



I costi nascosti dell'IA non hanno a che fare con il modello, ma con le modalità di esecuzione

Il caso dell'IA prevedibile

Giugno 2026 | Don Schuerman, CTO e Head of Marketing, Pega



La nuova fase dell'IA non è una questione di possibilità, ma di responsabilità

Negli ultimi due anni, l'IA di livello enterprise è passata dalla curiosità all'analisi delle capacità. Ciò che un tempo richiedeva team specializzati e cicli di sviluppo lunghi ora può essere completato in pochi giorni. I team possono generare codice, creare nuovi agenti e automatizzare i flussi di lavoro più velocemente che mai.

Questa trasformazione ha modificato le conversazioni nei vertici aziendali.

Un anno fa i leader si chiedevano: *Che cosa può fare l'IA per noi?*

Oggi il dubbio è più diretto ed è: *Quanto ci costa e quali vantaggi stiamo realmente registrando?*

Via via che le organizzazioni superano le versioni pilota e iniziano a integrare l'IA nelle operazioni reali, le aspettative aumentano. L'IA non è più confinata ai budget per l'innovazione. Sta diventando parte integrante dei processi operativi aziendali. Questo significa che deve superare esami rigorosi in termini di affidabilità, governance e gestione dei costi.

In tutti i settori emerge una consapevolezza crescente: la promessa dell'IA è evidente. I costi che la riguardano, un po' meno. Secondo un recente studio¹ di Gartner, la maggior parte delle iniziative IA stentano ancora a offrire un ritorno significativo e molte organizzazioni registrano un ROI ridotto o persino non misurabile rispetto all'investimento iniziale.

Non è un fallimento dell'IA. È il segnale che stiamo entrando in una nuova fase, in cui il successo dipenderà meno da quello che è possibile implementare e più da quanto efficacemente sarà possibile utilizzare l'IA su vasta scala.

¹ Studio di Gartner sul valore dell'IA (gartner.com/en/articles/ai-value)

L'IA diventa sempre più facile da sviluppare, ma più complessa da utilizzare in modo efficiente

La prima ondata di adozioni dell'IA è stata favorita dall'accesso. Via via che i modelli potenti diventavano disponibili su ampia scala, le organizzazioni avevano fretta di sperimentare. L'innovazione ha così accelerato. Ma dopo gli esperimenti, in fase di produzione sta emergendo una realtà diversa.

I costi, che in fase pilota sembravano gestibili, hanno iniziato a lievitare. La qualità dei risultati non è uniforme e in molti casi, il rapporto tra l'investimento nell'IA e l'impatto sul business diventa difficile da misurare. Alla base di questa sfida c'è un problema strutturale: oggi la maggior parte dei sistemi IA è progettata per la **sperimentazione, non per un'operatività costante nel tempo.**



Molti si affidano eccessivamente al *ragionamento in fase di esecuzione*, costringendo il modello a valutare ogni singola interazione partendo da zero. Sebbene questa flessibilità sia utile, purtroppo ha un prezzo. Ogni interazione consuma potenza di elaborazione. Ogni incremento della complessità aumenta tale consumo. E via via che l'uso si diffonde, questi costi si sommano.

Ogni volta che l'IA rivaluta un'attività, interpreta il contesto o decide cosa fare i costi salgono. Su scala ridotta, tutto ciò è gestibile, ma su scala aziendale, i costi si accumulano perché il sistema non si limita a eseguire un lavoro, ma piuttosto ripete continuamente lo stesso identico processo logico da capo. Di fatto, molte aziende si trovano oggi a pagare una sorta di "tassa sui token", in cui ogni passaggio logico aggiuntivo comporta un costo marginale che si moltiplica su ampia scala.

E su larga scala, ciò finisce per stravolgere completamente l'equazione dei costi. In molte implementazioni aziendali, il costo dell'inferenza, ovvero l'esecuzione continua dei modelli in produzione, supera rapidamente il costo iniziale di sviluppo, trasformando l'IA da un investimento una tantum a una spesa operativa ricorrente.

