



El costo oculto de la IA no es el modelo — es cómo se gestiona

El valor de una IA predecible

Junio de 2026 | Don Schuerman, CTO y director de
marketing de Pega



La siguiente fase de la IA no va de posibilidades. Va de responsabilidad.

En los últimos dos años, la IA empresarial ha pasado de ser una curiosidad a una capacidad real. Lo que antes requería equipos especializados y largos ciclos de desarrollo, ahora se puede hacer en cuestión de días. Los equipos pueden generar código, poner en marcha agentes y automatizar flujos de trabajo más rápido que nunca.

Ese cambio ha modificado el debate a nivel ejecutivo.

Hace un año, los directivos se preguntaban: *¿Qué podría hacer la IA por nosotros?* Ahora la pregunta es más directa: *¿Cuánto nos está costando y qué valor estamos obteniendo realmente?*

A medida que las organizaciones van más allá de los proyectos piloto y comienzan a implementar la IA a gran escala en sus operaciones reales, las expectativas aumentan. La IA ya no se limita a los presupuestos de innovación. Se está convirtiendo en parte del funcionamiento del negocio. Y eso significa que tiene que superar un examen minucioso en lo que respecta a fiabilidad, gobernanza y costos.

En todos los sectores, existe una creciente conciencia de que el potencial de la IA es evidente. Su rentabilidad no lo es. Un estudio reciente de Gartner ¹ muestra que la mayoría de las iniciativas de IA todavía tienen dificultades para ofrecer un retorno significativo, y muchas organizaciones ven poco o ningún retorno de la inversión medible de la inversión inicial.

Eso no es un fallo de la IA. Es una señal de que estamos entrando en una nueva fase: una en la que el éxito dependerá menos de lo que puedas construir y más de la eficacia con la que se pueda operar a gran escala.

¹. Gartner research on AI value (gartner.com/en/articles/ai-value)

La IA es cada vez más fácil de desarrollar, pero más difícil de operar de manera eficiente

La primera oleada de adopción de la IA vino impulsada por el acceso. A medida que los modelos potentes se hicieron más accesibles, las organizaciones se apresuraron a experimentar. La innovación se aceleró. Pero a medida que esos experimentos pasan a la fase de producción, surge una realidad diferente.

Los costos que parecían manejables a escala piloto empiezan a aumentar. Los resultados varían en cuanto a calidad. Y en muchos casos, la relación entre la inversión en IA y el impacto en el negocio resulta difícil de medir. En el fondo de este reto se encuentra un problema estructural: la mayoría de los sistemas de IA actuales están diseñados para la **experimentación, no para un funcionamiento continuado.**

Muchos dependen en gran medida del *razonamiento en tiempo de ejecución*, donde el modelo evalúa cada interacción desde cero. Si bien esa flexibilidad puede ser útil, hay que pagar un precio. Cada interacción consume recursos computacionales. Cada aumento en la complejidad incrementa ese consumo. Y a medida que aumenta el uso, esos costos se acumulan.

Cada vez que la IA tiene que reevaluar una tarea, interpretar el contexto o decidir qué hacer a continuación, incurre en un costo adicional. A pequeña escala, esto es manejable. A escala empresarial, el problema se agrava porque el sistema no solo ejecuta tareas, sino que repite el mismo razonamiento una y otra vez. De hecho, muchas organizaciones están experimentando lo que equivale a una especie de "impuesto de los tokens", donde cada paso adicional de razonamiento conlleva un costo marginal que se acumula a medida que aumenta la escala.

Y a gran escala, eso empieza a cambiar por completo la ecuación de costos. En muchas implementaciones empresariales, el costo de la inferencia (la ejecución continua de modelos en producción) supera rápidamente el costo inicial de su creación, convirtiendo la IA de una inversión puntual en un costo operativo continuo.

El verdadero costo de la IA no reside en lo que creas. Es algo que hay que mantener en funcionamiento continuamente. Las organizaciones están pagando continuamente para "averiguar" cómo realizar tareas que podrían haberse estructurado con antelación. Por eso, las soluciones que funcionan bien en un programa piloto controlado a menudo tienen dificultades para escalar eficazmente en el mundo real. **La restricción no es el potencial. Es la eficiencia.**



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

El verdadero costo de la IA reside en cómo se realiza el trabajo.

Cuando los directivos se proponen gestionar el gasto en IA, su primera reacción suele ser centrarse en los modelos: comparar proveedores, capacidades y estructuras de precios.

Pero hay una pregunta más importante: **¿En qué parte de su operativa se utiliza la IA? ¿Y cómo se estructura ese trabajo?**

En muchas organizaciones, la IA ha crecido de forma orgánica, con diferentes equipos implementando distintos modelos, agentes y flujos de trabajo con poca coordinación. El resultado no es solo complejidad. Es duplicación. Las mismas decisiones se toman varias veces, en varios lugares, a menudo con una lógica diferente.

