Ordine internazionale e diritti umani
International Legal Order and Human Rights
Ordenamiento Jurídico Internacional y Derechos Humanos
Ordre Juridique International et Droits de l'Homme
Diretta da Claudio Zanghì, Lina Panella, Carlo Curti Gialdino
EDITORIALE SCIENTIFICA

GIUSEPPINA PIZZOLANTE *

REGOLAMENTO SULL'INTELLIGENZA ARTIFICIALE, METODI DI GESTIONE DEL RISCHIO E TUTELA DEI DIRITTI UMANI

1. *Considerazioni introduttive. L'approccio* risk-based *dell'AI Act e la centralità della tutela dei diritti umani*

La progressiva integrazione dei sistemi di intelligenza artificiale (anche IA) in molteplici ambiti della vita sociale quali lavoro, sanità, istruzione, sicurezza e partecipazione democratica, ha determinato un significativo incremento delle preoccupazioni relative al loro potenziale impatto sulla tutela dei diritti fondamentali. Queste tecnologie, difatti, pur assicurando efficienza e innovazione, hanno accresciuto, sotto una parvenza di neutralità, le disuguaglianze sociali consolidando modelli di esclusione e discriminazione[1].

* Professoressa associata di Diritto dell'Unione europea, Università degli Studi di Bari Aldo Moro.

[1] I sistemi di intelligenza artificiale, infatti, non sono né sviluppati né implementati in un contesto neutrale ma si fondano su basi di dati che riflettono, e talvolta accrescono (può accadere ad esempio che una tecnologia, inizialmente imparziale, apprenda un errore per poi amplificarlo), disuguaglianze storiche, sociali ed economiche preesistenti. Ne consegue che l'assenza di un adeguato intervento regolatorio rischia di tradursi non già nella promozione di un progresso equo e inclusivo, bensì nella cristallizzazione e nella riproduzione automatizzata delle ingiustizie e delle disparità già radicate. Sul punto L. PANELLA (a cura di), *Nuove tecnologie e diritti umani: profili di diritto internazionale e di diritto interno*, Napoli, 2018; G.M. RUOTOLO, A. STIANO, *Intelligenza artificiale e (non) discriminazione nel diritto internazionale e dell'Unione europea: alcuni spunti di riflessione*, in G.M. RUOTOLO, A. STIANO (a cura di), *Discriminazioni automatiche. Come il diritto può aiutare l'intelligenza artificiale e viceversa*, Torino, 2026, p. 7 ss.

Il regolamento (UE) 2024/1689 (anche AI Act)[2] rappresenta un importante ed organico intervento legislativo volto a disciplinare lo sviluppo, l'impiego e la *governance* dell'IA ed è destinato a divenire un modello di riferimento a livello internazionale potendo esercitare un'influenza significativa su ordinamenti extra UE nei relativi processi normativi in punto di sicurezza, etica e trasparenza[3]. Esso introduce un approccio normativo fondato sul rischio, strutturato secondo una rappresentazione piramidale al cui vertice si collocano i sistemi che presentano un rischio inaccettabile, vietati in quanto ritenuti incompatibili con i diritti fondamentali ed i valori dell'Unione europea[4]. Il livello immediatamente inferiore comprende i sistemi ad alto rischio, soggetti a requisiti stringenti in termini di *governance*, gestione del rischio, trasparenza e controllo umano, in ragione del possibile impatto su sicurezza e diritti fondamentali. Seguono i sistemi a rischio limitato, per i quali il legislatore prevede obblighi di trasparenza circoscritti, finalizzati principalmente a garantire un'adeguata informazione degli utenti. Alla base della piramide si trovano infine i sistemi a rischio minimo o nullo, che costituiscono la maggior parte delle applicazioni di IA attualmente diffuse, e che non sono sottoposti a obblighi specifici[5].

Tale classificazione raffigura l'architrave del modello dell'AI Act poiché orienta l'intensità degli obblighi normativi al livello di rischio associato al sistema[6]. Sono così vietate alcune pratiche inaccettabili, tra cui, le tecniche subliminali (art. 5, par. 1, lett. a), i sistemi di

---

[2] Regolamento (UE) 2024/1689 del Parlamento europeo e del Consiglio del 13 giugno 2024 che stabilisce regole armonizzate sull'intelligenza artificiale e modifica i regolamenti (CE) n. 300/2008, (UE) n. 167/2013, (UE) n. 168/2013, (UE) 2018/858, (UE) 2018/1139 e (UE) 2019/2144 e le direttive 2014/90/UE, (UE) 2016/797 e (UE) 2020/1828 (regolamento sull'intelligenza artificiale), in *GUUE* L del 12 luglio 2024, p. 1 ss. Si veda, per l'Italia, la l. 23 settembre 2025, n. 132, *Disposizioni e deleghe al Governo in materia di intelligenza artificiale*, in *GU* Serie Generale n. 223 del 25 settembre 2025, il cui art. 1, par. 2, afferma che «[l]e disposizioni della presente legge si interpretano e si applicano conformemente al regolamento (UE) 2024/1689 del Parlamento europeo e del Consiglio, del 13 giugno 2024». Va anche segnalato che l'Unione europea ha firmato la Convenzione quadro del Consiglio d'Europa sull'intelligenza artificiale e i diritti umani, la democrazia e lo Stato di diritto, adottata il 17 maggio 2024 e aperta alla firma il 5 settembre 2024. V. decisione (UE) 2024/2218 del Consiglio del 28 agosto 2024 relativa alla firma, a nome dell'Unione europea, della convenzione quadro del Consiglio d'Europa sull'intelligenza artificiale e i diritti umani, la democrazia e lo Stato di diritto, in *GUUE* L del 4 settembre 2024, p. 1 ss., nonché F. ZORZI GIUSTINIANI, *La Convenzione quadro sull'intelligenza artificiale e i diritti umani, lademocrazia e lo stato di diritto* e *l'*international cyberspace and digital policy strategy *statunitense*, in *Nomos. Le attualità nel diritto*, 2/2024, p. 1 ss.; F. SEATZU, *Assessing the Council of Europe's AI Convention: Challenges and Prospects for Protecting Human Rights and Democracy in the Age of AI*, in *Studi sull'integrazione europea*, 2025, p. 6 ss. Sull'AI Act, si veda F. BORGIA, *Il modello europeo di regolazione dell'intelligenza artificiale: tra ambizione etica e ipertrofia normativa*, in *Ordine internazionale e diritti umani*, 2025, p. 1165 ss.; G. CONTALDI, *L'approccio regolatorio dell'Unione europea in materia di intelligenza artificiale*, in *Rivista Quaderni AISDUE*, anticipazione fascicolo *Intelligenza artificiale e nuove vulnerabilità*, 2026, reperibile *online*, p. 1 ss.