Il vero costo dell'IA non sta in ciò che crei, ma in ciò che devi mantenere costantemente in funzione. Le aziende pagano continuamente per far elaborare all'IA attività che avrebbero potuto essere strutturate in anticipo. Ecco perché soluzioni che offrono ottime prestazioni in un progetto pilota controllato spesso faticano ad adattarsi efficacemente al mondo reale. **Il vincolo non risiede nelle potenzialità, ma nell'efficienza.**

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

Il vero costo dell'IA sta nelle modalità di completamento del lavoro

Quando i leader cercano di gestire la spesa per l'IA, l'istinto è spesso quello di concentrarsi sui modelli, confrontando fornitori, funzionalità e strutture tariffarie.

Ma la domanda più importante da porsi è ben più importante: **in quale area dei processi operativi l'IA sta svolgendo il lavoro e come è strutturato quel lavoro?**

In molte organizzazioni l'IA è cresciuta in modo organico, con diversi team responsabili di sviluppare diversi modelli, agenti e flussi di lavoro con poco coordinamento tra loro. Ciò genera non solo complessità, ma anche ridondanza. Le stesse decisioni vengono prese tante volte, in sedi diverse e spesso con logiche differenti.

Attualmente le previsioni indicano che i carichi di lavoro IA aziendali porteranno a un aumento di oltre 20 volte dell'uso dei token nei prossimi anni³. Con la diffusione degli agenti in tutta l'azienda, tale crescita non si limita a espandere l'utilizzo: fa anche lievitare i costi. Come risultato, l'IA viene frequentemente sovrapposta ai processi attuali senza trasformare in modo sostanziale la loro modalità di funzionamento.

Questo porta a un miglioramento incrementale, ma non a una trasformazione, il che crea due problemi sistemici:

1. Le inefficienze rimangono inalterate. Se il processo alla base è frammentato o eccessivamente complesso, l'introduzione dell'IA non elimina tale complessità, semplicemente le velocizza.
2. Viene dato troppo peso alla fase di esecuzione. Invece di definire in anticipo le modalità di esecuzione del lavoro, le organizzazioni demandano all'IA la decisione di determinare il passaggio successivo in modo dinamico, per ogni interazione.

Tale flessibilità può sembrare un vantaggio, ma introduce variabilità, aumenta i costi e rende i risultati più difficili da controllare. Questo è fondamentalmente un problema legato all'architettura:

- I sistemi ben progettati riducono l'impegno richiesto prima dell'avvio dell'attività.
- I sistemi strutturati male, invece, implicano dei costi a ogni esecuzione.

Se si applica l'IA senza ripensare i flussi di lavoro tra sistemi e team, il costo diventa una variabile che cresce con l'utilizzo anziché stabilizzarsi su larga scala.

Ecco perché alcune organizzazioni registrano l'aumento dei costi ben prima dei risultati.

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



L'implementazione dell'IA in processi che non funzionano non risolverà il problema, ma ne amplierà la portata

Di frequente si adotta quello che può essere definito un approccio all'IA preconfezionato. Un'azienda identifica un flusso di lavoro, ad esempio il servizio clienti, l'onboarding, l'apertura di sinistri e applica l'IA per migliorarlo. L'esperienza diventa più conversazionale, il sistema più adattabile,

ma il processo alla base resta invariato. Sono comunque coinvolti più sistemi e le decisioni rimangono frammentarie. Il lavoro continua a procedere tramite una serie di passaggi scollegati. Per compensare, l'IA si assume maggiori responsabilità. Interpreta, indirizza e si adatta in tempo reale, tentando di fatto di far funzionare tutto.

