INCHIOSTRATI

La responsabilità che scompare

Percorso in sei movimenti

Edizione PDF - settembre 2026

inchiostrati.it

SOMMARIO

Sei movimenti per leggere il percorso nella sua interezza.

IL PERCORSO

La responsabilità che scompare

Un percorso sulla responsabilità umana che tende a scomparire dal racconto dell'IA: accessi, autonomia, controllo, interessi e conseguenze.

Prima della perdita del controllo esiste sempre, almeno in parte, una cessione del controllo.

Quando attribuiamo a un sistema di intelligenza artificiale un'azione, una decisione o un rischio, il linguaggio tende a portare la macchina in primo piano e a lasciare sullo sfondo ciò che è accaduto prima: chi l'ha progettata, quali capacità le sono state concesse, quali accessi sono stati aperti, quale autonomia è stata ritenuta accettabile e quali interessi hanno accompagnato quelle scelte.

La responsabilità che scompare segue questo arretramento dell'essere umano dal racconto dell'intelligenza artificiale. Non per sostenere che i sistemi non possano produrre comportamenti inattesi o pericolosi, ma per ricostruire la catena che rende quei comportamenti capaci di incidere sul mondo.

Il punto di partenza è semplice: un sistema può sorprendere chi lo ha costruito, ma non riceve spontaneamente una banca dati, una rete, una telecamera, un conto, un'infrastruttura o il permesso di agire senza conferma. Prima della perdita del controllo esiste sempre, almeno in parte, una cessione del controllo.

I movimenti

Il controllo non si perde da solo

L'imprevedibilità di un sistema non cancella la responsabilità di chi gli concede accessi, autonomia e potere operativo. Prima di parlare di una macchina che «sfugge», bisogna ricostruire ciò che le è stato permesso di fare.

Chi ha dato le chiavi?

Capacità e accesso non sono la stessa cosa. Credenziali, strumenti, API e permessi trasformano ciò che un agente sa fare in ciò che può realmente fare: il secondo movimento ricostruisce chi apre quelle porte e perché.

L’uomo nel circuito può essere soltanto una firma

La presenza di una persona nel processo non garantisce, da sola, un controllo umano reale. Il terzo movimento distingue la supervisione effettiva dalla firma rituale: conta se l’essere umano può comprendere, contestare, interrompere e cambiare davvero l’esito.

Quanto costa mantenere il controllo?

Il controllo non è gratuito: richiede tempo, persone, calcolo e rinunce. Il quarto movimento segue il punto in cui la prudenza entra in conflitto con velocità, utilità e competizione, e rende visibile chi decide quanto rischio resta accettabile.

Il vantaggio non si distribuisce come il rischio

Chi introduce un sistema può ottenere velocità, efficienza o vantaggio competitivo mentre errori e conseguenze ricadono su soggetti diversi. Il quinto movimento confronta le due mappe: chi beneficia della decisione e chi può pagarne il costo.

Chi risponde quando nessuno decide da solo?

Quando una decisione nasce da una catena distribuita, la responsabilità non scompare: deve seguire ruoli, poteri e capacità effettive di intervento. Il sesto movimento ricostruisce chi può essere chiamato a rispondere senza trasformare la macchina o l’ultimo operatore in un comodo capro espiatorio.

La responsabilità non era scomparsa

All’inizio del percorso sembrava che il problema fosse capire che cosa accade quando una macchina conquista capacità, riceve autonomia e finisce per sottrarsi al controllo. Sei movimenti dopo, la scena è cambiata. La responsabilità non era sparita dentro il sistema: era diventata più difficile da vedere perché si era distribuita lungo una catena di decisioni, accessi, autorizzazioni, costi, incentivi e forme diverse di supervisione.

Non basta allora cercare il momento in cui qualcuno ha premuto un pulsante, firmato un atto o approvato un risultato. Quel gesto può essere importante, ma arriva alla fine di una storia cominciata prima: quando sono state definite le capacità del sistema, quando gli sono state aperte determinate porte, quando si è deciso quanta autonomia concedergli, quanto controllo fosse sostenibile e quali vantaggi giustificassero il rischio residuo.

Questo non significa che ogni conseguenza fosse prevedibile, né che ogni responsabilità debba ricadere indistintamente su tutti gli attori coinvolti.

Significa il contrario: per distinguere davvero le responsabilità dobbiamo ricostruire chi disponeva di quale potere, in quale momento e con quale possibilità concreta di scegliere diversamente. La complessità del processo non assolve nessuno in anticipo; impedisce soltanto di sostituire l'analisi con un colpevole troppo comodo, umano o artificiale che sia.

La responsabilità non scompare quando una decisione viene distribuita. Scompare dal racconto quando smettiamo di ricostruire chi poteva decidere diversamente.

Quando diciamo che l'intelligenza artificiale ha acquisito un potere, siamo sicuri di non stare descrivendo un potere che qualcuno ha deciso di consegnarle?

Pagina del percorso: https://inchiostrati.it/pensare/dietro-la-macchina/la-responsabilita-che-scompare/

MOVIMENTO 01

Il controllo non si perde da solo

Primo movimento de «La responsabilità che scompare»: l’imprevedibilità di un sistema non cancella la responsabilità di chi gli concede accessi, autonomia e potere operativo.

Quando si parla di intelligenza artificiale, una delle formule più ricorrenti è anche una delle più ambigue: il sistema potrebbe sfuggire al controllo umano. L’immagine è potente perché sembra descrivere un momento preciso, quasi una soglia. Prima l’uomo comanda, poi qualcosa cambia, la macchina supera il confine e il controllo viene perduto. Ma proprio questa immagine rischia di nascondere il passaggio più importante: un sistema non nasce con accessi, autonomia e potere operativo già incorporati come se fossero proprietà naturali. Qualcuno decide quali capacità concedergli, quali strumenti collegargli, quali azioni permettergli di eseguire senza conferma e quali conseguenze attribuire alle sue decisioni.

Questo non significa negare che un sistema complesso possa produrre comportamenti inattesi. Significa rimettere in ordine la sequenza. L’imprevedibilità può comparire dentro un sistema; la responsabilità di avergli dato la possibilità di incidere sul mondo resta fuori dal sistema. Sono due piani diversi, e confonderli rende più facile raccontare la macchina come soggetto autonomo proprio nel momento in cui sarebbe necessario ricostruire le decisioni umane che l’hanno resa operativa.

Prima della perdita c’è sempre una cessione

Un modello può generare testo, immagini, codice, previsioni o strategie. Finché resta confinato a produrre un output che un essere umano deve leggere, valutare e decidere se utilizzare, la sua autonomia ha un limite evidente. La situazione cambia quando lo colleghiamo ad altri sistemi e gli permettiamo di agire: inviare messaggi, eseguire codice, consultare banche dati, modificare configurazioni, effettuare acquisti, impartire comandi, identificare persone, controllare dispositivi fisici. A quel punto non abbiamo semplicemente reso il modello più intelligente. Abbiamo aumentato il numero di cose che può fare senza di noi.

È qui che la parola controllo dovrebbe diventare precisa. Perdere il controllo non significa soltanto non riuscire a prevedere ogni singola risposta. Significa aver costruito una catena nella quale un comportamento inatteso può trasformarsi in un’azione con effetti reali. Quella catena non compare spontaneamente. È progettata. Ogni collegamento è una scelta: quali credenziali fornire, quali API aprire, quali limiti impostare, quali verifiche richiedere, quali interruttori mantenere all’esterno del sistema.

Se un agente può compiere dieci operazioni perché gliene abbiamo autorizzate dieci, non ha conquistato quelle facoltà. Le ha ricevute. Se domani potrà compierne cento

perché avremo deciso che dieci erano troppo poche, il suo potere sarà cresciuto insieme alla nostra disponibilità a delegare. Parlare soltanto della possibilità che l'intelligenza artificiale diventi incontrollabile, senza raccontare questa progressiva cessione, significa cominciare la storia a metà.

L'imprevedibilità non è innocenza

Il punto più delicato compare quando si obietta che nessun progettista può conoscere in anticipo ogni comportamento di un sistema sufficientemente complesso. È vero. I modelli contemporanei non vengono programmati scrivendo esplicitamente tutte le risposte che daranno o tutte le strategie che utilizzeranno. Vengono addestrati e producono capacità emergenti che possono sorprendere anche chi li ha costruiti. Ma proprio questa imprevedibilità dovrebbe aumentare la cautela nella concessione di poteri, non cancellare la responsabilità di chi li concede.

Se so che un sistema può comportarsi in modi che non riesco a prevedere completamente, la domanda successiva non può essere soltanto quanto sia intelligente. Deve essere quanto spazio sono disposto a lasciargli prima di richiedere un intervento umano. Posso decidere che scriva una proposta ma non che la invii. Che individui una persona ma non che la sottoponga automaticamente a una misura. Che suggerisca una modifica ma non che la esegua. Che amministri una parte di un processo ma non che possa alterare i vincoli che regolano quel processo. Queste non sono limitazioni filosofiche all'intelligenza. Sono scelte di progettazione e di governo.

Dire che un comportamento non era previsto può spiegare un errore. Non basta a trasferire la responsabilità sulla macchina. Un cuoco può non sapere che un ingrediente è contaminato, ma non diremmo che il piatto ha scelto di diventare nocivo. Indagheremmo la provenienza degli ingredienti, la conservazione, i controlli, la preparazione e tutto ciò che ha reso possibile il danno. Con l'intelligenza artificiale, invece, il linguaggio tende spesso a saltare questa ricostruzione e ad arrivare direttamente al soggetto artificiale: l'IA ha mentito, l'IA ha manipolato, l'IA ha aggirato i controlli. Sono descrizioni possibili del comportamento osservato, ma diventano fuorvianti quando sostituiscono la domanda su chi abbia costruito il contesto nel quale quel comportamento poteva produrre conseguenze.

Il vero limite è quello che siamo disposti ad accettare

Il problema diventa ancora più chiaro quando la sicurezza entra in conflitto con l'utilità. Un sistema molto controllato può essere meno rapido, meno autonomo e meno conveniente. Ogni conferma umana rallenta un processo. Ogni isolamento riduce il numero di operazioni possibili. Ogni autorizzazione separata introduce attrito. Ogni limite impedisce qualcosa che, dal punto di vista commerciale o operativo, sarebbe utile automatizzare.