[3] Cfr. M. ALMADA, *The EU AI Act in a Global Perspective*, 17 giugno 2025, reperibile *online* al sito https://papers.ssrn.com/, che per l'AI Act riporta, tra l'altro,il c.d. "effetto Bruxelles" in forza del quale gli attori privati, anche in giurisdizioni in cui si applicano regole meno stringenti, spesso si conformano a una normativa UE particolarmente rigorosa.

[4] Sui valori fondanti dell'Unione europea, v., per tutti, L. PANELLA, C. ZANGHÌ, *La protezione internazionale dei diritti dell'uomo*[4], Torino, 2019, p. 337 ss.; U. VILLANI, *Istituzioni di Diritto dell'Unione europea*[7], Bari, 2024, p. 42 ss.

[5] Nell'art. 3, punto 1, dell'AI Act, un sistema IA è definito come «un sistema automatizzato progettato per funzionare con livelli di autonomia variabili e che può presentare adattabilità dopo la diffusione e che, per obiettivi espliciti o impliciti, deduce dall'input che riceve come generare output quali previsioni, contenuti, raccomandazioni o decisioni che possono influenzare ambienti fisici o virtuali». Per la definizione di IA si rinvia a P. WANG, *On Defining Artificial Intelligence*, in *Journal of Artificial General Intelligence*, 2019, p. 1 ss.J. SCHUETT, *Defining the Scope of AI Regulations*, in *Law, Innovation and Technology*, 2023, p. 60 ss.

[6] Sull'approccio basato sul rischio G. DE GREGORIO, P. DUNN, *The European Risk-Based Approaches: Connecting Constitutional Dots in the Digital Age*, in *Common Market Law Review*, 2022, p. 473 ss.

categorizzazione biometrica che classificano gli individui per trarne deduzioni o inferenze in merito a razza, opinioni politiche, appartenenza sindacale, convinzioni religiose o filosofiche, vita sessuale o orientamento sessuale (art. 5, par. 1, lett. g) e l'uso di sistemi di identificazione biometrica remota in tempo reale in spazi accessibili al pubblico a fini di attività di contrasto, con alcune eccezioni (art. 5, par. 1, lett. h)[7].

Un sistema è *high-risk* invece se, ai sensi dell'art. 6, par. 1, è parte di un prodotto – o è il prodotto stesso – ed esso è già regolato da norme europee di armonizzazione sulla sicurezza, come nelle ipotesi di dispositivi medici, macchine industriali o automobili, e se, prima di essere immesso sul mercato, sia sottoposto ad una valutazione di conformità di terza parte. In questi casi evidentemente l'IA, essendo componente di un prodotto, può incidere sulla sicurezza fisica degli individui e se, dunque, la normativa UE di base richiede controlli rigorosi, anche l'IA relativa a quel prodotto diventerà automaticamente ad alto rischio, collegandosi l'IA, in forza dell'art. 6, par. 1, al regime di sicurezza dei prodotti.

Un sistema di IA è poi ad alto rischio, alla luce dell'art. 6, par. 2, quando è compreso in una delle fattispecie elencate nell'allegato III. Ci muoviamo, così, dalla sicurezza dei prodotti, all'impatto sui diritti fondamentali ove appunto i sistemi presentino un rischio significativo di danno per la salute, la sicurezza o i diritti fondamentali delle persone fisiche. Tale qualificazione riguarda un'ampia serie di applicazioni come i componenti di sicurezza supportati dall'IA impiegati nelle infrastrutture critiche, quali i trasporti (allegato III, punto 2), ed anche le soluzioni IA utilizzate in ambito formativo in quanto idonee a incidere sull'accesso all'istruzione e sul percorso professionale degli individui, qual è il caso dei sistemi di valutazione automatizzata dei risultati dell'apprendimento (allegato III, punto 3).

Ulteriori ambiti rilevanti includono gli strumenti adoperati nel settore dell'occupazione, della gestione dei lavoratori e dell'accesso al lavoro autonomo, come i *software* di preselezione destinati ad analizzare o filtrare le candidature o a valutare le prestazioni e il comportamento dei lavoratori (allegato III, punto 4). Sono inoltre considerati ad alto rischio i sistemi di IA adottati per determinare l'accesso a servizi, pubblici e privati, essenziali, quali l'assistenza sanitaria, anche di emergenza, l'affidabilità creditizia e il merito di credito (con l'eccezione dei sistemi volti ad individuare frodi finanziarie), la valutazione dei rischi e la determinazione dei prezzi in relazione ad assicurazioni sulla vita e ad assicurazioni sanitarie (allegato III, punto 5).

