NEURAL NEXUS Review
Rivista scientifica ufficiale di **ENIA**
Ente Nazionale per l'Intelligenza Artificiale
ISSN 3103-604X
https://neuralnexus-review.org/

I position paper della Fondazione ENIA

Data di ricezione: 11.12.25 / Data accettazione: 12.12.25 / Data di pubblicazione: 31.12.25
doi: 10.82015/NNR.2025.100116

Governare l'intelligenza artificiale: Il modello RICE

Governing Artificial Intelligence: The RICE model

di Stefano Gorla, Valeria Lazzaroli, Luigi Di Cataldo*

1. Introduzione

L'intelligenza artificiale (IA) sta assumendo un ruolo sempre più pervasivo come tecnologia abilitante per decisioni critiche in ambito sanitario, finanziario, industriale, lavorativo e infrastrutturale. In questi domini, la semplice eccellenza tecnica non può essere intesa come condizione sufficiente per l'adozione dell'IA.

Per lungo tempo, un insieme ristretto di metriche tecniche è stato ritenuto sufficiente per definire la qualità di un sistema di IA. Accuratezza, precisione, recall, velocità di inferenza e capacità di scalare su grandi volumi di dati hanno rappresentato l'orizzonte quasi esclusivo dell'innovazione algoritmica. In contesti sperimentali, di ricerca o a basso impatto decisionale, questo approccio ha prodotto risultati soddisfacenti, accelerando l'adozione dell'AI in numerosi settori. Tuttavia, quando l'IA esce dai

* Stefano Gorla, co-direttore del Dipartimento AI ETHICS di Fondazione Ente Nazionale per l'Intelligenza Artificiale (ENIA) / E-mail: stefano.gorla@acmcert.net; Valeria Lazzaroli, Presidente di Fondazione Ente Nazionale per l'Intelligenza Artificiale (ENIA) e Lead Auditor ISO IEC 42001 / E-mail: presidenza@enia.ai; Luigi Di Cataldo, componente del comitato scientifico-legale di Fondazione Ente Nazionale per l'Intelligenza Artificiale (ENIA) e sociologo dei processi economici e del lavoro, Dipartimento di Scienze Sociali e Politiche, Università di Milano / E-mail: luigi.dicataldo@unimi.it.

laboratori per essere integrata nei processi decisionali reali, soprattutto quelli che incidono direttamente su persone, diritti, risorse economiche e infrastrutture critiche, il paradigma "tecnico" mostra con chiarezza i propri limiti. Un sistema, infatti, può essere altamente performante dal punto di vista statistico e produrre, al tempo stesso, effetti indesiderati e persino dannosi sul piano umano, organizzativo e sociale.

Negli ultimi anni, numerosi casi hanno reso evidente questa frattura: errori sistemici difficili da individuare, bias incorporati nei dati di addestramento, processi decisionali non spiegabili, la diffusa assenza di meccanismi in grado di assicurare un intervento umano tempestivo. In tali situazioni, la performance tecnica passa in secondo piano e la legittimità operativa del sistema viene di fatto messa in discussione. Se il funzionamento dell'IA non fosse comprensibile, controllabile e giustificabile, quest'ultimo risulterebbe attaccabile sotto il profilo etico e giuridico, suscitando potenziali implicazioni reputazionali significative.

Emerge con forza, dunque, la necessità di un cambio di paradigma per assicurare, accanto alla centralità della performance, anche le dimensioni della compliance, della responsabilità sociale e della fiducia. Compliance, responsabilità e fiducia si costruiscono attraverso un funzionamento che risponde ai principi di affidabilità, prevedibilità, trasparenza e controllo umano, permettendo ai diversi stakeholder – come le organizzazioni, la cittadinanza e il legislatore stesso – di accettare l'uso delle tecnologie avanzate anche in contesti ad alto impatto, dacché nella condizione di riconoscerne processi, confini, garanzie e presidi di controllo.

L'integrazione dell'IA per l'esercizio dei pubblici poteri e processi decisionali che mettono in pericolo i diritti fondamentali, la salute e la sicurezza delle persone, nonché la democrazia, lo Stato di diritto e l'ambiente non richiede semplici misurazioni di rapidità e accuratezza statistica, ma strumenti e pratiche di governance per illuminarne i processi interni e le logiche corrispondenti, prevedere le situazioni in cui le probabilità di fallimento aumentano e assicurare la piena effettività del principio di responsabilità.

All'interno di questa fase, in cui risulta necessario un'evoluzione di paradigma che coinvolge la dimensione epistemologica e quella operativa, la Fondazione Ente Nazionale per l'Intelligenza Artificiale propone, e sviluppa ulteriormente, il Modello RICE (Ji J. et al., 2023) come cornice solida e integrata adatta a supportare la conformità dei sistemi di IA agli standard previsti dalla normativa europea[1], agevolando l'incorporazione delle preoccupazioni sociali ed ecologiche, per suscitare un atteggiamento di maggiore fiducia in coloro che sono direttamente sottoposti al funzionamento di queste tecnologie e nell'opinione pubblica.

2. Il Modello RICE: fondamenti concettuali

Nella fase di transizione attuale, in cui le nuove tecnologie iniziano ad essere impiegate in processi decisionali critici, che hanno il potenziale di impattare in modo significativo sui diritti fondamentali delle persone, emerge la necessità di creare consapevolezza e fiducia, ingredienti indispensabili per supportarne l'applicazione anche nei contesti più problematici in termini di rischi potenziali.

La proposta della Fondazione ENIA ha origine dall'esigenza di ricondurre le principali dimensioni che concorrono a definire la *trustworthiness* di un sistema di IA all'interno di una struttura concettuale unitaria. Il Modello RICE rappresenta una lente integrativa, che permette di rendere leggibili, coerenti e operativamente gestibili requisiti di carattere tecnico e conformità legale che vengono ancora percepiti come frammentati tra le dimensioni tecnica, giuridica e organizzativa. Esso, dunque, non si propone come ulteriore standard tecnico, né rappresenta un'alternativa ai già complessi quadri normativi esistenti.

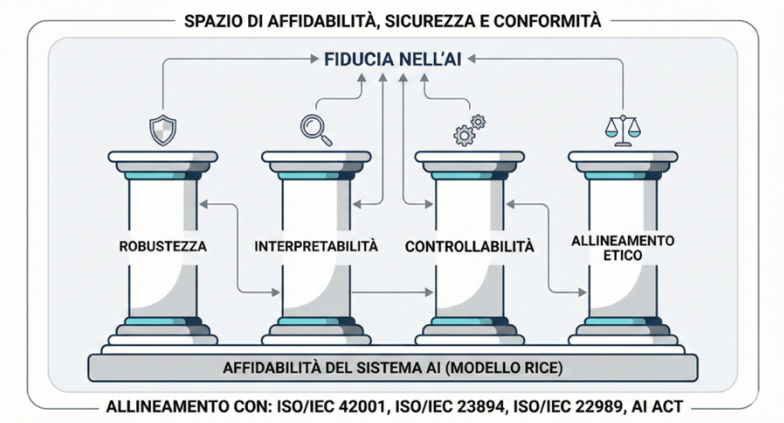
Il Modello RICE offre una grammatica comune da utilizzarsi nei processi di governance, audit, valutazione del rischio e certificazione dell'IA. Esso introduce una visione sistemica dell'affidabilità – non additiva, ma strutturale – che si articola in quattro

[1] Ji, J., et al. (2023). *Ai alignment: A comprehensive survey*. arXiv preprint arXiv:2310.19852.

pilastri distinti, ma interdipendenti (Fig. 1).

I quattro pilastri che compongono il modello sono: 1) Robustezza, intesa come la capacità di un sistema IA di mantenere le proprie prestazioni in un dato dominio di applicazione, in un'ampia varietà di circostanze (incluse quelle non previste) legate alle funzioni che sono state affidate al sistema e di resistere a impatti o attacchi esterni che sono di fatto specifici per ogni ambito di applicazione; 2) Interpretabilità, intesa come trasparenza operativa che rende visibili le correlazioni tra variabili di input e risultati di output, permettendo agli esseri umani di mantenere un controllo attivo, quindi di analizzare e comprendere in profondità il funzionamento di un sistema di IA; 3) Controllabilità, intesa come possibilità di intervento tempestivo sul funzionamento del sistema di IA da parte di un agente esterno (tipicamente un essere umano con responsabilità di supervisione e intervento) per arrestarne i processi, modificarne lo sviluppo e correggerne i risultati quando emergono condizioni inattese che potrebbero dare luogo a effetti indesiderati; 4) Allineamento etico, inteso come incorporazione dei valori, dei principi e degli obiettivi umani nel funzionamento delle IA, assicurando che queste tecnologie agiscano in modo conforme ai quadri etici e giuridici esistenti.

Figura 1 - I quattro pilastri del Modello RICE



Fonte: Elaborazione degli autori

Ciascuno dei quattro pilastri – che saranno dettagliatamente descritti nei sotto-paragrafi susseguenti – rappresenta una condizione necessaria per l'affidabilità complessiva del sistema, traducendosi in prerequisito per la formazione di un contesto sociale di consapevolezza e fiducia in modo del tutto indipendente dal livello di performance tecnica raggiunto. Congiuntamente, i quattro pilastri delineano lo spazio entro cui un sistema di IA può essere considerato compliance, affidabile e sicuro.

