

Guida alla progettazione agentica per i CIO: come sviluppare app aziendali alla velocità dell'IA senza compromettere l'affidabilità

Guida per COO, CXO e
leader operativi



La pressione è reale, tanto quanto il problema.

I consigli di amministrazione non si chiedono più se implementare l'IA, ma quando, quanto rapidamente e qual è il relativo ritorno. Per i COO e i responsabili operativi, tale pressione si è tradotta in un'ondata di progetti pilota con agenti IA autonomi: implementazioni in cui l'IA gestisce le interazioni con i clienti, instrada le pratiche, verifica le policy e amministra i flussi di lavoro con un intervento umano minimo.

I primi risultati erano incoraggianti: tempistiche di risoluzione ridotte, meno personale necessario, meno escalation manuali. Ma quando i progetti pilota sono passati in produzione, è emerso un quadro differente: i costi non corrispondevano a quanto previsto dai modelli, i risultati non erano coerenti con quanto le demo lasciavano intendere e il controllo operativo che i leader speravano di ottenere si è indebolito.

Il 56% dei CEO non riferisce al momento alcun vantaggio finanziario derivante dagli investimenti nell'IA ([29th Global CEO Survey di PwC, gennaio 2026](#)). Non si tratta di un problema tecnico, ma di un problema di architettura. La domanda non è se l'IA possa offrire l'efficienza operativa, ma se il modo in cui la maggior parte delle imprese la sta implementato sia strutturalmente in grado di farlo.

La risposta, nella maggior parte dei casi, è no.



I costi nascosti del ragionamento

L'IA ha sostanzialmente spostato il peso economico dell'automazione aziendale. I token, l'unità di misura del ragionamento dei sistemi IA, sono diventati la voce consequenziale del budget destinato alla trasformazione di un'impresa. E per la maggior parte delle aziende, quella voce di spesa si sta comportando in modi che nessuno aveva previsto.

Quando i leader aziendali hanno approvato l'implementazione degli agenti IA, il modello di costo era semplice: sostituire il lavoro manuale con il ragionamento automatico, ridurre i costi del personale e incassare il margine. Ciò di cui quel modello non ha tenuto conto è proprio l'imprevedibilità del consumo di token su larga scala aziendale. Due fattori alimentano questa imprevedibilità e, insieme, stanno compromettendo il ROI che era stato promesso ai consigli di amministrazione.

Il primo è la natura cumulativa dell'utilizzo dei token da parte degli agenti. Gli agenti IA che eseguono attività a lungo termine non si limitano a elaborare un pensiero una volta sola per poi passare oltre. Ripetono il processo logico a ogni passaggio, verificando il contesto, valutando le opzioni e generando risultati in modo continuo. La spesa in token necessaria per far funzionare questi sistemi si somma a quella già sostenuta inizialmente per crearli e svilupparli. Per le operazioni di un'impresa di medie dimensioni, questo fenomeno accade più velocemente rispetto a quanto avesse previsto la maggior parte dei modelli finanziari.

Noi chiamiamo questa dinamica la “tassa sui token”.

Il secondo fattore è determinato dall'instabilità dei prezzi. I primi fornitori di IA impostavano modelli di prezzo agevolati con consumi illimitati così le imprese avevano a disposizione una struttura di costi prevedibile. Quell'epoca è finita. Poiché l'economia dell'IA si trova ad affrontare una carenza di risorse sul fronte di chip, energia e capacità dei data center e data la tendenza dei fornitori di IA verso l'IPO, le aziende hanno dovuto assorbire continui cambi di tariffe che stanno trasformando il modello con canone fisso in una fatturazione basata sul consumo reale. Ora il costo dello stesso carico di lavoro non è più fisso. Sam Altman ha riconosciuto pubblicamente a giugno 2026 che i costi dei token di IA sono diventati "un enorme problema" persino per i fornitori stessi ([Axios, giugno 2026](#)).