---

[7] Ai sensi dell'art. 5 dell'AI Act, sono vietati tutti i sistemi di intelligenza artificiale considerati una chiara minaccia per la sicurezza e i diritti degli individui ovvero le tecniche volutamente manipolative o ingannevoli capaci di provocare un danno significativo (art. 5, par. 1, lett. a); lo sfruttamento delle vulnerabilità tale da provocare un danno significativo (art. 5, par. 1, lett. b); il punteggio sociale (art. 5, par. 1, lett. c); la valutazione o la previsione del rischio di reato (art. 5, par. 1, lett. d); lo *scraping* non mirato di materiale Internet o i filmati di telecamere a circuito chiuso per creare o espandere database di riconoscimento facciale (art. 5, par. 1, lett. e); il riconoscimento delle emozioni nei luoghi di lavoro e negli istituti di istruzione (art. 5, par. 1, lett. f); la categorizzazione biometrica per dedurre determinate caratteristiche protette (art. 5, par. 1, lett. g); l'identificazione biometrica remota in tempo reale per attività di contrasto in spazi accessibili al pubblico, salvo quando sia strettamente necessario per ricercare vittime specifiche o persone scomparse, prevenire una minaccia grave e imminente alla vita o un possibile attacco terroristico, identificare o localizzare sospettati di reati gravi (art. 5, par. 1, lett. h). Questa parte del regolamento è applicabile dal 2 febbraio 2025.
La Commissione ha pubblicato delle linee guida per supportare l'applicazione pratica delle pratiche vietate. V. comunicazione della Commissione del 29 luglio 2025, Orientamenti della Commissione relativi alle pratiche di intelligenza artificiale vietate ai sensi del regolamento (UE) 2024/1689 (regolamento sull'IA), C(2025) 5052 final. Il catalogo delle pratiche vietate è peraltro destinato ad ampliarsi: il c.d. Digital Omnibus on AI (proposta COM(2025) 836 del 19 novembre 2025) aggiunge all'art. 5 un nuovo divieto avente a oggetto i sistemi di IA che generano materiale sessuale e intimo non consensuale o materiale di abuso sessuale su minori.

Rientrano nella medesima categoria i sistemi di intelligenza artificiale impiegati per l'identificazione biometrica da remoto, il riconoscimento delle emozioni e la categorizzazione biometrica (allegato III, punto 1), nonché quelli adoperati dalle forze dell'ordine, qualora, evidentemente come in tutti i casi descritti, possano interferire con l'esercizio dei diritti fondamentali, ad esempio nella valutazione dell'affidabilità delle prove (allegato III, punto 6, lett. c). Analoga qualificazione è prevista per i sistemi di IA utilizzati nella gestione delle migrazioni, dell'asilo e del controllo delle frontiere, per l'esame automatizzato delle domande di asilo, di visto o di permesso di soggiorno e per i relativi reclami (allegato III, punto 7), e per le soluzioni adottate nell'amministrazione della giustizia e nei processi democratici, inclusi i sistemi volti a supportare l'autorità giudiziaria nella ricerca, nell'interpretazione e nell'applicazione della legge (allegato III, punto 8).

Se, in base all'art. 6, par. 3, terzo comma, un sistema di IA di cui all'allegato III è sempre considerato ad alto rischio qualora esso effettui profilazione di persone fisiche, in forza dell'art. 6, par. 3, primo comma, lo stesso sistema non è automaticamente *high-risk* qualora esso rechi un impatto minimo e il risultato del processo decisionale non sia stato condizionato dall'impiego dell'IA. A tal fine, deve essere soddisfatta almeno una qualsiasi delle condizioni richiamate ovvero che il sistema sia destinato a svolgere un compito procedurale circoscritto; che sia volto a migliorare l'esito di un'attività umana previamente svolta; che si limiti a individuare schemi decisionali o eventuali scostamenti rispetto a decisioni precedenti, senza sostituire né influenzare la valutazione umana; ovvero che sia impiegato per l'esecuzione di un'attività meramente preparatoria (art. 6, par. 3, secondo comma).

I sistemi di intelligenza artificiale qualificati come ad alto rischio sono sottoposti a un regime di obblighi particolarmente stringente volto a garantire appunto la tutela della sicurezza e dei diritti fondamentali ed il cui adempimento è necessario per il loro utilizzo. L'AI Act segue una logica simile a quella dei prodotti pericolosi in forza della quale non si vieta l'IA ma se ne permette l'uso solo se rispetti gli standard previsti[8]. Questi ultimi comprendono, tra l'altro, la gestione del rischio, la qualità dei dati, la documentazione tecnica, la trasparenza, la supervisione umana e la robustezza del sistema il cui rispetto costituisce una condizione preliminare per l'immissione sul mercato o la messa in servizio. In mancanza di valutazione di conformità, il sistema non può essere legittimamente commercializzato o utilizzato nell'Unione europea.