In questo documento, il Modello RICE è stato ulteriormente elaborato in modo pienamente coerente con i principali riferimenti internazionali in materia di AI governance e trustworthiness, tra cui l'ISO/IEC 42001 in materia di sistemi di gestione dell'IA, l'ISO/IEC 23894 in materia di fiducia verso i sistemi di IA, l'ISO/IEC 22989 in materia di tassonomia dell'IA, nonché con l'impianto del Regolamento UE 2024/1689 del Parlamento europeo e del Consiglio del 13 giugno 2024, che stabilisce regole armonizzate sull'intelligenza artificiale.

Il Modello RICE, dunque, rappresenta un dispositivo per l'allineamento delle IA rispetto al complessivo panorama dei requisiti etici, legali, organizzativi, politici e sociali.

2.1. Il primo pilastro: la robustezza

Per robustezza si intende la capacità di un sistema IA di mantenere le proprie prestazioni in un dato dominio di applicazione, in un'ampia varietà di circostanze (incluse quelle non previste) legate alle funzioni che sono state affidate al sistema e di resistere a impatti o attacchi esterni che sono di fatto specifici per ogni ambito di applicazione.

La robustezza si manifesta nella capacità del sistema di IA di mantenere comportamenti coerenti, prevedibili e sicuri anche quando opera in condizioni non ideali. Non si tratta di raggiungere un certo standard di accuratezza media, si tratta piuttosto di resistere alla variabilità, all'incertezza e alle perturbazioni che caratterizzano gli ambienti reali. La robustezza coinvolge proprietà tecniche, come la

gestione di dati rumorosi, incompleti o distorti, la gestione di situazioni limite, nonché la resistenza ad eventuali attacchi informatici. Valutare la robustezza di un sistema significa andare oltre la semplice correttezza formale-computazionale dei processi interni al sistema e spostare lo sguardo sui rischi operativi. In questo senso, occorre un approccio di governance strutturato e continuo, che inizia con la fase di sviluppo ma che richiede verifiche e interventi di mantenimento nel tempo attraverso test dedicati, simulazioni di scenario e procedure di monitoraggio continuo.

La robustezza dei sistemi di intelligenza artificiale (IA) è intrinsecamente legata alla qualità dei set di addestramento, rendendo la *data robustness* un requisito metodologico fondamentale per garantire l'affidabilità predittiva. In questo ambito, la gestione rigorosa di *outlier*, bias e dati mancanti trascende la semplice fase di pre-elaborazione, configurandosi come una criticità strutturale per la corretta generalizzazione dei modelli di *machine learning*. Gli *outlier*, introducendo rumore e deviazioni statistiche, rischiano di alterare i gradienti durante la fase di training, imponendo l'adozione di architetture resilienti o funzioni di costo penalizzanti. Parallelamente, la mitigazione sistematica dei bias nei dati è imperativa non solo per la validità del modello, ma per prevenire la propagazione di asimmetrie decisionali e garantire l'equità degli algoritmi. Infine, la gestione dei dati mancanti richiede strategie avanzate che preservino le distribuzioni latenti senza generare artefatti informativi. Un approccio olistico a queste tre dimensioni permette di sviluppare sistemi IA sicuri, trasparenti e realmente applicabili in scenari operativi caratterizzati da incertezza.

Oltre alla validazione preventiva tramite stress test, un aspetto cruciale nell'implementazione delle soluzioni di IA nel mondo reale è rappresentato propriamente dalla robustezza operativa, strettamente legata alla stabilità del sistema in ambienti dinamici. A differenza degli scenari statici e controllati tipici della fase di *training*, gli ambienti operativi reali sono intrinsecamente caratterizzati da elevata entropia, rumore non stazionario e continua evoluzione delle distribuzioni di input (fenomeno noto in letteratura come *concept drift*).

In tale contesto, la stabilità si traduce nella capacità dell'algoritmo di mantenere prestazioni coerenti e sicure nel tempo, adattando le proprie inferenze alle nuove condizioni al contorno senza subire crolli predittivi. Ciò rende imprescindibile la progettazione di architetture dotate di *loop*, feedback adattivi o meccanismi di *continual learning*, capaci di assimilare le fluttuazioni ambientali senza incorrere in fenomeni di *catastrophic forgetting*. Garantire e certificare questa forma di robustezza dinamica costituisce il prerequisito fondamentale per il dispiegamento affidabile di sistemi IA in contesti critici e non strutturati.

Accanto alla resilienza verso le fluttuazioni ambientali, un paradigma fondamentale nella valutazione dei modelli di intelligenza artificiale è costituito dalla robustezza alla sicurezza, relativa proprio alla resistenza del sistema ad attacchi intenzionali. In contesti applicativi aperti, agenti malevoli possono infatti sfruttare le vulnerabilità intrinseche degli algoritmi attraverso tecniche di manipolazione mirata, quali il *data poisoning* durante la fase di addestramento o gli attacchi di *adversarial evasion* in fase di inferenza. Queste perturbazioni ingegnerizzate, pur risultando spesso impercettibili all'operatore umano, sono disegnate matematicamente per forzare il modello a produrre classificazioni errate o output catastrofici. Di conseguenza, l'architettura deve essere blindata integrando protocolli di *AI cybersecurity*, come l'addestramento avversariale (*adversarial training*) e l'impiego di funzioni di regolarizzazione avanzate. Solo attraverso una difesa proattiva contro tali minacce è possibile preservare l'integrità decisionale dell'algoritmo e prevenire la compromissione sistematica dell'infrastruttura intelligente.

Un'ulteriore dimensione di fondamentale rilevanza nell'ingegnerizzazione e nel dispiegamento dei sistemi di IA è la robustezza temporale, ovvero la capacità di mantenere prestazioni e comportamenti accettabili nel tempo. A differenza dei software deterministici tradizionali, i modelli *data-driven* sono fisiologicamente soggetti a un progressivo decadimento prestazionale, noto in letteratura come *model aging*. Questo fenomeno è spesso innescato dalla graduale obsolescenza dei pattern

appresi durante la fase di addestramento rispetto all'inesorabile evoluzione del dominio applicativo reale. Per garantire che il sistema preservi un'elevata affidabilità a lungo termine, la validazione statica pre-rilascio risulta pertanto insufficiente. È invece imperativo implementare architetture integrate con *pipeline* di monitoraggio longitudinale delle metriche di errore e dei comportamenti del modello in produzione. L'adozione di rigorose strategie di ri-addestramento e l'impiego di pratiche avanzate di *Machine Learning Operations* (MLOps) consentono all'algoritmo di ricalibrarsi dinamicamente, assicurando che le decisioni automatizzate restino accurate e pertinenti nonostante il trascorrere del tempo.

La valutazione operativa della robustezza di un sistema di intelligenza artificiale necessita di metriche rigorose e protocolli di validazione dinamici. Tra gli indicatori tipici di tale robustezza assumono un ruolo metodologico centrale l'esecuzione strutturata di stress test e l'analisi di scenario. Queste tecniche permettono di sondare i limiti prestazionali dell'algoritmo in condizioni ambientali estreme o impreviste. Parallelamente, l'integrazione di test *adversarial* mirati risulta indispensabile per misurare e mitigare la vulnerabilità del modello di fronte a perturbazioni o attacchi esterni. Un ulteriore pilastro è rappresentato dal monitoraggio continuo degli scostamenti tra gli input ricevuti e gli output generati, fondamentale per intercettare tempestivamente eventuali derive nei dati (*data drift*). Infine, un'architettura olisticamente robusta richiede l'adozione di meccanismi di *graceful degradation*. Tali paradigmi *fail-safe* consentono al sistema, qualora esposto ad anomalie severe, di ridimensionare le proprie funzionalità o la propria autonomia operativa in modo controllato, scongiurando il rischio di generare effetti critici o collassi improvvisi.

La robustezza non è una proprietà che si acquisisce con l'eliminazione dell'errore, ma con la riduzione dei suoi impatti, anche in condizioni avverse rimanendo "allineato" al proprio scopo e ai propri vincoli anche quando il contesto cambia.

2.2. Il secondo pilastro: l'interpretabilità

Per interpretabilità si intende la possibilità di rendere le decisioni, i comportamenti e le logiche interne di un sistema di IA comprensibili per gli esseri umani. L'interpretabilità consiste nella trasparenza operativa che rende visibili le correlazioni tra variabili di input e risultati di output, permettendo agli esseri umani di mantenere un controllo attivo, quindi di analizzare e comprendere in profondità il funzionamento di un sistema di IA. Essa rappresenta, dunque, un pre-requisito rilevante per la costruzione di un contesto di fiducia, per la conduzione di audit approfonditi e per l'adempimento degli obblighi normativi previsti.

La mancanza di interpretabilità comporta la degenerazione dell'IA in "scatola nera", ossia in un sistema in grado di produrre un output senza che però sia possibile comprenderne appieno le ragioni e, quindi, senza che sia possibile spiegare eventuali errori, contestare le decisioni prese nei propri confronti, associare le responsabilità agli effetti prodotti e dimostrare la conformità del sistema a requisiti legali e organizzativi.