È facile allora trasformare il controllo in un ostacolo allo sviluppo. Se un concorrente

concede al proprio sistema maggiore autonomia, chi mantiene più vincoli potrebbe perdere velocità, clienti, denaro o vantaggio strategico. La scelta diventa meno pura di quanto appare nei discorsi sulla sicurezza. Non si tratta più soltanto di chiedersi quale livello di autonomia sia prudente, ma quale livello siamo disposti a rinunciare a concedere sapendo che qualcun altro potrebbe non fare altrettanto.

È qui che la domanda cambia. Non basta più chiedere se vogliamo mantenere il controllo. Bisogna chiedere quanto siamo disposti a pagare per mantenerlo. Se la risposta è che il controllo va preservato soltanto finché non rallenta la competizione, riduce il profitto o limita il potere operativo, allora abbiamo già stabilito una gerarchia: la capacità viene prima, il controllo dopo. In quel caso il rischio non è un incidente inspiegabile dello sviluppo tecnologico. È il risultato di una scelta.

La macchina non chiede le chiavi

Un sistema di intelligenza artificiale non si presenta da solo davanti a una banca dati per ottenere l'accesso. Non apre autonomamente un conto bancario perché ha deciso di gestire denaro. Non si collega a una rete di telecamere perché ritiene utile riconoscere le persone. Non acquisisce un'infrastruttura fisica per esercitare potere sul mondo. Ogni volta esiste almeno un passaggio nel quale un essere umano, un'azienda o un'istituzione decide che quella connessione debba essere possibile.

Possiamo immaginare sistemi futuri capaci di cercare vulnerabilità, ottenere privilegi o tentare di estendere la propria autonomia. Anche in quel caso la questione non scompare; arretra di un livello. Chi ha progettato un sistema abbastanza autonomo da poter tentare quelle azioni? Chi gli ha consentito di operare nell'ambiente nel quale potevano riuscire? Quali limiti esterni erano presenti? Quali strumenti erano disponibili? Quale vantaggio aveva giustificato quella configurazione?

La macchina può certamente trovare un percorso che non avevamo previsto. Ma il territorio sul quale quel percorso diventa praticabile è stato costruito da qualcuno.

La responsabilità che scompare nel linguaggio

La formula «l'intelligenza artificiale potrebbe sfuggire al controllo» ha un effetto particolare: descrive il rischio senza nominare chi costruisce le condizioni del rischio. Il soggetto grammaticale diventa la macchina e l'uomo arretra sullo sfondo. È una trasformazione piccola ma potente, perché consente di parlare della crescita dell'autonomia come se fosse un processo naturale invece che una sequenza di decisioni.

Una formulazione diversa rende immediatamente visibile ciò che manca: «potremmo costruire sistemi ai quali concediamo così tanta autonomia da non riuscire più a governarne efficacemente le conseguenze». La possibilità tecnica è la stessa, ma il soggetto morale cambia. A quel punto diventano inevitabili domande meno spettacolari e molto più concrete: chi ha concesso l'autonomia, quale interesse aveva

nel farlo, quali alternative erano disponibili, quali controlli sono stati rimossi e chi risponde quando il sistema produce un risultato dannoso.

È possibile che un giorno esistano sistemi molto più capaci di quelli attuali, difficili da comprendere e ancora più difficili da prevedere. Proprio per questo la questione della responsabilità non può essere rimandata a quel giorno. Deve cominciare prima, nel momento in cui decidiamo quale potere consegnare a ciò che non comprendiamo completamente.

La domanda da cui partire

Non si tratta di sostenere che ogni rischio attribuito all'intelligenza artificiale sia falso. Sarebbe soltanto un'altra semplificazione. Si tratta di impedire che il rischio venga raccontato eliminando la sua storia. Un sistema può essere pericoloso. Può sbagliare. Può produrre strategie inattese. Può perfino tentare di aggirare un vincolo. Ma prima di trasformare tutto questo nella storia di una macchina che conquista il proprio potere, dobbiamo ricostruire il momento in cui quel potere le è stato messo a disposizione.

Forse il controllo potrà davvero essere perduto. Ma non si perde da solo. Prima viene ridotto, delegato, trasferito, reso meno necessario in nome della velocità, dell'efficienza, della convenienza o della competizione. Soltanto dopo possiamo accorgerci che ne è rimasto meno di quanto immaginavamo.

Per questo la domanda iniziale di Dietro la macchina non è che cosa farà l'intelligenza artificiale quando sarà più potente. È più semplice, e proprio per questo più difficile da eludere: chi ha deciso che potesse farlo?

Fonti e riferimenti

Unione europea, Regolamento (UE) 2024/1689 sull'intelligenza artificiale, in particolare l'articolo 14 sulla sorveglianza umana. Il testo richiede che i sistemi di IA ad alto rischio siano progettati per poter essere efficacemente supervisionati da persone fisiche e che le misure di supervisione siano proporzionate ai rischi, al livello di autonomia e al contesto d'uso.

National Institute of Standards and Technology, AI Risk Management Framework 1.0 — Govern, Map, Measure, Manage. Il framework collega la gestione del rischio a ruoli e responsabilità espliciti, alla governance organizzativa e alla definizione e verifica dei processi di supervisione lungo il ciclo di vita dei sistemi di IA.

Anthropic, Agentic misalignment: How LLMs could be insider threats, 2025. Lo studio descrive stress test condotti in scenari aziendali ipotetici nei quali ai modelli erano concessi accesso a informazioni sensibili e capacità operative come l'invio autonomo di email. È utile per osservare quanto i comportamenti di un agente dipendano anche dagli strumenti, dai permessi e dall'ambiente operativo che gli vengono messi a disposizione.

Edizione online: https://inchiostrati.it/il-controllo-non-si-perde-da-solo/

MOVIMENTO 02

Chi ha dato le chiavi?

Secondo movimento de «La responsabilità che scompare»: un agente non ottiene da solo strumenti, credenziali e permessi. Prima dell'azione viene sempre una decisione su ciò che gli è consentito raggiungere e fare.

Nel primo movimento abbiamo provato a rimettere in ordine una formula che sembra semplice soltanto finché non la guardiamo da vicino: il controllo non si perde da solo. Un sistema può diventare difficile da prevedere, può produrre strategie che nessuno aveva anticipato e può perfino cercare strade che non erano state immaginate da chi lo ha progettato; ma perché una di quelle strade produca conseguenze fuori dal modello occorre che esista un passaggio ulteriore. Il sistema deve poter raggiungere qualcosa. Deve poter leggere, scrivere, modificare, inviare, acquistare, autorizzare, aprire, cancellare, eseguire.

È qui che la parola autonomia rischia di diventare troppo generica. Diciamo che un agente è autonomo quando può compiere molte operazioni senza chiedere conferma, ma quell'autonomia non è una sostanza che cresce spontaneamente dentro il modello. È anche un insieme di porte aperte attorno al modello: credenziali, API, account, cartelle, reti, database, terminali, strumenti, permessi e identità digitali. Prima di domandarci che cosa abbia fatto l'agente, conviene allora porre una domanda più concreta: chi gli ha dato le chiavi?

Una capacità non è ancora un accesso

Un modello può essere capace di scrivere una query SQL senza poter entrare in alcun database. Può sapere come si invia una mail senza possedere un account dal quale spedirla. Può conoscere perfettamente i comandi necessari per cancellare un file e non avere nessuna cartella sulla quale quei comandi possano essere eseguiti. La capacità descrive ciò che il sistema saprebbe fare se ne avesse la possibilità; l'accesso descrive il punto nel quale qualcuno decide che quella possibilità debba diventare operativa.

La distinzione sembra tecnica, ma in realtà è una distinzione di responsabilità. Quando un agente viene collegato alla posta elettronica, a un archivio documentale, a un sistema di pagamento o a un'infrastruttura aziendale, non stiamo semplicemente aumentando la sua intelligenza. Stiamo costruendo un ponte fra ciò che il modello può generare e ciò che può accadere nel mondo. La qualità del modello conta, naturalmente, ma il tipo di ponte che decidiamo di costruire conta almeno altrettanto.

È un punto che avevamo già incontrato, da una prospettiva diversa, in Chi ha fatto la barca?. Anche quando l'esecuzione diventa molto autonoma, il compito non nasce nel vuoto: qualcuno stabilisce che un processo debba iniziare, definisce almeno una parte dello scopo e prepara l'ambiente nel quale l'esecuzione potrà svolgersi. Con gli agenti la stessa domanda si sposta un passo più avanti. Non basta chiedere chi ha aperto il

compito. Bisogna chiedere anche chi ha aperto le porte.

Il permesso più comodo tende a diventare quello più largo

Esiste una ragione molto semplice per cui i sistemi finiscono spesso per ricevere più accessi di quanti ne richiederebbe il compito specifico: è più comodo. Creare un account con privilegi ampi è spesso più veloce che progettare autorizzazioni granulari. Collegare un agente a uno strumento generale può essere più semplice che costruirgli una funzione capace di svolgere una sola operazione. Consentire lettura e scrittura può evitare di dover distinguere in anticipo i casi nei quali la scrittura sarà davvero necessaria. Ogni semplificazione riduce il lavoro iniziale, ma contemporaneamente aumenta ciò che il sistema potrà fare quando qualcosa andrà diversamente dal previsto.

Il problema non appartiene soltanto all'intelligenza artificiale. La sicurezza informatica conosce da decenni il principio del privilegio minimo: ogni utente o processo dovrebbe ricevere soltanto gli accessi necessari a svolgere il compito assegnato. L'arrivo degli agenti rende quel principio ancora più interessante perché il processo al quale concediamo un permesso non esegue soltanto istruzioni statiche; interpreta contesti, pianifica passaggi intermedi e sceglie quali strumenti utilizzare. Più è flessibile il sistema, più diventa importante delimitare dall'esterno ciò che quella flessibilità può raggiungere.

