Dalla colpa al codice. Ripensare la responsabilità nell'era dell'intelligenza artificiale*

Enrico Maestri, Carlotta Mognato

Sommario

1. Introduzione

La rivoluzione algoritmica ha investito in profondità non solo i settori produttivi, cognitivi e comunicativi, ma anche le fondamenta teoriche del diritto, in particolare quelle relative alla struttura della responsabilità giuridica. L'intelligenza artificiale (IA), specie nella sua declinazione più recente e pervasiva – quella basata su modelli di apprendimento automatico e reti neurali profonde – produce effetti giuridicamente rilevanti senza agire secondo le categorie classiche dell'intenzionalità, della volontarietà o della consapevolezza. L'IA non è una semplice estensione strumentale dell'agire umano, bensì un ibrido sistemico: è al contempo agente tecnico e ambiente normativo, fattore causale e infrastruttura abilitante. In altre parole, essa non si limita a eseguire comandi, ma co-costruisce scenari d'azione, influenzando dinamicamente i comportamenti individuali e collettivi attraverso *affordances*, *feedback*, automatismi adattivi e operazioni predittive.
Questo mutamento ontologico e funzionale rende sempre più evidente l'inadeguatezza del modello giuridico tradizionale della responsabilità, fondato su presupposti antropocentrici e meccanicistici: colpa, imputabilità, nesso di causalità lineare. Tali categorie, consolidate nella modernità

* Sebbene il presente scritto costituisca il frutto di riflessioni condivise degli autori, Enrico Maestri ha redatto i paragrafi 1, 2, 3, 4, 5, 6, 7; Carlotta Mognato i paragrafi 8, 9, 10. Su determinazione della direzione, il contributo è stato sottoposto a referaggio anonimo in conformità all'art. 15 del regolamento della Rivista

giuridica, presuppongono un soggetto pienamente intenzionale, razionale e dotato di *agency* consapevole. Ma di fronte ad architetture algoritmiche capaci di apprendere, adattarsi, operare in autonomia e generare decisioni opache anche per i loro stessi creatori, questo paradigma va in crisi. Non è più possibile attribuire la responsabilità attraverso un'analisi lineare dell'evento lesivo, né isolare un "autore" come fonte unica del danno.
Si impone, dunque, un ripensamento radicale della funzione stessa della responsabilità giuridica. Da strumento sanzionatorio e distributivo *ex post*, essa deve trasformarsi in dispositivo regolativo *ex ante*, capace di incidere sulle condizioni di possibilità dell'azione algoritmica. Il passaggio cruciale è quello dalla persona ragionevole – fulcro dell'imputazione nel diritto civile – al sistema responsabile: una configurazione normativa che distribuisce il dovere di cura non più soltanto tra individui, ma tra architetture tecniche, procedure decisionali, modelli di design e logiche computazionali. In questa nuova cornice, la responsabilità si configura come proprietà emergente di ecosistemi sociotecnici complessi, e il diritto è chiamato a operare non solo come reazione, ma come progettazione: non tanto per rispondere al danno, quanto per prevenirlo, tracciarlo, contestarlo, e, quando necessario, redistribuirlo secondo criteri equi e funzionali.
Sul piano normativo, il quadro europeo si è assestato sul trittico AI Act, *Cyber Resilience Act* e nuova *Product Liability Directive*, che spostano l'asse verso obblighi di conformità, sicurezza del software e difetto da prodotto applicato anche al software/IA[1].
Il paragrafo 2 ricostruisce la struttura della colpa e ne evidenzia le aporie nell'interazione con sistemi di IA; il paragrafo 3 mostra la crisi del modello tradizionale di imputazione e il contestuale spostamento europeo verso strumenti di conformità *ex ante*. Il paragrafo 4 mappa sei vie teoriche della responsabilità. Il paragrafo 5 rilegge la responsabilità vicaria; il paragrafo 6 discute il modello "da sistema"; il paragrafo 7 critica l'ipotesi del *reasonable robot*. Il paragrafo 8 sviluppa il caso clinico sanitario con le tendenze regolatorie UE/USA; il paragrafo 9 offre la comparazione giuridica UE–USA. Il paragrafo 10 conclude sul passaggio "dalla colpa al codice".

2. La struttura della colpa e le sfide introdotte dall'IA

Nel diritto civile, la responsabilità per colpa si fonda tradizionalmente su quattro requisiti: (1) l'esistenza di un dovere di diligenza; (2) la sua violazione, valutata rispetto allo standard della persona ragionevole; (3) la produzione di un danno giuridicamente rilevante; (4) il nesso causale tra condotta e pregiudizio. La giurisprudenza ha da sempre sottolineato come la colpa consista in negligenza, imprudenza o imperizia, forme che delineano un quadro di prevedibilità e controllo dell'azione.
L'emergere dei sistemi di intelligenza artificiale, complessi, opachi e adattivi, ha incrinato questo modello. La difficoltà di ricostruire il nesso causale

[1] Regolamento (UE) 2024/1689 (*AI Act*), G.U. 12 luglio 2024; direttiva (UE) 2024/2853 (*Product Liability Directive*).

– a causa della non linearità degli algoritmi e della natura auto-apprendente dei processi – rende incerto lo stesso criterio di prevedibilità. Ne consegue che lo schema della negligenza, fondato sulla razionalità e volontarietà dell'agire umano, mostra una sostanziale inadeguatezza. Per questo motivo sono state avanzate soluzioni alternative, come la responsabilità oggettiva per rischio, la responsabilità vicaria e l'*accountability by design*.