L'interpretabilità non è un concetto monolitico, ma si declina su diversi livelli. Esiste un livello tecnico, rivolto a data scientist e sviluppatori, che consente di comprendere il funzionamento interno del sistema. Questa dimensione dell'interpretabilità, intrinsecamente legata all'ambito della *Explainable AI* (XAI), risulta essenziale per decostruire la logica decisionale latente delle architetture complesse. Attraverso l'impiego di metodi di spiegazione matematici sia globali che locali – quali i valori SHAP (*Shapley Additive exPlanations*), gli algoritmi LIME o la mappatura dei gradienti di attivazione – i progettisti possono quantificare il contributo delle singole *feature* rispetto all'output finale. Le ispezioni granulari costituiscono lo strumento diagnostico primario non solo per il *debugging* avanzato dell'algoritmo, ma soprattutto per intercettare l'eventuale apprendimento di correlazioni spurie, garantendo che le inferenze siano fondate.

L’interpretabilità dovrà necessariamente tradursi anche in un livello operativo, destinato a utenti esperti che utilizzano il sistema nei processi quotidiani. In questo stadio, l'obiettivo primario non è svelare l'ottimizzazione matematica dei pesi o la topologia della rete, bensì fornire giustificazioni contestuali e semanticamente intellegibili calate nel dominio di applicazione. Per i decisori finali – quali medici, giuristi o analisti finanziari – le spiegazioni erogate dal modello devono configurarsi come *actionable insights*, capaci di supportare attivamente il processo decisionale umano limitando il rischio di un affidamento acritico (*automation bias*). Tale paradigma risulta fondamentale per costruire un solido rapporto di fiducia (*trust*) tra l'operatore e la macchina, abilitando una reale e sicura sinergia *human-in-the-loop*. Di conseguenza, un'interpretabilità operativa efficace richiede interfacce che traducano le inferenze latenti in regole di dominio comprensibili, permettendo all'utente di validare o rigettare consapevolmente la raccomandazione algoritmica.

Il requisito dell’interpretabilità coinvolge anche il piano decisionale e strategico, strettamente connesso ai paradigmi di governance e di conformità normativa. In questa veste, la trasparenza algoritmica trascende la singola interazione utente-macchina per rivolgersi a *stakeholder* organizzativi, *auditor* e *policymaker*, i quali necessitano di garanzie sistemiche circa l'equità, la sicurezza e la legalità delle soluzioni adottate. Questo livello risulta imprescindibile per rispondere ai dettami delle normative di recente introdotte a livello europeo, che impongono la dimostrabilità delle logiche automatizzate qualora queste impattino sui diritti fondamentali o su processi aziendali critici. L'interpretabilità decisionale fornisce pertanto il substrato documentale e logico per l'*algorithmic auditing*, permettendo di mappare le catene di responsabilità (*accountability*) e di valutare l'impatto etico-sociale delle predizioni su larga scala. Solo strutturando la spiegabilità a questo livello apicale un'organizzazione può mitigare il rischio reputazionale e istituzionalizzare l'IA come leva strategica affidabile e pienamente rendicontabile.

Per concludere, l'interpretabilità si declina a livello individuale rendendo effettivo il

diritto dell'interessato a ricevere spiegazioni comprensibili in merito alle decisioni automatizzate subite. Questa dimensione pone al centro il cittadino o l'utente in quanto soggetto passivo dell'inferenza – si pensi, ad esempio, al diniego di una linea di credito o all'esclusione da un processo di selezione. In simili contesti, la spiegabilità non risponde a finalità di *debugging* tecnico o di profilazione aziendale, bensì a un'istanza giuridica e democratica di primaria importanza, come previsto dal quadro normativo europeo a tutela dei diritti fondamentali. Tradurre la complessa computazione matematica in motivazioni chiare, contestuali e prive di tecnicismi inaccessibili diviene un imperativo legale per abilitare la contestabilità della decisione. Solo attraverso questa democratizzazione dell'intelligibilità è possibile preservare l'autodeterminazione della persona, colmando le fisiologiche asimmetrie informative e offrendo una reale tutela contro opacità e potenziali discriminazioni algoritmiche.

La concretizzazione del principio di interpretabilità nelle sue molteplici stratificazioni richiede un vasto arsenale di metodologie complementari. Le tecniche che possono favorire questo risultato includono modelli interpretabili *by design*, come alberi decisionali o sistemi a regole, architetture (*white-box*) trasparenti fin dalla loro concezione, consentendo di seguire passo dopo passo il flusso logico, i nodi e le soglie di attivazione che conducono all'output finale. Laddove, invece, l'elevata dimensionalità dei dati o la complessità del *task* imponessero l'impiego di algoritmi intrinsecamente opachi ma altamente performanti (come le reti neurali profonde o gli *ensemble methods*), si potranno applicare tecniche di *explainable AI* post-hoc, come LIME, SHAP o l'analisi dell'importanza delle variabili, che operano a posteriori sul modello già addestrato, permettono di estrarre e approssimare le logiche latenti senza doverne alterare la struttura.

Dal punto di vista della granularità informativa, a queste metodologie si affiancano spiegazioni locali, focalizzate sulla singola decisione, e spiegazioni globali, che descrivono il comportamento complessivo del modello. Le prime risultano essenziali per decostruire l'inferenza su una specifica istanza (ad esempio, isolando le ragioni

matematiche che hanno portato al rifiuto di un particolare mutuo bancario), mentre le seconde offrono una panoramica macroscopica e strutturale sui pesi e sulle correlazioni apprese dall'algoritmo sull'intero *dataset* di addestramento. Alla luce di questa vasta eterogeneità di strumenti, risulta evidente come non esista una soluzione universale: la scelta delle tecniche dipende strettamente dal contesto applicativo e dal tipo di stakeholder coinvolti. Se un *data scientist* prediligerà le metriche globali per finalità di *debugging* e validazione parametrica del sistema, un operatore clinico o un cittadino necessiteranno di interfacce basate su spiegazioni squisitamente locali, le uniche in grado di giustificare in modo intelligibile la singola diagnosi o la specifica classificazione. Calibrare correttamente l'adozione di queste tecniche costituisce pertanto la chiave di volta per trasformare l'astratto concetto di intelligibilità in uno strumento concretamente funzionale alle esigenze sia operative che etico-normative.

2.3. Il terzo pilastro: la controllabilità

Parallelamente ai requisiti strutturali di robustezza e di trasparenza, l'affidabilità di un sistema di intelligenza artificiale postula un rigoroso paradigma di controllabilità. Per controllabilità si intende la possibilità di intervento tempestivo sul funzionamento del sistema di IA da parte di un agente esterno (tipicamente un essere umano con responsabilità di supervisione e intervento) per arrestarne i processi, modificarne lo sviluppo e correggerne i risultati quando emergono condizioni inattese che potrebbero dare luogo a effetti indesiderati.

Affinché un'architettura intelligente sia considerata realmente sicura per il dispiegamento operativo, dovrà garantirsi che le sue azioni rimangano costantemente ancorate all'agenzia umana (*human agency*). Un sistema pienamente controllabile si qualifica pertanto per la sua propensione a essere non solo passivamente osservato, ma anche attivamente guidato e limitato nel suo raggio d'azione qualora si discosti dalle metriche di sicurezza prefissate. Questo livello di *governance* tecnologica richiede

l'implementazione nativa di interfacce di *override* e meccanismi di blocco di emergenza (*kill switch*) che consentano agli operatori responsabili di intervenire tempestivamente per correggere, sospendere o arrestare in via definitiva l'esecuzione del modello. Inquadrare l'IA all'interno di logiche *human-in-the-loop* (HITL) o *human-on-the-loop* (HOTL) assicura quindi che la delega decisionale alla macchina non si traduca in un'abdicazione della responsabilità umana, preservando l'autorità dell'operatore specialmente in contesti critici.

Dal punto di vista della macro-governance aziendale e istituzionale, la controllabilità di un sistema algoritmico trascende la semplice implementazione di interfacce tecniche, risultando strettamente legata all'adozione di rigorosi modelli di responsabilità e alla chiara separazione dei ruoli. L'integrazione dell'IA nei processi operativi critici porta con sé il rischio intrinseco di diluire la *liability* all'interno di architetture computazionali complesse, generando opacità organizzativa. Per contrastare questo fenomeno, i framework di *risk management* impongono di mappare esplicitamente la catena di comando, individuando con precisione gli attori umani che possiedono l'autorità ultima per validare, modificare o rigettare le raccomandazioni prodotte dalla macchina. La segregazione delle funzioni – distinguendo nettamente le competenze di chi sviluppa il modello, di chi ne verifica le metriche e di chi lo utilizza per deliberare – diviene pertanto uno snodo procedurale fondamentale per mitigare eventuali derive. In questo scenario, la controllabilità funge da perimetro normativo e ontologico: essa garantisce in modo sistemico che l'IA rimanga incardinata nel proprio ruolo di mero strumento di supporto alle decisioni, potenziando l'analisi umana senza mai esautorarla. L'obiettivo strategico ed etico di tale impostazione è impedire categoricamente che la delega tecnologica si spinga fino al punto in cui l'algoritmo si trasformi in un soggetto decisionale pienamente autonomo, operante in un vuoto di *accountability*. Mantenere l'operatore umano al vertice della piramide decisionale (*human-in-command*) assicura infatti che ogni scelta ad alto impatto sia sottoposta a un insindacabile vaglio cognitivo e morale, permettendo all'organizzazione di assumersi la

piena responsabilità giuridica e sociale degli esiti generati.