Il presente contributo intende soffermarsi specificamente sulla gestione del rischio, esaminandola in relazione alla tutela dei diritti umani poiché essa non si esaurisce in un adempimento di natura meramente tecnica ma si configura quale presidio essenziale per la prevenzione e la mitigazione dei rischi che i sistemi di intelligenza artificiale ad alto rischio possono arrecare ai diritti fondamentali. La gestione del rischio assume una funzione sistematica nell'ambito dell'impianto regolatorio delineato dall'AI Act, ponendosi quale

[8] Si veda il capo III del regolamento (UE) 2023/988 del Parlamento europeo e del Consiglio del 10 maggio 2023 relativo alla sicurezza generale dei prodotti, che modifica il regolamento (UE) n. 1025/2012 del Parlamento europeo e del Consiglio e la direttiva (UE) 2020/1828 del Parlamento europeo e del Consiglio, e che abroga la direttiva 2001/95/CE del Parlamento europeo e del Consiglio e la direttiva 87/357/CEE del Consiglio, in *GUUE* L 135 del 23 maggio 2023, p. 1 ss, nonché il considerando 166 dell'AI Act, secondo cui «[è] importante che i sistemi di IA collegati a prodotti che non sono ad alto rischio in conformità del presente regolamento e che pertanto non sono tenuti a rispettare i requisiti stabiliti per i sistemi di IA ad alto rischio siano comunque sicuri al momento dell'immissione sul mercato o della messa in servizio. Per contribuire a tale obiettivo, sarebbe opportuno applicare come rete di sicurezza il regolamento (UE) 2023/988 del Parlamento europeo e del Consiglio».

strumento attraverso cui va assicurata l'effettività delle garanzie sostanziali previste; l'analisi si concentrerà, pertanto, sul ruolo e sulle modalità di attuazione di tale gestione con riguardo, tra l'altro, agli ulteriori requisiti richiesti[9].

2. *L'ambito di applicazione della gestione del rischio nell'AI Act: profili oggettivi e soggettivi*

Con riferimento all'ambito di applicazione sostanziale della gestione del rischio, i sistemi di IA indipendenti, diversi da quelli che sono componenti di sicurezza dei prodotti o che sono essi stessi prodotti, sono ad alto rischio se, alla luce della loro finalità prevista, presentino un elevato rischio di pregiudicare la salute, la sicurezza o i diritti fondamentali delle persone, tenendo conto, come chiariremo meglio, sia della probabilità che il danno si verifichi sia della sua gravità. La gestione del rischio disciplinata dall'AI Act si applica esclusivamente ai sistemi che sono appunto contemplati nell'art. 6, par. 2 e nell'allegato III[10]. I sistemi dell'art. 6, par. 1, parti di prodotti regolamentati UE, non rientrano nel sistema di gestione del rischio *ex* AI Act, essendo già disciplinati dai rispettivi regolamenti di sicurezza del prodotto.

Muovendo dalla necessità di prevenire e mitigare i potenziali rischi che i sistemi di intelligenza artificiale possono comportare per salute, sicurezza o diritti fondamentali, l'AI Act, come chiarisce il considerando 65, impone ai fornitori di istituire un sistema di gestione del rischio per identificare, valutare, mitigare, verificare e monitorare continuamente i rischi (ad esempio malfunzionamenti o *bias* nei dati[11] e relativi impatti sui diritti fondamentali) lungo tutto il ciclo di vita del sistema, introducendo un complesso di requisiti specifici applicabili, in forza dell'art. 8, ai sistemi ad alto rischio[12].

Tali prescrizioni assumono un'importanza centrale alla luce della crescente integrazione dell'IA in ambiti ad elevata incidenza sociale ed economica e dunque fortemente impattanti sui diritti degli individui, come le infrastrutture critiche e il settore sanitario, nei quali una gestione inadeguata dei rischi potrebbe determinare conseguenze particolarmente gravi.

La gestione dei rischi è disciplinata dall'art. 9 dell'AI Act che, per quanto riguarda l'ambito di applicazione personale, si indirizza ai fornitori. Va preliminarmente evidenziato che i fornitori di sistemi di IA ad alto rischio devono osservare i requisiti previsti nella sezione 2 del regolamento, pur nella consapevolezza che possono comunque persistere rischi residui. L'art. 8 dell'AI Act svolge, in questo contesto, una funzione strutturale e trasversale, in

---

[9] V. *infra* par. 4.

[10] Dal punto di vista temporale, i fornitori devono istituire i sistemi di gestione del rischio a decorrere dal 2 agosto 2026 per i sistemi indipendenti di cui all'allegato III, mentre per i sistemi ad alto rischio collegati a prodotti di cui all'allegato I l'art. 113 AI Act fissa il termine al 2 agosto 2027. Tali scadenze sono peraltro destinate a mutare per effetto del citato Digital Omnibus on AI, approvato dal Parlamento europeo il 16 giugno 2026 e, al momento in cui si scrive, in attesa dell'adozione formale da parte del Consiglio. Il testo modifica l'art. 113 AI Act prevedendo che gli obblighi relativi ai sistemi ad alto rischio si applichino dal 2 dicembre 2027 per i sistemi indipendenti (allegato III) e dal 2 agosto 2028 per quelli integrati in prodotti (allegato I).

[11] I *bias* sono distorsioni che si verificano nei sistemi digitali, che "imparano" da dati o regole che contengono già errori o squilibri e che conducono a risultati non equi, inaccurati o influenzati da pregiudizi.

[12] Si veda anche N.A. SMUHA, E. RENGERS, A. HARKENS, W. LI, J. MACLAREN, R. PISELLI, K. YEUNG, *How the EU Can Achieve Legally Trustworthy AI: A Response to the European Commission's Proposal for an Artificial Intelligence Act*, 5 agosto2021, reperibile *online*: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3899991.

quanto stabilisce il principio secondo cui i sistemi di IA ad alto rischio devono rispettare un insieme di requisiti obbligatori per poter essere introdotti sul mercato od immessi in servizio; in particolare, l'art. 8, par. 1, dispone che i sistemi di IA ad alto rischio si conformano ai requisiti prefissati tenendo conto sia delle differenti finalità sia dello "stato dell'arte" generalmente riconosciuto in materia di tecnologie, per un verso, applicando un principio di proporzionalità molto utile, come vedremo, nella gestione del rischio, per altro verso, introducendo un criterio dinamico e tecnologicamente aggiornato di conformità.