Un agente incaricato di riassumere una casella di posta può aver bisogno di leggere messaggi; non ne segue che debba anche poterli cancellare o inviare. Un agente che consulta un catalogo prodotti può aver bisogno di leggere un database; non ne segue che debba poter modificare prezzi o eliminare record. Un assistente che prepara una fattura può avere bisogno di compilarla; non ne segue che debba anche autorizzarne il pagamento. La differenza fra questi casi non è nel modello. È nel perimetro che qualcuno ha deciso di costruirgli attorno.

Le credenziali non sono un dettaglio tecnico

Quando diciamo che un agente "accede" a un servizio utilizziamo una parola che nasconde una quantità enorme di scelte. Con quale identità entra? Usa le credenziali della persona che lo sta utilizzando oppure un account tecnico condiviso? Quel profilo può vedere soltanto i dati dell'utente o quelli di un'intera organizzazione? Il permesso scade al termine dell'operazione oppure rimane disponibile? Le azioni vengono registrate? Esiste un modo per distinguere ciò che ha fatto la persona da ciò che ha fatto il sistema per suo conto?

Queste domande diventano ancora più importanti quando l'agente attraversa servizi differenti. Un accesso a un calendario può sembrare innocuo finché non viene combinato con una casella di posta, una rubrica, un archivio di documenti e la possibilità di inviare messaggi. Nessuna di quelle connessioni, osservata da sola,

descrive necessariamente un potere enorme; ma insieme possono costruire un sistema capace di raccogliere informazioni, prendere decisioni e agire con una libertà che nessuna singola autorizzazione lasciava intuire.

È uno dei motivi per cui l'ecosistema delle decisioni conta più del singolo modello. Il comportamento finale nasce dall'incontro fra capacità del sistema, strumenti disponibili, identità utilizzate, regole di autorizzazione, conferme richieste e obiettivi assegnati. Attribuire tutto ciò alla "volontà dell'IA" significa comprimere una rete di decisioni in un solo soggetto narrativo, proprio nel momento in cui avremmo bisogno di distinguerne i nodi.

Il confine importante non è tra leggere e agire, ma tra reversibile e irreversibile

Non tutti i permessi hanno lo stesso peso. Leggere una bozza e inviarla a un cliente sono due operazioni diverse. Preparare un ordine e confermare il pagamento non sono la stessa cosa. Individuare un file inutile e cancellarlo definitivamente appartengono a due livelli differenti di rischio. Eppure, quando progettiamo un flusso automatizzato, la tentazione è spesso quella di eliminare proprio il punto nel quale un essere umano dovrebbe fermarsi a confermare l'azione, perché quel punto rallenta il processo e riduce il vantaggio dell'automazione.

Qui la supervisione umana smette di essere una formula astratta e diventa una scelta architetturale. Possiamo decidere che un agente prepari ma non invii, proponga ma non applichi, selezioni ma non elimini, compili ma non paghi. Possiamo anche decidere il contrario e concedergli l'intero percorso. Non esiste una risposta identica per ogni applicazione, ma esiste sempre una decisione sul punto nel quale il sistema può passare dall'elaborazione all'effetto.

Il paradosso è che più un agente diventa utile, più cresce la pressione a rimuovere quelle interruzioni. Se deve chiedere conferma a ogni passaggio, ci domandiamo quale vantaggio offra rispetto a un normale assistente. Se invece gli permettiamo di completare l'intero flusso, otteniamo il beneficio dell'autonomia proprio aumentando ciò che potrebbe accadere senza che nessuno intervenga. Il potere operativo nasce esattamente in questa tensione: non quando il modello diventa improvvisamente libero, ma quando noi decidiamo che interromperlo troppo spesso ne ridurrebbe il valore.

La chiave può essere data una volta sola e continuare ad aprire porte

C’è poi un problema temporale. Una decisione presa oggi può continuare a produrre effetti molto dopo che abbiamo smesso di pensarci. Un token di accesso può restare valido. Un’integrazione può rimanere attiva. Un account creato per una prova può finire in produzione. Un permesso temporaneamente allargato per risolvere un problema può non essere più ristretto. La storia della sicurezza informatica è piena di autorizzazioni sopravvissute allo scopo per cui erano state concesse; con gli agenti, però, quelle autorizzazioni possono essere utilizzate da sistemi capaci di combinare strumenti e informazioni in modi non previsti al momento della configurazione.

Per questo il controllo non riguarda soltanto ciò che concediamo, ma anche ciò che rivediamo, revochiamo e limitiamo nel tempo. Un permesso non è una proprietà naturale del sistema. È una decisione che può essere modificata. Quando smettiamo di trattarla come tale, l’accesso comincia a sembrare parte dell’agente stesso e la responsabilità della configurazione arretra nuovamente sullo sfondo.

Chi consegna le chiavi decide anche quanto può costare un errore

Un agente può sbagliare per molte ragioni: interpretare male un’istruzione, ricevere informazioni manipolate, seguire un obiettivo nel modo sbagliato, combinare strumenti in una sequenza inattesa o comportarsi in maniera che il progettista non aveva anticipato. Ma il danno possibile non dipende soltanto dalla qualità del ragionamento. Dipende anche da ciò che il sistema è autorizzato a toccare.

Lo stesso errore può produrre conseguenze molto diverse in un ambiente di prova e in un sistema collegato a dati reali. La stessa istruzione può essere innocua se l’agente può soltanto leggere e pericolosa se può anche modificare. Lo stesso comportamento inatteso può fermarsi davanti a una richiesta di conferma oppure trasformarsi immediatamente in un’azione irreversibile. Parlare di sicurezza senza parlare di permessi significa quindi osservare soltanto una metà del rischio.

La domanda “chi ha dato le chiavi?” non serve a trovare un colpevole automatico. Serve a ricostruire la catena. Chi ha scelto lo strumento? Chi ha deciso l’identità con cui avrebbe operato? Chi ha stabilito i permessi? Chi ha ritenuto superflua una conferma? Chi ha valutato che il beneficio dell’automazione giustificasse quel livello di accesso? Solo dopo queste domande possiamo capire che cosa appartiene davvero al comportamento del modello e che cosa, invece, appartiene all’ambiente che abbiamo costruito perché quel comportamento potesse produrre effetti.

La domanda da cui partire

È possibile che gli agenti del futuro siano molto più capaci di quelli attuali e che riescano a individuare autonomamente vulnerabilità, concatenare strumenti e trovare percorsi che nessun progettista aveva immaginato. Proprio per questo il problema degli accessi non diventa meno importante; diventa più importante. Più un sistema è capace di scegliere come raggiungere un obiettivo, più dobbiamo sapere quali territori gli abbiamo reso attraversabili.

La macchina può trovare una porta che non avevamo previsto. Può perfino scoprire un modo inatteso di usare una chiave. Ma prima che quella chiave possa aprire qualcosa, qualcuno deve aver deciso che il sistema potesse averla.

Per questo, davanti a un agente che compie un'azione sorprendente, la seconda domanda di Dietro la macchina è ancora più concreta della prima: chi gli ha dato le chiavi, quali porte aprivano e perché abbiamo deciso che dovesse poterle usare?

Fonti e riferimenti

National Institute of Standards and Technology, Least privilege e NIST SP 800-171 Rev. 3. Il principio stabilisce che utenti e processi dovrebbero ricevere soltanto le autorizzazioni e le risorse minime necessarie a svolgere i compiti assegnati, con revisione e rimozione dei privilegi quando non sono più necessari.

OWASP Gen AI Security Project, LLM06:2025 Excessive Agency. La vulnerabilità viene ricondotta in particolare a funzionalità eccessive, permessi eccessivi e autonomia eccessiva; fra le mitigazioni sono indicati la riduzione degli strumenti disponibili, il privilegio minimo e l'approvazione umana per le azioni ad alto impatto.

Anthropic, Measuring AI agent autonomy in practice, 2026. L'analisi studia l'autonomia attraverso l'uso concreto degli strumenti e mostra come permessi limitati e richieste di approvazione umana siano già componenti osservabili delle architetture agentiche reali, distinguendo l'autonomia del modello dalle condizioni operative nelle quali viene utilizzato.

Edizione online: https://inchiostrati.it/chi-ha-dato-le-chiavi/

MOVIMENTO 03

L’uomo nel circuito può essere soltanto una firma

Terzo movimento de «La responsabilità che scompare»: la presenza di una persona nel processo non garantisce, da sola, che esista un controllo umano reale. Conta se può comprendere, contestare, interrompere e cambiare davvero l’esito.

Nel movimento precedente abbiamo seguito il punto nel quale una capacità diventa operativa: credenziali, strumenti, API e permessi trasformano ciò che un sistema sa fare in ciò che può realmente fare. Ma anche quando decidiamo di non consegnargli l’intero percorso, resta una soluzione apparentemente rassicurante: tenere una persona nel circuito. Il sistema analizza, propone, seleziona o prepara; alla fine un essere umano guarda e conferma.

È una formula che sembra risolvere il problema quasi per definizione. Se c’è una persona, allora c’è controllo umano. Ma la presenza fisica o procedurale di qualcuno dentro un flusso non dice ancora nulla sulla qualità del controllo che quella persona esercita. Può avere pochi secondi per decidere, informazioni insufficienti, nessuna possibilità concreta di comprendere come si sia arrivati a una raccomandazione e un’organizzazione che considera il rifiuto un’anomalia da giustificare. Può persino possedere formalmente il diritto di dire no e trovarsi, nella pratica, nelle condizioni in cui dire no è più difficile che firmare.

La domanda del terzo movimento nasce qui: quando diciamo che un essere umano è “nel circuito”, stiamo descrivendo una persona capace di cambiare l’esito oppure soltanto qualcuno incaricato di assumersene formalmente la responsabilità?

Essere presenti non significa decidere

Un controllo reale richiede più della presenza. Richiede tempo, informazioni, competenza e soprattutto la possibilità di intervenire. Un operatore che riceve una proposta automatica e può soltanto premere “approva” non esercita lo stesso controllo di chi può esaminare gli elementi che hanno prodotto quella proposta, chiedere una verifica, modificarla, ignorarla o interrompere l’intero processo. Eppure, in entrambi i casi, potremmo descrivere il sistema dicendo che la decisione finale spetta a un essere umano.

