This document was translated from English by machine and may contain mistakes. The English original is at https://www.novaracg.com/2026/04/18/when-control-systems-inherit-the-same-uncertainty-they-are-meant-to-govern/.

Quando i sistemi di controllo ereditano la stessa incertezza che dovrebbero governare

Heather Grizzle · April 18, 2026

When AI guardrails are themselves probabilistic models, governance inherits the uncertainty it is meant to control. What real control requires.

La governance dell'IA viene attualmente descritta e venduta in base a ciò che è più facile da visualizzare, piuttosto che a ciò che è più importante da analizzare.

La conversazione è dominata da artefatti visibili: dashboard di policy, diagrammi di controllo stratificati, guardrail etichettati e flussi di esecuzione passo-passo che suggeriscono che una richiesta venga valutata rigorosamente prima che avvenga qualsiasi azione. Queste rappresentazioni creano l'impressione di controllo perché assomigliano alla supervisione tradizionale dei sistemi. Tuttavia, oscurano la questione più rilevante di quale tipo di meccanismo stia effettivamente svolgendo la valutazione all'interno di questi livelli.

Nella maggior parte delle architetture attuali, un sistema di IA genera un'azione proposta, e tale proposta viene instradata attraverso una sequenza di controlli che culminano in un altro modello che determina se l'azione debba procedere. Questo livello di valutazione può essere descritto in modi diversi a seconda del fornitore o del framework. Può essere presentato come un motore di policy, un meccanismo di valutazione del rischio o un sistema di classificazione della confidenza integrato da regole. Nonostante queste variazioni terminologiche, la struttura sottostante rimane coerente. A un modello probabilistico viene affidato il compito di valutare l'appropriatezza di un output generato da un altro modello probabilistico.

Questa scelta progettuale ha implicazioni significative se esaminata attraverso una lente di rischio e affidabilità. I sistemi probabilistici, per definizione, non producono

risultati deterministici e ripetibili in condizioni identiche. I loro output sono plasmati da schemi statistici derivati dai dati di addestramento, e le loro valutazioni sono analogamente influenzate da inferenze probabilistiche piuttosto che da criteri fissi. Quando un sistema di questo tipo viene utilizzato per governare un altro sistema dello stesso tipo, la funzione di governo non introduce una forma di affidabilità fondamentalmente diversa. Al contrario, replica lo stesso profilo di incertezza a un livello di astrazione superiore.

Dal punto di vista della governance, questo crea una struttura circolare. Il sistema destinato a imporre il controllo è soggetto alla stessa variabilità, instabilità nei casi limite e sensibilità al contesto del sistema che supervisiona. Di conseguenza, la presenza di livelli, controlli o wrapper aggiuntivi non si traduce necessariamente in una riduzione significativa del rischio. Può aumentare la complessità e può creare l'apparenza di diligenza, ma non modifica il comportamento sottostante del sistema in modo tale da produrre un confine di controllo distinto.

> Introduce ricorsione, non controllo.

Questa distinzione diventa più chiara se confrontata con il modo in cui la governance funziona in altri ambiti ad alto rischio. Nei sistemi finanziari, ad esempio, i processi di approvazione delle transazioni si basano su regole predefinite, registri di audit e criteri redatti da esseri umani che possono essere tracciati e ricostruiti. Nei contesti di ingegneria critica per la sicurezza, i sistemi di controllo sono progettati per operare con proprietà di affidabilità diverse rispetto ai sistemi che monitorano. Il meccanismo di governo è intenzionalmente separato sia nella funzione che nel comportamento, in modo da poter fornire un punto di riferimento stabile quando si verificano incertezza o guasti.

I pattern attuali di governance dell'IA spesso non stabiliscono questa separazione. Al contrario, sovrappongono un giudizio probabilistico a una generazione probabilistica e trattano la struttura risultante come un sistema di controllo. Questo approccio presuppone che l'aggiunta di ulteriori passaggi di valutazione all'interno della stessa classe di affidabilità produca risultati di governance. In pratica, introduce ricorsione anziché controllo, perché ogni livello dipende dallo stesso tipo di processo di ragionamento ed è vulnerabile agli stessi tipi di guasto.

La conseguenza non è semplicemente teorica. Quando un sistema produce un risultato che causa un danno, la domanda che segue non è se nella decisione fossero coinvol-

ti più livelli. La domanda è se uno qualsiasi di quei livelli abbia funzionato come un vero controllo, ossia se abbia operato con un livello diverso di determinismo, tracciabilità e responsabilità rispetto al sistema che governava. Se ogni livello condivide lo stesso fondamento probabilistico, allora il sistema manca di un punto chiaro in cui la responsabilità possa essere ancorata o le decisioni possano essere ricostruite in modo affidabile.

> **IL TEST DECISIVO**
>
> La governance è definita dall'introduzione di un meccanismo che opera con un profilo di affidabilità diverso rispetto al sistema governato.

La posizione che deriva da questa analisi è diretta. La governance non è definita dal numero di controlli presenti in un sistema, né dalla complessità visiva della sua architettura. La governance è definita dall'introduzione di un meccanismo che opera con un profilo di affidabilità diverso rispetto al sistema governato. Senza questa distinzione, i livelli aggiuntivi non costituiscono controllo. Estendono la stessa incertezza a più componenti.

Quando un livello di controllo si comporta allo stesso modo del sistema che dovrebbe regolare, l'architettura non raggiunge una governance significativa. Produce una struttura ricorsiva in cui l'incertezza valuta l'incertezza, e il confine tra azione e supervisione collassa. È in quel collasso che risiede il rischio reale, indipendentemente da quanto strutturato appaia il sistema in superficie.

> **TIPS**
>
> Know something about a sign language AI system that the public should know? Send it to tips@novaracg.com. Every tip is read by me and kept confidential. If you work in the UK, writing to us is not by itself a protected disclosure under the Public Interest Disclosure Act 1998, so before you share employer information, check your options with Protect (protect-advice.org.uk), the free whistleblowing advice line.

Come citare

Grizzle, H. M. (2026, April 18). Quando i sistemi di controllo ereditano la stessa incertezza che dovrebbero governare. Novara Consulting Group. https://www.novaracg.com/it/2026/04/18/when-control-systems-inherit-the-same-uncertainty-they-are-meant-to-govern/