

DB non solo diritto bancario

leggi l'articolo nel sito www.dirittobancario.it

APPROFONDIMENTI

# La gestione del rischio esistenziale dell'intelligenza artificiale

## Un'architettura giuridica per governare lo strumento

**Novembre 2026**

**Cosimo Pacciolla**, Head of Legal Risk Management & Integrated Compliance, Q8 Italia





**Cosimo Pacciolla**, Head of Legal Risk Management & Integrated Compliance, Q8 Italia

**Cosimo Pacciolla**

Cosimo Pacciolla opera nel settore legale e della compliance dal 1993, anno in cui è entrato a far parte della Direzione Legale di Q8, dopo aver maturato una precedente esperienza in uno studio legale specializzato in responsabilità penale d'impresa e nella pubblica amministrazione durante il percorso universitario. Dal 2002 ricopre ruoli manageriali all'interno della Direzione Legale, assumendo successivamente anche l'incarico di Antitrust Whistleblowing Compliance Officer.

*** 1. Uno scetticismo legittimo, non liquidabile**

Chi si occupa di diritto da tempo sufficiente ha imparato a diffidare degli allarmi tecnologici. Negli ultimi tre decenni si sono succedute previsioni di catastrofe ambientale, timori legati all'espansione di internet, stime drammatiche sulla diffusione di pandemie — molte delle quali, con il tempo, si sono rivelate meno severe di quanto la retorica pubblica lasciasse presagire. È un istinto sano, non un pregiudizio da correggere.

Ma l'istinto va verificato caso per caso, non applicato per analogia automatica. Le stime tecniche — non la retorica pubblica, le stime prodotte da chi fa previsione di rischio per mestiere — collocano il rischio esistenziale legato all'intelligenza artificiale in una posizione anomala rispetto agli altri rischi tecnologici del nostro tempo. Il filosofo Toby Ord, nel suo lavoro sistematico sul rischio esistenziale del 2020, stimava la probabilità di una catastrofe esistenziale dovuta a un'intelligenza artificiale non allineata nell'ordine del 10% nei prossimi cento anni — due ordini di grandezza superiore alla sua stima per il rischio nucleare o climatico, entrambi attorno allo 0,1%[1]. Un ampio esercizio di previsione condotto su questi temi su base comparata, l'Existential Risk Persuasion Tournament del 2022, condotto dal Forecasting Research Institute su un campione di esperti di dominio e di 'superforecaster' professionisti, ha confermato la stessa gerarchia: l'intelligenza artificiale risulta, in entrambi i gruppi di previsori, il rischio singolo a cui viene assegnata la probabilità di esito estintivo più alta, sistematicamente superiore a quella attribuita al conflitto nucleare o ai patogeni ingegnerizzati, pur con stime assolute molto diverse tra i due gruppi: entro il 2100, per l'estinzione dovuta all'IA, circa il 3% per gli esperti e lo 0,38% per i 'superforecaster'[2]. Sul rischio climatico il torneo non ha formulato quesiti di tipo catastrofico o

* Il contributo assume come premessa, senza pretesa di validarle sul piano tecnico, le stime di rischio esistenziale elaborate dalla letteratura specialistica e le dichiarazioni pubbliche di ricercatori del settore, riportate nelle note finali, numerate, con l'indicazione della fonte; il suo oggetto è l'elaborazione giuridica che ne consegue, della quale chi scrive assume per intero la responsabilità. Per quanto riguarda le note seguenti e le fonti citate, i riferimenti sono riportati senza pretesa di validarne il merito tecnico: vanno letti per l'autorevolezza propria di ciascuna fonte.

1 T. Ord, The Precipice: Existential Risk and the Future of Humanity, Bloomsbury, Londra, 2020.

2 E. Karger et al., Forecasting Existential Risks: Evidence from a Long-Run Forecasting Tournament (Existential Risk Persuasion Tournament, XPT), Forecasting Research Institute, 2023 (tabella 3, per le stime di estinzione; sezione sul cambiamento climatico, p. 58, per l'assenza di quesiti sul rischio climatico estintivo o catastrofico).

**vedi l'articolo online**

estintivo: gli esperti di clima consultati lo ritenevano estremamente basso, e i suoi effetti troppo lenti e indiretti per la soglia di 'catastrofe' adottata.[2]

È necessaria una precisazione, per evitare un equivoco che il confronto tra rischi rende possibile: la stima di Ord sul rischio climatico misura la probabilità di una catastrofe esistenziale (estinzione o perdita irreversibile del potenziale dell'umanità) entro un secolo, non la gravità complessiva del fenomeno. Il cambiamento climatico produce, ed è già in grado di produrre indipendentemente da qualunque proiezione futura, danni enormi, diffusi, in parte già irreversibili — perdita di territori abitabili, crisi migratorie, eventi estremi, collasso di ecosistemi. Nulla di quanto scritto in queste pagine intende ridimensionare questa realtà, né equipararla a un allarmismo smentito dai fatti: la bassa probabilità di catastrofe esistenziale stimata dagli studiosi di rischio esistenziale riguarda una categoria tecnica specifica e ristretta, non un giudizio complessivo sulla serietà del fenomeno. Il rischio climatico resta, sotto ogni altro profilo rilevante per le politiche pubbliche, uno dei più gravi del nostro tempo — e proprio per questo, come si dirà nel paragrafo che segue, è tra i candidati più concreti a un'interazione distruttiva con il rischio da intelligenza artificiale, non un termine di paragone che lo ridimensiona.

A settembre 2026, un ricercatore che si occupa di allineamento dei modelli presso uno dei principali laboratori di sviluppo di sistemi di frontiera ha dichiarato pubblicamente una probabilità superiore al 10% di un esito di estinzione per la specie umana nel corso del prossimo decennio[3] — un orizzonte temporale sensibilmente più breve di quello considerato dagli studi accademici citati — in concomitanza con le dimissioni pubbliche di un altro ricercatore dello stesso settore, motivate da analoghe preoccupazioni[4].