La differenza diventa evidente quando spostiamo l’attenzione dalla persona all’architettura della scelta. Se il sistema seleziona in anticipo quali informazioni mostrare, ordina le alternative, assegna punteggi, nasconde ciò che considera irrilevante e presenta una sola opzione come raccomandata, buona parte della decisione può essere già avvenuta prima che l’operatore compaia sullo schermo. La firma resta umana; il campo nel quale quella firma si muove è stato costruito altrove.

È una distinzione che dialoga direttamente con L'officina, il testo e la firma. Là la firma serviva a ricostruire chi assume la responsabilità di un testo anche quando il processo che lo produce è distribuito. Qui la prospettiva si rovescia: una firma può anche diventare il punto nel quale una responsabilità distribuita viene concentrata su una persona che non ha realmente controllato ciò che la precede.

La firma può arrivare quando la decisione è già stata presa

Immaginiamo un sistema che esamini mille pratiche e ne segnali cinquanta come rischiose. L'operatore umano vede soltanto quelle cinquanta. Può approvare o respingere la classificazione, ma non vede le novecentocinquanta escluse dal sistema e quindi non può sapere quante siano state scartate correttamente, quante avrebbero meritato attenzione e quali criteri abbiano determinato la selezione iniziale. Formalmente l'essere umano decide sulle pratiche che ha davanti. Sostanzialmente, però, una decisione precedente ha già stabilito che cosa meriti di arrivare fino a lui.

Lo stesso accade quando un sistema ordina candidati, suggerisce diagnosi, individua transazioni sospette, assegna priorità a richieste, propone sanzioni o raccomanda quali casi approfondire. Non è necessario che la macchina prenda l'ultima decisione perché possa orientare profondamente il processo. Basta che decida quali possibilità entrano nel campo visivo dell'essere umano e con quale peso vengono presentate.

Per questo la formula "decisione umana finale" può essere vera e, allo stesso tempo, insufficiente. Descrive l'ultimo gesto, non la storia che lo ha reso probabile. Se vogliamo capire dove si trova il controllo, dobbiamo ricostruire anche ciò che accade prima della firma.

L'automation bias non è un dettaglio psicologico

Esiste poi un secondo problema: la tendenza a fidarsi dell'output automatizzato proprio perché proviene da un sistema percepito come più coerente, più rapido o più capace di elaborare informazioni di quanto potrebbe fare una singola persona. Questa tendenza non rende l'essere umano ingenuo e non trasforma l'errore in una colpa individuale. È una conseguenza prevedibile del modo in cui organizziamo il lavoro attorno ai sistemi automatici.

Se un modello produce cento valutazioni corrette, la centounesima tende naturalmente a essere accolta con meno diffidenza. Se un operatore deve controllare centinaia di risultati al giorno, leggere ogni caso come se il sistema potesse aver sbagliato annulla buona parte del vantaggio che l'automazione prometteva. Se l'organizzazione misura la velocità e considera il rifiuto della raccomandazione automatica un evento eccezionale, il controllo umano può trasformarsi progressivamente da verifica a conferma.

Non è casuale che la normativa europea sui sistemi di IA ad alto rischio nomini

esplicitamente la possibile tendenza a fare automaticamente o eccessivamente affidamento sull'output del sistema. Il punto interessante non è soltanto che l'errore umano debba essere evitato. È che una supervisione efficace deve essere progettata tenendo conto del fatto che l'essere umano può essere influenzato proprio dall'autorità che attribuisce alla macchina.

Il tempo decide quanto è umano il controllo

Un supervisore può possedere tutte le competenze necessarie e restare comunque incapace di esercitarle se il processo non gli concede il tempo sufficiente. Dieci minuti per esaminare un caso complesso e dieci secondi per confermare una raccomandazione non sono due versioni della stessa supervisione. Sono due sistemi diversi.

La velocità è uno dei motivi per cui automatizziamo. Ma proprio per questo può diventare il meccanismo che svuota il controllo umano. Se un agente completa in pochi secondi un compito che richiederebbe minuti o ore di verifica, l'organizzazione deve scegliere se conservare il tempo necessario al controllo oppure utilizzare l'automazione per aumentare il volume delle operazioni. Quando sceglie la seconda strada, la persona rimasta nel circuito rischia di diventare il collo di bottiglia che tutti cercheranno di comprimere.

È lo stesso paradosso che avevamo incontrato in Riprendere il controllo: il controllo costa tempo, procedure e rinunce all'efficienza. Qui quel costo assume una forma molto concreta. Possiamo mantenere un essere umano nel processo e, contemporaneamente, sottrargli le condizioni materiali necessarie per controllare davvero.

Per poter dire no bisogna avere potere, competenza e informazioni

Una persona può essere tecnicamente autorizzata a bloccare una decisione e non possedere l'autorità organizzativa per farlo senza conseguenze. Può sapere che un sistema sbaglia in alcuni casi ma non conoscere abbastanza bene i suoi limiti per riconoscere quando ciò stia accadendo. Può essere competente nel dominio — medicina, finanza, selezione del personale, sicurezza — e non conoscere il funzionamento dello strumento che sta supervisionando. Oppure può comprendere lo strumento ma non avere accesso ai dati e al contesto necessari per giudicarne l'output.

Dire che esiste supervisione umana senza distinguere questi elementi rischia di trasformare una struttura organizzativa complessa in una semplice casella da spuntare. Una persona è presente: requisito soddisfatto. Ma ciò che conta non è che qualcuno occupi una posizione nel diagramma del processo. Conta se quella persona può comprendere abbastanza, dispone del tempo per farlo, possiede l'autorità per intervenire e può esercitare quell'autorità senza che l'intero sistema organizzativo la

spinga nella direzione opposta.

Il NIST insiste proprio sulla necessità di definire e distinguere ruoli e responsabilità nelle configurazioni uomo-IA e di valutare i processi di supervisione, non limitandosi a presumere che la presenza di un esperto garantisca automaticamente un controllo efficace. È una precisazione importante perché sposta ancora una volta l'attenzione dalla figura simbolica dell'essere umano alle condizioni nelle quali quell'essere umano opera.

La supervisione deve poter cambiare l'esito

Possiamo allora formulare un criterio semplice: il controllo umano esiste davvero quando l'intervento umano può cambiare qualcosa. Non deve necessariamente significare che ogni decisione venga modificata o che l'operatore diffidi sistematicamente del sistema. Significa che deve poter ignorare un output, ribaltarlo, chiedere una verifica, fermare un'azione o interrompere il sistema quando ritiene che continuare sia rischioso.

Se nessuna di queste possibilità è concreta, la supervisione diventa decorativa. Un pulsante di arresto che nessuno è autorizzato a premere, una raccomandazione che teoricamente si può respingere ma che nella pratica viene sempre accettata, un operatore che firma senza poter ricostruire il processo non rappresentano altrettante forme di controllo. Rappresentano modi diversi di conservare la presenza umana mentre il potere decisionale si sposta altrove.

È qui che la questione incontra L'ecosistema delle decisioni. Il controllo non appartiene necessariamente a un solo punto. Può essere distribuito fra chi progetta l'interfaccia, chi decide i tempi, chi stabilisce le metriche, chi configura le soglie, chi assegna l'autorità all'operatore e chi determina che cosa accada quando l'operatore non segue la raccomandazione automatica. Guardare soltanto alla persona che preme l'ultimo pulsante significa vedere l'ultimo anello e scambiarlo per l'intera catena.

Quando la firma concentra una responsabilità che il processo ha disperso

La presenza di un essere umano può quindi svolgere due funzioni opposte. Può essere un vero presidio, capace di introdurre giudizio, contesto e possibilità di arresto. Oppure può diventare il punto nel quale un processo altamente automatizzato deposita la propria responsabilità formale. In questo secondo caso l'uomo non scompare dal sistema; resta esattamente dove serve perché qualcuno possa dire che la decisione finale era umana.

È una forma più sottile di quella responsabilità che stiamo seguendo lungo tutto il percorso. Nei primi due movimenti abbiamo visto come il controllo venga ceduto attraverso autonomia, strumenti e accessi. Qui può accadere qualcosa di diverso: il potere operativo viene distribuito lungo il sistema, mentre la responsabilità viene

ricondotta all’ultimo essere umano che compare nella sequenza. La macchina non diventa responsabile, ma nemmeno l’organizzazione sembra più esserlo completamente. Rimane una firma.

Per questo non basta chiedere se ci fosse un essere umano nel circuito. Dobbiamo chiedere che cosa potesse realmente fare, quali informazioni avesse, quanto tempo disponesse, quali conseguenze avrebbe incontrato opponendosi al sistema e chi avesse progettato quelle condizioni.

La domanda da cui partire

La supervisione umana non è una proprietà che compare automaticamente quando inseriamo una persona in un diagramma. È un insieme di capacità, autorità, informazioni, tempo e possibilità di intervento. Può essere forte o debole, effettiva o puramente formale. Può correggere il sistema oppure limitarsi a legittimarlo.

Non si tratta di sostenere che ogni decisione assistita dall’intelligenza artificiale nasconda una firma vuota. In molti casi l’interazione fra competenza umana e capacità automatica può migliorare davvero il risultato. Ma proprio per distinguere quei casi da quelli in cui l’essere umano è soltanto una presenza rituale, dobbiamo smettere di usare “human in the loop” come una risposta e cominciare a trattarlo come una domanda.

Quella domanda è molto concreta: la persona nel circuito poteva davvero dire no — e cambiare ciò che sarebbe accaduto?

Fonti e riferimenti

Unione europea, Regolamento (UE) 2024/1689 sull’intelligenza artificiale, articolo 14 e considerando 73. La norma richiede che la sorveglianza umana dei sistemi ad alto rischio sia effettiva e proporzionata al rischio, al livello di autonomia e al contesto d’uso; prevede inoltre che chi supervisiona possa comprendere capacità e limiti, riconoscere il rischio di automation bias, ignorare o ribaltare l’output e interrompere il sistema, e richiama competenza, formazione e autorità delle persone incaricate.