3. La crisi del modello tradizionale di imputazione della responsabilità

Nel paradigma giuridico moderno, la responsabilità è tradizionalmente inquadrata all'interno di un sistema di modelli normativi autonomi, ciascuno fondato su principi propri e su presupposti specifici di imputazione: responsabilità civile per colpa, responsabilità oggettiva, responsabilità contrattuale, responsabilità penale. Questo impianto si rivela sempre meno adeguato quando si tratta di sistemi tecnologici complessi come quelli basati sull'intelligenza artificiale. Come osservato da Bertolini (2025)[2], i modelli classici della responsabilità civile — fondati su imputazione individuale, linearità causale e violazione di regole predeterminate — risultano inadeguati di fronte alla complessità dei sistemi di intelligenza artificiale. Tali modelli, efficaci in contesti a bassa complessità, faticano a operare quando l'azione è il risultato di interazioni sistemiche tra architetture algoritmiche, ambienti digitali e attori umani e non umani. L'azione non è più il prodotto lineare di una volontà individuale, ma il risultato di un'interazione continua e retroattiva tra molteplici livelli: dati, architetture algoritmiche, ambienti computazionali, utenti, progettisti, operatori economici, piattaforme. L'evento lesivo, in questi casi, non è riconducibile a un unico attore, né a un'unica condotta giuridicamente rilevante: è il frutto di un sistema di relazioni dinamiche e non sempre osservabili, che rendono estremamente difficile, se non impossibile, attribuire la responsabilità secondo i criteri tradizionali[3].

Il sistema compartimentale, in questo contesto, rischia di produrre effetti distorsivi: o si attribuisce la responsabilità in modo arbitrario, per assecondare l'illusione di una coerenza giuridica formale, oppure si lascia scoperto l'anello più rilevante della catena causale, quello sistemico, che sfugge alle categorie giuridiche convenzionali. In entrambi i casi, il diritto fallisce la sua funzione regolativa.

È dunque necessario superare l'idea che la responsabilità sia un'operazione di "etichettatura giuridica" *ex post*, fondata su griglie predeterminate di qualificazione normativa.

Per superare l'impasse, occorre ripensare la responsabilità come infrastruttura in grado di cogliere la natura relazionale e ambientale dei sistemi tecnologici: dalla linearità alla rete, dalla colpa all'*accountability*, dalla sog-

[2] A. Bertolini, *Intelligenza artificiale e responsabilità civile. Problema, sistema, funzioni*, Bologna, 2025.

[3] V. Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*, New York, 2018.

gettività individuale all'infrastruttura sociotecnica[4]. In Europa, la proposta di *AI Liability Directive* (COM(2022) 496) aveva tentato di riequilibrare gli oneri probatori (presunzione di causalità, disclosure), ma è stata ritirata nel 2025. La traiettoria si è consolidata sul combinato disposto AI Act – *Cyber Resilience Act* – nuova *Product Liability Directive*, che configura un approccio ibrido, tra responsabilità *ex post* e conformità *ex ante* e che trasla l'enfasi verso un regime di conformità tecnico-documentale e una responsabilità da prodotto aggiornata al software/IA, agevolando la prova attraverso documentazione, *audit* e tracciabilità. Il ritiro nel 2025 della proposta di *AI Liability Directive* conferma questo spostamento: la funzione riequilibratrice degli oneri probatori è oggi perseguita indirettamente tramite documentazione obbligatoria e tracciabilità tecnico-organizzativa[5].

La trasformazione è anche filosofica: Hildebrandt[6] propone la responsabilità by design, ossia l'incorporazione di tracciabilità e contestabilità nei sistemi già in fase di progettazione; Floridi[7] insiste sulla natura ambientale dell'IA, che impone di distribuire la responsabilità lungo tutti i livelli della rete. Sul piano epistemologico si produce un doppio scarto: anzitutto, la causalità invocata dal diritto non è più ricostruibile in termini deterministici, ma — quando pertinente — solo probabilistici; inoltre, di fronte ai sistemi generativi anche il lessico della causalità probabilistica risulta insufficiente, perché tali sistemi non stimano relazioni causali tra variabili del mondo, bensì mettono in scena coerenze linguistiche e operative, formulando ipotesi pragmatiche a partire dal contesto. È una logica performativa: il modello non spiega il reale, ma propone congetture funzionanti qui-ed-ora; le cosiddette "allucinazioni" non sono mero rumore statistico, bensì falsi positivi inferenziali, l'altra faccia della capacità congetturale quando l'informazione è incompleta.

Se prendiamo sul serio questa non-causalità interna, cambia l'oggetto della responsabilità. Non ha senso chiedere al giudice di estrarre dalla *black box* una catena causa/effetto che il sistema non rappresenta simbolicamente. La prova va spostata all'esterno del modello, verso l'ambiente sociotecnico che ne abilita l'azione: dati, regole di aggiornamento, tracciabilità, *human oversight*, avvertenze e procedure d'uso. La causalità diventa situata: si accerta mostrando che una governance ragionevolmente praticabile avrebbe impedito la congettura fuorviante o ne avrebbe neutralizzato gli effetti. La colpa si traduce in deficit di progettazione/uso rispetto a cautele esigibili, il nesso in coerenza strutturale ricostruita tramite *log* e *change control*, l'imputazione segue i nodi di controllo che potevano interrompere la traiettoria del danno[8].

---

[4] Z. Porter-P.R. Conmy-P. Morgan-J. Al-Qaddoumi-B. Twomey-J. McDermid-I. Habli, *Unravelling Responsibility for AI*, York Research, 2023 (working paper).

[5] COM(2022) 496**,** Proposta di direttiva sulla responsabilità civile per l'IA (*AI Liability Directive*), ritirata nel 2025; European Parliamentary Research Service (EPRS)**,** *Artificial Intelligence and Civil Liability*, Bruxelles, 2025.

[6] M. Hildebrandt, *Smart Technologies and the End(s) of Law*, Cheltenham, 2015.

[7] L. Floridi, *Etica dell'intelligenza artificiale. Sviluppi, opportunità, sfide*, Milano, 2022.

[8] E. Maestri, *Contro la riduzione bayesiana dell'intelligenza artificiale*, in *Pandora Rivista*, 2025.

