respuestas mediante inferencia probabilística en lugar de reglas deterministas. En contextos de poca importancia, una alucinación es solo un inconveniente. En una disputa con un cliente, un flujo de trabajo de reclamaciones, una interacción con servicios financieros o una comunicación en una industria regulada, constituye un riesgo.

Las principales predicciones de Gartner para 2026 incluyen una previsión de que las demandas legales por "muerte por IA" superarán las 2000 a nivel mundial, impulsadas por la insuficiente gobernanza y los controles de riesgo en los resultados generados por la IA ( [Gartner, octubre de 2025](#)). Estos no son riesgos abstractos para futuras implementaciones. Representan una exposición activa para cualquier organización que haya otorgado a agentes de IA autónomos una autoridad significativa sobre flujos de trabajo regulados o de cara al cliente en la actualidad.

La respuesta más eficaz está en la arquitectura: las reglas de negocio rigen lo que los agentes de IA pueden y no pueden hacer. La IA nunca debería ser el sistema de referencia para su lógica de cumplimiento normativo. Debe funcionar dentro de ese marco. La acción correcta para decisiones claras y sin ambigüedades viene determinada por la lógica de negocio, aplicada en la capa de orquestación, y no delegada a un modelo que podría llegar a una respuesta diferente en otro momento. Cuando se establece esa separación, el cumplimiento normativo pasa a ser estructural en lugar de meramente de supervisión. La IA no puede salirse del guion porque el guion no admite interpretaciones.

# La arquitectura que ofrece ambos

El reto operativo al que se enfrentan la mayoría de las empresas no es que tengan demasiada IA. El problema es que tienen la IA operando en las partes equivocadas de su arquitectura de procesos.

El argumento económico a favor de la IA en las operaciones es real: reduce el costo del servicio, aumenta el rendimiento y ofrece experiencias de cliente más consistentes a gran escala. Estos son resultados que se pueden alcanzar. Pero requieren una visión clara de dónde la IA aporta valor y dónde introduce costo y riesgos sin un retorno correspondiente.

La arquitectura que ofrece estos resultados de forma consistente consta de dos capas que trabajan conjuntamente. La primera es una capa de orquestación determinista: un motor basado en reglas que gestiona el flujo de procesos, aplica la lógica de negocio, realiza un seguimiento del estado y enruta el trabajo en función de criterios definidos. Esta capa opera con un costo marginal por transacción prácticamente nulo. No razona. Se ejecuta. Es totalmente predecible, completamente auditable y no produce alucinaciones.

La segunda capa es la inferencia de IA, que se invoca precisamente en los puntos del proceso donde la lógica determinista resulta insuficiente. Análisis de la opinión de los clientes. Interpretación de documentos no estructurados. Generación de respuestas en lenguaje natural. Resumen de un caso complejo. Se trata de tareas en las que la naturaleza probabilística de la IA es una ventaja en lugar de un inconveniente, porque la tarea en sí no tiene una única respuesta correcta que las reglas puedan calcular.

Cuando estas dos capas están debidamente separadas, el impuesto de los tokens se reduce drásticamente. El gasto de tokens se concentra en tareas que realmente requieren inferencia. Todo lo demás se ejecuta de forma determinista. Al mismo tiempo, el riesgo de cumplimiento normativo disminuye porque la capa de orquestación aplica las reglas de negocio independientemente de lo que produzca la capa de IA. La IA opera dentro de limitaciones que no pueden eludirse mediante el razonamiento.

Actualmente, solo el 12% de los directores ejecutivos afirman que la IA ofrece ventajas cuantificables tanto en costos como en ingresos simultáneamente ( [PwC, enero de 2026](#) ). Las empresas que forman parte de ese 12% no están implementando más IA que sus competidores. Lo están implementando con mayor disciplina estructural.



# El control es la ventaja competitiva

Los responsables de operaciones que obtendrán un retorno de la inversión duradero gracias a la IA no son los que la implementaron más rápidamente ni los que otorgaron mayor autonomía a sus agentes. Son aquellos que mantuvieron el control de la arquitectura del proceso al tiempo que integraban la IA precisamente donde generaba valor.

Esto significa tratar la orquestación como un activo estratégico, no como una limitación heredada. Significa definir con claridad qué decisiones son deterministas y cuáles requieren inferencia, y construir una arquitectura que haga respetar ese límite. Esto significa medir el gasto en IA a nivel de transacción, no solo a nivel de plataforma, para que el impuesto de los tokens sea visible antes de que se convierta en un problema presupuestario.

La junta directiva tiene razón al afirmar que la IA puede reducir los costos de servicio y mejorar la experiencia del cliente. El mecanismo para lograrlo no consiste en que agentes autónomos descifren sus flujos de trabajo. Se trata de una arquitectura de procesos gobernada y orquestada donde la IA se aplica con precisión.

¿Está listo para tomar el control? Use la [calculadora de TCO](#) para cuantificar la diferencia entre su gasto actual en tokens y un enfoque de orquestación estructurada en sus flujos de trabajo de mayor volumen.





# Gracias

## Acerca de Pega

Pega proporciona la plataforma líder impulsada por IA para la transformación empresarial. Las organizaciones más influyentes del mundo confían en nuestra tecnología para reimaginar cómo se hace el trabajo automatizando los flujos de trabajo, personalizando la experiencia del cliente y modernizando los sistemas heredados. Desde 1983, nuestra arquitectura escalable y flexible ha impulsado la innovación continua, ayudando a los clientes a acelerar su camino hacia la empresa autónoma.

**[PEGA.COM/ES](https://www.pega.com/es)**