Laddove un prodotto sia alimentato da un sistema di IA cui si applicano i requisiti sia del regolamento sia della normativa di armonizzazione dell'Unione, si dispone, nell'art. 8, par. 2, la responsabilità dei fornitori che devono garantire che il loro prodotto sia pienamente conforme a tutti i requisiti previsti altresì dalla normativa di armonizzazione applicabile. Al fine di assicurare coerenza ed evitare duplicazioni, essi possono integrare i processi di prova, comunicazione e documentazione nelle procedure già disciplinate dalla normativa europea pertinente.

I fornitori, come definiti dall'art. 3, includono i soggetti responsabili dell'immissione di tali sistemi di IA sul mercato dell'UE, indipendentemente dalla circostanza che abbiano sede nell'Unione o in Paesi terzi (art. 22, par. 1). Difatti l'AI Act si applica a tutti i fornitori che operano nell'Unione europea, nonché a quelli i cui sistemi di IA producono effetti all'interno del territorio dell'UE. Dunque, anche i fornitori extra-UE devono conformarsi agli standard in parola quando i loro sistemi siano destinati a produrre effetti giuridicamente rilevanti nello spazio UE. Nell'ambito di applicazione sono generalmente ricomprese le autorità pubbliche e le organizzazioni internazionali salvo che utilizzino i sistemi di IA nel quadro della cooperazione, anche tramite accordi internazionali, delle autorità di contrasto e giudiziarie con l'UE o con uno o più Stati membri, sempreché che forniscano «garanzie adeguate per quanto riguarda la protezione dei diritti e delle libertà fondamentali delle persone» (art. 2, par. 4).

I rischi connessi all'utilizzo dell'IA eccedono il mero malfunzionamento tecnico, ricomprendendo altresì quelli derivanti dalle modalità operative, dalle implicazioni etiche[13] e dagli usi impropri[14]. In considerazione di tali rischi, il regolamento predispone un articolato sistema di gestione del rischio, finalizzato a garantire un impiego responsabile dell'IA e a condizionarne la distribuzione all'adozione di adeguate garanzie. Negli ultimi anni, la gestione dei rischi è divenuta una priorità in particolare per gli operatori del settore privato. Ad esempio, il National Institute of Standards and Technology (NIST) ha sviluppato negli Stati Uniti un *framework* volto a gestire i rischi per individui, organizzazioni e società promuovendo un utilizzo più etico dell'IA[15]. Tuttavia, il NIST AI Risk Management Framework (AI RMF)

---

[13] Cfr. J. WHITTLESTONE, R. NYRUP, A. ALEXANDROVA, K. DIHAL, S. CAVE, *Ethical and Societal Implications of Algorithms, Data, and Artificial Intelligence: a Roadmap for Research*, London, 2019, p. 1 ss.; L. VESNIC-ALUJEVIC, S. NASCIMENTO, A. POLVORA, *Societal and Ethical Impacts of Artificial Intelligence: Critical Notes on European Policy Frameworks*, in *Telecommunications Policy*, 2020.

[14] Si veda M. VEALE, F. ZUIDERVEEN BORGESIUS, *Demystifying the Draft EU Artificial Intelligence Act. Analysing the Good, the Bad, and the Unclear Elements of the Proposed Approach*, in *Computer Law Review International*, 2021, p. 97 ss.

[15] NIST, Artificial Intelligence Risk Management Framework (AI RMF 1.0), January 2023, reperibile *online*: https://doi.org/10.6028/NIST.AI.100-1. Si tratta di un quadro volontario concepito a favore degli attori dell'IA per rendere i sistemi sicuri, protetti e affidabili. V. anche il *Risk Management Profile for Artificial Intelligence and Human Rights*, una guida del Dipartimento di Stato americano pubblicata nel luglio 2024 allo scopo di allineare il quadro di gestione del rischio dell'AI RMF del NIST agli standard internazionali in materia di diritti umani.

presenta base volontaria laddove l'AI Act pone in capo agli operatori un obbligo giuridico qualificato di *governance* del rischio che si sostanzia in vincoli permanenti di vigilanza, controllo e aggiornamento delle misure di mitigazione adottate[16].

Tale obbligo si inscrive in una più ampia logica di responsabilizzazione degli attori coinvolti, su cui torneremo oltre, imponendo loro l'adozione di assetti organizzativi e tecnici idonei a garantire la gestione sistematica e continuativa dei rischi. Ed, infatti, la portata innovativa del regolamento si coglie nella sua capacità di coniugare la promozione dell'innovazione tecnologica con l'obiettivo di affermare un modello strutturato per la regolamentazione dell'intelligenza artificiale[17].

3. *La gestione del rischio* ex *art. 9 AI Act tra articolazione del procedimento e tipologia delle misure*

Il principio generale della gestione del rischio per i sistemi di IA ad alto rischio prevede un approccio organizzato, contenuto nel par. 1 dell'art. 9, in forza del quale i fornitori devono istituire, attuare, documentare e mantenere un sistema di gestione dei rischi, con l'obiettivo di identificare, valutare e mitigare i rischi derivanti dall'uso dell'IA. L'istituzione comporta la creazione e l'approvazione di procedure specifiche, mentre l'attuazione richiede la messa in pratica efficace di tali misure entro il sistema. La documentazione consiste nella descrizione sistematica dei registri, laddove il mantenimento richiede revisioni e aggiornamenti periodici per garantirne costantemente l'efficacia.