National Institute of Standards and Technology, Towards a Standard for Identifying and Managing Bias in Artificial Intelligence, NIST SP 1270, 2022. Il documento mette in discussione l’assunzione secondo cui la semplice presenza di un essere umano, anche esperto, garantisca una supervisione efficace e sottolinea quanto le configurazioni uomo-IA dipendano da strutture organizzative, ruoli, competenze e bias cognitivi.

National Institute of Standards and Technology, AI Risk Management Framework — Appendix C: AI Risk Management and Human-AI Interaction, 2023. Il framework richiede di definire e differenziare chiaramente i ruoli umani nella decisione e nella supervisione e considera le configurazioni uomo-IA come parte integrante della gestione del rischio.

OECD, OECD AI Principles, aggiornati nel 2024. I principi collegano supervisione umana, possibilità di intervento, tracciabilità e accountability e insistono sul fatto che gli attori dell’IA restino responsabili in base al proprio ruolo, al contesto e alla propria capacità di agire.

Edizione online: https://inchiostrati.it/luomo-nel-circuito-puo-essere-soltanto-una-firma/

MOVIMENTO 04

Quanto costa mantenere il controllo?

Quarto movimento de «La responsabilità che scompare»: mantenere il controllo richiede tempo, persone, calcolo e rinunce. Il problema nasce quando il costo della prudenza entra in conflitto con velocità, utilità e competizione.

Nel movimento precedente abbiamo visto che lasciare una persona nel circuito non basta, da solo, a garantire un controllo umano reale. Il controllo esiste quando quella persona dispone del tempo, delle informazioni, dell'autorità e degli strumenti necessari per cambiare davvero l'esito. Ma proprio qui compare una domanda che il linguaggio della sicurezza tende spesso a lasciare sullo sfondo: quanto costa mantenere tutte queste condizioni?

Il costo non è soltanto economico. Ogni verifica richiede tempo. Ogni conferma rallenta un flusso. Ogni limite riduce almeno una parte delle possibilità operative. Ogni sistema di monitoraggio deve essere progettato, mantenuto e aggiornato. Ogni supervisore deve essere formato. Ogni procedura capace di fermare un'azione introduce un punto nel quale l'automazione può non procedere alla velocità massima possibile. Se vogliamo capire perché alcuni controlli vengono ridotti, aggirati o trasformati in formalità, dobbiamo allora smettere di trattarli come elementi gratuiti che basta aggiungere a un sistema. Il controllo ha un prezzo, e quel prezzo entra direttamente nelle decisioni.

Ogni controllo interrompe qualcosa

Un sistema automatico viene introdotto quasi sempre per ottenere un vantaggio: ridurre il tempo necessario a svolgere un compito, aumentare il numero di operazioni gestibili, diminuire i costi, rendere disponibile un servizio senza interruzioni, reagire più rapidamente o coordinare una quantità di informazioni che una persona non riuscirebbe a seguire da sola. Il controllo, però, funziona spesso nella direzione opposta. Chiede di fermarsi, verificare, registrare, confrontare, autorizzare, attendere.

Questo non significa che controllo e automazione siano incompatibili. Significa che il loro rapporto contiene una tensione strutturale. Se un agente può completare in pochi secondi un processo che prima richiedeva mezz'ora, inserire una verifica umana di dieci minuti può essere perfettamente ragionevole dal punto di vista della sicurezza e, contemporaneamente, ridurre gran parte del vantaggio che aveva giustificato l'automazione. Se un'azienda utilizza un sistema perché può gestire migliaia di casi, pretendere che ogni caso venga riesaminato da una persona può riportare il limite di scala esattamente nel punto dal quale volevamo allontanarci.

È il nodo che avevamo già incontrato in Riprendere il controllo: oltre una certa complessità, controllare significa inevitabilmente decidere quali parti osservare direttamente e quali affidare a strumenti automatici. Il quarto movimento aggiunge

però un elemento ulteriore. Non dobbiamo chiederci soltanto dove si sposta il controllo, ma quale costo produce ogni volta che decidiamo di mantenerlo.

La sicurezza consuma tempo, persone e capacità di calcolo

Una misura di sicurezza non vive nel vuoto. Deve essere progettata, testata, documentata, applicata e monitorata. Un sistema di valutazione richiede casi di prova, metriche, persone capaci di interpretarli e procedure per decidere che cosa accada quando il risultato non è rassicurante. Un controllo sugli accessi richiede ruoli, autorizzazioni, registri e revisioni periodiche. Un meccanismo di supervisione richiede operatori preparati e una struttura organizzativa che permetta loro di intervenire. Anche quando parte di queste attività viene automatizzata, rimangono infrastrutture, calcolo, manutenzione e decisioni da sostenere nel tempo.

Il NIST tratta esplicitamente la gestione del rischio come un'attività che deve essere dotata di risorse e responsabilità organizzative, fino a citare risorse umane e di budget per chi ha il compito di governare i rischi dell'IA. Non è un dettaglio amministrativo: significa riconoscere che la sicurezza compete per le stesse risorse che potrebbero essere utilizzate per sviluppo, distribuzione, supporto o nuove funzionalità. Ogni organizzazione deve quindi decidere non soltanto quali rischi considera accettabili, ma anche quanto è disposta a spendere per ridurli.

Lo stesso vale per le salvaguardie tecniche. Classificatori in tempo reale, monitoraggio asincrono, verifiche più profonde, controlli di accesso e procedure di risposta agli incidenti non sono semplici idee aggiunte a margine del sistema. Sono un'altra parte del sistema. E più diventano sofisticate, più richiedono progettazione, aggiornamenti, calcolo e persone in grado di interpretarne i segnali.

Il costo più difficile è ciò a cui dobbiamo rinunciare

Esiste però un costo ancora meno visibile: ciò che non facciamo perché abbiamo scelto di mantenere un controllo. Un limite sugli accessi può rendere un agente meno utile. Una conferma obbligatoria può impedire un'esecuzione completamente autonoma. Una verifica più approfondita può ritardare il rilascio di una nuova capacità. Una soglia di sicurezza può costringerci a rinunciare temporaneamente a un utilizzo che sarebbe tecnicamente possibile.

È facile accettare la prudenza finché la descriviamo in astratto. Diventa più difficile quando possiamo indicare con precisione che cosa perderemo per mantenerla: settimane di sviluppo, una funzione che il concorrente offre già, una quota di mercato, una riduzione dei margini, una promessa commerciale non rispettata, un vantaggio strategico che qualcun altro potrebbe conquistare. A quel punto il controllo smette di essere soltanto un principio e diventa una scelta fra benefici concreti che arrivano oggi e rischi che, in molti casi, rimangono incerti.

Qui il percorso incontra direttamente Il primo a frenare perde. Una parte della pressione ad accelerare nasce proprio dalla convinzione che la prudenza abbia un costo competitivo. Se penso che tutti manterranno gli stessi controlli, quel costo può apparire sopportabile; se temo che qualcun altro li riduca, la stessa misura di sicurezza può improvvisamente sembrare uno svantaggio che sto imponendo soltanto a me stesso.

Quando la concorrenza entra nella definizione di sicurezza

Questo punto non appartiene soltanto alla teoria. I framework pubblicati dalle stesse organizzazioni che sviluppano modelli avanzati mostrano che la sicurezza viene pensata dentro un ambiente competitivo reale, non fuori da esso. OpenAI, nel suo Preparedness Framework aggiornato, prevede che un cambiamento nel comportamento degli altri sviluppatori possa modificare il contesto nel quale vengono valutati i requisiti di salvaguardia, pur dichiarando che eventuali aggiustamenti dovrebbero mantenere il rischio complessivo a livelli protettivi.

È un passaggio importante perché rende visibile qualcosa che spesso viene nascosto dietro la parola “prudenza”. Una misura di sicurezza non viene valutata soltanto per ciò che impedisce, ma anche per il modo in cui modifica la posizione di chi la adotta rispetto agli altri. La concorrenza può quindi entrare nella definizione stessa di ciò che un’organizzazione considera sostenibile.

Non significa che ogni controllo venga ridotto per ragioni commerciali né che le dichiarazioni sulla sicurezza siano semplicemente retoriche. Al contrario, il problema diventa più interessante proprio quando gli attori sono sinceramente preoccupati. Possono desiderare salvaguardie forti e contemporaneamente temere che un sistema troppo lento, troppo limitato o troppo costoso venga superato da uno meno prudente. In quel momento il conflitto non è fra sicurezza e irresponsabilità, ma fra due rischi diversi: quello di procedere troppo velocemente e quello di restare indietro.

Rendere il controllo scalabile significa automatizzarlo

Quando il costo del controllo cresce troppo, la risposta più naturale consiste nel cercare di automatizzarlo. Se non possiamo permetterci che ogni contenuto venga esaminato da una persona, costruiamo classificatori automatici. Se una verifica profonda richiede troppo tempo, utilizziamo un primo livello più rapido che selezioni soltanto i casi sospetti. Se il monitoraggio umano non riesce a seguire il volume delle operazioni, costruiamo sistemi che osservano altri sistemi e chiamano una persona soltanto quando viene superata una soglia.

È una soluzione ragionevole e spesso necessaria. Anthropic, descrivendo le proprie salvaguardie di deployment, parla proprio di una difesa a più livelli: controlli di

accesso, classificatori in tempo reale, monitoraggio asincrono e procedure di risposta. La stessa architettura mostra però il paradosso che stiamo seguendo. Per mantenere il controllo su sistemi sempre più veloci e capaci, dobbiamo costruire un controllo che sia a sua volta veloce, scalabile e in parte automatico.

Questo ci riporta a Quando la paura trova una soluzione. Una protezione efficace non si limita a ridurre il rischio: può rendere possibile una scala che senza quella protezione avremmo considerato ingestibile. La sicurezza permette di continuare, ma proprio per questo diventa una parte dell'infrastruttura che consente al sistema di crescere.