Nell'ottica della gestione del rischio tecnologico, l'efficienza predittiva di un algoritmo non può in alcun modo compensare o giustificare un deficit nella sua supervisione. Al contrario, un sistema di IA che si dimostri altamente efficace ma non adeguatamente controllabile rappresenta un rischio intrinseco inaccettabile per qualsiasi ecosistema operativo. Una simile configurazione, infatti, sottrae all'organizzazione la capacità fondamentale di intervenire in modo tempestivo al verificarsi di anomalie, bias emergenti o imprevisti ambientali. L'impossibilità di modulare o arrestare un processo decisionale automatizzato trasforma la pura efficienza in un potenziale danno sistemico. Per scongiurare tale deriva, la controllabilità deve tradursi strutturalmente all'interno del design algoritmico, esprimendosi attraverso diversi modelli formali di interazione uomo-macchina. Questi paradigmi progettuali definiscono l'esatto grado di autonomia del sistema: spaziano da un controllo preventivo stringente, in cui la macchina necessita di una validazione umana per ogni step critico, a forme di supervisione più distaccate in cui l'operatore monitora un processo fluido, mantenendo però la facoltà inibitoria assoluta.

La controllabilità si esprime operativamente attraverso la strutturazione di diversi modelli di interazione uomo-macchina. Il livello di salvaguardia più restrittivo all'interno di questo spettro è delineato dai sistemi definiti *Human-in-the-Loop*. Il paradigma *Human-in-the-Loop* descrive un sistema in cui l'intervento umano è strutturalmente richiesto in ogni ciclo decisionale. In questa configurazione, l'IA non possiede un’autonomia esecutiva completa, agendo invece come strumento di supporto avanzato: elabora dati e formula raccomandazioni o bozze di decisione che l'operatore umano deve obbligatoriamente e attivamente validare, modificare o respingere. All'interno di queste specifiche architetture, la delega tecnologica è volutamente mantenuta parziale, poiché l'intervento umano risulta essere un passaggio obbligatorio per la validazione delle decisioni. Tale approccio "bloccante" si rende indispensabile nei domini ad alto rischio per impedire inneschi automatici

potenzialmente critici, confinando l'algoritmo al ruolo di avanzato sistema di raccomandazione. Un più fluido bilanciamento tra automazione ed efficienza è invece offerto dai sistemi *Human-on-the-Loop*. Il paradigma *Human-on-the-Loop* definisce un'architettura in cui il sistema IA opera in modo semi-autonomo, eseguendo le decisioni in tempo reale senza richiedere un'approvazione umana preventiva per ogni singolo *step*. Tuttavia, l'operatore mantiene un ruolo di supervisione continua e proattiva sull'andamento del processo, possedendo la capacità tecnica, procedurale e l'autorità per intervenire *ex post* al fine di alterare, correggere o interrompere forzatamente l'esecuzione dell'algoritmo qualora ne rilevi deviazioni critiche. Il sistema di IA riceve un mandato esecutivo e opera con maggiore autonomia, mentre l'essere umano esercita una funzione di supervisione continua. Il fulcro metodologico e di sicurezza di questo assetto risiede nella costante garanzia di un intervento *ex post*: l'operatore mantiene l'accesso a interfacce di *override* che gli consentono di intercettare, correggere o inibire l'azione della macchina qualora si discosti dalle metriche tollerate.

Infine, a coronamento di queste dinamiche strettamente operative, si colloca il livello strategico e organizzativo delineato dai modelli *Human-in-Command*. L'approccio *Human-in-Command* si colloca a un macro-livello strategico e di *governance*, riferendosi alla capacità e al diritto umano di supervisionare l'impatto complessivo del sistema IA (incluso quello etico e sociale). Implica il potere assoluto e incondizionato di decidere se, quando e come utilizzare l'algoritmo in una specifica situazione. Rientra in questo costrutto la prerogativa di non affidarsi alla macchina, di limitarne preventivamente il raggio d'azione o di scavalcarne l'autorità in qualsiasi momento per salvaguardare l'agenzia umana. In questo scenario di *governance* apicale, l'autorità umana rimane piena e incondizionata rispetto all'infrastruttura intelligente. Tale preminenza si sostanzia nella prerogativa esclusiva dei decisori umani di poter ridefinire attivamente gli obiettivi generali, imporre nuovi limiti operativi e alterare radicalmente le modalità operative del sistema stesso. L'adozione consapevole di

questa tassonomia relazionale assicura che il dispiegamento dell'IA avvenga sempre all'interno del saldo perimetro della responsabilità umana.

Affinché i modelli di interazione uomo-macchina funzionino come previsto e siano utili a garantire una *governance* affidabile, risulta necessario integrare specifici elementi strutturali all'interno dell'architettura tecnologica di riferimento. Tra i requisiti tecnici e procedurali primari spicca l'implementazione di *kill switch* e di procedure strutturate di *fallback* manuale. Queste salvaguardie estreme fungono da interruttori di emergenza, consentendo agli operatori di arrestare o degradare in sicurezza le funzionalità del sistema in presenza di comportamenti anomali o derive critiche. Per garantire la tempestività e il necessario coordinamento dell'intervento umano, l'infrastruttura deve essere supportata da *policy* di *escalation* rigorosamente definite e da limitazioni esplicite dell'autonomia decisionale del modello, circoscrivendo a priori il perimetro delle azioni consentite alla macchina. A completamento di questo apparato di sicurezza, assumono un ruolo cardine i meccanismi di *logging* avanzato e di tracciabilità granulare delle singole inferenze. Questi registri non costituiscono un mero strumento tecnico, bensì il substrato probatorio essenziale per l'*auditing*: essi permettono di ricostruire retrospettivamente il comportamento della macchina, isolare la genesi degli errori e, in ultima analisi, esercitare un controllo effettivo e rendicontabile lungo l'intero ciclo di vita del sistema intelligente.

2.4. Il quarto pilastro: l'allineamento etico

Il concetto di "allineamento etico" viene impiegato per indicare l'incorporazione dei valori, dei principi e degli obiettivi umani nel funzionamento delle IA, assicurando che queste tecnologie agiscano in modo conforme ai quadri etici e giuridici esistenti. L'IA sarà allineata sul piano etico laddove nella condizione di operare in modo coerente con i diritti fondamentali e i principi etici condivisi.

Anche rispetto al quarto pilastro del Modello RICE, l'adozione di misure meramente

tecniche non può assicurare che le inferenze producano effetti discriminatori, impatti asimmetrici o impatti lesivi dei diritti fondamentali. Un modello computazionale, infatti, può paradossalmente rivelarsi tecnicamente ineccepibile, dimostrandosi altamente robusto alle perturbazioni e finanche interpretabile nelle sue logiche, ma risultare al contempo socialmente dannoso. È chiaro che efficienza statistica e giustizia sociale non sono concetti necessariamente coincidenti. Per colmare questo divario, l'allineamento etico presuppone una progettazione in grado di trascendere i confini dell'informatica per abbracciare il coinvolgimento strutturale di competenze multidisciplinari, integrando sociologi, giuristi ed esperti di dominio capaci di contestualizzare l'azione della macchina. In secondo luogo, l'ipostatizzazione di questo quarto pilastro richiede la formulazione di *policy* organizzative esplicite e l'esecuzione periodica di rigorose valutazioni di impatto. In ultima istanza, l'allineamento etico ci impone di misurare e giustificare non limitatamente ciò che l'IA realizza in termini di mera esecuzione materiale o di accuratezza predittiva, ma soprattutto le ragioni profonde per cui agisce e, primariamente, a vantaggio o a discapito di chi le sue decisioni vengono implementate.

L'allineamento etico di un sistema di IA si assicura attraverso l'adesione a un nucleo di principi universali che ne guidano lo sviluppo e l'applicazione. Tra i pilastri ricorrenti in letteratura emerge innanzitutto la *fairness*, intesa come l'impegno attivo a mitigare bias e discriminazioni, garantendo equità nei risultati algoritmici. A questa si affiancano i doveri di beneficienza e non maleficenza, che impongono al sistema di operare per il benessere degli individui, evitando categoricamente di arrecare danno o di esacerbare vulnerabilità esistenti. Un ruolo centrale è rivestito dall'autonomia umana, che sancisce la necessità per l'utente di non essere soggiogato da decisioni automatizzate opache, preservando la propria libertà di scelta. L'intero apparato deve poi essere sorretto dall'*accountability*, ovvero una chiara attribuzione di responsabilità per le azioni intraprese dalla macchina, e dal principio di proporzionalità, che esige un bilanciamento rigoroso tra i mezzi impiegati e le finalità perseguite.

Per tradurre questi valori astratti in prassi operativa, l'organizzazione deve dotarsi di strumenti di allineamento etico concreti e verificabili. In primo luogo, assumono un'importanza fondamentale le valutazioni di impatto, tra cui spicca la *Fundamental Rights Impact Assessment* (FRIA), procedura essenziale per mappare preventivamente i rischi per i diritti umani e definire strategie di mitigazione adeguate. Parallelamente, l'istituzione di comitati etici indipendenti garantisce una supervisione multidisciplinare e costante, capace di dirimere i dilemmi morali derivanti dall'uso dell'IA in contesti sensibili. Infine, l'adozione formale di *policy* interne e codici di condotta stringenti permette di istituzionalizzare questi requisiti, trasformando l'etica da un mero esercizio teorico a una componente strutturale della governance aziendale. Solo attraverso questa combinazione di principi guida e strumenti procedurali sarà possibile garantire che l'intelligenza artificiale rimanga una tecnologia orientata al valore umano e pienamente rendicontabile di fronte alla società.