In conformità del par. 2, il processo di gestione del rischio è inteso come un processo iterativo, continuo, pianificato ed eseguito nel corso dell'intero ciclo di vita del sistema di IA ad alto rischio. Esso inizia con l'individuazione e l'analisi dei rischi per i diritti fondamentali, in particolare quelli prevedibili derivanti dall'utilizzo conforme alla finalità prevista. Tale

---

Le iniziative del governostatunitense in materia di IA e dirittiumani, oltre all'AI RMF del NIST, comprendono, tra l'altro: Executive Order 13859, 11 February 2019, *Maintaining American Leadership in Artificial Intelligence*; Executive Order 14110, 30 October 2023, *Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*; *Blueprint for an AI Bill of Rights. Making Automated Systems Work For The American People*, October 2022; *Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI*, 21 July 2023 (gli impegni volontari sono stati assunti da Amazon, Anthropic, Google, Inflection, Meta, Microsoft, OpenAI). V., inoltre, Office of Management and Budget's (OMB), *Memorandum on Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence*, 28 March 2024. Cfr. C. SBAILÒ, *Executive Order 14110 Governing Artificial Intelligence: Technological Leadership and Regulatory Challenges in an Era of Exponential Growth*, in *DPCE online*, 2024, p. 275 ss.; G.M. RUOTOLO, *Intelligenza Artificiale e trattamento dei dati tra diritto internazionale e ordinamenti interni*, in *AI LAW*, 2025, p. 236 ss.

[16] In argomento E. HINE, *Artificial Intelligence Laws in the US States are Feeling the Weight of Corporate Lobbying*, in *Nature*, 2024, p. 633 ss. Va, d'altro canto, messo in luce il progressivo interesse dei privati verso la regolamentazione. Si rinviasul punto a T. WHEELER, *The Three Challenges of AI Regulation*, in *Brookings*, 15 June 2023, reperibile online: https://www.brookings.edu/articles/the-three-challenges-of-ai-regulation/, chesottolinea in particolarequanto segue: «[t]he drum beat of artificial intelligence corporate chieftains calling for government regulation of their activities is mounting (…) As Senate Judiciary Committee Chairman Richard Durbin (R-IL) observed, it is "historic" to have "people representing large corporations… come before us and plead with us to regulate them» e che la sfida è «how to protect the public interest in a race that promises to be the fastest ever run yet is happening without a referee».

[17] Sul punto J. LAUX, S. WACHTER, B. MITTELSTADT, *Trustworthy Artificial Intelligence and the European Union AI Act: On the Conflation of Trustworthiness and Acceptability of Risk*, in *Regulation & Governance*, p. 3 ss.

individuazione si basa su metodi sistematici, come tassonomie del rischio e analisi di scenario, per rilevare potenziali pericoli e scongiurare esiti dannosi[18].

La fase successiva all'identificazione è la stima dei rischi che possono emergere quando il sistema di IA ad alto rischio è usato conformemente alla sua finalità – prevista – o in condizioni di uso improprio – ragionevolmente prevedibile –. La stima fornisce così al fornitore uno strumento per misurare il potenziale impatto del rischio qualora si concretizzi e, in base alla sua entità, per stabilire le priorità. Valutati i rischi, i fornitori attivano misure di gestione al fine di affrontarli e mitigarli. Queste includono modifiche progettuali, attivazione di controlli o altre azioni preventive in grado di eliminare ragionevolmente i rischi o comunque ridurli efficacemente[19]. In quest'ultimo caso, il rischio viene documentato secondo le procedure indicate all'art. 9, par. 1, e il processo di gestione viene concluso.

Il modello accolto dall'art. 9 è strutturato secondo un processo di gestione del rischio propriamente detto (paragrafi 3-5) e procedure di prova e verifica (paragrafi 6-8), con ulteriori disposizioni per categorie particolarmente vulnerabili come i minori[20].

L'art. 9, par. 4 stabilisce che le misure di gestione dei rischi da adottare devono tenere adeguatamente conto degli effetti e delle possibili interazioni tra i vari requisiti tecnici previsti dalla sezione relativa alla normativa sui sistemi ad alto rischio così da ridurre i rischi in modo efficace e garantire un equilibrio nella loro applicazione. Dunque, nel decidere quale misura di mitigazione assumere, al fine di raggiungere questo obiettivo, è necessario considerare come esse si combinino tra di loro e con gli altri obblighi in materia di dati e *governance* dei dati, trasparenza, documentazione tecnica, accuratezza, robustezza e cybersicurezza, sorveglianza umana.

In forza dell'art. 9, par. 5, comma 1, le misure di gestione del rischio «sono tali che i pertinenti rischi residui associati a ciascun pericolo nonché il rischio residuo complessivo dei sistemi di IA ad alto rischio sono considerati accettabili»; ne consegue che, dopo aver applicato tutte le misure pertinenti, i rischi residui, che dunque non è possibile eliminare del tutto, devono essere sostenibili e compatibili con gli obiettivi di tutela dei diritti fondamentali[21].

L'art. 9, par. 5, comma 2, suddivide le misure di gestione dei rischi in tre tipologie: sulla progettazione, di attenuazione e di informazioni agli utilizzatori (anche *deployers*). Così, i fornitori devono eliminare o ridurre i rischi, individuati ai sensi del par. 2, il più possibile durante la fase di progettazione e sviluppo del sistema di IA ad alto rischio ovvero *ex ante*. Nei casi in cui l'eliminazione non sia possibile, intervengono, in fase *ex post*, le misure di attenuazione in forza delle quali i fornitori possono implementare i controlli per gestire i rischi residui, utilizzando ad esempio filtri sui contenuti. Infine, devono essere fornite informazioni e formazione adeguate ai *deployers* affinché siano consapevoli delle capacità e dei limiti del sistema.