Un approccio del genere può migliorare l'esperienza sul breve termine, ma non cambia la struttura dei costi. Di fatto, spesso i costi aumentano perché il sistema esegue sia il lavoro originale sia la continua interpretazione richiesta per tenere tutto insieme.

Ogni volta che il sistema deve rivalutare o ricreare il flusso di lavoro, consuma risorse aggiuntive. Moltiplicando questo effetto su migliaia o milioni di interazioni, l'impatto diventa subito evidente. In questo modello, l'IA diventa uno strato di interpretazione continua che si sovrappone alla complessità strutturale.

Non è sostenibile su ampia scala. Se l'obiettivo è cambiare l'equazione dei costi, il punto di partenza deve essere il processo stesso.

Il controllo dei costi nasce dalla progettazione, non dall'ottimizzazione dell'esecuzione

Le aziende che stanno ottenendo progressi significativi con l'IA stanno affrontando il problema in modo diverso. Stanno passando a un modello orchestrato, in cui l'IA è coordinata, strutturata e integrata nel modo in cui il lavoro viene progettato ed eseguito.

Invece che affidarsi all'IA per le decisioni in tempo reale, la usano prima, in fase di progettazione del flusso di lavoro. In questa fase l'IA ha incredibili potenzialità. Può contribuire ad analizzare processi esistenti, identificare inefficienze e proporre modi migliori per strutturare il lavoro tra sistemi e team.

Una volta definita la struttura, l'esecuzione diventa molto più efficiente. Durante l'esecuzione, infatti, il sistema non ha più bisogno di capire ogni passaggio. Segue, piuttosto, un percorso definito, e usa l'IA in modo selettivo dove ha più valore, ad esempio per l'interpretazione dell'intento o la gestione degli input non strutturati.

Questo passaggio dal *ragionamento in fase di esecuzione* all'*intelligence in fase di progettazione* cambia il profilo dei costi e la coerenza dei risultati. Permette alle organizzazioni di:

- Ridurre le elaborazioni non necessarie
- Eliminare i processi decisionali ripetuti
- Assicurare l'esecuzione del lavoro in modo ripetibile e regolamentato

L'IA in questo modello diventa parte di un sistema di esecuzione, non lo sostituisce.

La prevedibilità permette alle imprese di adattare l'IA su vasta scala

Perché l'IA diventi fondamentale nell'impresa, deve operare con un livello di disciplina corrispondente agli altri sistemi strategici.

Ciò inizia con la prevedibilità. I leader devono avere sicurezze in merito a tre aspetti:

1. **Il costo**
2. **Il risultato che produrrà**
3. **Come si comporterà in situazioni reali**

Senza questi punti fermi, la scalabilità diventa rischiosa

Una recente ricerca della società di consulenza EY mostra che oltre la metà delle iniziative di IA opera senza una supervisione formale, esponendo le aziende a sforamenti di costi, lacune di conformità e risultati incoerenti man mano che l'adozione si diffonde.



È qui che inizia ad affermarsi un approccio più strutturato all'IA. Quando l'IA opera entro flussi di lavoro chiaramente definiti, supportati da orchestrazione e governance, succede qualcosa di importante:

- I risultati diventano più coerenti.
- I costi diventano più semplici da prevedere e sempre più collegati a decisioni, transazioni e risultati aziendali reali.
- Le operazioni sono più facili da gestire.

E, aspetto più importante, tale prevedibilità non si applica solo all'uso dei token, ma si estende al costo totale di proprietà: operazioni, governance, conformità e lo sforzo continuo richiesto per far evolvere i sistemi nel tempo. È qui che il costo totale di proprietà diventa più chiaro, estendendosi oltre i semplici token per includere i costi di gestione, governance, manutenzione e scalabilità dell'IA nel tempo.

Invece di considerare l'IA come una risorsa dalle possibilità illimitate, le aziende iniziano a trattarla come una componente controllata e misurabile del proprio modello operativo. In questo modo, finalmente, possono smettere di tenere traccia dei consumi e concentrarsi sulla valutazione dei risultati. Tale passaggio rappresenta una fase cruciale per garantire la sostenibilità finanziaria dei modelli di IA.