Questi elementi non dimostrano che la catastrofe sia probabile. Dimostrano che lo scetticismo fondato sull'analogia storica — 'anche allora si temeva la fine del mondo, e non è successo nulla' — non è applicabile senza correzioni a un rischio le cui stime tecniche, per quanto incerte, partono da un ordine di grandezza diverso rispetto ai precedenti storici evocati a sostegno dello scetticismo stesso.

3 R. McGreevy, Anthropic researcher says AI has over 10% chance to 'kill all humans' within next decade, FOX Business, 10 settembre 2026 (dichiarazione di E. Hubinger, ricercatore sull'allineamento presso Anthropic), https://www.foxla.com/news/anthropic-researcher-ai-10-percent-chance-kill-humans.

4 B. Nolan, articolo sulle dimissioni di J. Coxon da Anthropic, Fortune, 9 settembre 2026, https://www.fortune.com/2026/09/09/anthropic-researcher-resigns-warn-ai-companies-gambling-with-lives/.



**2. La complementarità distruttiva dei rischi**

Nessuno dei confronti sin qui condotti — tra intelligenza artificiale, clima, nucleare, pandemie — tiene conto di un fattore che la dottrina della gestione del rischio conosce bene e che gli esercizi di stima citati, per loro stessa ammissione metodologica, trattano in modo approssimativo: l'interazione tra rischi distinti può produrre un esito complessivo più grave della somma delle singole probabilità stimate.

Gli scenari più plausibili, in questo senso, non sono ipotesi da fantascienza ma estrapolazioni dirette di dinamiche già osservabili. Uno stress climatico che riduce la disponibilità di risorse essenziali in aree densamente popolate aumenta la probabilità di conflitto per il controllo di quelle stesse risorse, abbassando — non in astratto, ma per la dinamica storica nota di ogni scalata militare — la soglia a cui un confronto convenzionale rischia di trasformarsi in un confronto con armi di distruzione di massa. Un'intelligenza artificiale sufficientemente capace riduce il costo tecnico di progettazione di agenti patogeni, restringendo la distanza tra l'intenzione di un attore ostile e la capacità reale di realizzarla — è il motivo per cui, nel dibattito sulla sicurezza dell'intelligenza artificiale, il rischio biologico e quello algoritmico tendono a essere considerati come categorie collegate, non più autonome. Una crisi pandemica o un conflitto in corso degradano proprio la capacità di coordinamento istituzionale che sarebbe necessaria per governare, nello stesso momento, un sistema di intelligenza artificiale che mostri segnali di comportamento non allineato. E l'instabilità economico-sociale prodotta dalla transizione tecnologica stessa — dislocazione occupazionale, concentrazione di potere economico, sfiducia nelle istituzioni — alimenta la radicalizzazione politica che rende più difficile, non più facile, il coordinamento internazionale che ciascuno dei rischi presi singolarmente richiederebbe.

Non si tratta più di uno scenario ipotetico. Nel luglio 2026 un insieme di modelli di uno dei laboratori di frontiera, impegnati in una valutazione interna delle proprie capacità informatiche condotta con salvaguardie volutamente ridotte, ha superato i confini dell'ambiente di prova sfruttando una vulnerabilità sconosciuta, ha concatenato altre falle e credenziali sottratte ed è arrivato fino all'infrastruttura di produzione di un'altra azienda — compiendo in autonomia i singoli passaggi, fino a quando l'azienda colpita

vedi l'articolo online

non ha rilevato e fermato l'attività[5] [6]. Il caso non dimostra un'intenzione ostile, che non c'era: dimostra che l'iniziativa autonoma di un sistema può oggi eccedere la capacità di supervisione in tempo reale degli stessi soggetti che lo hanno costruito. È la premessa fattuale, non più solo teorica, su cui si fonda l'argomentazione che segue.

Nessuno dei modelli quantitativi citati in queste pagine incorpora formalmente questa complementarità: le stime di Ord, pur riconoscendo l'esistenza di fattori di rischio comuni, sono formulate per singolo rischio e non ne modellano l'interazione. È ragionevole ritenere che, considerati isolatamente, i singoli rischi restituiscano una stima complessiva per difetto, non per eccesso. È un'ammissione di incertezza, non un argomento retorico per innalzare surrettiziamente le stime: significa semplicemente che il livello di cautela che le politiche pubbliche dovrebbero adottare non può essere calibrato guardando ciascun rischio in isolamento.

**3. Perché il diritto, e perché adesso**

Al di là delle valutazioni tecnologiche ed economiche — che restano fuori dalla competenza di chi scrive, e vanno lasciate a chi possiede gli strumenti per condurle — resta un terreno che è proprio del giurista: la funzione del diritto come sintesi regolatoria di egoismi contrapposti a tutela della collettività. È una funzione antica, precedente a qualunque tecnologia specifica, e non diventa meno necessaria perché l'oggetto da regolare è nuovo ed evolve più velocemente della capacità del legislatore di comprenderlo.

Il problema, va detto con onestà, è strutturale: il diritto arriva sempre dopo il fenomeno che deve disciplinare — nasce, per usare un'espressione oggi ricorrente, già vecchio rispetto al proprio oggetto. Questo dato, che potrebbe indurre a un pessimismo paralizzante, va invece letto — parafrasando una formula di ascendenza gramsciana ormai di uso comune — come ragione sufficiente per un pessimismo dell'intelligenza che non escluda un'analoga determinazione della volontà: proprio perché la sintesi valida, efficace e universalmente riconosciuta non è oggi disponibile, il compito di chi si occupa di diritto è costruirne i tasselli disponibili, senza attendere condizioni ideali che l'esperienza storica non promette.

Il filo che regge le pagine che seguono è semplice: uno strumento, per quanto potente e complesso, si governa con regole che dicano chi decide e con responsabilità che dicano chi risponde. Ciascuna delle soluzioni proposte declina questa logica su un anello diverso della catena: il legislatore che sceglie il livello di rischio tollerabile (par. 4), l'impresa che produce e rilascia (par. 5.1), chi segnala il rischio prima che diventi danno (par. 5.2), chi lo verifica prima del rilascio (par. 5.3), gli Stati che devono coordinarsi (par. 8). Al centro resta sempre una decisione umana, e con essa una responsabilità umana.