---

[18] J. NEWMAN, *A Taxonomy of Trustworthiness for Artificial Intelligence*, UC Berkeley Center for Long-Term Cybersecurity (CLTC), January 2023.

[19] Questo processo richiede monitoraggio e aggiornamenti continui; in particolare, i fornitori devono riesaminare e perfezionare le misure di gestione laddove diventino disponibili nuovi dati provenienti dai sistemi di monitoraggio post-commercializzazione.

[20] J. SCHUETT, *Risk Management in the Artificial Intelligence Act*, in *European Journal of Risk Regulation*, 2024, p. 367 ss.

[21] Sul punto torneremo meglio *infra* ai paragrafi 8 e 9.

4. *Procedure di prova, documentazione tecnica e regime di responsabilità del fornitore*

Le procedure di prova definite dall'art. 9, paragrafi 6-8, hanno lo scopo di assistere i fornitori ad individuare le misure di gestione dei rischi più mirate e appropriate garantendo che il sistema di IA funzioni in modo coerente per la finalità prevista e, dunque, se, da un lato, dispongono la verifica del sistema in relazione a parametri predefiniti e a soglie probabilistiche, al fine di assicurarne l'accuratezza e la robustezza in contesti applicativi reali caratterizzati da elevata eterogeneità, dall'altro, sono volte ad affrontare criticità quali l'eventuale variazione delle prestazioni del sistema in conseguenza di mutamenti, rispetto alla fase di preparazione, delle condizioni operative.

Sebbene i fornitori siano obbligati a svolgere attività di controllo lungo l'intero ciclo di sviluppo del sistema di intelligenza artificiale, tali procedure devono essere completate, in ogni caso, prima che il sistema di IA sia immesso sul mercato. Il regolamento non chiarisce se esse debbano essere eseguite in base a processi interni o da soggetti terzi; tuttavia viene attribuita al fornitore la responsabilità ultima sulla conformità, impostando la responsabilità quale elemento centrale dell'assetto del regolamento (art. 8).

L'art. 11, par. 1, comma 1, dal quale emerge un modello in cui la conformità non si esaurisce in un fatto tecnico e formale, afferma che «[l]a documentazione tecnica di un sistema di IA ad alto rischio è redatta prima dell'immissione sul mercato o della messa in servizio di tale sistema ed è tenuta aggiornata» ai sensi dell'allegato IV (art. 11, par. 1, comma 2). L'art. 9 AI Act rappresenta l'elemento cardine dell'impianto regolatorio previsto poiché questa norma esprime l'obbligo, in capo al fornitore, di organizzare un sistema di gestione del rischio che guidi l'intero ciclo di vita del sistema di IA ad alto rischio. L'adempimento non riveste natura formale inserendosi in una logica "operativa" in cui il rischio va individuato, analizzato, mitigato e continuamente riesaminato. Conseguentemente, la conformità nasce a monte, all'interno del processo di progettazione e sviluppo.

La previsione della documentazione ha lo scopo di rendere visibile e verificabile ciò che altrimenti resterebbe interno all'organizzazione del fornitore, infatti, il sistema di gestione del rischio richiesto dall'art. 9 deve essere declinato in un insieme organizzato di informazioni, tra cui, le metodologie adottate, le scelte progettuali, le misure di mitigazione, i risultati delle verifiche. La documentazione diventa il momento in cui il rischio, da fenomeno tecnico, assume forma giuridica ed è attraverso di essa che le autorità, gli organismi notificati e, più in generale, il sistema di vigilanza possono comprendere la logica interna del sistema di IA.

L'art. 11, par. 1 permette alla gestione del rischio di distaccarsi dalla dimensione puramente tecnica e di accogliere quella della valutazione, della responsabilità e dell'*enforcement*. Dunque, la conformità, nel modello dell'AI Act, esiste pienamente solo quando è insieme realizzata e documentata confermandosi che la gestione del rischio accolta nel regolamento è funzionalmente integrata con tutti i requisiti richiesti dalla sezione 2 per i sistemi ad alto rischio[22].

Rileva in questo contesto anche l'art. 77, par. 1 dell'AI Act, in forza del quale le autorità pubbliche nazionali che «controllano o fanno rispettare gli obblighi previsti dal diritto dell'Unione a tutela dei diritti fondamentali, compreso il diritto alla non discriminazione, in relazione all'uso di sistemi di IA ad alto rischio» hanno il potere di richiedere e accedere alla

[22] V. *supra* par. 1.

documentazione rilevante necessaria per adempiere ai propri compiti; qualora la documentazione disponibile «non sia sufficiente per accertare un'eventuale violazione degli obblighiprevisti dal diritto dell'Unione a tutela dei diritti fondamentali», l'autorità pubblica può richiedere alla autorità di vigilanza del mercato l'organizzazione di verifiche tecniche sui sistemi di IA ad alto rischio (art. 77, par. 3); le informazioni ottenute devono essere trattate «in conformità degli obblighi di riservatezza stabiliti all'articolo 78» (art. 77, par. 3).

Possiamo dunque cogliere, tra le altre, la caratteristica dell'AI Act di realizzare una normativa incentrata sui processi organizzativi e sulla responsabilità dell'operatore, circostanza quest'ultima rilevante, come vedremo, anche nella prevenzione del rischio. Il fornitore non è chiamato solo a rispettare regole ma a costruire un sistema interno capace di governare il rischio e, al contempo, a darne conto in modo strutturato e controllabile.