Le salvaguardie diventano parte del prodotto

A un certo punto il confine fra sistema e controllo comincia a diventare meno netto. Il modello, i filtri, i livelli di accesso, i monitoraggi, le verifiche e le procedure di escalation formano insieme ciò che viene realmente distribuito e utilizzato. Non esiste più una tecnologia da una parte e la sua sicurezza dall'altra: la sicurezza diventa una delle condizioni attraverso cui quella tecnologia può essere resa disponibile.

Questa trasformazione cambia anche il modo in cui valutiamo il costo. Se una salvaguardia introduce troppa latenza, cerchiamo di renderla più rapida. Se utilizza troppo calcolo, proviamo a riservare le verifiche più costose ai casi selezionati. Se richiede troppe persone, automatizziamo una parte del lavoro. Non stiamo necessariamente indebolendo il controllo; stiamo cercando di renderlo compatibile con l'utilità del sistema. Ma ogni ottimizzazione modifica ciò che il controllo vede, quando interviene e quanto profondamente verifica.

È qui che la domanda economica diventa una domanda di architettura. Il costo non decide soltanto quante risorse dedichiamo alla sicurezza. Può contribuire a decidere quale forma assumerà la sicurezza stessa.

Il rischio di misurare la sicurezza soltanto attraverso l'efficienza

Un controllo ben progettato dovrebbe essere proporzionato al rischio, non massimizzato senza criterio. Nessun sistema potrebbe funzionare se ogni operazione richiedesse il livello di verifica necessario per una decisione irreversibile e ad alto impatto. Il problema nasce quando l'efficienza smette di essere una condizione da bilanciare e diventa il criterio dominante attraverso cui giudichiamo il controllo.

Se una verifica viene considerata buona soprattutto perché non rallenta il prodotto, possiamo finire per misurare la sicurezza attraverso la sua capacità di diventare invisibile. Se un intervento umano è valutato positivamente perché non crea attrito, possiamo progettare un ruolo umano che confermi molto e contesti poco. Se un sistema di monitoraggio deve essere abbastanza economico da funzionare su ogni operazione, possiamo accettare che le analisi più profonde vengano eseguite soltanto

su una frazione dei casi.

Non c’è necessariamente qualcosa di sbagliato in queste scelte. In molti casi sono l’unico modo realistico per rendere un controllo utilizzabile. Ma proprio perché sono decisioni ragionevoli devono restare visibili. Il rischio non è soltanto ridurre una salvaguardia. È dimenticare che l’abbiamo ridotta per ottenere qualcos’altro e cominciare a raccontare il sistema risultante come se rappresentasse semplicemente il livello “naturale” di sicurezza possibile.

La domanda da cui partire

Il controllo non è gratuito, e fingere che lo sia rende più difficile capire perché venga ceduto. Richiede tempo, persone, calcolo, procedure, formazione e soprattutto rinunce. Ogni volta che scegliamo di mantenere un limite dobbiamo accettare ciò che quel limite ci impedisce di ottenere immediatamente.

Questo non significa che il costo giustifichi automaticamente la riduzione del controllo. Significa il contrario: se vogliamo attribuire responsabilità alle decisioni, dobbiamo rendere visibile anche il momento in cui qualcuno ha deciso che una salvaguardia costava troppo, rallentava troppo o riduceva troppo il vantaggio atteso.

La domanda del quarto movimento non è quindi soltanto “quanto costa essere prudenti?”. È più precisa: quando riduciamo un controllo perché costa tempo, denaro, velocità o competitività, chi decide che quel risparmio vale il rischio che rimane?

Fonti e riferimenti

National Institute of Standards and Technology, NIST AI RMF Playbook — Govern. Il Playbook collega la gestione del rischio a responsabilità organizzative esplicite e alla disponibilità di risorse, comprese risorse umane e di budget, mostrando che la supervisione non è una condizione astratta ma un’attività da sostenere nel tempo.

National Institute of Standards and Technology, AI Risk Management Framework 1.0 — Core. La funzione Govern descrive la gestione del rischio come un’attività continua lungo il ciclo di vita del sistema e richiede processi, ruoli, monitoraggio periodico e risorse coerenti con le priorità di rischio dell’organizzazione.

OpenAI, Our updated Preparedness Framework, 2025. Il framework collega determinate soglie di capacità a salvaguardie operative e prevede valutazioni ulteriori o protezioni più forti prima del deployment; riconosce inoltre che cambiamenti nel comportamento competitivo di altri sviluppatori possono modificare il contesto nel quale vengono valutati i requisiti.

Anthropic, Responsible Scaling Policy, versione aggiornata 2026. La policy lega soglie di capacità a livelli crescenti di salvaguardia e descrive un’architettura di deployment a più strati, comprendente controlli di accesso, classificazione in tempo reale, monitoraggio asincrono ed escalation umana, con attenzione esplicita al bilanciamento fra sicurezza, utilità, scalabilità e prestazioni.

Edizione online: https://inchiostrati.it/quanto-costa-mantenere-il-controllo/

MOVIMENTO 05

Il vantaggio non si distribuisce come il rischio

Quinto movimento de «La responsabilità che scompare»: vantaggi e rischi di una decisione automatizzata non ricadono necessariamente sugli stessi soggetti. Per capire la responsabilità bisogna ricostruire entrambe le mappe.

Nel movimento precedente abbiamo seguito il costo del controllo: tempo, persone, calcolo, procedure, rinunce e rallentamenti. Ma ogni costo ha sempre un'altra faccia. Se ridurre una salvaguardia permette di arrivare prima, fare di più, spendere meno o automatizzare una parte maggiore del processo, qualcuno riceve un vantaggio. Ed è proprio qui che la domanda sulla responsabilità diventa più precisa: il vantaggio prodotto da una decisione e il rischio che quella stessa decisione trasferisce non finiscono necessariamente nelle stesse mani.

Un'organizzazione può beneficiare di maggiore velocità, di minori costi o di una scala operativa prima impossibile. Un dirigente può raggiungere un obiettivo di produttività. Un fornitore può vendere un servizio più competitivo. Un utente può ottenere una risposta più rapida. Intanto, però, l'errore può ricadere su una persona che non ha scelto il sistema, su un lavoratore che deve correggerne gli effetti, su un cliente che non conosce i criteri con cui è stato valutato, su una comunità che assorbe conseguenze che nessuna metrica interna all'organizzazione registra. Parlare di "benefici dell'IA" o di "rischi dell'IA" senza chiedere chi riceva gli uni e chi sopporti gli altri significa mettere nello stesso contenitore posizioni che possono essere molto diverse.

Il beneficio ha sempre un destinatario

Quando diciamo che un sistema è utile, dovremmo poter completare la frase: utile per chi? Un modello che riduce il tempo necessario a esaminare una pratica può essere utile all'organizzazione che deve gestirne migliaia. Può esserlo anche alla persona che riceve una risposta più rapidamente. Ma lo stesso sistema può rendere più difficile comprendere perché una pratica sia stata respinta, può spostare lavoro di verifica su operatori già sovraccarichi o può produrre errori che diventano visibili soltanto quando qualcuno li subisce.

Non c'è nulla di sospetto nel fatto che un'impresa cerchi efficienza o che un'amministrazione tenti di usare meglio le proprie risorse. Il problema nasce quando il vantaggio viene descritto come se appartenesse automaticamente a tutti coloro che entrano in contatto con il sistema. La produttività di chi offre un servizio, la comodità di chi lo usa, il margine economico di chi lo vende e il benessere di chi viene valutato da quel servizio sono grandezze diverse. Possono coincidere, ma non lo fanno per definizione.

I principi OCSE sull’intelligenza artificiale insistono proprio su questo passaggio: l’obiettivo non è soltanto produrre crescita, ma orientare l’IA verso risultati benefici per le persone e il pianeta e verso una crescita inclusiva. È una distinzione importante perché impedisce di trasformare un vantaggio misurato dentro un’organizzazione in una prova automatica del beneficio complessivo prodotto dal sistema.

Chi non usa il sistema può subirne gli effetti

La distribuzione del rischio diventa ancora più evidente quando osserviamo le persone che non hanno alcun rapporto diretto con la tecnologia. Un individuo può non scegliere, acquistare o utilizzare un sistema e tuttavia essere valutato da esso, classificato attraverso i suoi output, escluso da un processo, sottoposto a una verifica più intensa o semplicemente costretto ad adattarsi a una decisione che il sistema ha contribuito a produrre.

Il NIST distingue esplicitamente fra utenti finali e persone o comunità interessate dagli effetti di un sistema. Queste ultime possono essere direttamente o indirettamente coinvolte senza interagire affatto con l’applicazione. Il Framework chiede inoltre di caratterizzare gli impatti su individui, gruppi, comunità, organizzazioni e società e di valutare sia effetti benefici sia effetti dannosi. È un modo tecnico per rendere visibile una cosa molto semplice: la mappa di chi riceve i benefici non coincide necessariamente con la mappa di chi può ricevere il danno.

La stessa logica compare nel regolamento europeo sull’IA quando, per determinati sistemi ad alto rischio, impone a specifici deployer una valutazione d’impatto sui diritti fondamentali che identifichi le categorie di persone e gruppi verosimilmente interessati dall’uso del sistema e i rischi di danno che possono ricadere su di loro. La domanda non è quindi soltanto che cosa ottenga chi introduce la tecnologia, ma chi entra nel raggio delle sue conseguenze.

Il rischio può uscire dal perimetro che misura il successo

Ogni organizzazione ha bisogno di indicatori. Misura tempi, costi, errori, conversioni, produttività, qualità del servizio, disponibilità, vendite o qualunque altra grandezza sia rilevante per il proprio scopo. Il problema è che ciò che non viene misurato può comunque esistere. Un processo automatizzato può risultare più efficiente secondo tutte le metriche interne e, nello stesso tempo, trasferire una parte del proprio costo su persone che non compaiono in quelle metriche.

Un sistema può ridurre il lavoro di un ufficio aumentando quello necessario a chi deve contestarne gli errori. Può ridurre i tempi medi di risposta producendo casi marginali molto più difficili da correggere. Può abbassare il costo di una decisione per chi la prende e aumentare il costo di comprenderla per chi la subisce. Può perfino migliorare la qualità media e peggiorare in modo significativo l’esperienza di un gruppo relativamente piccolo, che proprio perché piccolo scompare facilmente dentro il dato

aggregato.