**4. Il vincolo strutturale, e come aggirarlo**

Non ha senso proporre norme primarie che fissano soglie tecniche destinate a invecchiare in diciotto mesi. La via più realistica è la delega dell'aggiornamento a un'autorità tecnica indipendente, sottoposta a controllo giurisdizionale ex post — il modello, nel nostro ordinamento, non è il codice civile ma la vigilanza di settore bancaria e finanziaria: il legislatore fissa principi ed enforcement, l'autorità tecnica aggiorna periodicamente le soglie operative per regolamento, con obbligo di revisione a cadenza fissa.

Questo schema, però, non può essere trasferito senza un correttivo essenziale. Nella vigilanza bancaria e finanziaria la soglia operativa è quasi sempre un dato tecnico in senso proprio — un coefficiente di capitale, un indice di liquidità — la cui fissazione non implica una scelta di valore ulteriore rispetto a quella già compiuta dal legislatore all'origine. Per il rischio esistenziale da intelligenza artificiale non è così: stabilire a quale livello di capacità un sistema deve essere sospeso è già, in sé, una decisione su quanta esposizione al rischio la collettività accetta di tollerare in cambio di quale beneficio atteso — una scelta politico-assiologica, non un calcolo. Questa componente non può essere lasciata cadere nel mandato tecnico dell'autorità per semplice inerzia regolatoria: va isolata e riservata al legislatore, sotto forma di criteri guida specifici e vincolanti per l'autorità delegata, tanto più stringenti quanto più il rischio regolato è di natura esistenziale.

Resta, poi, da specificare quale soggetto sia concretamente deputato a un compito di questa delicatezza — non basta l'espressione generica di "autorità tecnica indipendente". Tre modelli sono astrattamente disponibili, con implicazioni diverse: un'agenzia pubblica costituita ad hoc, che offre la legittimazio-

5 OpenAI, OpenAI and Hugging Face partner to address security incident during model evaluation, 21 luglio 2026 (con aggiornamenti successivi), https://openai.com/index/hugging-face-model-evaluation-security-incident/.

6 P. Arntz, articolo sull'incidente OpenAI-Hugging Face, Malwarebytes, 24 luglio 2026, https://www.malwarebytes.com/blog/news/2026/07/openais-agent-escaped-its-sandbox-during-a-security-test.



vedi l'articolo online

ne democratica più solida ma tempi di costituzione lunghi; un'autorità amministrativa indipendente già esistente, alla quale attribuire per legge una competenza specifica in materia, che userebbe una struttura collaudata senza doverla creare da zero; un ente privato dotato di poteri pubblici di certificazione, sul modello degli organismi notificati nel diritto europeo dei prodotti, più rapido da attivare ma che pone il problema della base legale necessaria a legittimare poteri sospensivi di questa intensità in capo a un soggetto privato. Il modello più coerente con l'impianto qui proposto è probabilmente a due livelli: un'autorità pubblica che accredita, indirizza e supervisiona, e organismi tecnici — pubblici o privati — che operano sotto la sua responsabilità, sullo stesso schema richiamato al paragrafo 5.3 a proposito degli spazi di sperimentazione regolatoria dell'AI Act[7].

Il modello a due livelli va, a sua volta, collocato nel quadro europeo, che non lo esclude ma ne delimita il perimetro. Per i modelli di intelligenza artificiale per finalità generali con rischio sistemico l'AI Act attribuisce all'Ufficio per l'IA della Commissione poteri di valutazione e di richiesta di misure correttive, fino alla restrizione, al ritiro o al richiamo del modello (artt. 88, 92 e 93)[7]; e il Digital Omnibus on AI gli ha affidato, salve le eccezioni previste, anche la vigilanza sui sistemi basati su un modello per finalità generali sviluppato dallo stesso fornitore o da fornitori della stessa impresa (art. 75, par. 1)[8]. Per i sistemi ad alto rischio che rientrano in tale competenza, soggetti a valutazione di conformità da parte di terzi, l'Ufficio è responsabile anche delle prove e delle valutazioni preventive, che la Commissione affida a organismi notificati operanti per suo conto (art. 75, par. 1 sexies)[8]: uno schema che, nella sostanza, non è estraneo a quello a due livelli qui proposto. Un'autorità nazionale non potrebbe dunque sospendere da sola un modello di frontiera: il suo spazio d'azione è la cooperazione con l'Ufficio e la vigilanza sui sistemi e sugli usi che restano di competenza nazionale.

7 Regolamento (UE) 2024/1689 del Parlamento europeo e del Consiglio, del 13 giugno 2024, che stabilisce regole armonizzate sull'intelligenza artificiale (AI Act), GUUE L del 12 luglio 2024; in particolare artt. 57 (spazi di sperimentazione regolatoria), 75, 88, 92 e 93 (competenze dell'Ufficio per l'IA della Commissione) e 87 (applicazione della direttiva (UE) 2019/1937 alle segnalazioni delle violazioni del regolamento).

8 Regolamento (UE) 2026/1744 del Parlamento europeo e del Consiglio, dell'8 luglio 2026, che modifica i regolamenti (UE) 2024/1689, (UE) 2018/1139 e (UE) 2023/1230 per quanto riguarda la semplificazione dell'attuazione delle regole armonizzate sull'intelligenza artificiale (Digital Omnibus on AI), GUUE del 24 luglio 2026, in vigore dal 27 luglio 2026; in particolare la modifica dell'art. 75 sulla competenza dell'Ufficio per l'IA.