3. Esempi di fallimento per ciascun pilastro del Modello RICE

L'osservazione dei fallimenti algoritmici attraverso la lente dei quattro pilastri del Modello RICE consente di comprendere, con immediatezza operativa, le dinamiche attraverso cui la fiducia in un sistema di IA possa collassare, persino laddove le metriche di *performance* tecnico-statistiche rimangono apparentemente soddisfacenti. Esaminando il primo pilastro, inerente alla *robustezza*, si rileva come uno dei malfunzionamenti più insidiosi e ricorrenti coincida con il degrado prestazionale dei sistemi di IA una volta messi in esercizio. Tale fenomeno è tipicamente innescato da mutamenti ambientali silenti: da un lato il *data drift*, che altera la distribuzione statistica dei dati di input, e dall'altro il *concept drift*, che modifica la relazione intrinseca tra le variabili e l'outcome atteso. Il risultato è una pericolosa deriva del comportamento dell'algoritmo: l'output continua a conservare una forma apparentemente plausibile, ma perde progressivamente aderenza rispetto alla realtà

operativa, trasformandosi in una fonte inosservata di errore sistemico.

Spostando il focus sul versante dell'*interpretabilità*, le criticità prendono spesso la forma di sistemi di *scoring* o classificazione opachi, incapaci di fornire spiegazioni comprensibili sia agli interessati che agli stakeholder. In tali condizioni, l'algoritmo assume una postura di totale chiusura che rende di fatto impraticabile l'attività di *audit*, impedisce la contestazione informata delle decisioni e indebolisce irrimediabilmente l'accountability organizzativa, specialmente in quei domini applicativi dove la trasparenza costituisce un requisito imprescindibile per la legittimità dell'intero processo.

Per quanto concerne la *controllabilità*, la fragilità strutturale del sistema emerge palesemente allorché la spinta verso l'automazione esautora del tutto la componente umana. *Workflow* decisionali completamente automatizzati, privi di reali presidi di supervisione e sprovvisti di meccanismi di *override* umano, convertono le tecnologie moderne in decisori di fatto. Questo scenario genera un rischio non solo tecnico, ma profondamente organizzativo: l'assenza di leve per un intervento correttivo spezza la catena della responsabilità e amplifica in modo significativo l'impatto di eventuali anomalie.

Per concludere, il fallimento sul piano dell'*allineamento etico* si manifesta in modo particolarmente subdolo attraverso il fenomeno delle discriminazioni indirette. Sistemi di IA formalmente "corretti" rispetto alle tradizionali metriche di accuratezza possono, nel tentativo di ottimizzare l'obiettivo misurabile, incorporare e riprodurre correlazioni spurie, *proxy* sensibili o asimmetrie strutturali storicamente presenti nei dati. L'algoritmo genera così esiti iniqui per specifici gruppi sociali, pur mantenendo una performance globale di alto livello. La sintesi di queste vulnerabilità evidenzia un punto critico trasversale: un'IA può continuare a "funzionare" sul piano strettamente computazionale, fallendo tuttavia in modo rovinoso sotto il profilo delle implicazioni sociali associate.

La rassegna di malfunzionamenti ricorrenti e problematici qui proposta per ciascun

pilastro dimostra chiaramente perché il Modello RICE debba essere trattato come una vera e propria architettura di affidabilità strutturale, e non come un mero corollario etico accessorio alla prestazione tecnologica. ogni componente di questo framework deve infatti essere integrato fin dalle prime fasi di sviluppo per garantire che i modelli siano resilienti agli imprevisti e trasparenti nelle loro logiche decisionali. L'adozione rigorosa di questi standard rappresenta, in definitiva, il prerequisito tecnico imprescindibile per un'implementazione sicura e verificabile su larga scala.

4. Trade-off tra RICE e performance

L'adozione strutturale del framework RICE impone una profonda riflessione teorica e operativa sul fisiologico *trade-off* intercorrente tra la *performance* di un sistema di IA e i requisiti sistemici che ne garantiscono l'affidabilità, la governabilità e la legittimità sociale. Nello sviluppo di architetture algoritmiche complesse – in particolar modo con riferimento a quelle caratterizzate da un'elevata capacità espressiva e non linearità – l'incremento dell'accuratezza statistica viaggia spesso di pari passo con un deterioramento della trasparenza, una maggiore criticità nelle dinamiche di controllo e una pervasiva dipendenza dall'automazione. Per converso, la deliberata scelta in fase di *design* di prediligere modelli intrinsecamente interpretabili, o l'integrazione di rigorosi presidi di supervisione umana, può tradursi in una flessione apparente delle prestazioni misurate secondo le metriche convenzionali, nonché in una fisiologica riduzione della velocità di inferenza e dell'autonomia decisionale del sistema.

La necessità di bilanciare tra esigenze di performance e necessità di governcance non può essere interpretata come espressione di un conflitto insanabile tra la spinta all'innovazione e il dovere di responsabilità. Specialmente in domini applicativi ad alto rischio – caratterizzati da stringenti vincoli regolatori o da un impatto severo sui diritti fondamentali, sulle persone e sugli *asset* critici – la mera massimizzazione della *performance* predittiva, in assenza di adeguati presidi di *Robustezza*, *Interpretabilità*,

Controllabilità e *Allineamento etico*, innesca un'amplificazione non lineare dei rischi operativi, legali e morali. Un modello che si dimostri estremamente accurato nei collaudi tecnici ma che risulti al contempo opaco, incontrollabile o eticamente disallineato nella realtà, è destinato a generare costi complessivi superiori ai benefici marginali ricavati in termini di precisione algoritmica.

In questa prospettiva, il Modello RICE cessa di essere percepito come un vincolo inibitorio all'adozione dell'IA, configurandosi piuttosto come una precondizione strutturale per abilitare un'innovazione realmente sostenibile nel medio-lungo periodo. L'apparente riduzione delle prestazioni osservabile in fase di validazione è spesso l'artefatto di una metrica valutativa parziale e riduzionista, focalizzata in via esclusiva su parametri puramente tecnici, che omette di quantificare l'effettiva qualità e tenuta del sistema all'interno del suo complesso ecosistema operativo. Nei contesti altamente regolamentati, infatti, la rincorsa ossessiva alla massima *performance* tecnica, priva di un solido ancoraggio ai principi RICE, espone le organizzazioni a una vulnerabilità sistemica, con rischi reputazionali e di *governance* che possono neutralizzare i vantaggi competitivi derivanti dall'ottimizzazione dell'algoritmo. Concludendo, l'architettura RICE impone un fondamentale salto concettuale: essa ridefinisce alla radice il concetto stesso di *performance*, traslandolo da una logica prettamente computazionale e quantitativa a una metrica integrata che incorpora indissolubilmente il valore, l'affidabilità e la sostenibilità decisionale dell'Intelligenza Artificiale.

5. Il Radar RICE

Il *Radar RICE* si configura come uno strumento avanzato di *governance* algoritmica, ideato per tradurre il complesso costrutto multidimensionale della *trustworthiness* (affidabilità) dell'IA in una rappresentazione visiva immediatamente intelligibile, comparabile e, soprattutto, funzionale al dibattito critico tra le diverse anime di

un'organizzazione, dallo sviluppo tecnico al *risk management*, fino all'*audit* e al settore legale. Più che un mero espediente grafico, questo strumento funge da dispositivo diagnostico in grado di mappare con precisione quella che può essere definita la geometria del rischio di un modello computazionale.

La premessa epistemologica e metodologica alla base del Radar RICE è che l'affidabilità costituisca una proprietà intrinsecamente sistemica e non compensativa: l'eccellenza prestazionale o un'elevata maturità su un singolo asse, come ad esempio la robustezza tecnica, non può in alcun modo controbilanciare o neutralizzare una fragilità strutturale su un altro fronte, quale l'assenza di controllabilità o un grave disallineamento etico. Di conseguenza, il livello di maturità complessiva non scaturisce da una semplice sommatoria, poiché una singola vulnerabilità critica su uno dei quattro pilastri è sufficiente a far collassare l'intera architettura della fiducia, compromettendo l'*accountability* istituzionale e la conformità normativa del progetto.

Per scongiurare il rischio che l'analisi si riduca a un mero esercizio impressionistico o a valutazioni autoreferenziali, la quantificazione degli assi postula un rigoroso processo di autovalutazione guidata. Tale *framework* vincola l'assegnazione dei punteggi a tre principi metodologici inderogabili

1. **Criteri osservabili:** criteri ancorati a pratiche operative, metriche e risultati verificabili (test, log, procedure, report, decisioni formalizzate);
2. **Scala di maturità standardizzata:** sequenza di grandezze (ad esempio 0–5 o 1–5) dotata di descrittori univoci, volta a garantire riproducibilità e coerenza valutativa anche tra team eterogenei;
3. **Vincolo dell'evidenza:** l'attribuzione di punteggi superiori a determinate soglie esige la produzione di documenti o output tecnici che lo supportino (es. report di drift monitoring, policy di escalation, verbali di review etica, model card, audit trail).

Scendendo nella granularità analitica, affinché il Radar RICE risulti esaustivo, ciascuna

dimensione viene decostruita in specifiche famiglie di indicatori che spaziano dai controlli procedurali ai *Key Performance/Risk Indicators* (KPI/KRI), fino alle evidenze probatorie.