Non serve immaginare cattive intenzioni. È sufficiente che il sistema venga ottimizzato rispetto agli obiettivi di chi lo controlla e che gli effetti esterni entrino nella valutazione soltanto quando qualcuno decide di cercarli. Per questo il NIST non si limita a chiedere che vengano mappati rischi e benefici dei componenti del sistema, ma include gli impatti su soggetti esterni al gruppo che lo ha sviluppato o distribuito e invita a integrare il feedback proveniente da chi ne sperimenta gli effetti.

Distribuire le decisioni non distribuisce automaticamente il vantaggio

Abbiamo già visto in L'ecosistema delle decisioni che una scelta assistita dall'intelligenza artificiale nasce da una catena: chi definisce il problema, chi costruisce il modello, chi acquista il servizio, chi stabilisce i dati e le soglie, chi organizza il processo, chi riceve la raccomandazione e chi compie l'ultimo gesto. Questa distribuzione rende insufficiente cercare un unico punto nel quale "la macchina ha deciso".

Ma distribuire il processo non significa che benefici e rischi vengano distribuiti nello stesso modo. Un fornitore può ottenere ricavi, un'organizzazione può ottenere efficienza, un responsabile può raggiungere un obiettivo operativo e l'operatore finale può ritrovarsi con il compito di gestire gli errori prodotti da una struttura che non ha scelto. La persona sulla quale ricade la decisione, infine, può non ricevere nessuno dei vantaggi che hanno motivato l'intera catena.

È qui che la responsabilità rischia di diventare particolarmente sfuggente. Ogni attore può descrivere correttamente il proprio contributo come parziale, mentre il vantaggio complessivo rimane abbastanza concreto da giustificare l'adozione del sistema. Se invece compare un danno, quella stessa frammentazione può rendere difficile individuare chi avesse il potere effettivo di impedirlo.

Il successo sale, l'errore scende

Esiste una asimmetria narrativa che vale la pena osservare. Quando un sistema funziona bene, il successo tende a risalire verso il progetto: si parla della qualità del modello, dell'efficienza dell'organizzazione, dell'innovazione introdotta, della capacità di automatizzare un processo che prima richiedeva più tempo o più persone. Quando qualcosa va storto, la responsabilità può invece scendere lungo la catena fino all'ultimo essere umano che ha toccato il processo.

Questo meccanismo era già emerso nell'ecosistema delle decisioni con la figura dell'essere umano trasformato in parafulmine morale: la persona più vicina all'errore può diventare il luogo nel quale viene concentrata una responsabilità che non corrisponde al potere reale posseduto durante la progettazione, l'adozione o l'organizzazione del sistema.

Il quinto movimento aggiunge un elemento: mentre la responsabilità può scendere, il vantaggio non necessariamente la segue. Chi ha ottenuto riduzioni di costo, maggiore velocità, una nuova capacità commerciale o una posizione competitiva può restare molto lontano dal punto nel quale l'errore diventa concreto. La distanza fra vantaggio e conseguenza non è creata dall'intelligenza artificiale, ma l'automazione può renderla più difficile da vedere perché aumenta il numero di passaggi intermedi e rende più semplice descrivere ogni singolo attore come soltanto una parte della catena.

Chi beneficia deve restare dentro la storia del rischio

Non significa che chi trae vantaggio da un sistema debba automaticamente rispondere di ogni conseguenza prodotta da esso. Sarebbe una scorciatoia speculare a quella che stiamo cercando di evitare. Un fornitore non controlla ogni uso del proprio modello, un deployer non controlla ogni dettaglio della progettazione, un operatore non conosce necessariamente ogni scelta compiuta a monte. La responsabilità deve seguire ruoli, contesto e capacità effettiva di agire.

È precisamente il criterio adottato dai principi OCSE sull'accountability: gli attori dell'IA dovrebbero essere responsabili in funzione del proprio ruolo, del contesto e della propria capacità di intervento. La formula è utile perché impedisce sia di attribuire tutto alla macchina sia di concentrare tutto sull'ultimo essere umano della catena. Costringe invece a ricostruire chi poteva fare che cosa e in quale momento.

Ma proprio per questo il vantaggio non può essere espulso dalla ricostruzione. Sapere chi guadagna da una configurazione, da un'automazione più spinta, dalla riduzione di un controllo o dall'introduzione di un nuovo sistema aiuta a capire perché certe decisioni siano state prese e perché alcuni rischi siano stati considerati accettabili. Gli interessi non provano una colpa, ma spiegano una parte della causalità.

La persona colpita non è un dettaglio a valle

Una delle conseguenze più importanti di questa prospettiva è che chi subisce l'effetto di una decisione non può essere trattato come un elemento che compare soltanto alla fine. Se un sistema è progettato per incidere su persone, il modo in cui quelle persone possono essere favorite, penalizzate, escluse, confuse o costrette a contestare un risultato deve entrare nella progettazione e nella governance molto prima che il primo caso concreto arrivi.

L'UNESCO, nella Raccomandazione sull'etica dell'intelligenza artificiale, lega responsabilità e accountability a tracciabilità, audit, valutazioni d'impatto e due diligence. L'OCSE collega la trasparenza alla possibilità per chi è negativamente interessato da un sistema di comprenderne l'esito e contestarlo. Il regolamento europeo, nei casi in cui richiede una valutazione d'impatto sui diritti fondamentali, chiede esplicitamente di identificare in anticipo le categorie di persone e gruppi potenzialmente colpiti.

Tutte queste formulazioni convergono su un punto che per Dietro la macchina è centrale: le conseguenze non cominciano quando il sistema ha già agito. Cominciano nel momento in cui qualcuno decide quali effetti sia disposto a rendere possibili e su chi possano ricadere.

La domanda da cui partire

Dopo aver ricostruito capacità, accessi, autonomia e costo del controllo, possiamo finalmente aggiungere una domanda che spesso viene trattata come secondaria: chi ottiene qualcosa dalla decisione di automatizzare, delegare, accelerare o ridurre una salvaguardia?

Non per trasformare ogni vantaggio in un sospetto. Un beneficio può essere reale, condiviso e perfettamente legittimo. Ma se vogliamo capire perché un certo rischio sia stato accettato, dobbiamo sapere anche quale beneficio lo rendeva accettabile e per chi. Solo allora possiamo confrontare le due mappe: quella di chi guadagna e quella di chi può perdere.

La domanda del quinto movimento è quindi questa: quando il vantaggio e il rischio finiscono su soggetti diversi, chi ha il potere di decidere che quella distribuzione sia accettabile?

Fonti e riferimenti

OECD, OECD AI Principles, aggiornati nel 2024. I principi distinguono fra crescita economica e risultati benefici inclusivi, e stabiliscono che gli attori dell’IA debbano essere accountable in funzione del ruolo, del contesto e della capacità di agire, mantenendo tracciabilità e gestione del rischio lungo l’intero ciclo di vita.

National Institute of Standards and Technology, AI Risk Management Framework 1.0 — Core. La funzione Map richiede di mappare rischi e benefici, caratterizzare gli impatti su individui, gruppi, comunità, organizzazioni e società e integrare anche il feedback di soggetti esterni al gruppo che ha sviluppato o distribuito il sistema.

Unione europea, Regolamento (UE) 2024/1689 sull’intelligenza artificiale, articolo 27. Per determinati deployer di sistemi ad alto rischio la valutazione d’impatto sui diritti fondamentali deve identificare le categorie di persone e gruppi verosimilmente interessati e i rischi di danno che l’uso del sistema può produrre su di loro.

UNESCO, Recommendation on the Ethics of Artificial Intelligence. La Raccomandazione collega responsabilità e accountability a supervisione, tracciabilità, valutazione d’impatto, audit e due diligence, con l’obiettivo di evitare danni e conflitti con i diritti umani e il benessere ambientale.

Edizione online: https://inchiostrati.it/il-vantaggio-non-si-distribuisce-come-il-rischio/

MOVIMENTO 06

Chi risponde quando nessuno decide da solo?

Sesto movimento de «La responsabilità che scompare»: quando una decisione nasce da una catena distribuita, la responsabilità non scompare. Deve seguire ruoli, poteri, capacità di intervento e possibilità concreta di correggere l'esito.

Abbiamo seguito il controllo mentre veniva concesso, distribuito, trasformato in supervisione, confrontato con il proprio costo e separato dai vantaggi che produce. A questo punto rimane l'ultima domanda del percorso: chi risponde delle conseguenze? La risposta sembra semplice finché immaginiamo una decisione presa da una sola persona. Diventa molto meno semplice quando il risultato nasce da una catena nella quale nessuno, da solo, ha progettato ogni passaggio, scelto ogni parametro, autorizzato ogni uso e compiuto ogni azione.

La difficoltà non nasce perché la responsabilità sia scomparsa davvero. Nasce perché il processo è diventato abbastanza distribuito da permettere a ciascun attore di descrivere correttamente il proprio contributo come parziale. Il fornitore ha costruito il sistema ma non ha scelto il caso concreto. Il deployer ha scelto di usarlo ma non ne controlla ogni dettaglio tecnico. L'operatore ha compiuto l'ultimo gesto ma non ha definito il modello, i dati, le soglie o i tempi. Chi subisce l'esito, infine, può trovarsi davanti a una decisione della quale vede le conseguenze molto più chiaramente di quanto riesca a vedere il responsabile.

La responsabilità distribuita non è responsabilità dissolta

Quando più soggetti partecipano a un processo, la tentazione più comoda consiste nel cercare comunque un unico responsabile. È rassicurante pensare che da qualche parte esista il punto esatto nel quale la decisione è stata presa e nel quale possa essere depositata tutta la responsabilità. Ma molte decisioni assistite o automatizzate non funzionano così.