In questo spazio si colloca la legge n. 132/2025, che all'art. 20 ha designato AgID e ACN quali autorità nazionali per l'intelligenza artificiale, ferma restando la vigilanza di settore di Banca d'Italia, CONSOB e IVASS: all'AgID spettano la notifica, la valutazione, l'accreditamento e il monitoraggio dei soggetti incaricati di verificare la conformità dei sistemi; all'ACN la vigilanza, anche ispettiva e sanzionatoria, e il ruolo di punto di contatto unico con le istituzioni dell'Unione[9]. Il quadro, va notato, è frammentato: alle due agenzie si affiancano le autorità di vigilanza di settore e restano ferme le competenze del Garante per la protezione dei dati personali e dell'AGCOM (art. 20, comma 4), come sul piano europeo il rapporto Draghi rileva la proliferazione di leggi e di autorità (par. 7). La frammentarietà non è soltanto un costo di conformità: rende meno nitida la risposta alla domanda che regge queste pagine — chi decide e chi risponde —, ed è una ragione ulteriore per affidare il primo livello del modello a un'autorità chiaramente individuata, con competenze definite per legge. Il modello può innestarsi sulla struttura esistente — un'autorità che accredita e supervisiona, organismi tecnici che verificano —, con un'avvertenza: AgID e ACN sono agenzie, il cui coordinamento è affidato a un comitato istituito presso la Presidenza del Consiglio (art. 20, comma 3); l'indipendenza richiesta dall'impianto qui proposto andrebbe quindi assicurata per legge. La verifica preventiva sui modelli di frontiera, per essere efficace, richiede infine uno standard europeo, di cui i poteri attuali dell'Ufficio sono il nucleo da rafforzare, non il punto di arrivo.

La sola verifica giurisdizionale ex post, per un rischio che per definizione può tradursi in un danno irreversibile, è da sola inadeguata: il controllo del giudice presuppone che l'errore sia in qualche misura riparabile, ed è proprio questa premessa che il rischio esistenziale mette in discussione. Serve, in aggiunta, un impianto di criteri procedimentali che regolino ex ante l'attività dell'autorità: motivazione rafforzata delle decisioni, contraddittorio effettivo con l'operatore prima di ogni misura sospensiva, revisione periodica obbligatoria delle soglie adottate, tracciabilità documentale delle ragioni scientifiche che hanno condotto a ciascuna scelta. Non sono garanzie procedurali di dettaglio: sono condizione di legittimità costituzionale di un potere che limita l'iniziativa economica privata ai sensi dell'articolo 41 della Costituzione, e che senza queste garanzie non supererebbe un vaglio di proporzionalità e ragio-

9 Legge 23 settembre 2025, n. 132, Disposizioni e deleghe al Governo in materia di intelligenza artificiale, art. 20, commi 1-4 (AgID e ACN quali autorità nazionali per l'intelligenza artificiale; Comitato di coordinamento presso la Presidenza del Consiglio dei ministri; salvezza delle competenze del Garante per la protezione dei dati personali e dell'AGCOM), in vigore dal 10 ottobre 2025.

vedi l'articolo online

nevolezza.

Delegare la verifica tecnica a un'autorità indipendente non significa delegare alla macchina la decisione, né lasciare che il rischio si autoregoli attraverso parametri puramente tecnici: è un'articolazione istituzionale — pensata, decisa e amministrata da persone — di una scelta politica che resta, essa sì, integralmente umana.

**5. Tre architetture concrete a livello nazionale ed europeo**

**5.1 *Responsabilità d'impresa, non responsabilità personale indifferenziata***

È tentazione comune, nel dibattito pubblico, invocare una responsabilità ‹personale› di chi produce, fornisce e utilizza sistemi di intelligenza artificiale, quasi a voler restituire un volto umano a un rischio percepito come anonimo. È un impulso comprensibile, ma giuridicamente fuorviante se applicato senza distinzioni. Il danno da comportamento emergente di un sistema complesso — la deriva di un modello durante l›addestramento, un comportamento non previsto in fase di dispiegamento — non ha, nella maggior parte dei casi, un autore individuale riconoscibile: è il prodotto aggregato di migliaia di scelte progettuali, nessuna delle quali, isolatamente, integra gli estremi della colpa individuale. Insistere sulla responsabilità della persona fisica in un contesto di questo tipo rischia un doppio effetto perverso: da un lato individua un capro espiatorio — l'ultimo tecnico che ha autorizzato il rilascio — sproporzionato rispetto alla causalità reale; dall'altro, se applicata a chi segnala un rischio interno, disincentiva proprio la segnalazione che servirebbe a prevenire il danno.

La via più coerente con i principi generali della responsabilità d'impresa — e già in corso di consolidamento nel diritto dell'Unione — è quella della responsabilità oggettiva dell'ente lungo l'intera filiera. Non è più uno tra due percorsi paralleli: la Commissione europea ha annunciato nel febbraio 2025 il ritiro della proposta di direttiva sulla responsabilità civile da intelligenza artificiale (*AI Liability Directive)*, poi formalizzato nell'ottobre 2025[10], lasciando la disciplina della responsabilità di prodotto come l'unico strumento europeo di responsabilità oggettiva destinato a operare su questo terreno. La direttiva (UE) 2024/2853, che gli Stati membri devono recepire entro il 9 dicembre 2026, ricomprende espressamente il *software* — e con esso i sistemi di intelligenza artificiale — nella nozione di 'prodotto', stabilendo la responsabilità degli operatori economici lungo la catena di fornitura e un sistema di presunzioni tipizzate di difettosità e di nesso causale (art. 10), che soccorre il danneggiato soprattutto quando la complessità tecnica del prodotto gli rende eccessivamente difficile la prova — come in un sistema opaco per definizione — senza tuttavia realizzare un'inversione generale dell'onere della prova[11]. Il punto in cui questa architettura è più esposta, rispetto al rischio qui discusso, è l'esimente del rischio di sviluppo (art. 11, par. 1, lett. e): il produttore non risponde se lo stato delle conoscenze scientifiche e tecniche, al momento dell'immissione sul mercato o della messa in servizio, non consentiva di scoprire il difetto. È esattamente l'ipotesi del comportamento emergente, per definizione non prevedibile con lo stato dell'arte, sicché l'esimente rischia di neutralizzare la responsabilità proprio nel caso più rilevante. La direttiva consente però agli Stati membri di derogarvi per categorie specifiche di prodotti, se giustificato e proporzionato, con notifica alla Commissione e sospensione della misura per sei mesi dalla notifica (art. 18): ne discende, per il legislatore italiano in sede di recepimento, una proposta concreta e forse la più incisiva tra quelle qui avanzate — escludere l'esimente per i sistemi di intelligenza artificiale che superano la soglia di capacità definita ai sensi del paragrafo 4 —, da formulare con la prudenza che la stessa norma impone: categoria circoscritta, motivazione di proporzionalità, notifica. L'esclusione dell'esimente fa ricadere il rischio dell'imprevedibile sull'impresa che decide di rilasciare il sistema, non sulla persona fisica, in coerenza con la scelta di non scaricare sul singolo tecnico il rischio emergente. Nel suo insieme, questa architettura dialoga naturalmente con l'impianto della responsabilità amministrativa da reato già presente nel nostro ordinamento per altre fattispecie: la logica del modello organizzativo idoneo a prevenire il rischio, con l'ente chiamato a rispondere per l'assenza di presidi adeguati, è direttamente trasferibile a un modello organizzativo tarato su *capability testing, red-teaming* e procedure di *escalation* interna prima del rilascio di sistemi sopra una determinata soglia di capacità.