5.1. Il Radar RICE e il pilastro della robustezza

All'interno dell'architettura diagnostica del Radar RICE, il pilastro della robustezza quantifica la resilienza strutturale e algoritmica del sistema di IA dinanzi alla variabilità ambientale, al rumore stocastico, ai mutamenti di contesto e alle potenziali minacce esterne (Tab. 1). Lungi dall'essere una mera misurazione statica dell'accuratezza in ambiente di laboratorio, questa dimensione valuta la capacità del modello di mantenere un'erogazione prestazionale sicura e costante nelle imprevedibili dinamiche del mondo reale.

Tabella 1 – Il Radar RICE: tecniche di misurazione della robustezza

	Misurazione della robustezza
Controlli tipici	Stress test, scenario analysis, test su outlier, adversarial testing (se pertinente), valutazione robustezza su segmenti critici, validazione in condizioni degradate.
KPI/KRI	Drift rate, frequenza di retraining, tasso di degradazione prestazioni nel tempo, incidenti dovuti a dati anomali, copertura dei test su corner cases.
Evidenze	Report di validazione, registri di monitoraggio drift, risultati di robustness testing, procedure di rollback o safe-mode.

Fonte: Elaborazione degli autori

Per sostanziare questo asse, il framework impone l'implementazione di rigorosi controlli tecnici che trascendono la validazione standard.

La maturità del sistema viene sondata attraverso metodologie di stress testing e

scenario analysis, volte a simulare condizioni operative estreme. A queste si aggiungono la valutazione sistematica sugli outlier e, laddove il dominio applicativo lo richieda in termini di sicurezza, l'adversarial testing, essenziale per immunizzare l'algoritmo contro input malevoli o deliberatamente perturbati. La robustezza, inoltre, si dimostra attraverso la stabilità su segmenti di dati critici e la capacità di superare cicli di validazione in condizioni operative degradate, assicurando che un'anomalia parziale non inneschi un collasso a cascata.

Sotto il profilo del monitoraggio quantitativo, il posizionamento sul Radar RICE è ancorato a specifici indicatori (KPI/KRI), tra cui spiccano il *drift rate* (la deviazione statistica dei dati nel tempo), la frequenza di *retraining* necessaria per riallineare il modello e il tasso di degradazione delle prestazioni. Cruciale risulta anche il tracciamento degli incidenti causati da dati anomali e la percentuale di copertura dei test sui corner cases.

Infine, in stretto ossequio al vincolo dell'evidenza, un punteggio elevato su questo asse esige la produzione di un inoppugnabile apparato probatorio: report di validazione formali, registri di monitoraggio continuo del *drift* e risultati documentati dei test di robustezza. La predisposizione di procedure ingegnerizzate di *rollback* o l'attivazione automatica di una *safe-mode* dinanzi ad anomalie critiche costituiscono la prova empirica definitiva che l'architettura è non solo resiliente, ma strutturalmente concepita per governare il rischio di fallimento.

5.2. Il Radar RICE e il pilastro dell'interpretabilità

Nel contesto del Radar RICE, il pilastro dell'interpretabilità valuta il grado di trasparenza algoritmica e la capacità del sistema di tradurre la propria complessa logica computazionale in spiegazioni intellegibili e giustificabili secondo diverse dimensioni (tecnico, operativo, management, interessato) e in modo adeguato al destinatario (Tab. 2). Tale requisito sfugge a un approccio generico, imponendo invece

che la narrazione delle decisioni e del comportamento del modello sia rigorosamente modulata in base all'asimmetria informativa del destinatario: dallo sviluppatore tecnico all'operatore di business, fino all'alta direzione (management) e al soggetto interessato dall'output (data subject).

Tabella 2 – Il Radar RICE: tecniche di misurazione dell'interpretabilità

	Misurazione dell'interpretabilità
Controlli tipici	Strategie XAI, spiegazioni locali e globali, documentazione delle feature e delle assunzioni, analisi di sensibilità, "reason codes" per gli output, verifiche di comprensibilità.
KPI/KRI	Percentuale di decisioni spiegabili con qualità sufficiente, tasso di contestazioni non gestibili per mancanza di spiegazione, metriche di stabilità delle spiegazioni (consistenza).
Evidenze	Model card / system card, report SHAP/LIME (o alternative), guideline interpretative per utenti, template di spiegazione per l'interessato.

Fonte: Elaborazione degli autori

Per presidiare efficacemente questo asse, il framework impone l'adozione di una strutturata strategia di *eXplainable AI* (XAI). I controlli tipici richiedono l'estrazione di spiegazioni sia globali (per comprendere l'architettura generale) sia locali (per motivare la singola inferenza), affiancate da un'esaustiva documentazione delle *feature* e delle assunzioni di base. La validazione dell'interpretabilità passa inoltre attraverso analisi di sensibilità, l'assegnazione di *reason codes* puntuali per ogni output e l'espletamento di rigorose verifiche empiriche di reale comprensibilità umana.

Sul fronte quantitativo, il posizionamento sul Radar si affida a specifici KPI e KRI: la metrica primaria è la percentuale di decisioni spiegabili con un livello qualitativo ritenuto sufficiente, unita alle metriche di stabilità e consistenza delle spiegazioni stesse dinanzi a input simili. Un indicatore di rischio particolarmente critico da monitorare è il tasso di contestazioni o reclami esterni che risultano inammissibili o

non gestibili proprio a causa dell'opacità del sistema (il cosiddetto effetto black-box). Infine, in stretta ottemperanza al vincolo dell'evidenza, l'eccellenza su questo asse postula la produzione di un solido impianto documentale: model card o system card formalizzate, report tecnici generati tramite metodologie XAI consolidate (come SHAP, LIME o framework equivalenti), linee guida interpretative a uso interno per gli operatori e, tassativamente, template di spiegazione redatti in linguaggio chiaro e accessibile a tutela dell'interessato finale.

5.3. Il Radar RICE e il pilastro della controllabilità

Nel quadro diagnostico del Radar RICE, il pilastro della controllabilità presidia una prerogativa inalienabile della governance algoritmica: la salvaguardia della reale agentività umana sul sistema. Questa dimensione misura l'effettiva capacità degli operatori di supervisionare, limitare, deviare o interrompere i processi automatizzati, tracciando confini netti e invalicabili tra l'autonomia computazionale e la responsabilità decisionale umana (Tab. 3).

Tabella 3 – Il Radar RICE: tecniche di misurazione della controllabilità

	Misurazione della controllabilità
Controlli tipici	Definizione del modello HITL/HOTL/HIC, kill switch, fallback manuale, policy di escalation, segregazione dei ruoli, autorizzazioni e change management.
KPI/KRI	Tempo medio di intervento, frequenza di override, tasso di incidenti senza possibilità di intervento, copertura del logging, qualità dell'audit trail.
Evidenze	Procedure operative, diagrammi di workflow con punti di controllo, registri di override e incident response, audit log, tracciabilità end-to-end..

Fonte: Elaborazione degli autori

Per garantire che l'algoritmo non sfugga al dominio dell'organizzazione, il framework esige l'implementazione di rigorosi controlli strutturali.

Il primo passo è la codifica formale del paradigma di interazione, stabilendo tecnicamente se il modello debba operare in un regime di *Human-in-the-Loop* (HITL), *Human-on-the-Loop* (HOTL) o *Human-in-Command* (HIC). A livello operativo, la controllabilità si traduce nell'infrastrutturazione di meccanismi di tutela tangibili: la predisposizione di un kill switch di emergenza, l'agilità di un fallback manuale fluido e l'attivazione di precise policy di escalation. Tale architettura deve inoltre essere blindata da una netta segregazione dei ruoli, da rigide autorizzazioni e da un formale processo di *change management*.

Spostando il focus sul cruscotto di monitoraggio (KPI/KRI), l'efficacia del controllo umano viene quantificata attraverso il tempo medio di intervento, la frequenza di utilizzo delle interfacce di override e l'ampiezza della copertura del logging. Un indicatore di rischio dirimente, capace di far collassare la valutazione dell'asse, è rappresentato dal tasso di incidenti in cui l'architettura ha tecnicamente precluso o ritardato la possibilità di un intervento umano correttivo.

Infine, in piena aderenza al vincolo dell'evidenza, l'assegnazione di punteggi di maturità elevati è subordinata all'esibizione di prove inconfutabili: procedure operative standardizzate, diagrammi di workflow che mappino inequivocabilmente i punti di controllo, e registri dettagliati degli *override* e delle operazioni di *incident response*. La qualità ineccepibile dell'*audit log* e una totale tracciabilità *end-to-end* certificano, in ultima analisi, che il sistema rimane un solido strumento a disposizione dell'uomo, e mai viceversa.

5.4. Il Radar RICE e il pilastro dell'allineamento etico

A completamento dell'architettura diagnostica del Radar RICE, il pilastro dell'allineamento etico rappresenta la dimensione trasversale e sistemica dell'intero framework. Esso non si limita a valutare le metriche prestazionali dell'algoritmo, ma

ne quantifica la rigorosa coerenza con il contesto istituzionale, garantendo equità, proporzionalità e il perseguimento esclusivo di finalità lecite.