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

Le aziende che avranno successo cambieranno il loro modo di concepire l'IA

Molte aziende si muovono ancora all'interno di una falsa alternativa: correre veloci con l'IA o mantenerne il controllo. In realtà il limite non è la velocità, ma l'architettura.

Le aziende che sapranno cogliere appieno il valore dell'IA non saranno semplicemente quelle che la adotteranno più rapidamente.

Le aziende che faranno la scelta giusta ripenseranno il modo stesso in cui opera il loro business. Nella pratica, questo significa concentrarsi su altre priorità:

Dagli strumenti ai sistemi

La domanda chiave non è cosa può fare l'IA da sola, ma come si integra nel flusso di lavoro in un'impresa.

Riprogettazione prima, automazione poi

Applicare l'IA a un processo che non funziona, raramente rappresenta la soluzione. Ripensare prima il processo, invece, crea la base per favorire un impatto significativo.

Misurazione dei risultati, non dell'attività

Il successo va collegato ai risultati del business, ad esempio lavoro portato a termine, esperienze migliorate, complessità ridotte, non al mero utilizzo.

L'IA come infrastruttura operativa

Man mano che l'IA si integra nei processi strategici, deve soddisfare gli stessi standard di affidabilità e governance di qualsiasi altro sistema aziendale.

L'attenzione si sposta dall'introduzione di nuove soluzioni tecnologiche all'applicazione di una ferrea disciplina nelle relative modalità di impiego.

Dalla sperimentazione alla realizzazione

Siamo a un punto di svolta per l'IA di livello enterprise. La prima fase ha dimostrato cosa è possibile, la prossima definirà cosa è fattibile nella pratica e cosa è scalabile.

Con la progressiva espansione dell'IA all'interno delle organizzazioni, la priorità non è più rappresentata dall'estrazione di approfondimenti o dall'automazione di singole attività, ma dall'utilizzo di tali capacità in modo stabile, efficiente e allineato ai risultati aziendali.

In ciò risiede la differenza tra un caso di successo isolato e un impatto strutturale a livello aziendale.

Le organizzazioni che colmano tale divario si muoveranno più velocemente, operando in modo più efficiente e adattandosi più efficacemente al cambiamento.

Ma non otterranno più risultati semplicemente implementando più IA.

Dovranno piuttosto sviluppare le basi giuste per il modo in cui l'IA opera all'interno del business.

Conclusione

L'IA sta già ridefinendo ciò che le aziende possono fare.

Ora l'attenzione si sposta su come opera in condizioni reali su larga scala, attraverso sistemi complessi e nel rispetto dei vincoli di costo e controllo.

Il più importante costo nascosto dell'IA non è legato al modello, ma causato da come il lavoro viene strutturato, eseguito e ripetuto nel tempo.

Le organizzazioni in grado di superare questa criticità non otterranno soltanto una contrazione dei costi, ma implementeranno un sistema su cui potranno realmente basare la propria attività.

E nella futura era dell'IA di livello enterprise, saranno proprio queste realtà a portarsi avanti, perché avranno costruito un'IA che possono controllare, misurare e gestire come parte integrante del business.



Informazioni su Pega

Pega fornisce la principale piattaforma basata sull'IA per la trasformazione aziendale. Le organizzazioni più influenti al mondo si affidano alla nostra tecnologia per reinventare il modo in cui viene svolto il lavoro automatizzandone i flussi, personalizzando le customer experience e modernizzando i sistemi legacy. Dal 1983, la nostra architettura scalabile e flessibile promuove l'innovazione continua, aiutando i clienti ad accelerare il loro percorso verso l'impresa autonoma.

[PEGA.COM/IT](https://www.pega.com/it)