4. Le vie della responsabilità nell'era dell'IA

Il dibattito contemporaneo ha proposto una costellazione di approcci tra loro complementari — dall'allocazione efficiente del rischio[9] all'accountability computazionale[10], dall'etica dell'infosfera[11] alla responsabilità multilivello[12], fino alle presunzioni riequilibratrici[13] e alle tassonomie funzionali e alla responsabilità "da sistema"[14] — che, letti in sintesi, spostano l'asse dalla centralità della colpa individuale a una concezione distribuita e preventiva della responsabilità, sostituendo alla causalità lineare una governance ambientale dell'azione algoritmica; anche il filone sul reasonable robot e sull'ipotesi di uno standard di negligenza specifico per l'IA conferma che il problema non è "soggettivare" la macchina, ma ridefinire gli standard applicabili all'umano e all'organizzazione che la progetta, integra e utilizza[15]. Questa riconfigurazione resta però pienamente traducibile negli istituti classici: il dovere di diligenza si amplia in senso organizzativo e documentale (selezione di sistemi idonei e conformi, human oversight effettivo, tracciabilità, gestione del drift e degli aggiornamenti; lato fornitore: risk management, data governance, validazioni e avvertenze), con l'operatività di proxy tecnico-giuridici quali il NIST AI RMF 1.0 e, in sanità, i principi GMLP[16]; la violazione/colpa non scompare ma si specifica come deficit verificabile di progettazione o di uso rispetto a regole tecniche e regolatorie (*log, audit*, change control, data sheets); la causalità si accerta mediante tracce tecniche (audit trail, *versioning*, registri di conformità) e confronto con alternative progettuali/operative ragionevoli, mentre i *Predetermined Change Control Plans* (PCCP) rendono "auditabile" ex post la catena causale delle modifiche di modello[17]; l'imputazione si riallinea alla posizione sistemica di controllo/beneficio, valorizzando vicarietà e responsabilità d'impresa lungo la filiera; l'onere della prova si riorganizza perché una quota della prova è "incorporata" nella documentazione di conformità (accountability by design), con presunzioni ragionevoli a riequilibrare l'asimmetria informativa senza scivolare in strict liability generalizzata, e con il riconoscimento—in sede europea—del software come "prodotto" che irrobustisce l'azione per difetto senza colpa[18]; i rimedi si orientano a combinare compensazione e deterrenza, calibrando ordini di adeguamento e

[9] G. Calabresi, *The Costs of Accidents: A Legal and Economic Analysis*, New Haven, 1970.

[10] M. Hildebrandt, *Law for Computer Scientists and Other Folk*, Oxford, 2020.

[11] L. Floridi, *The Ethics of Information*, Oxford, 2013.

[12] V. Dignum, *Responsibility and Artificial Intelligence*, in M. Dubber-F. Pasquale-S. Das (eds.), *The Oxford Handbook of Ethics of AI*, Oxford, 2020, 215 ss.

[13] A. Vasudevan, *Addressing the Liability Gap in Artificial Intelligence*, CIGI Policy Brief n. 177, Waterloo (ON), luglio 2023.

[14] Z. Porter et al., *Unravelling Responsibility for AI*, cit.; A. Beckers-G. Teubner, *Three Liability Regimes for Artificial Intelligence: Algorithmic Actants, Hybrids, Crowds*, London, 2022.

[15] M.E. Diamantis, *Reasonable AI: A Negligence Standard*, in *Vanderbilt Law Review*, 78(2), 2025, 573 ss.

[16] FDA, *Predetermined Change Control Plans (PCCP) – Final Guidance*, 2024 (SaMD AI/ML); MDCG 2025-6, *FAQ on the Interplay between MDR/IVDR and the AI Act*, 2025.

[17] Regolamento (UE) 2025/327 (*European Health Data Space*), 11 febbraio 2025.

[18] Direttiva (UE) 2024/2853 (*Product Liability Directive*).

post-market fixes sulla capacità di prevenzione; infine, nel concorso multi-attore la ripartizione segue i canoni tradizionali (gravità e causalità efficiente) ma è informata da un criterio funzionale: chi, in quale punto della catena, poteva interrompere la traiettoria del danno (un alert ignorato, un aggiornamento non implementato, un'avvertenza non trasmessa).
In questo quadro, l'impostazione funzionalista di Calabresi[19] offre la bussola allocativa: attribuire i costi del danno a chi può prevenire meglio il rischio e internalizzarne gli effetti (*prevention/insurance/deterrence*). Questo criterio però tende a rimanere cieco alla dimensione epistemico-architetturale dei sistemi di IA, dove l'opacità dei modelli e la non-linearità delle catene causali rendono difficile una ricostruzione *ex post*. Da qui la mossa di Hildebrandt, in opere come *Smart Technologies and the End(s) of Law*[20] e *Law for Computer Scientists*[21], dove sposta la responsabilità *ex ante* nelle architetture: tracciabilità, contestabilità e spiegabilità devono essere vincoli di progetto (*log*, *audit trail*, registri di versione, model/data sheets), non meri oneri postumi, così che una quota della prova e della motivazione sia "incorporata" nel sistema. In parallelo, la torsione ontologica di Floridi che in testi come The Philosophy of Information[22] e The Fourth Revolution[23] ricolloca la responsabilità come cura dell'ambiente informazionale (distributed morality)[24]: evitare l'antropomorfizzazione normativa dell'IA e trattare i sistemi come pazienti morali entro ecosistemi da progettare fair-by-design (equità informazionale, *agency* distribuita, umiltà ontologica). Su questo sfondo, Dignum[25] stratifica i livelli della responsabilità (individuale, collettivo, sistemico) e li rende governabili tramite processi partecipativi, formazione e trasparenza, collegando le scelte di design a pratiche organizzative verificabili. Per colmare il *gap* imputativo che l'opacità lascia aperto, le presunzioni riequilibratrici di Vasudevan[26] invertono pragmaticamente l'onere probatorio (chi trae beneficio deve dimostrare di aver agito con diligenza), in coerenza con le proposte dell'AI Liability Directive: così si riequilibra l'asimmetria informativa senza scivolare in una strict liability generalizzata. Infine, la tassonomia funzionale di Porter[27] scompone la responsabilità nei suoi assi (causale, di ruolo, di answerability/*accountability*, morale) e consente di allineare scopi e strumenti: deterrenza e compensazione, distribuzione del rischio e chiarificazione epistemica. In breve, queste "nuove" teorie non scardinano la dogmatica: ne re-ingegnerizzano i contenuti, riportando dovere, colpa, causalità, imputazione, prova e rimedi dentro l'ambiente computazionale che oggi struttura l'azione e la responsabilità (anche attraverso *standard* e piani

[19] G. Calabresi, *The Costs of Accidents: A Legal and Economic Analysis*, cit.