Las previsiones apuntan ahora a que las cargas de trabajo de IA empresarial impulsarán un aumento de más de 20 veces en el uso de tokens en los próximos años ³. A medida que el número de agentes aumenta en toda la empresa, ese crecimiento no solo incrementa el uso, sino también los costos. Como resultado, la IA a menudo se incorpora a los procesos existentes sin cambiar fundamentalmente la forma en que funcionan dichos procesos.

Esto conduce a una mejora gradual, pero no a una transformación, lo que crea dos problemas sistémicos:

1. Perpetúa la ineficiencia. Si el proceso subyacente está fragmentado o es excesivamente complejo, la introducción de la IA no elimina esa complejidad. Simplemente lo acelera.
2. Traslada demasiada responsabilidad al tiempo de ejecución. En lugar de definir de antemano cómo debe realizarse el trabajo, las organizaciones confían en la IA para determinar el siguiente paso de forma dinámica, interacción tras interacción.

Esa flexibilidad puede parecer beneficiosa. Pero introduce variabilidad, aumenta los costos y dificulta el control de los resultados. En última instancia, se trata de un problema de arquitectura:

- Los sistemas bien diseñados reducen el esfuerzo antes de que comience el trabajo.
- Los sistemas mal estructurados generan costos cada vez que se ejecutan.

Cuando la IA se aplica sin replantearse cómo fluye el trabajo entre sistemas y equipos, el costo se convierte en algo que aumenta con el uso en lugar de estabilizarse con la escala.

Por eso, muchas organizaciones ven cómo el gasto aumenta más rápido que los resultados.

³ Goldman Sachs, "Se prevé que los agentes de IA impulsen el flujo de caja de las empresas tecnológicas a medida que se dispara su uso"



Agregar IA a procesos defectuosos no los arregla, sino que agrava sus problemas.

Un patrón que se observa con frecuencia es lo que podría describirse como un enfoque de "añadido complementario" a la IA. Una empresa identifica un flujo de trabajo, como el servicio al cliente, la incorporación o la gestión de reclamos, e inserta una capa de IA para mejorarlo. La experiencia se vuelve más conversacional. El sistema se vuelve más adaptativo.

Pero el proceso subyacente permanece inalterado. Siguen involucrados varios sistemas. Las decisiones siguen estando fragmentadas. El trabajo aún se desarrolla a través de una serie de pasos inconexos. Para compensar, la IA asume una mayor responsabilidad. Interpreta, enruta y se adapta en tiempo real, intentando mantener el proceso unido eficazmente.

Ese enfoque puede mejorar la experiencia a corto plazo, pero no cambia la estructura de costos. De hecho, a menudo lo aumenta. Porque ahora el sistema realiza tanto el trabajo original como la interpretación continua necesaria para mantener la coherencia de dicho trabajo.

Cada vez que el sistema tiene que reevaluar o reconstruir el flujo de trabajo, consume recursos adicionales. Multiplique eso por miles o millones de interacciones, y rápidamente podrá visualizar el impacto. En este modelo, la IA se convierte en una capa de interpretación continua que se sitúa sobre la complejidad estructural.

Eso no es sostenible a gran escala. Si el objetivo es cambiar la ecuación de costos, el punto de partida debe ser el proceso en sí mismo.

El control de costos empieza por el diseño, no por la optimización en tiempo de ejecución.

Las organizaciones que están logrando avances significativos con la IA están abordando el problema de manera diferente. Están pasando a un modelo orquestado, donde la IA se coordina, estructura e integra en el diseño y la ejecución del trabajo.

En lugar de confiar en la IA para tomar decisiones en el momento, la están utilizando en una fase anterior, durante el diseño de cómo se debe realizar el trabajo. En esta etapa, la IA puede ser increíblemente potente. Puede ayudar a analizar los procesos existentes, identificar ineficiencias y proponer mejores maneras de estructurar el trabajo en todos los sistemas y equipos.

Una vez definida esa estructura, la ejecución se vuelve mucho más eficiente. En tiempo de ejecución, el sistema ya no necesita "averiguar" cada paso. Sigue una ruta definida, utilizando la IA de forma selectiva allí donde aporta mayor valor, como por ejemplo al interpretar la intención o al procesar datos de entrada no estructurados.

Este cambio, que pasa del *razonamiento en tiempo de ejecución* a la *inteligencia en tiempo de diseño*, modifica tanto el perfil de gasto como la coherencia de los resultados. Esto permite a las organizaciones:

- Reducir los cálculos innecesarios
- Eliminar la toma de decisiones repetida
- Asegurar que el trabajo se ejecute de forma repetible y controlada.

En este modelo, la IA pasa a formar parte de un sistema de ejecución, no a sustituirlo.

La previsibilidad hace posible una IA a escala empresarial.