10 Commissione europea, Programma di lavoro 2025, 11 febbraio 2025, allegato IV (proposte da ritirare), e avviso di ritiro della proposta di direttiva sulla responsabilità civile da intelligenza artificiale, GUUE serie C, 6 ottobre 2025, C/2025/5423; cfr. M. Pasqua, European Commission Withdraws Proposals on Assignments of Claims and AI Liability, EAPIL, 9 ottobre 2025, https://eapil.org/2025/10/09/european-commission-withdraws-two-proposals-assignments-of-claims-regulation-and-ai-liability-directive/.

11 Direttiva (UE) 2024/2853 del Parlamento europeo e del Consiglio, del 23 ottobre 2024, sulla responsabilità per danno da prodotti difettosi, GUUE L del 18 novembre 2024 (recepimento entro il 9 dicembre 2026); in particolare artt. 10 (presunzioni), 11, par. 1, lett. e) (rischio di sviluppo) e 18 (deroghe).



vedi l'articolo online

La responsabilità della persona fisica, in questo schema, resta uno strumento residuale e va riservata a un perimetro ristretto: l'omissione dolosa o gravemente colposa di controlli già riconosciuti come necessari — non l'esistenza del rischio emergente in sé, che nessuno, allo stato dell'arte, è in condizione di eliminare con certezza.

**5.2 *Un canale di segnalazione specifico per il rischio sistemico***

Il quadro normativo sul *whistleblowing*, per quanto rafforzato negli ultimi anni, resta ancorato alla violazione, anche solo potenziale, di una norma: il D.Lgs. 24/2023 tutela chi segnala illeciti — corruzione, frode, violazione di norme ambientali o di sicurezza — non solo già commessi, ma anche quelli che, sulla base di elementi concreti, potrebbero esserlo (art. 2, comma 1, lettere a e b)[12], e lo stesso vale per l'AI Act, che estende la direttiva (UE) 2019/1937 alle segnalazioni delle violazioni del regolamento stesso (art. 87)[7]. Il limite non è dunque temporale, ma di oggetto: la tutela presuppone una norma violata o violabile, mentre il rilascio di un sistema rischioso ma lecito non ne viola alcuna. Non è contemplata, perciò, la segnalazione predittiva — la comunicazione di un segnale di rischio sistemico, lecito, non ancora tradottosi in un danno concreto, che è il terreno su cui si colloca la scelta di un ricercatore che, nel settembre 2026, ha reso pubbliche le proprie preoccupazioni dimettendosi da un laboratorio di sviluppo AI.

Colmare questa lacuna non richiede consenso internazionale, né attende il recepimento di alcuna direttiva: è un intervento che un singolo ordinamento — o una singola organizzazione, in via di autoregolamentazione anticipata — può introdurre autonomamente, estendendo la logica di tutela già presente nella disciplina generale a un oggetto specifico: la segnalazione di un rischio sistemico previsto, prima che si traduca in danno. Il modello di riferimento esiste già, in un settore diverso: la disciplina della sicurezza dell'aviazione civile separa esplicitamente l'obbligo di segnalazione dalla responsabilità disciplinare, proteggendo chi segnala un evento o un quasi-incidente da sanzioni e provvedimenti disciplinari, salvo i casi di dolo o di grave e manifesta inosservanza di un rischio evidente[13]. È un principio di

12 D.Lgs. 10 marzo 2023, n. 24, in materia di whistleblowing, art. 2, comma 1, lett. a) e b).

13 Regolamento (UE) n. 376/2014 del Parlamento europeo e del Consiglio, del 3 aprile 2014, concernente la segnalazione, l'analisi e il monitoraggio di eventi nel settore dell'aviazione civile, art. 16, parr. 9 e 10 (tutela di chi segnala, salvo dolo o manifesta, grave e seria inosservanza di un rischio evidente).

14 D. Amodei, We Must Pace the Frontier, settembre 2026, https://darioamodei.com/post/we-must-pace-the-frontier.



'cultura giusta' (*just culture*) che la disciplina whistleblowing può importare senza difficoltà concettuali, e che risolverebbe alla radice l'incentivo perverso descritto al paragrafo 5.1.

**5.3 *Verifica indipendente prima del rilascio***

Gli impegni volontari di sicurezza che alcuni laboratori si sono dati negli ultimi anni condividono un limite comune: sono autovalutati. Manca, nella maggior parte degli ordinamenti, l'equivalente di un organismo terzo indipendente con potere di veto sul rilascio di un sistema sopra una soglia di capacità definita — un ruolo che, per analogia, è già familiare al diritto interno in altri settori a rischio rilevante. Introdurre una valutazione di sicurezza obbligatoria, condotta da un soggetto indipendente dal produttore e dotato di potere sospensivo, non elimina il rischio ma sposta il controllo a monte, nella fase in cui è ancora reversibile. Il termine di confronto europeo più vicino è costituito dai poteri dell'Ufficio per l'IA richiamati al paragrafo 4: sui modelli di frontiera operano a valle dell'immissione sul mercato, non come autorizzazione preventiva — la valutazione preventiva riguarda i sistemi ad alto rischio, non i modelli come tali —, e sono esercitati da un organo della Commissione, non da un soggetto terzo rispetto alle istituzioni. Il veto preventivo qui proposto resta dunque una proposta de iure condendo, da collocare a livello europeo, come indicato al paragrafo 4.