[20] M. Hildebrandt, *Smart Technologies and the End(s) of Law*, cit.

[21] M. Hildebrandt, *Law for Computer Scientists and Other Folk*, Oxford, 2020.

[22] L. Floridi, *The Ethics of Information*, cit.

[23] L. Floridi, *The Fourth Revolution: How the Infosphere is Reshaping Human Reality*, Oxford, 2014.

[24] L. Floridi, *Etica dell'intelligenza artificiale. Sviluppi, opportunità, sfide*, Milano, 2022.

[25] V. Dignum, *Responsible Artificial Intelligence. How to Develop and Use AI in a Responsible Way*, Cham, 2019.

[26] A. Vasudevan, *Addressing the Liability Gap in Artificial Intelligence*, CIGI Policy Brief n. 177, Waterloo (ON), 2023.

[27] Z. Porter et al., *Unravelling Responsibility for AI*, cit.

di modifica predeterminata, come i PCCP in ambito medical-AI).

5. La responsabilità vicaria dell'umano per l'agire artificiale: continuità e discontinuità nel diritto civile

Nel dibattito contemporaneo sulla responsabilità per danno da intelligenza artificiale, si è affacciata con rinnovato interesse la possibilità di reinterpretare in chiave tecnologica il paradigma della responsabilità vicaria. Questo modello, noto nel diritto civile per la responsabilità dei genitori (art. 2048 c.c.), dei padroni e committenti (art. 2049 c.c.), si fonda su una logica relazionale e fiduciaria: non si punisce chi agisce direttamente, ma chi – per posizione, controllo o beneficio – deve rispondere degli effetti altrui.
L'interesse per questo schema deriva dalla constatazione che l'IA non è soggetto di diritto, né può essere imputata in senso tecnico. In assenza di volontà soggettiva e di colpevolezza, la soluzione potrebbe consistere nel traslare la responsabilità sull'umano "dominante": il produttore, il fornitore, il professionista, l'utilizzatore. Il titolare della responsabilità è identificato in base alla posizione sistemica: risponde chi controlla o beneficia del sistema, con una logica di custodia del rischio che evita finzioni antropomorfiche e consente tutela effettiva. La responsabilità diventa allora una funzione del potere: risponde chi esercita il controllo o trae vantaggio, indipendentemente dalla prova di una propria colpa diretta.
In questo senso, la responsabilità non è più solo imputazione *ex post*, ma dovere di presidio preventivo, costruzione dell'*habitat* operativo dell'IA. Il danno non è tanto un evento imprevedibile quanto il segnale di un fallimento progettuale o organizzativo. L'introduzione nella proposta di direttiva UE di una presunzione di causalità per sistemi di IA ad alto rischio va esattamente in questa direzione: più alta è la pericolosità potenziale del sistema, maggiore è il dovere di architettura normativa e tecnica.
Questo modello solleva alcune tensioni: da un lato, preserva la coerenza sistemica con i principi classici del diritto civile, evitando l'introduzione di entità artificiali come soggetti giuridici; dall'altro, rischia di scaricare sul solo umano una responsabilità non più controllabile *ex ante*, data la complessità e l'autonomia operativa dei sistemi di *machine learning*.
A differenza dell'approccio allocativo di Calabresi[28], che vede nella responsabilità uno strumento per redistribuire in modo efficiente i costi sociali del danno, il modello vicario si fonda su una presunzione etico-giuridica di custodia. Ma come mostrano Hildebrandt[29] e Floridi[30], l'azione dell'IA si configura sempre più come evento emergente da una rete distribuita di agenti, regole e ambienti. In tal senso, la responsabilità non può essere ridotta a una relazione binaria tra soggetto dominante e oggetto dannoso. Occorre pensare in termini di accountability sistemica, dove la responsabilità è un attributo collettivo, organizzativo e progettuale.

[28] G. Calabresi, *The Costs of Accidents: A Legal and Economic Analysis*, cit.

[29] M.Hildebrandt, *Law for Computer Scientists and Other Folk*, cit.

[30] L. Floridi, *Etica dell'intelligenza artificiale. Sviluppi, opportunità, sfide*, cit.

Se dunque la responsabilità vicaria rappresenta una soluzione transitoria e pragmatica, capace di garantire tutela senza scardinare l'architettura codicistica, essa non può esaurire il problema teorico. In prospettiva, sarà necessario passare da una responsabilità per intermediazione soggettiva a una responsabilità per disegno ambientale, nella quale la configurazione stessa dell'ecosistema digitale diventa luogo di imputazione e di prevenzione.
In questo senso, la responsabilità vicaria non può più essere letta come una mera delega di imputazione, ma come una forma specifica di responsabilità ambientale, in cui l'umano è responsabile della co-costruzione del sistema, dei suoi output e dei suoi *failure points*[31].