Le aziende stanno esaurendo rapidamente i fondi allocati per la trasformazione in assenza di una concreta prospettiva di ritorno sull'investimento. Secondo Gartner, oltre il 40% dei progetti di IA agentica verrà annullato entro la fine del 2027, a causa dell'aumento e dell'imprevedibilità dei costi operativi, considerati uno dei fattori principali ([Gartner, giugno 2025](#)).

Ora i dirigenti non si chiedono più se utilizzare l'IA, ma come raggiungere un ROI positivo quando la struttura dei costi sottostante continua a cambiare. La risposta richiede un cambio di prospettiva: passare dal pagamento dei token al pagamento dei risultati. Ciò significa creare un'architettura che concentri il consumo di token dell'IA solo sulle operazioni che necessitano effettivamente di ragionamento, demandando le restanti attività a una struttura di costo fissa, prevedibile e priva di costi cumulativi.

Quando l'IA agisce in modo imprevisto

L'imprevedibilità dei costi è solo un aspetto del problema. L'altro aspetto è ciò che accade quando agli agenti IA viene concessa l'autorità su decisioni che, fin dal principio, non avrebbero mai dovuto prendere

I grandi modelli linguistici sono addestrati per essere utili e fluidi, non per rispettare le policy aziendali sulla conformità. Non comprendono SLA, obblighi normativi, regole di escalation o l'impegno specifico assunto dal brand nei confronti dei clienti. Quando agenti autonomi operano in flussi di lavoro rivolti ai clienti, senza limiti invalicabili, il divario tra ciò che l'IA ritiene sia una risposta ragionevole e cosa approvarebbe il team di conformità può essere enorme.

Non è una questione di tolleranza, ma di architettura.

Le decisioni che pongono i principali rischi legali e per la reputazione (determinazione dell'idoneità, risoluzione delle controversie, eccezioni alle policy, comunicazioni regolamentate) sono quelle che prevedono una risposta corretta ben definita. Non sono ambigue e non traggono alcun vantaggio dall'inferenza probabilistica. L'instradamento di una pratica in base allo stato dell'account, la conferma del fatto che un assegno o un sinistro soddisfi un elenco di controllo documentato e l'applicazione di un'eccezione alle policy secondo criteri definiti sono operazioni deterministiche. Concedere all'agente IA l'autorità di ragionare su tali aspetti non aggiunge valore, ma introduce variabilità nei processi che invece richiedono coerenza e responsabilità nelle interazioni che necessitano di precisione.

Eventuali allucinazioni dell'IA non sono casi limite in ambienti di produzione su vasta scala. Rappresentano piuttosto il risultato prevedibile di qualsiasi sistema che genera risposte mediante inferenza probabilistica e non in base a regole deterministiche. Quando la posta in gioco è bassa, un'allucinazione è solo un fastidio. Ma per la contestazione di un cliente, in un flusso di lavoro per sinistri, in un'interazione relativa a servizi finanziari o nelle comunicazioni che riguardano un settore regolamentato, è una responsabilità.

Le principali previsioni di Gartner per il 2026 includono una stima secondo cui le cause legali globali per incidenti di tipo "death by AI" supereranno quota 2.000, a causa di una governance e di controlli del rischio insufficienti sui risultati generati dall'intelligenza artificiale ([Gartner, ottobre 2025](#)). Non si tratta di rischi astratti per le future implementazioni, ma di un'esposizione attiva per qualsiasi organizzazione che oggi concede ad agenti IA autonomi l'autorità sui flussi di lavoro rivolti ai clienti o regolamentati.

La risposta più produttiva è di tipo strutturale: le regole aziendali stabiliscono ciò che gli agenti IA possono e non possono fare. L'IA non dovrebbe mai rappresentare il sistema di riferimento per la logica di conformità aziendale, ma funzionare entro i suoi confini. L'azione corretta per le decisioni nette è determinata dalla logica aziendale, applicata a livello di orchestrazione, non delegata a un modello che potrebbe generare una risposta diversa in un altro momento. Quando tale separazione è attiva, la conformità diventa strutturale anziché supervisionale. L'IA non può deviare dalle linee guida perché le linee guida non possono essere messe in discussione.