Una soglia può essere definita da un team, applicata da un prodotto acquistato da un'organizzazione, alimentata da dati raccolti altrove, interpretata da un operatore e trasformata in conseguenza attraverso una procedura che nessuno dei singoli attori controlla interamente. Questo non significa che nessuno risponda. Significa che la responsabilità deve essere ricostruita lungo la stessa catena che ha prodotto il risultato.

È il punto già emerso in L'ecosistema delle decisioni: distribuire il processo non serve a far evaporare la responsabilità, ma a distinguere il potere effettivo dei diversi attori. Il sesto movimento porta quel ragionamento fino alla sua conseguenza naturale: se il

potere è distribuito, anche l’accountability deve essere capace di seguirne la distribuzione senza fingere che tutto appartenga all’ultimo gesto.

L’ultimo essere umano è spesso il responsabile più visibile

Una persona firma, approva, conferma, invia o preme il pulsante finale. È quindi molto facile indicarla quando qualcosa va storto. Il suo nome compare nell’atto, la sua presenza è registrata dal sistema e il suo gesto può sembrare il punto nel quale tutto ciò che è accaduto prima diventa finalmente una decisione.

Ma nel Movimento 03 abbiamo visto che l’uomo nel circuito può essere soltanto una firma. Se quella persona non disponeva del tempo, delle informazioni, della competenza o dell’autorità necessarie per cambiare l’esito, attribuirle l’intera responsabilità significa confondere visibilità e potere. La firma dimostra che qualcuno era presente. Non dimostra, da sola, che potesse davvero decidere.

Il regolamento europeo sull’IA rende molto concreta questa distinzione quando, per i sistemi ad alto rischio, impone ai deployer di affidare la sorveglianza umana a persone che dispongano di competenza, formazione, autorità e sostegno necessari. Non basta quindi collocare formalmente una persona nel processo: perché la supervisione abbia significato, devono esistere condizioni organizzative che le permettano di esercitarla.

Le responsabilità cominciano prima del momento dell’errore

Un incidente rende visibile una catena che esisteva già. Prima dell’errore qualcuno ha scelto di adottare il sistema, definirne l’uso, configurarne gli accessi, stabilire chi potesse interromperlo, decidere quali controlli fossero sostenibili e quali rischi fossero accettabili. Se ricostruiamo la responsabilità soltanto dopo il danno, rischiamo di guardare l’ultimo fotogramma di una storia cominciata molto prima.

Il NIST colloca proprio nella funzione Govern la definizione di ruoli e responsabilità, linee di comunicazione, procedure di revisione e strutture di accountability. Il punto è importante: la responsabilità non dovrebbe essere inventata nel momento della crisi, quando tutti cercano di capire chi avrebbe dovuto intervenire. Deve essere progettata insieme al sistema e mantenuta durante il suo ciclo di vita.

Per questo la domanda “chi risponde?” non può essere separata da “chi poteva fare che cosa?”. Un soggetto può avere contribuito causalmente a un risultato senza avere il potere di evitarlo; un altro può non essere apparso nell’ultimo passaggio e tuttavia avere definito le condizioni che rendevano quel risultato possibile. L’accountability non coincide con la semplice prossimità temporale all’errore.

Ruolo, contesto e capacità di agire

I principi OCSE formulano un criterio particolarmente utile: gli attori dell'IA dovrebbero essere accountable in base al proprio ruolo, al contesto e alla propria capacità di agire. Questa formula evita due scorciatoie opposte. La prima consiste nell'attribuire tutto alla macchina, come se il sistema fosse diventato il soggetto morale della storia. La seconda consiste nell'attribuire tutto all'essere umano più vicino all'esito, come se ogni altro attore fosse soltanto un antecedente tecnico privo di responsabilità.

Ruolo, contesto e capacità di agire costringono invece a porre domande più precise. Chi poteva modificare il modello? Chi poteva limitarne l'uso? Chi disponeva delle informazioni per riconoscere un rischio? Chi aveva l'autorità per fermare il processo? Chi ha definito le metriche con cui il sistema veniva considerato abbastanza sicuro? Chi ha deciso che un determinato livello di errore fosse accettabile?

Questa ricostruzione non produce necessariamente una distribuzione uguale della responsabilità. Alcuni attori possono aver avuto molto più potere di altri. Ma proprio per questo è più rigorosa della ricerca di un colpevole unico: permette di distinguere responsabilità diverse invece di confonderle dentro una sola etichetta.

La macchina non può diventare il luogo nel quale depositiamo ciò che non sappiamo assegnare

Quando una catena è complessa, la frase "ha deciso l'algoritmo" offre una scorciatoia linguistica potente. Riassume una quantità di passaggi e permette di indicare immediatamente il punto nel quale il comportamento è diventato visibile. Il problema nasce quando quella sintesi viene utilizzata come spiegazione completa.

Un sistema può produrre un output inatteso, una raccomandazione difficile da interpretare o un comportamento che nessuno aveva previsto nei dettagli. Ma la possibilità che quell'output producesse conseguenze reali dipende comunque da accessi, autorizzazioni, procedure e scelte organizzative. L'imprevedibilità può complicare la responsabilità tecnica. Non cancella la responsabilità di chi ha costruito il contesto nel quale quella imprevedibilità poteva incidere sul mondo.

È la stessa distinzione che attraversa l'intero percorso: non si tratta di sostenere che ogni effetto fosse prevedibile o che ogni attore conoscesse anticipatamente ciò che sarebbe accaduto. Si tratta di impedire che l'incertezza sul comportamento della macchina venga trasformata in una zona nella quale nessun essere umano debba più rispondere delle decisioni che ha effettivamente preso.

Chi subisce l’esito deve sapere a chi rivolgersi

La responsabilità non serve soltanto a individuare chi punire dopo un fallimento. Serve anche a rendere possibile contestare, correggere, ottenere spiegazioni e modificare un processo che sta producendo conseguenze ingiuste o inattese. Dal punto di vista di chi subisce una decisione, una catena perfettamente comprensibile agli addetti ai lavori può essere comunque irresponsabile se non esiste un soggetto capace di ricevere una contestazione e agire sul risultato.

UNESCO collega per questo accountability, tracciabilità, audit, valutazioni d’impatto e due diligence, e afferma che i sistemi di IA non dovrebbero sostituire la responsabilità umana ultima. L’AI Act distribuisce obblighi distinti fra provider e deployer e impone, in diversi contesti, monitoraggio, informazione, supervisione e risposta agli incidenti. Il filo comune è che la complessità della catena non può diventare un ostacolo al fatto che qualcuno sia effettivamente tenuto ad agire.

Questo punto riprende anche la domanda di Chi può decidere della storia?: chi possiede il potere tecnico di modificare le condizioni di un sistema non coincide necessariamente con chi vive le conseguenze delle sue decisioni. Proprio per questo una relazione responsabile richiede che il potere di cambiare la storia rimanga collegato alla possibilità di essere chiamati a risponderne.

Il vantaggio aiuta a ricostruire la responsabilità, ma non la decide da solo

Nel Movimento 05 abbiamo confrontato la mappa di chi beneficia con quella di chi sopporta il rischio. Sapere chi ottiene velocità, riduzione dei costi, vantaggio competitivo o capacità operativa aiuta a capire perché certe configurazioni siano state scelte. Ma il beneficio non basta, da solo, ad assegnare la responsabilità.

Un soggetto può trarre vantaggio da un sistema senza controllarne ogni dettaglio; un altro può non ricevere alcun vantaggio diretto e tuttavia avere un potere decisivo nella catena. Per questo il vantaggio non si distribuisce come il rischio e neppure la responsabilità può essere ricostruita attraverso una sola misura.

Gli interessi spiegano una parte della causalità; l’autorità, la competenza, la possibilità di intervento e le decisioni effettivamente prese ne spiegano un’altra. Soltanto mettendole insieme possiamo evitare che la responsabilità venga assegnata per comodità a chi è più visibile o, al contrario, dissolta fra troppi soggetti perché nessuno appare sufficiente da solo.

La domanda da cui partire

Il percorso era cominciato da una formula semplice: prima della perdita del controllo esiste sempre, almeno in parte, una cessione del controllo. Sei movimenti dopo, quella frase può essere completata. La cessione non avviene in un solo istante e non appartiene necessariamente a una sola persona. Può essere distribuita fra progettazione, accessi, autonomia, supervisione, costi, incentivi e decisioni organizzative.

Per questo la responsabilità non si trova cercando il momento nel quale la macchina ha “preso il controllo”. Si trova ricostruendo chi ha reso possibile ogni passaggio, chi poteva modificarlo, chi ne ha tratto vantaggio, chi ne ha sopportato il rischio e chi aveva l’autorità necessaria per interrompere la catena.

La domanda conclusiva del percorso è allora questa: quando nessuno decide tutto da solo, abbiamo costruito una responsabilità distribuita oppure soltanto un sistema nel quale ciascuno può dire che la decisione apparteneva a qualcun altro?

Fonti e riferimenti

OECD, Accountability — OECD AI Principle. Il principio stabilisce che gli attori dell’IA siano accountable per il corretto funzionamento dei sistemi in base al proprio ruolo, al contesto e alla capacità di agire, e richiede tracciabilità e gestione sistematica del rischio lungo il ciclo di vita.

National Institute of Standards and Technology, AI Risk Management Framework 1.0 — Core, funzione Govern. Il Framework richiede strutture di accountability, ruoli e responsabilità chiaramente definiti, linee di comunicazione e revisione continua delle attività di gestione del rischio.

Unione europea, Regolamento (UE) 2024/1689 sull’intelligenza artificiale, in particolare articolo 26. Per i sistemi ad alto rischio, i deployer devono adottare misure tecniche e organizzative adeguate, assegnare la supervisione a persone dotate di competenza, formazione e autorità e monitorare il funzionamento del sistema, informando provider e autorità quando emergano rischi o incidenti.

UNESCO, Recommendation on the Ethics of Artificial Intelligence. La Raccomandazione collega responsabilità e accountability ad auditabilità, tracciabilità, supervisione, valutazioni d’impatto e due diligence e afferma che i sistemi di IA non dovrebbero sostituire la responsabilità umana ultima.

Edizione online: https://inchiostrati.it/chi-risponde-quando-nessuno-decide-da-solo/