6. I tre regimi funzionali della responsabilità nell'era algoritmica

Il recente lavoro di Anna Beckers e Gunther Teubner[32] propone una tipologia innovativa delle forme di responsabilità giuridica nell'era dell'intelligenza artificiale, superando le categorie classiche della responsabilità soggettiva. La loro analisi distingue tra tre configurazioni operative – attori algoritmici, sistemi ibridi e folle interconnesse – ognuna delle quali richiede un regime normativo specifico, fondato su logiche di imputazione funzionale piuttosto che su attribuzioni colpose.
Nel caso degli attori algoritmici, l'approccio suggerito si ispira alla figura della responsabilità vicaria, assimilando il rapporto tra progettista o titolare del sistema e comportamento dell'IA al modello tra *dominus* e agente. Sebbene sprovvista di intenzionalità cosciente, l'intelligenza artificiale viene considerata un attante socio-digitale, capace di produrre effetti giuridicamente rilevanti imputabili a chi ne determina l'autonomia operativa.
Per i sistemi ibridi, che vedono l'interazione strutturale tra esseri umani e IA, Beckers e Teubner propongono una forma di responsabilità d'impresa reticolare, fondata su criteri proporzionali di imputazione. Tale modello considera la rete sociotecnica come un insieme composito in cui ogni nodo partecipa causalmente e decisionisticamente all'esito finale, rendendo necessario un criterio distribuito di responsabilità.
Infine, di fronte alle folle interconnesse – sistemi distribuiti, auto-organizzati e altamente interconnessi – il diritto tradizionale non riesce più a individuare un singolo soggetto responsabile. Per queste configurazioni si invoca una forma di responsabilità collettiva o da fondo, ispirata ai meccanismi di socializzazione del rischio, come i fondi compensativi pubblici, che garantiscono tutela in assenza di imputazioni individuali identificabili.
Questa articolazione si integra con la visione sistemica della responsabilità delineata nel presente saggio. Gli attanti algoritmici, descritti come soggettività emergenti co-costruite in ambienti digitali, trovano ora una sistematizzazione normativa coerente con le architetture tecniche dell'azione computazionale. La responsabilità, dunque, non va più intesa come

[31] R. Abbott, *The Reasonable Robot: Artificial Intelligence and the Law*, Cambridge, 2020.

[32] A. Beckers-G. Teubner, *Three Liability Regimes for Artificial Intelligence,* cit., 2022.

etichetta post-fattuale applicata *ex post*, bensì come dispositivo di governance *ex ante*, che si articola lungo l'asse che va dalla colpa individuale alla progettazione responsabile, dalla causalità lineare all'architettura sistemica del rischio.

7. Il paradosso del robot ragionevole

Nel suo volume *The Reasonable Robot*[33], Ryan Abbott formula una proposta provocatoria quanto ambiziosa: estendere all'intelligenza artificiale lo stesso *standard* giuridico applicabile alla persona umana. In particolare, egli suggerisce che, laddove un sistema automatizzato si comporti come farebbe una "persona ragionevole", esso non dovrebbe essere considerato responsabile per un eventuale danno, e che, simmetricamente, anche gli esseri umani dovrebbero essere valutati, in certi casi, in base a ciò che un'IA sufficientemente avanzata avrebbe fatto nella medesima situazione. Questo principio – definito di "neutralità legale" – mira a ridurre le asimmetrie normative che oggi gravano sull'adozione di tecnologie intelligenti, offrendo così un incentivo all'innovazione: se l'IA si comporta come (o meglio di) un essere umano, non dovrebbe essere trattata più severamente solo in quanto artefatto.

Questo approccio ha un certo fascino intuitivo, perché sembra estendere in modo lineare gli *standard* giuridici preesistenti. Tuttavia, presenta gravi criticità teoriche e pratiche. In primo luogo, il concetto di "ragionevolezza" è radicato in una semantica umana, che presuppone capacità di giudizio, equilibrio tra interessi e valutazione di contesto. Trasferirlo a un algoritmo rischia di essere una mera finzione, incapace di cogliere la logica statistico-probabilistica del *machine learning*.

In secondo luogo, l'adozione del "robot ragionevole" potrebbe produrre una normalizzazione pericolosa: invece di ripensare le categorie giuridiche, si ridurrebbe l'IA a una caricatura antropomorfica. In tal modo, si finirebbe per occultare i veri nodi, ossia la distribuzione delle responsabilità tra sviluppatori, fornitori, utenti e istituzioni, e la necessità di *accountability* strutturale.

In questa prospettiva, l'agente ragionevole dell'IA resta una finzione euristica più che uno *standard* operativo.

Il paradosso del "robot ragionevole" è quindi duplice: appare come un ponte tra vecchio e nuovo diritto, ma in realtà rischia di bloccare l'evoluzione del pensiero giuridico, rinchiudendolo in un antropocentrismo obsoleto. Più che costruire un "robot ragionevole", il diritto deve elaborare concetti adeguati alla specificità dei sistemi algoritmici, accettando la sfida di un paradigma post-antropomorfico.

[33] R. Abbott, *The Reasonable Robot: Artificial Intelligence and the Law*, cit.

8. Sanità digitale e responsabilità: un caso clinico

La sanità digitale è il banco di prova più esigente per la responsabilità nell'era algoritmica. L'esempio – una diagnosi oncologica mancata per falso negativo IA – mostra come l'evento lesivo emerga dall'interazione tra modello, dati, contesto clinico e condotta professionale. Per evitare astrazioni, l'esempio va collocato nelle tendenze regolatorie in atto.

Il quadro regolatorio e comparato di riferimento è illustrato al paragrafo 9; qui assumiamo soltanto gli *standard* minimi oggi esigibili (tracciabilità, supervisione effettiva, gestione degli aggiornamenti e della deriva del modello, validazione clinicamente rilevante; per l'ambito medicale: principi GMLP e piani di cambiamento predeterminato per i modelli che evolvono)[34].

In ambito medico, consideriamo il falso negativo oncologico prodotto da un sistema di supporto alla decisione.

Lo *standard* di diligenza, nel contesto clinico, non è una formula astratta ma la risultante di tre piani convergenti: progetto e fornitura (qualità e rappresentatività dei dati, validazioni tecniche e cliniche pertinenti allo scopo d'uso, limiti e avvertenze comprensibili, tracciabilità delle versioni); integrazione ospedaliera (procedure per l'uso appropriato, *human oversight* sostanziale, gestione del *drift*, controllo degli aggiornamenti, canali di segnalazione); condotta professionale (uso critico dell'output, confronto con il quadro clinico, motivazione delle deviazioni). È, in altre parole, uno *standard* organizzativo-documentale prima che cognitivo: la diligenza si misura su ciò che è stato reso verificabile tramite *log, versioning e audit*[35].