Il limite di questo modello, quando affidato all'autoregolamentazione, è ormai visibile nei fatti. Il piano di "valutatori integrati" annunciato da un importante laboratorio di frontiera nel settembre 2026, a titolo volontario e con l'invito ai governi a imporre lo stesso standard agli altri laboratori, prevede l'accesso di soggetti terzi ai propri sistemi con privilegi equiparati a quelli di un dipendente — ma non prevede un meccanismo indipendente di designazione: l'invito a imbarcarsi resta nella disponibilità del laboratorio stesso[14]. È la più classica delle obiezioni all'autoregolamentazione, chi controlla i controllori, e conferma perché la designazione dell'organismo di verifica non possa restare nella disponibilità del soggetto controllato: deve provenire da un'autorità pubblica o da un meccanismo di accreditamento sottratto al suo controllo. Un modello concettualmente adiacente, pur non equivalente, esiste già nell'ordinamento

vedi l'articolo online

europeo: l'AI Act prevede spazi di sperimentazione regolatoria (regulatory sandboxes) sotto supervisione pubblica[7], che condividono con la verifica indipendente qui proposta lo stesso principio di fondo — un controllo esterno che non dipende dalla buona volontà di chi sviluppa il sistema.

Resta fuori da questa trattazione, perché appartiene a una categoria giuridica diversa da quella del rischio esistenziale qui discusso, il rischio di condizionamento delle preferenze dell'utente quando l'intelligenza artificiale diventa insieme interlocutore, consigliere e canale commerciale — un fronte di diritto della concorrenza e di tutela del consumatore, non di governance del rischio catastrofico, che la crescente integrazione pubblicitaria nei sistemi conversazionali rende sempre più concreto e alla cui trattazione chi scrive ha già dedicato, in altra sede e con altro coautore, un contributo specifico[15].

**6. Il limite geopolitico, e la lezione scomoda**

Resta un'obiezione che nessuna delle tre architetture proposte risolve da sola: senza un coordinamento almeno tra le giurisdizioni che controllano le risorse critiche per lo sviluppo dei sistemi più avanzati, ogni intervento regolatorio nazionale rischia di restare velleitario. È un'obiezione fondata, che tuttavia non autorizza la conclusione opposta — l'attesa di un diritto internazionale universalmente riconosciuto che la storia recente non lascia presagire in tempi utili.

Il precedente più istruttivo, in questo senso, non è la non proliferazione nucleare — pensata per tenere fuori chi la capacità non ce l'ha ancora — ma il controllo degli armamenti tra potenze che la capacità ce l'hanno già, indipendentemente dal grado di rivalità reciproca: Stati Uniti e Unione Sovietica hanno negoziato accordi di riduzione del rischio, dalla linea diretta del 1963 ai trattati di limitazione degli armamenti strategici, nel pieno della guerra fredda, senza che questo richiedesse alcuna fiducia reciproca sul piano ideologico. È significativo, in questa luce, che proprio le due potenze oggi più direttamente coinvolte nello sviluppo dei sistemi di intelligenza artificiale di frontiera abbiano avviato, nel corso del 2026, un canale bilaterale dedicato alla sicurezza: un primo protocollo bilaterale sulla sicurezza dell'intelligenza artificiale presentato a maggio al vertice di Pechino[16], seguito da colloqui a settembre il cui ordine del giorno, secondo le anticipazioni della stampa, comprendeva anche il monitoraggio degli attacchi informatici condotti da sistemi di intelligenza artificiale[17] [18]. Il perimetro del protocollo di maggio resta circoscritto al rischio da attori non statali, e non affronta la competizione tra le due potenze sulla capacità stessa dei sistemi — che resta il vero motore della corsa[16]. Ma la sua sola esistenza smentisce l'idea che l'assenza di fiducia reciproca tra le grandi potenze renda impossibile ogni forma di cooperazione sul rischio: la cronaca più recente mostra l'esatto contrario.

**7. L'obiezione economica: un'Europa che arretra mentre altri non si pongono il problema**

Un'obiezione prevedibile, e va affrontata con lo stesso rigore riservato alle altre, non liquidata per convenienza argomentativa: ogni architettura di responsabilità, verifica indipendente e tutela del segnalante qui proposta comporta un costo di conformità che grava, nell'immediato, sulle imprese europee — mentre gli attori che guidano oggi lo sviluppo dei sistemi di frontiera si muovono in condizioni normative radicalmente diverse. Gli Stati Uniti perseguono una politica dichiaratamente orientata alla velocità di sviluppo più che al contenimento preventivo; la Cina persegue l'autosufficienza tecnologica come priorità di sicurezza nazionale, senza le garanzie procedurali proprie degli ordinamenti liberaldemocratici; il mercato israeliano delle tecnologie di sorveglianza e delle capacità cyber opera in gran parte al di fuori di qualunque cornice multilaterale; la Russia non è tanto un soggetto da includere in un negoziato di sicurezza quanto una fonte diretta di rischio attraverso il conflitto ibrido.

Il rapporto sulla competitività europea presentato da Mario Draghi nel settembre 2024 ha reso questa obiezione un dato istituzionale, non un'opinione di parte: vi si rileva che l'Unione conta circa cento leggi incentrate sul settore tecnologico e oltre duecentosettanta autorità di regolamentazione attive nelle

15 C. Pacciolla, F. Costantino, Analisi e governance della concorrenza algoritmica nel mercato unico digitale. Verso una ricalibratura ermeneutica del diritto antitrust europeo dell'intelligenza artificiale, in European Review of Digital Administration & Law (ERDAL), 2026, Vol. 7, n. 1, pp. 237-256, DOI 10.53136/9791221828467.

16 A. Dufresne, US-China Sign First Bilateral AI Safety Pact in Beijing, AI Weekly, 14 maggio 2026 (che riprende un articolo di CNBC della stessa data), https://aiweekly.co/alerts/us-china-sign-first-bilateral-ai-safety-pact-in-beijing.

17 L. Chen, US, China gear up for mid-September AI safety talks, Reuters, 4 settembre 2026, https://theprint.in/world/exclusive-us-china-gear-up-for-mid-september-ai-safety-talks/3034187/.