Com'è stato già scritto, questo asse funge da vero e proprio ponte epistemologico e operativo, collegando inestricabilmente le scelte tecniche di sviluppo alle policy organizzative e ai reali impatti socio-economici, con l'obiettivo primario di prevenire danni e consolidare l'accountability istituzionale. Per presidiare questa dimensione del Modello Rice e realizzare una misurazione quantitativa dell'allineamento etico, saranno necessari controlli che trascendono la mera compliance formale (Tab. 4).

Tabella 4 – Il Radar RICE: tecniche di misurazione dell'allineamento etico

	Misurazione dell'allineamento etico
Controlli tipici	AI Impact Assessment / FRIA, valutazioni di fairness (diretta e indiretta), governance etica (board, review, escalation), privacy e data governance, valutazioni ex ante ed ex post.
KPI/KRI	Metriche di disparity per gruppi rilevanti, segnalazioni di danno/impatti, tasso di reclami e contenziosi, esiti di audit etici, conformità a policy interne.
Evidenze	report di impatto, verbali e decisioni del review board, misure di mitigazione adottate, tracciamento delle non conformità etiche e delle azioni correttive.

Fonte: Elaborazione degli autori

Risulta imprescindibile l'esecuzione di strutturati AI Impact Assessment – in materia di diritti fondamentali, trattamento dei dati personali e implicazioni per la salute e la sicurezza dei destinatari – integrati da profonde valutazioni di fairness volte a intercettare e disinnescare eventuali forme di discriminazione indiretta. A livello organizzativo, l'eticità esige una robusta governance incardinata su comitati dedicati (*review board*), chiare procedure di escalation e una rigorosa data governance a tutela della privacy, supportata da valutazioni continue *ex ante* ed *ex post*.

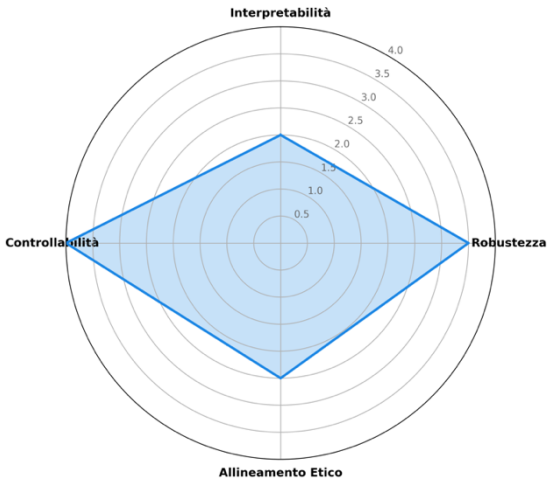
Sul fronte del monitoraggio quantitativo, il posizionamento sul Radar è ancorato a KPI e KRI strettamente focalizzati sull'impatto umano: metriche di disparity per quantificare i bias su gruppi vulnerabili, segnalazioni di danno effettivo, tasso di reclami e potenziali contenziosi, nonché i risultati formali degli audit etici.

Infine, in rigorosa osservanza del vincolo dell'evidenza, l'eccellenza su questo asse non ammette petizioni di principio, ma postula un apparato probatorio documentato: report di impatto analitici, verbali e decisioni del board, un'anagrafica delle misure di mitigazione adottate e, soprattutto, il tracciamento puntuale delle non conformità etiche e delle relative azioni correttive intraprese.

6. Valutare attraverso il Radar RICE

Per scongiurare il rischio che la validazione di un sistema di IA si riduca a valutazioni arbitrarie o meramente auto-assolutorie, l'assegnazione dei punteggi all'interno del Radar RICE postula l'adozione di una rigorosa rubrica di maturità (Fig. 2).

Figura 2 – La rubrica di maturità del Radar RICE



Fonte: Elaborazione degli autori

La rubrica di maturità si articola tipicamente su una scala incrementale a cinque livelli, progettata per sterilizzare la soggettività dei valutatori. La progressione ha inizio da un livello *Iniziale* (1), caratterizzato da prassi operative episodiche e documentazione minima, avanzando verso un livello *Gestito* (2), dove i presidi di controllo sussistono ma risultano deficitari in termini di sistematicità. Il vero consolidamento organizzativo si registra a partire dal livello *Definito* (3), che certifica l'esistenza di processi formalizzati, ripetibili e supportati da chiari confini di responsabilità. L'architettura evolve ulteriormente nel livello *Misurato* (4), fondato sul monitoraggio continuo e quantitativo di specifici indicatori (KPI e KRI), per culminare infine nel livello *Ottimizzato* (5), l'apice della maturità in cui il sistema si auto-regola attraverso processi di automazione dei controlli, revisione multidisciplinare e integrazione nativa con la più ampia gestione del rischio aziendale.

Il pilastro portante e ineludibile di tale rubrica è il principio dell'evidenza probatoria: l'avanzamento lungo questa scala assiologica è precluso in assenza di riscontri oggettivi e documentali. Sotto il profilo procedurale, non risulta ammissibile rivendicare un'elevata robustezza algoritmica senza che siano prodotti parallelamente registri continuativi di *drift monitoring* o chiare *policy* di risposta alle derive dei dati.

Ancorato a questi requisiti stringenti, il Radar RICE trascende la sua natura di cruscotto grafico per assolvere a tre precise funzioni strategiche. In primo luogo, abilita una comparazione analitica tra ecosistemi eterogenei, consentendo il *benchmarking* oggettivo tra diversi fornitori, iterazioni dello stesso sistema (ad esempio tra una versione *v1* e *v2*) in diversi contesti applicativi. In secondo luogo, esso agisce come dispositivo diagnostico per l'individuazione delle priorità organizzative: le asimmetrie o le contrazioni (le cosiddette "aree scavate") nel poligono visivo rivelano infatti un latente debito di *governance*, la cui colmatura assume spesso un'urgenza ben superiore all'inseguimento di frazionali incrementi nell'accuratezza statistica. Infine, costituisce l'infrastruttura logica essenziale per delineare una chiara *roadmap* di miglioramento, vincolando l'organizzazione ad azioni correttive tracciabili,

all'assegnazione di precise *ownership* e al raggiungimento di *milestone* verificabili nei cicli di *audit* successivi.

Affinché questo strumento esplichi appieno il proprio potenziale cautelativo, la sua misurazione non deve configurarsi come un adempimento *una tantum*, bensì deve dispiegarsi longitudinalmente lungo l'intero ciclo di vita del sistema di IA. L'operazionalizzazione del framework impone la tracciatura di una *baseline* iniziale in fase di *Design/Build*, una rigorosa verifica delle evidenze al traguardo del *Go-live*, seguita da un sistematico monitoraggio durante le fasi di *Operation* – per intercettare *incident* o *override* – e da un ricalcolo integrale in occasione di attività di *Change* o *Re-training*.

Attraverso questa iterazione perenne, la misurazione del Radar RICE non si esaurisce nella mera restituzione grafica, ma si sublima in output documentali cruciali per la conformità aziendale. Il completamento del processo genera infatti *asset* inestimabili: un profilo testuale che decodifica i rischi sistemici, un inventario meticoloso delle evidenze, un piano temporizzato di *remediation* e, fondamentale tassello finale, la deliberazione esplicita della *governance* – il cosiddetto "go/no-go" – che statuisce formalmente le limitazioni d'uso, i requisiti di supervisione umana e le soglie di allarme vincolanti per il dispiegamento in produzione dell'IA.

7. Il Modello RICE come strumento di governance

Il valore strategico del Modello RICE si manifesta appieno quando esso viene implementato come un vero e proprio dispositivo di *governance*. Questa architettura si rivela fondamentale per raccordare in modo sistemico e coerente le molteplici dimensioni dell'IA: quella puramente ingegneristica, quella organizzativa e quella etico-regolatoria.

Attraverso l'applicazione del paradigma RICE, il concetto di "affidabilità" dell'IA cessa di essere un'astrazione teorica per trasformarsi in una struttura intelligibile,

quantificabile e pienamente governabile. In questa cornice, la *Robustezza*, l'*Interpretabilità*, la *Controllabilità* e l'*Allineamento etico* abbandonano la veste di requisiti isolati o accessori per configurarsi come componenti profondamente interdipendenti all'interno di un ecosistema unificato, orientato alla generazione di fiducia, sicurezza e conformità. L'affidabilità, pertanto, non è intesa come una proprietà intrinseca o statica del modello computazionale, bensì come il prodotto di un equilibrio dinamico che bilancia costantemente i presidi tecnici, i flussi decisionali e l'assunzione delle responsabilità organizzative.

Dal punto di vista strettamente operativo, la griglia valutativa del framework RICE permette di mappare e classificare i sistemi di IA in tre macro-scenari distinti. Il primo scenario riguarda i modelli che presentano un'altissima *performance* computazionale ma che risultano non *RICE-compliant*. In queste casistiche, l'efficienza predittiva funge da pericoloso paravento per vulnerabilità strutturali legate all'opacità, all'ingovernabilità o al disallineamento etico, generando in definitiva un profilo di rischio inaccettabile per l'organizzazione. Un secondo scenario è rappresentato dai sistemi pienamente *RICE-compliant* ma caratterizzati da prestazioni puramente tecniche sub-ottimali. Tali architetture, pur essendo perfettibili attraverso mirati interventi di ottimizzazione, possiedono il vantaggio inestimabile di non compromettere in alcun modo la solida infrastruttura di *governance* già in essere. Il terzo scenario, individua infine quei sistemi capaci di coniugare una performance adeguata con una totale aderenza ai quattro pilastri RICE. Solo queste soluzioni possono dirsi pienamente affidabili, giuridicamente difendibili e, in prospettiva, formalmente certificabili.