La violazione si manifesta come deficit verificabile rispetto a tale *standard*: dataset non idonei allo scopo; validazioni cliniche non rappresentative; *log* lacunosi; *oversight* di facciata; avvertenze non trasmesse o ignorate; uso fuori indicazione; aggiornamenti non eseguiti o non tracciati. Nel rapporto medico–paziente, l'eventuale responsabilità del professionista — quando sussiste — resta responsabilità per colpa: tipicamente negligenza nella supervisione dell'output algoritmico quando il clinico si limita a recepirlo malgrado segnali dissonanti presenti nel caso[36].

Il nesso causale si accerta come coerenza strutturale: se i *log* mostrano che il sistema operava fuori dal perimetro previsto; se la documentazione rivela un governo inadeguato del modello o dati non pertinenti allo scopo clinico; se mancano i controlli dichiarati, è ragionevole inferire che il danno discenda da quel difetto di governance. Il controfattuale non interpella «come avrebbe dovuto pensare la macchina», ma quale governance praticabile (dati/validazioni, *oversight*, avvertenze, aggiornamenti) avrebbe

[34] FDA, *Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions*, 2025; FDA – MHRA – Health Canada, *Good Machine Learning Practice (GMLP) – Guiding Principles*, 2021/2025; FDA, *Transparency for Machine Learning-Enabled Medical Devices – Guiding Principles*, 2021/2024.

[35] Y. Jia-J. McDermid-T. Lawton-I. Habli, *The Role of Explainability in Assuring Safety of Machine Learning in Healthcare*, in *arXiv*: 2109.00520, 2022.

[36] E.S. Andersen et al., *Monitoring performance of clinical artificial intelligence in deployment: a systematic review*, in *NPJ Digital Medicine*, 2024; C. Terranova et al., *AI and professional liability assessment in healthcare: a medico-legal perspective*, in *Frontiers in Artificial Intelligence*, 2024.

impedito o interrotto la traiettoria del danno[37].
L'imputazione segue i nodi di controllo della filiera: al produttore/fornitore per difetti di progetto, validazioni inadeguate, avvertenze insufficienti; alla struttura per carenze di integrazione (procedure, formazione, gestione aggiornamenti, canali di segnalazione); al professionista per l'errore clinico residuo. Il riparto è proporzionale alla capacità effettiva di interrompere la traiettoria del danno nel punto di controllo di ciascuno.
Quanto ai rimedi, il risarcimento va accompagnato da misure correttive che trasformino l'esito contenzioso in apprendimento istituzionale: piani di *change control* e ri-validazione, rafforzamento dei flussi di tracciabilità e, nei casi-limite, protocolli di *second reading*/consulto; l'evidenza empirica (mammografia e screening prospettici) mostra che la doppia lettura calibrata sull'uso dell'IA riduce errori e incertezze[38]. Così la decisione giudiziaria diventa leva di prevenzione, riallineando *ex ante* progetto, integrazione e uso.

9. Comparazione UE–USA: responsabilità da IA e sanità digitale

Una comparazione utile non si esaurisce nel contrasto, ormai divenuto cliché, tra approcci *ex ante* ed ex post. La vera frattura riguarda "dove" ciascun ordinamento colloca il baricentro della responsabilità e come organizza gli strumenti della prova. L'Unione europea tende a istituzionalizzare la prevenzione: la responsabilità privata è incardinata dentro un'architettura di conformità tecnico-documentale che produce, già in fase di progettazione e di impiego, le tracce necessarie a ricostruire gli eventi lesivi[39]. Gli Stati Uniti, al contrario, *giudizializzano* la prevenzione[40]: lo *standard* di diligenza matura soprattutto ex post nella dialettica del contenzioso, in cui *discovery*, periti ed esiti negoziali (*settlement*) stabiliscono, caso per caso, che cosa fosse ragionevolmente esigibile[41]. Ne derivano incentivi, oneri probatori e stili decisionali profondamente diversi[42].
In Europa, l'AI Act non sostituisce l'*EU Medical Device Regulation* (MDR) e il *In Vitro Diagnostic Regulation* (IVDR): i software medicali restano nelle classi di rischio dei Reg. (UE) 2017/745 e 2017/746, con sorveglianza *post-market* e organismi notificati. L'"alto rischio" dell'AI Act attiene alla

[37] U.S. Government Accountability Office (GAO), *Medical Devices: FDA Has Begun Building an Active Surveillance System*, July 2024.

[38] S. Taylor-Phillips-C. Stinton. *Double reading in breast cancer screening: considerations for policy-making*; in *The British journal of radiology*, vol. 93,1106, 2020, 20190610; Y.W. Chang et al., *Artificial intelligence for breast cancer screening in population-based practice (Screen Trust CAD)*, in *Nature Communications*, 2025, 2248.

[39] D. Vogel, *The Politics of Precaution: Regulating Health, Safety, and Environmental Risks in Europe and the United States*, Princeton, 2012.

[40] S. I. Strong, *Regulatory Litigation in the European Union: Does the U.S. Class Action Have a New Analogue?* in *Notre Dame Law Review*, 88, 2012, 899 ss.

[41] R.A. Kagan, *Adversarial Legalism: The American Way of Law*, Harvard, 2019.

[42] J.B. Wiener-M.D. Rogers-J.K. Hammitt-P.H. Sand, (eds.), *The Reality of Precaution: Comparing Risk Regulation in the United States and Europe*, London, 2011.