L'architettura che garantisce entrambe

La sfida operativa che la maggior parte delle aziende si trova ad affrontare non è quella di avere troppa IA, ma che l'IA sia stata introdotta nelle parti sbagliate dell'architettura dei processi.

Le ragioni economiche a favore dell'IA nelle attività operative sono concrete: riduzione del costo del servizio, aumento della produttività e offerta di customer experience più coerenti su ampia scala. Si tratta di risultati raggiungibili, ma richiedono di capire chiaramente dove l'IA aggiunge valore e dove invece introduce rischi e costi senza offrire un ritorno.

L'architettura che offre tali risultati in modo coerente dispone di due livelli che interagiscono. Il primo è un livello deterministico di orchestrazione, ovvero un motore basato su

regole in grado di gestire il flusso del processo, di applicare la logica aziendale, di monitorare lo stato e di instradare il lavoro in base a criteri definiti. Questo livello opera a un costo marginale per transazione prossimo allo zero. Non deve ragionare, ma solo eseguire. È totalmente prevedibile, verificabile e non produce allucinazioni.

Il secondo livello è l'inferenza dell'IA, invocata precisamente nei punti del processo in cui la logica deterministica è insufficiente: analisi del sentiment del cliente, interpretazione dei documenti non strutturati, generazione di risposte in linguaggio naturale, sintesi di pratiche complesse. Questi sono compiti in cui la natura probabilistica dell'IA costituisce un vantaggio anziché un limite, poiché il compito stesso non presenta un'unica risposta corretta che le regole potrebbero elaborare.

Quando questi due livelli sono opportunamente separati, la tassa sui token si riduce in modo significativo. Il costo in token si concentra sulle attività che richiedono veramente l'inferenza. Tutto il resto viene eseguito in modo deterministico. Al contempo, il rischio per la conformità si riduce perché il livello di orchestrazione applica le regole aziendali a prescindere da quanto produce il livello dell'inferenza dell'IA. L'IA opera entro limiti che non possono essere valicati, nonostante il ragionamento.

Solo il 12% dei CEO al momento segnala che l'IA offre simultaneamente risultati misurabili circa costi e ricavi ([PwC, gennaio 2026](#)). Le imprese che rientrano in questa percentuale non stanno implementando più IA di altre dello stesso settore, ma semplicemente stanno applicando una disciplina più strutturata.



Il controllo come reale vantaggio competitivo

I leader operativi che otterranno un ROI duraturo dall'IA non sono quelli che hanno implementato più velocemente o che hanno concesso la massima autonomia ai propri agenti, ma quelli che hanno mantenuto il controllo dell'architettura del processo, integrando l'IA esattamente dove crea valore.

Ciò significa considerare l'orchestrazione come una risorsa strategica, non come un vincolo ereditato. Significa definire chiaramente quali decisioni siano deterministiche e quali richiedano inferenza e costruire un'architettura che faccia rispettare tale confine. Significa misurare la spesa per l'IA a livello di transazione, non solo a livello di piattaforma, in modo che la tassa sui token sia visibile prima di diventare un problema di budget.

Il consiglio di amministrazione ha ragione sul fatto che l'IA possa ridurre il costo del servizio e migliorare la customer experience. Ma il meccanismo per raggiungere questi obiettivi non è rappresentato da agenti che ragionano sui flussi di lavoro delle aziende: occorre un'architettura orchestrata e regolamentata dei processi, in cui l'IA viene applicata con precisione.

Vuoi prendere il controllo? Usa lo [strumento di calcolo del TCO](#) per quantificare la differenza tra l'attuale spesa in token e un approccio strutturato all'orchestrazione nei flussi di lavoro a maggiore volume.





Ringraziamento

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)