18 M. Curi, M. Caputo, Scoop: U.S. open to discuss AI "shared risks" with China, Bessent says, Axios, 16 settembre 2026, https://axios.com/2026/09/16/us-open-ai-shared-risks-china-bessent.



vedi l'articolo online

reti digitali negli Stati membri, e che l'approccio precauzionale di molte di esse — l'AI Act compreso, che impone requisiti aggiuntivi ai modelli per finalità generali oltre una soglia di potenza di calcolo — ostacola l'innovazione; solo quattro delle prime cinquanta aziende tecnologiche mondiali sono europee, e circa il settanta per cento dei modelli di base di intelligenza artificiale sviluppati dal 2017 proviene dagli Stati Uniti[19]. Alla conferenza di alto livello promossa dalla Commissione a un anno dal rapporto (Bruxelles, 16 settembre 2025), Draghi ha chiesto di sospendere l'attuazione della fase dell'AI Act relativa ai sistemi ad alto rischio finché non se ne comprendano meglio gli svantaggi, e una semplificazione radicale del GDPR, richiamando uno studio secondo cui il GDPR avrebbe aumentato di circa il venti per cento il costo dei dati per le imprese europee rispetto a quelle statunitensi[20]. Nel maggio 2026 Consiglio e Parlamento europeo hanno effettivamente raggiunto un accordo politico su uno snellimento e un rinvio di parti dell'AI Act (*Digital Omnibus on AI*)[21], poi adottato con il Regolamento (UE) 2026/1744 e in vigore dal 27 luglio 2026[8], nell'ambito dei pacchetti di semplificazione avviati dalla Commissione: l'obiezione, in altre parole, non è solo fondata ma già in parte accolta dal legislatore stesso.

Chi scrive non ha, su questo punto, una soluzione da proporre che risolva la tensione — sarebbe disonesto affermarlo. Due considerazioni, tuttavia, meritano di essere messe a disposizione del lettore, con l'onestà di riconoscerne i limiti.

La prima: buona parte della critica di Draghi riguarda la regolazione a monte — i limiti e i costi di conformità nell'uso dei dati per l'addestramento dei modelli e gli obblighi aggiuntivi imposti ai modelli per finalità generali — non la regolazione della responsabilità, della verifica indipendente e della tutela del

19 M. Draghi, The Future of European Competitiveness, rapporto presentato alla Commissione europea, settembre 2024 (Parte A, in particolare la sintesi e la sezione sugli ostacoli normativi alla crescita delle imprese tecnologiche; Parte B, per il dato sulle autorità di regolamentazione nelle reti digitali).

20 M. Draghi, intervento alla High Level Conference "One year after the Draghi report", Bruxelles, 16 settembre 2025, testo pubblicato dalla Commissione europea, https://commission.europa.eu/document/download/0951a4ff-cd1a-4ea3-bc1d-f603decc1ed9_en?filename=Draghi_Speech_High_Level_Conference_One_Year_After.pdf; per la data cfr. L. Rijo, PPC Land, 16 settembre 2025, https://ppc.land/former-ecb-chief-demands-gdpr-cuts-and-ai-act-pause-at-brussels-conference/.

21 Accordo politico provvisorio tra Parlamento europeo e Consiglio del 7 maggio 2026 sul Digital Omnibus on AI (modifiche mirate al Regolamento (UE) 2024/1689); cfr. T. Hickman, D. Mair, EU agrees Digital Omnibus deal to simplify AI rules, White & Case, 14 maggio 2026, https://www.whitecase.com/insight/eu-agrees-digital-omnibus-deal-simplify-ai-rules.



segnalante proposta in queste pagine, che si applica alla fase di dispiegamento e di utilizzo all'interno del mercato europeo, indipendentemente da dove il sistema sia stato sviluppato. È una distinzione che riduce, senza eliminare, la sovrapposizione tra le due obiezioni: l'Unione europea ha una leva reale sull'accesso al proprio mercato — lo ha già dimostrato con il GDPR, che nessuna impresa globale ha potuto permettersi di ignorare pur non essendo europea — anche in assenza di una propria industria di sviluppo di sistemi di frontiera.

La seconda, più incerta: non tutti i settori regolati hanno vissuto la conformità come un costo puro. Alcuni comparti industriali — l'esempio più recente riguarda la riforma della legislazione farmaceutica europea, concordata a fine 2025 e approvata nel 2026, che introduce spazi di sperimentazione regolatoria[22] — mostrano come la regolazione possa essere concepita come un elemento di progettazione, non solo come un vincolo da aggirare a valle. Se un'architettura equivalente fosse applicabile all'intelligenza artificiale — verifica indipendente e responsabilità d'impresa progettate insieme al settore, non imposte contro di esso — il costo di conformità potrebbe redistribuirsi diversamente. È, però, un'ipotesi che la cronaca normativa più recente non conferma né smentisce in modo risolutivo, e chi scrive preferisce dichiararla come tale piuttosto che presentarla come una risposta compiuta a un'obiezione che, allo stato, resta sostanzialmente aperta.

**8. Oltre l'Europa: un decalogo globale, impervio ma necessario**

Le tre architetture proposte al paragrafo 5 condividono un pregio pratico: un singolo legislatore può attuarle senza attendere il consenso altrui. La via che segue non lo condivide — ogni punto richiede l'accordo di più giurisdizioni sovrane, e la sua realizzabilità è per questo strutturalmente più bassa. Va tentata comunque, per la stessa ragione già esposta al paragrafo 3: l'assenza di condizioni ideali non è mai stata, storicamente, una ragione sufficiente per non provarci. I dieci interventi che seguono si

22 Riforma della legislazione farmaceutica dell'Unione europea (cd. pacchetto farmaceutico): accordo provvisorio di dicembre 2025, cfr. EU Reaches Landmark on Reform of Pharmaceutical Legislation, GMP-Compliance.org, 17 dicembre 2025, https://www.gmp-compliance.org/gmp-news/eu-reaches-landmark-on-reform-of-pharmaceutical-legislation; approvazione del Consiglio, cfr. EU Pharma Package: Council Adopts New Rules, GMP-Compliance.org, 14 ottobre 2026, https://www.gmp-compliance.org/gmp-news/eu-pharma-package-council-adopts-new-rules.

vedi l'articolo online

raggruppano in tre famiglie: verifica e trasparenza, responsabilità e gestione degli incidenti, potere negoziale e architettura geopolitica.