governance dell'IA, non alla classe del dispositivo, come chiarito nella prassi del *Medical Device Coordination Group* (MDCG)[43]. A completare il quadro, il *Cyber Resilience Act* estende al software obblighi di sicurezza lungo il ciclo di vita (applicazione principale dall'11 dicembre 2027), mentre la *Product Liability Directive* (2024/2853)[44] chiarisce che anche il software è "prodotto" ai fini della responsabilità senza colpa. L'*European Health Data Space* (EHDS, Reg. 2025/327)[45] apre canali legali e tecnici per l'uso secondario dei dati sanitari (addestramento, *testing*, valutazione)[46], elevando lo *standard* atteso sulla qualità dei dati e la forza probatoria della governance[47]. L'effetto è sistemico: la responsabilità non "scompare" a favore della regolazione, ma si riallinea alla conformità. Non conformità documentali (*log* assenti, *risk management* lacunoso, validazioni insufficienti, *drift* non governato, *human oversight* fittizio) diventano snodi giuridicamente rilevanti, capaci di sostenere la prova sia del difetto sia della violazione dello *standard* di diligenza. È significativo, in questo quadro, che la conformità non operi come scudo assoluto: l'adempimento regolatorio rileva, ma non elide la responsabilità quando l'evento rivela che il rischio era comunque mal governato. Il vantaggio evidente di questa impostazione è la ricostruibilità: i fascicoli tecnici, i registri, i *change log* e gli *audit* richiesti *ex lege* permettono una "prova strutturale" del nesso che non dipende dal penetramento della *black box*, bensì dalla coerenza dell'ecosistema di governance. Il rovescio della medaglia è il rischio di conformità rituale: la promessa di prevenzione può appiattirsi in adempimenti formali se mancano controlli sostanziali; inoltre, l'elevato costo organizzativo può pesare in modo sproporzionato su operatori medio-piccoli[48].

Negli Stati Uniti, il quadro resta policentrico e reattivo: la trama è fatta di *tort law* statale e *product liability* (secondo le categorie del *Restatement (Third)*: difetto di fabbricazione, di progetto, di avvertenze), integrata da una regolazione federale selettiva — in sanità, la FDA per il *Software as a Medical Device* (SaMD) — e orientata da *soft law* tecnico come il NIST AI RMF e, più di recente, il *Generative AI Profile*[49]. La vera novità statunitense è la gestione del modello che apprende nel tempo: i *Predetermined Change Control Plans* (PCCP) consentono di pre-autorizzare traiettorie di aggiornamento; in tal modo, l'evoluzione dell'algoritmo non è un evento opaco e postumo, ma un processo sorvegliato e verificabile. Sul piano della responsabilità civile, l'effetto resta ambivalente: da un lato, la disponibilità di *design history file*, piani PCCP e validazioni arricchisce la prova; dall'altro, il sistema rima-

[43] MDCG 2025-6, FAQ on the interplay between MDR/IVDR and the AI Act (Juny 2025).

[44] Direttiva (UE) 2024/2853 (*Product Liability Directive*).

[45] Regolamento (UE) 2025/327 *(European Health Data Space)*, 11 febbraio 2025.

[46] K.G. van Leeuwen-L. Doorn-E. Gelderblom, *The AI Act: responsibilities and obligations for healthcare professionals and organizations*, in *Diagnostic and Interventional Radiology*, 2025.

[47] E.P. Vardas et al., *Medicine, healthcare and the AI Act: gaps, challenges and opportunities*, in *European Heart Journal – Digital Health*, 2025, 833 ss.

[48] M. Power, *The Audit Society: Rituals of Verification*, Oxford, 1997.

[49] NIST, *AI Risk Management Framework 1.0*, Gaithersburg, 2023; NIST, *Generative AI Profile*, Gaithersburg, 2024; FDA, *Predetermined Change Control Plans (PCCP) – Final Guidance*, dicembre 2024 (SaMD AI/ML).

ne esposto a frammentazione e incertezza — l'esito dipende dal foro, dal test adottato nel *design defect* (*risk-utility* vs *consumer expectations*), dall'orientamento del giudice sulle avvertenze e dalla qualità della *expert testimony*[50].
In sanità, due snodi delimitano ulteriormente il campo: la *pre-emption* per i dispositivi PMA (*Riegel v. Medtronic*)[51] e la *learned intermediary doctrine*, che canalizza il dovere di avvertimento verso il medico. La conseguenza è che la responsabilità del produttore tende a giocarsi sul progetto e sulle avvertenze, mentre quella dell'ospedale e del clinico si misura sul modo d'uso e sull'integrazione organizzativa del sistema IA.
La sanità digitale mostra con particolare chiarezza la divergenza funzionale tra i due modelli. In Europa, l'ospedale e il fornitore sono valutati come nodi di una catena di conformità: selezione del sistema idoneo allo scopo, *human oversight* effettivo, protocolli di segnalazione, gestione del *drift*, *post-market* e aggiornamenti tracciati. La non conformità di uno solo di questi anelli può costituire il fulcro dell'imputazione, anche quando l'errore clinico individuale non è macroscopico[52]. Negli Stati Uniti, la traiettoria è diversa: il caso giudiziario "costruisce" lo *standard* attraverso il confronto tra ciò che era stato progettato, avvertito e validato e ciò che, secondo periti e precedenti, un operatore ragionevole avrebbe dovuto fare. Il punto di forza è la capacità adattiva del sistema; il punto debole è la selettività: la protezione dipende dalla possibilità (e dal costo) di portare il caso in giudizio, con esiti non uniformi e margini di *forum shopping* [53].
Se si guarda alla prova del nesso, i regimi svelano la propria natura. Nell'UE, la causalità si affronta con una prova di coerenza: se i *log* dicono che il modello era fuori dal perimetro autorizzato, se la qualità dei dati non rispetta lo *standard*, se l'*oversight* era solo nominale, il giudice può inferire ragionevolmente che il danno discende dal difetto di governance; la questione non è penetrare l'algoritmo, ma verificare la struttura che lo rende affidabile. Negli USA, la causalità rimane controfattuale e peritale: si chiede se un progetto alternativo fattibile avrebbe ridotto il rischio, se avvertenze più incisive avrebbero mutato la condotta del clinico, se l'uso prevedibile del dispositivo ricadeva entro le condizioni dichiarate dal produttore. L'una e l'altra via hanno costi e benefici: l'UE internalizza *ex ante* i costi di prevenzione lungo la filiera[54]; gli USA li esternalizzano *ex post* attraverso il contenzioso, che può essere deterrente ma anche (e non di rado) inefficiente.
Non mancano i punti di convergenza. Da entrambi i lati si afferma una gestione di ciclo di vita (EU: obblighi di sorveglianza e *post-market*; USA: TPLC FDA con PCCP), l'idea di tracciabilità come condizione di legalità

[50] S.B. Burbank-S. Farhang, *Private Enforcement of Public Law*, in *Columbia Law Review*, 123, 2013, 1473 ss.