Proiettando il modello RICE a livello *corporate*, esso si candida a divenire la struttura portante per l'assimilazione dell'IA all'interno dei preesistenti sistemi di gestione aziendale, integrandosi con le funzioni di *risk management*, *compliance*, controllo qualità e sicurezza delle informazioni. La forza metodologica del framework risiede nella sua capacità di tradurre ogni singolo asse in requisiti organizzativi tangibili,

controlli operativi stringenti, indicatori di *performance* e di rischio (KPI e KRI) e in un apparato di evidenze documentali rigorosamente verificabili in sede di *audit*.

La trasposizione operativa del Modello RICE consente di arginare la frammentazione che tipicamente affligge le iniziative di *AI governance*, fornendo una grammatica istituzionale comune che fluidifica il coordinamento tra i reparti tecnici (sviluppatori e *data scientist*), gli uffici legali, i comitati etici e gli organi di controllo interno. In ultima istanza, il modello RICE agisce come un solido ponte strutturale che salda l'ingegneria dei sistemi intelligenti con i quadri normativi vigenti che ne regolano l'adozione. Lungi dall'introdurre inutili livelli di complessità burocratica, esso razionalizza i processi di revisione, garantisce la trasparenza delle valutazioni di conformità e agevola percorsi di certificazione allineati ai nuovi standard internazionali emergenti. Il framework fornisce, in definitiva, la chiave di lettura indispensabile per governare l'IA come un complesso sistema socio-tecnico, assicurando che la massima prestazione ingegneristica, la responsabilità giuridica e il valore sociale rimangano sempre indissolubilmente e strutturalmente allineati.

L'implementazione del modello RICE esprime il suo massimo potenziale strategico allorché viene applicata come logica di governance persistente lungo l'intero ciclo di vita dell'IA, superando i riduttivi limiti di una validazione puramente statica. Nella delicata fase di design, il framework orienta preventivamente le scelte architetturali: impone l'adozione di paradigmi di *explainability by design*, la mappatura rigorosa delle fonti dati e dei potenziali bias, definendo i limiti di validità del modello e abbattendo così il debito di governance sin dalla sua genesi. Durante l'esercizio operativo (operation), i pilastri RICE si traducono nella gestione empirica e quotidiana del rischio aziendale. In questo stadio assumono rilevanza il monitoraggio puntuale del data drift e del concept drift, la gestione strutturata degli incidenti e l'applicazione tempestiva delle policy di escalation, rendendo la controllabilità un perimetro di sicurezza tangibile e non un mero requisito astratto. Il ciclo si compie infine con la fase di monitoring, incentrata sul riesame e sul miglioramento continuo: attraverso audit

periodici e l'aggiornamento iterativo del Radar RICE, l'organizzazione trasforma le evidenze raccolte in apprendimento procedurale, governando di fatto l'IA come un complesso ecosistema socio-tecnico in perenne evoluzione che deve essere costantemente mantenuto e dimostrato.

8. Il Modello RICE: un'infrastruttura di compliance

Uno degli elementi di maggiore pregio analitico e operativo del modello RICE risiede nella sua vocazione a fungere da struttura di raccordo tra i requisiti regolatori del Regolamento UE 2024/1689 (c.d. AI Act) e i framework di gestione delineati dagli standard internazionali ISO/IEC. Questa mappatura non si esaurisce in una semplice corrispondenza terminologica, elevandosi invece a vera e propria traduzione operativa che permette alle organizzazioni di inscrivere la progettazione, la gestione e il controllo dei sistemi intelligenti all'interno di un unico impianto di *governance* coerente.

L'asse della *Robustezza* trova un riscontro speculare nei dettami del Regolamento europeo summenzionato inerenti all'accuratezza, alla resilienza e alla sicurezza dei sistemi, requisiti imperativi soprattutto per le applicazioni classificate come ad alto rischio. Vincoli normativi che convergono in modo sinergico con lo standard ISO/IEC 42001, in particolare nella sezione 6.1, che prescrive la sistematica identificazione, valutazione e trattamento dei rischi legati ai sistemi di IA lungo il loro intero ciclo di vita. In questa convergenza, la robustezza cessa di essere un mero attributo tecnico per divenire una proprietà olistica, governata tramite processi strutturati di *risk assessment* e monitoraggio continuo.

Rispetto all'asse dell'Interpretabilità, il Modello RICE si allinea rigorosamente ai requisiti di trasparenza imposti dalla normativa europea, che esigono la piena comprensibilità del funzionamento, delle capacità e delle limitazioni dei sistemi a beneficio degli *stakeholder*. A livello di standardizzazione gestionale, questa

dimensione si radica nel fondamento della sezione 7. dell'ISO/IEC 42001, dedicata alla competenza, alla consapevolezza e alla comunicazione. L'interpretabilità si configura pertanto non solo come una scelta algoritmica, ma come un requisito organizzativo che necessita di documentazione adeguata, processi di comunicazione strutturati e strumenti di spiegazione calibrati sul contesto decisionale.

L'asse della *Controllabilità* costituisce il punto di tangenza più esplicito con il principio cardine di *human oversight* (supervisione umana) sancito dal Regolamento europeo, indispensabile affinché i sistemi siano progettati per consentire un intervento efficace capace di prevenire o mitigare potenziali danni. Tale principio trova la sua operazionalizzazione nella sezione § 8. dell'ISO/IEC 42001, che disciplina le attività operative, il controllo dei processi e la gestione delle modifiche. Entro questa prospettiva, la controllabilità si traduce strutturalmente in chiare attribuzioni di ruoli, procedure di intervento predefinite e inoppugnabile tracciabilità delle decisioni.

Infine, l'asse dell'*Allineamento etico* si salda in modo diretto alla tutela dei diritti fondamentali, che rappresenta l'architrave giuridico dell'impostazione europea e il criterio dirimente per la legittimità delle applicazioni di IA in Europa. Questo livello trova una sua formalizzazione avanzata e trasversale nello standard ISO/IEC 23894, focalizzato sulla *trustworthiness* (affidabilità) dell'IA, il quale impone di integrare le considerazioni etiche, sociali e di impatto all'interno dei processi decisionali. L'allineamento etico connette così il rispetto normativo alla responsabilità sociale d'impresa, esigendo rigorose valutazioni di impatto, *policy* esplicite e saldi meccanismi di *accountability*.

Nella sua visione d'insieme, questa mappatura evidenzia come il framework RICE agisca da insostituibile chiave di lettura unificante: da un lato esso rende immediatamente operativi i principi legislativi del Regolamento europeo e dall'altro consente di strutturare un *AI Management System* perfettamente conforme all'ISO/IEC 42001, altresì coerente con gli standard globali di *trustworthiness*. In conclusione, il Modello RICE agevola il superamento degli attriti organizzativi tra la *compliance*

normativa, la gestione del rischio aziendale e la progettazione tecnica dei sistemi intelligenti.

Considerazioni conclusive

Come evidenziato in apertura di questa trattazione, la crescente pervasività dell'IA nei processi decisionali critici – dalla sanità alla finanza, fino alle infrastrutture e all'esercizio dei pubblici poteri – impone un definitivo superamento della mera eccellenza tecnica come criterio sufficiente da tenere in considerazione per la sua adozione. L'era in cui lo sviluppo algoritmico era guidato in via quasi esclusiva da metriche statistiche (quali *accuracy*, *recall* o velocità di inferenza) si scontra oggi con l'evidenza dei limiti di sistemi che, pur rivelandosi altamente performanti sul piano computazionale, producono effetti indesiderati, opachi o lesivi sul piano umano e sociale. Da questa frattura emerge l'urgenza del cambio di paradigma formalizzato dal Modello RICE e dal Radar di misurazione RICE: la transizione da un approccio focalizzato sulla potenza di calcolo a uno fondato sulla legittimità operativa, sulla *compliance* e sulla responsabilità.

La sintesi ragionata di questo framework dimostra che la fiducia in una tecnologia non si decreta, ma si costruisce metodicamente attraverso un'infrastruttura di presidi interdipendenti. La vera e profonda utilità del Modello RICE e dEl Radar di misurazione RICE risiede nella formidabile capacità di questi strumenti di operazionalizzare i riferimenti etici e normativi, fornendo alle organizzazioni una grammatica istituzionale e visiva comune (il Radar RICE) capace di far dialogare in modo sinergico le unità di sviluppo tecnico con le divisioni legali, di *risk management* e di *audit*. Il modello funge da ponte strutturale che traduce i complessi requisiti europei dell'AI Act e le prescrizioni dei sistemi di gestione ISO/IEC (come la 42001) in indicatori tangibili, evidenze documentali, procedure di *override* e valutazioni di impatto sistematiche.

In conclusione, l'adozione del Modello RICE si configura come un presidio strategico e

democratico ineludibile, garantendo che la spinta all'automazione non si traduca in un'abdicazione della supervisione umana, permettendo ad aziende e istituzioni di rispondere con certezza e trasparenza alle domande cardine dell'innovazione contemporanea: non solo quanto un modello sia veloce, ma in quali condizioni possa fallire, chi ne debba rispondere e quali logiche valoriali incorpori.