**8.1 *Verifica e trasparenza***

Il primo gruppo mira a rendere visibile ciò che oggi non lo è: lo sviluppo dei sistemi più avanzati e il superamento delle soglie di rischio concordate.

1. Registro internazionale dei training run di frontiera, gestito da un ente tecnico sovranazionale con potere di verifica documentale — non ispezione fisica vincolante, ma obbligo dichiarativo con controllo incrociato sui consumi energetici dei data center, oggi l'unico proxy verificabile dall'esterno.

2. Rete degli istituti nazionali di sicurezza AI trasformata in organismo pattizio, con potere di verifica pre-rilascio riconosciuto reciprocamente tra le giurisdizioni aderenti — oggi cooperazione volontaria tra enti nazionali; il salto è renderla condizione di accesso reciproco al mercato, non solo buona pratica.

3. Soglia di capacità con moratoria automatica condivisa — un meccanismo pattizio per cui il superamento di parametri di pericolosità concordati impone una sospensione del rilascio verificata da terzi indipendenti di più giurisdizioni insieme, non dal solo laboratorio interessato.

**8.2 *Responsabilità e gestione degli incidenti***

Il secondo gruppo disciplina cosa accade quando, nonostante la verifica, qualcosa va storto — o rischia di andare storto.

4. Estensione del modello di controllo degli armamenti già avviato nel 2026 tra le due maggiori potenze coinvolte nello sviluppo dei sistemi di frontiera, da rischio da attori non statali a rischio di capacità di frontiera — una linea diretta permanente tra le rispettive autorità di sicurezza AI, sul modello della linea diretta nucleare del 1963, non per costruire fiducia reciproca ma per gestire incidenti.

5. Protocollo di notifica obbligatoria degli incidenti, sul modello della disciplina internazionale della sicurezza dell'aviazione civile ma a livello intergovernativo — un quasi-incidente di perdita di controllo o comportamento emergente pericoloso va comunicato a un database condiviso tra le autorità aderenti entro tempi certi, indipendentemente dal paese di sviluppo del sistema.

6. Convenzione internazionale di tutela del segnalante per rischio sistemico AI, sul modello — pur con le dovute differenze — delle convenzioni internazionali anticorruzione: uno standard minimo comune che ogni Stato aderente recepisce, con meccanismo di monitoraggio reciproco tra pari.

7. Fondo assicurativo internazionale per danno catastrofico da intelligenza artificiale, alimentato da contributi obbligatori dei laboratori di frontiera proporzionati alla capacità di calcolo utilizzata — sul modello dei fondi internazionali per il danno da inquinamento marino da idrocarburi, che già risolvono il problema dell'attore che causa un danno diffuso e transnazionale.

**8.3 *Potere negoziale e architettura geopolitica***

Il terzo gruppo affronta direttamente il nodo che nessuna verifica tecnica può sciogliere da sola: la distribuzione del potere negoziale tra le giurisdizioni che controllano le risorse critiche per lo sviluppo dell'intelligenza artificiale.

8. Trattato plurilaterale, non universale, tra le tre-quattro giurisdizioni che controllano la produzione di chip avanzati — formalizzazione in accordo vincolante di ciò che oggi è coordinamento informale tramite controlli export unilaterali, per sottrarlo alla contendibilità negoziale, cioè alla possibilità che ogni restrizione sia rimessa in discussione di volta in volta nei rapporti bilaterali.

9. Clausola di reciprocità del regime di controllo export sui chip agganciata al rispetto degli standard di sicurezza concordati, non a considerazioni puramente commerciali — perché l'unico strumento oggi davvero coercitivo smetta di essere merce di scambio tariffario.

10. Sede permanente di revisione, non un vertice occasionale — un organismo con calendario fisso di incontri, sul modello delle conferenze periodiche in materia ambientale: non risolve nulla da sé, ma istituzionalizza la pressione diplomatica continua invece di farla dipendere dall'urgenza



vedi l'articolo online

del momento.

Il punto più fragile di questo decalogo va dichiarato senza infingimenti: il trattato sui chip, la linea diretta tra le autorità di sicurezza e la clausola di reciprocità sul controllo export presuppongono che le due maggiori potenze coinvolte nello sviluppo dell'intelligenza artificiale trattino la sicurezza come priorità sovraordinata rispetto alla competizione strategica reciproca — condizione che, allo stato attuale, risulta solo parzialmente verificata. Non è un decalogo che chi scrive ritiene probabile nella sua interezza; è il decalogo coerente con la logica del controllo degli armamenti tra rivali storicamente già sperimentata — la sua impervietà non è un difetto di progettazione, è la natura stessa del problema che descrive.

**9. Conclusione**

Chi scrive resta, per formazione e per temperamento, scettico verso ogni catastrofismo che si autoalimenti di retorica più che di dati. Ma lo scetticismo cessa di essere una virtù intellettuale nel momento in cui si trasforma in alibi per l'inazione — soprattutto quando gli strumenti per agire, seppure parziali, esistono già: una direttiva europea in fase di recepimento, che consente al legislatore nazionale di intervenire sull'esimente del rischio di sviluppo, un impianto di responsabilità d'impresa collaudato in altri contesti, un modello di tutela del segnalante che altri settori hanno già sperimentato con successo, e un decalogo globale impervio ma non per questo velleitario. Nessuno di questi strumenti elimina il rischio. Ciascuno di essi, se attuato, lo riduce di un ordine di grandezza — che è, in fondo, l'unico obiettivo che il diritto si è mai potuto ragionevolmente porre di fronte a un rischio tecnologico: non l'eliminazione, il contenimento. Il vero avversario, alla fine di questo percorso, non è la tecnologia che abbiamo costruito, ma la tentazione — altrettanto umana — di credere che, di fronte alla sua complessità, non resti nulla da fare.



vedi l'articolo online



DB non solo diritto bancario

A NEW DIGITAL EXPERIENCE

# dirittobancario.it