[51] *Riegel v. Medtronic*, 552 U.S. 312 (2008).

[52] D. Vogel, *The Politics of Precaution,* cit.

[53] R.A. Kagan, *Adversarial Legalism,* cit.

[54] L. Bergkamp-L. Kogan, *Trade, the Precautionary Principle, and Post-Modern Regulatory Process. Regulatory Convergence in the Transatlantic Trade and Investment Partnership*, in *European Journal of Risk Regulation,* 4, 2013, 493 ss.

e una valorizzazione, almeno di principio, dell'intervento umano[55]. Ma le assonanze non cancellano la differenza di fondo sul vettore di enforcement: amministrativo e documentale in Europa, giudiziario e competitivo negli Stati Uniti. In termini di politica del diritto, ciò riflette due concezioni della giustizia del rischio: per l'UE la giustizia passa per *standard* pubblici e verificabili; per gli USA passa per responsabilità private e contendibili.
Quale modello funziona meglio per i sistemi che apprendono? L'UE offre certezza *ex ante* e infrastrutture probatorie che abbassano le barriere per il danneggiato, ma rischia di irrigidire l'innovazione se la conformità diventa *checklist* e non *learning compliance*[56]. Gli USA offrono plasticità e una regolazione *case by case* capace di selezionare i comportamenti desiderabili nel mercato, ma pagano in disomogeneità e in costi di accesso alla tutela. Una via matura potrebbe essere ibrida: adottare in Europa strumenti "agili" alla PCCP che riconoscano il valore del cambiamento controllato; valorizzare negli USA, in sede probatoria, la documentazione di governance come *prima facie evidence* di diligenza, senza trasformarla in immunità.
Se letta in questa chiave, la comparazione non consegna un vincitore, ma una agenda di riforma incrociata. All'Europa si chiede di evitare la trappola del formalismo, riconoscendo margini di apprendimento ai modelli entro binari verificabili; agli Stati Uniti si chiede di consolidare *ex ante* gli *standard* minimi di documentazione e tracciabilità, così da non affidare tutto al contenzioso[57]. In sanità, dove il rischio è eminentemente personale, la responsabilità da IA diventa credibile solo se questi due mondi — conformità e contestabilità — si parlano davvero: la prima costruisce condizioni e prove; la seconda verifica, pubblicamente, che quelle condizioni non restino sulla carta.

10. Conclusione: dalla colpa al codice

La responsabilità giuridica nell'era dell'IA non può più essere ancorata unicamente al paradigma della colpa. Tre passaggi risultano decisivi: (i) il tramonto della causalità lineare e della soggettività intenzionale come unici cardini; (ii) la trasformazione della responsabilità da imputazione *ex post* a progettazione ecologica *ex ante* (accountability computazionale, responsabilità vicaria e ambientale); (iii) l'apertura comparatistica: l'Europa costruisce un regime di prevenzione e conformità, gli Stati Uniti affidano gran parte della regolazione ai giudici e al contenzioso. Il caso medico mostra la necessità di un approccio distribuito e multilivello. Il passaggio "dalla colpa al codice" è la trasformazione della responsabilità da categoria individuale a proprietà emergente di sistemi: il diritto deve progettare ambienti normativi che rendano l'azione algoritmica contestabile, trasparente e giusta.

---

[55] Regolamento (UE) 2025/327 (*European Health Data Space*), 11 febbraio 2025; FDA, *Predetermined Change Control Plans (PCCP) – Final Guidance*, dicembre 2024 (SaMD AI/ML).

[56] M. Power, *The Audit Society,* cit.

[57] C.R. Sunstein, *Laws of Fear: Beyond the Precautionary Principle*, Cambridge, 2005.

Abstract

La rivoluzione algoritmica ha investito anche le fondamenta della responsabilità giuridica. I sistemi di intelligenza artificiale, in particolare quelli basati su apprendimento profondo, agiscono senza essere soggetti morali o agenti intenzionali, ma producono effetti concreti e talvolta dannosi. Questo saggio analizza la crisi del paradigma classico della responsabilità (causalità, colpa, imputabilità) e ne propone una rifondazione teorica. Muovendo da alcuni tra i più rilevanti approcci contemporanei, si ricostruisce una concezione postumana della responsabilità: non più imputazione soggettiva *ex post*, ma architettura normativa e ambientale *ex ante*. Al centro della riflessione si colloca la responsabilità vicaria e sistemica dell'umano, inteso come architetto del contesto e custode del rischio. Il contributo discute criticamente i principali modelli contemporanei, approfondisce il settore sanitario come caso paradigmatico e mette a confronto l'evoluzione europea (AI Act, *Cyber Resilience Act*, nuova *Product Liability Directive*) con il quadro nordamericano. Ne emerge la proposta di una responsabilità distribuita e ambientale, capace di regolare l'azione algoritmica senza ridurla al calcolo.

The algorithmic revolution has also reshaped the foundations of legal responsibility. Artificial intelligence systems — particularly those based on deep learning — do not act as moral subjects or intentional agents, yet they produce tangible and sometimes harmful effects. This paper examines the crisis of the classical paradigm of responsibility (causality, fault, imputability) and proposes its theoretical re-foundation. Drawing on some of the most significant contemporary approaches, it reconstructs a posthuman conception of responsibility: no longer a matter of *ex post* subjective attribution, but a matter of ex prior normative and environmental design. At the heart of this reflection lies the notion of systemic and vicarious human responsibility, where the human is framed as the architect of context and guardian of risk. The paper critically discusses key contemporary models, develops healthcare as a paradigmatic case, and compares European developments (AI Act, Cyber Resilience Act, new Product Liability Directive) with the North American framework. It advances the idea of a distributed and environmental responsibility fit for governing algorithmic action beyond mere computation.

Keywords

responsabilità giuridica – intelligenza artificiale – imputazione – architettura normativa – sanità digitale