Para que la IA se convierta en un elemento fundamental de la empresa, debe operar con un nivel de disciplina que esté a la altura de los demás sistemas centrales.

Y eso comienza con la previsibilidad. Los directivos necesitan tener confianza en tres áreas:

1. **Cuánto va a costar**
2. **Qué resultado producirá**
3. **Cómo se comportará en condiciones reales**

Sin eso, la escalabilidad supone un riesgo.

Un estudio reciente de EY muestra que más de la mitad de las iniciativas de IA operan sin una supervisión formal, lo que expone a las organizaciones a sobrecoste, deficiencias en el cumplimiento normativo y resultados inconsistentes a medida que se amplía su implementación.

Aquí es donde comienza a afianzarse un enfoque más estructurado de la IA. Cuando la IA opera dentro de flujos de trabajo claramente definidos, respaldados por la orquestación y la gobernanza, sucede algo importante:

- Los resultados se vuelven más consistentes.
- Resulta más fácil prever el gasto, que además están cada vez más vinculados a decisiones, transacciones y resultados de negocio reales.
- Las operaciones se vuelven más fáciles de gestionar.

Y, lo que es más importante, esta previsibilidad no solo se aplica al uso de tokens. Abarca el costo total de la propiedad: desde las operaciones, la gobernanza, y el cumplimiento normativo hasta el esfuerzo continuo necesario para que los sistemas evolucionen con el tiempo. Aquí es donde el costo total de propiedad se vuelve más claro, extendiéndose más allá de los tokens para incluir el gasto de operar, gobernar, mantener y escalar la IA a lo largo del tiempo.

En lugar de tratar la IA como una capacidad sin límites, las organizaciones comienzan a considerarla como una parte controlada y medible de la forma en que se realiza el trabajo. Ese cambio les permite pasar de hacer un seguimiento del consumo a hacer un seguimiento de los resultados. Y ese es un paso fundamental para la IA sea económicamente viable.



⁴ EY, "La adopción de la IA autónoma se dispara mientras la supervisión se queda atrás" (2026)

Las organizaciones que tengan éxito cambiarán su forma de pensar sobre la IA.

Muchas organizaciones siguen operando bajo una falsa disyuntiva: avanzar rápidamente con la IA o mantener el control. En realidad, la limitación no es la velocidad, sino la arquitectura.

Las empresas que comprendan el verdadero valor de la IA no serán simplemente aquellas que la adopten rápidamente.

Las que acierten replantearán cómo funciona su negocio. En términos prácticos, eso significa centrarse en un conjunto diferente de prioridades:

Cambiar el enfoque de las herramientas a los sistemas	Rediseñar antes de automatizar
La cuestión clave no es qué puede hacer la IA de forma aislada, sino cómo se integra en el flujo de trabajo de toda la empresa.	Aplicar IA a un proceso defectuoso rara vez lo soluciona. Reimaginarlo sienta las bases para lograr un impacto significativo.
Medir los resultados, no la actividad	Tratar la IA como infraestructura operativa
El éxito debe estar vinculado a los resultados de negocio (trabajo completado, experiencias mejoradas, reducción de la fricción), no al uso en sí mismo.	A medida que la IA se integra cada vez más en los procesos clave, debe cumplir con los mismos estándares de fiabilidad y gobernanza que cualquier otro sistema empresarial.

No se trata tanto de adoptar nuevas tecnologías, sino de aplicar disciplina a la forma en que se utilizan.

De la experimentación a la ejecución

Nos encontramos en un punto de inflexión en la IA empresarial. La fase inicial demostró lo que era posible. La siguiente fase definirá qué es práctico y a qué escala.

A medida que las organizaciones amplían el uso de la IA, el reto ya no reside en generar información valiosa o automatizar tareas individuales. Se trata de operar esas capacidades de una manera estable, eficiente y alineada con los resultados del negocio . Esa es la diferencia entre el éxito aislado y el impacto en toda la empresa.

Las organizaciones que logren cerrar esa brecha podrán moverse con mayor rapidez, operar de manera más eficiente y adaptarse con mayor eficacia al cambio.

Pero no lo conseguirán solamente implementando más IA.

Lo lograrán sentando las bases adecuadas sobre cómo opera la IA dentro de su negocio.

Conclusiones

La IA ya está transformando lo que las organizaciones pueden hacer.

Ahora la atención se centra en cómo se comporta en condiciones reales a gran escala, en sistemas complejos y dentro de las limitaciones de gasto y control.

El mayor costo oculto en la IA no está ligado al modelo en sí. Proviene de cómo se estructura, se ejecuta y se repite el trabajo a lo largo del tiempo.

Las organizaciones que resuelvan ese problema no solo reducirán los costos. Construirán un sistema con el que realmente puedan gestionar su negocio .

Y en la próxima era de la IA empresarial, esas son las organizaciones que toman la delantera, porque han creado una IA que pueden controlar, medir y operar como parte del negocio.



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)