Il principio di responsabilità del fornitore è richiamato anche nel momento in cui si dispone che l'art. 9 debba essere applicato attraverso sanzioni effettive, proporzionate e dissuasive disposte dagli Stati membri. L'art. 99, par. 7 dell'AI Act chiarisce che nel decidere se infliggere una sanzione amministrativa pecuniaria e nel determinarne l'importo deve tenersi conto di tutte le circostanze relative allo specifico caso e, casomai, di alcune circostanze tipizzate come il grado di responsabilità dell'operatore, rispetto ai parametri delle misure tecniche e organizzative attuate, nonché dell'eventuale azione intrapresa per attenuare il danno subito dalle persone interessate[23].

La disposizione, pur fissando soglie massime per le sanzioni, concede agli Stati membri un ampio margine di discrezionalità nella determinazione delle modalità applicative, con il rischio di una frammentazione nell'*enforcement* e di disparità di trattamento tra operatori attivi in diversi ordinamenti nazionali. Da un lato, dunque, il regolamento definisce le fattispecie sanzionabili, i limiti edittali – anche molto elevati – ed una cornice procedurale, dall'altro, è lasciato agli Stati membri il compito di individuare le autorità competenti, disciplinare i procedimenti sanzionatori, coordinare le nuove sanzioni con i principi generali come quelli di legalità, del contraddittorio e di proporzionalità.

Va altresì indicata la questione delle capacità operative delle autorità di vigilanza che potrebbero non disporre sempre delle competenze tecniche e delle risorse necessarie per accertare le violazioni in un settore altamente complesso e in rapida evoluzione come quello dell'intelligenza artificiale.

Ulteriori criticità emergono sul piano probatorio poiché la natura spesso opaca dei sistemi di IA, specialmente quelli basati su modelli avanzati, rende difficile ricostruire i processi decisionali e, conseguentemente, attribuire responsabilità in modo chiaro e tempestivo, con un possibile indebolimento della funzione deterrente delle sanzioni. Inoltre, il regime sanzionatorio si innesta su un modello regolatorio che si fonda, fra l'altro, su obblighi di autovalutazione e segnalazione in capo agli operatori. La significativa dipendenza

[23] Si segnala, in proposito, la proposta di direttiva del Parlamento europeo e del Consiglio relativa all'adeguamento delle norme in materia di responsabilità civile extracontrattuale all'intelligenza artificiale (direttiva sulla responsabilità da intelligenza artificiale), 28 settembre 2022, (COM(2022) 496 final), volta a introdurre meccanismi probatori agevolati per il risarcimento dei danni causati da sistemi di IA. Tale proposta, tuttavia, è stata successivamente ritirata dalla Commissione europea nel 2025, lasciando allo stato agli ordinamenti nazionali la disciplina della responsabilità civile, pur nel quadro degli obblighi preventivi delineati dall'AI Act. La decisione di abbandonare la proposta è stata riportata nel programma di lavoro della Commissione per il 2025, adottato l'11 febbraio e presentato al Parlamento europeo il 12 febbraio. Tale improvvisa inversione di rotta pare riconducibile, almeno in parte, alle pressioni provenienti dagli operatori del settore, che hanno manifestato preoccupazioni circa l'impatto di un regime di responsabilità sui modelli di *business* esistenti.

dalla loro collaborazione che ne consegue potrebbe non essere sempre garantita, specie quando la *disclosure* di violazioni comporti costi economici o reputazionali.

Non meno rilevante è il profilo della proporzionalità giacché le sanzioni, parametrate anche al fatturato globale, possono risultare particolarmente onerose per le piccole e medie imprese e per le *startup*, con il rischio di produrre effetti disincentivanti sull'innovazione e sull'accesso al mercato.

In Italia, l'attuazione del regolamento 2024/1689 è disciplinata dalla citata l. n. 132 del 23 settembre 2025, entrata in vigore il 10 ottobre 2025, il cui art. 24, par. 1 dispone che il governo è delegato ad adottare, entro dodici mesi dalla data di entrata in vigore della stessa legge, «uno o più decreti legislativi per l'adeguamento della normativa nazionale al regolamento (UE) 2024/1689», volti, tra l'altro, ad «attribuire alle autorità di cui all'articolo 20 [Autorità nazionali per l'intelligenza artificiale] il potere di imporre le sanzioni e le altre misure amministrative previste dall'articolo 99 del regolamento (UE) 2024/1689 per la violazione delle norme del regolamento stesso e degli atti di attuazione, nel rispetto dei limiti edittali e delle procedure previsti dal medesimo articolo 99 e dalle disposizioni nazionali che disciplinano l'irrogazione delle sanzioni e l'applicazione delle altre misure amministrative da parte delle autorità anzidette» (art. 24, par. 2, lett. d della l. 132/2025). Il rinvio espresso alle «disposizioni nazionali che disciplinano l'irrogazione delle sanzioni» suggerisce che il legislatore delegante intende evitare la creazione di un sistema completamente autonomo, favorendo invece un innesto del regime AI Act nei modelli procedurali già esistenti.

5. *La valutazione d'impatto sui diritti umani* ex *art. 27 AI Act e il modello multilivello di gestione del rischio*

L'art. 27 AI Act introduce, nell'architettura del regolamento 2024/1689, un obbligo di particolare rilievo in materia di valutazione d'impatto sui diritti fondamentali (anche FRIA, *Fundamental Rights Impact Assessment*) in capo agli utilizzatori di determinati sistemi di IA ad alto rischio.

Sul piano funzionale, la disposizione si inserisce nella più ampia logica di prevenzione e gestione dei rischi che permea il regolamento, ma si distingue per il suo focus specifico sulla dimensione dei diritti umani, oltre salute e sicurezza, configurandosi come uno strumento autonomo e complementare rispetto agli obblighi gravanti sui fornitori *ex* art. 9. L'art. 27 contribuisce a colmare il divario tra la progettazione del sistema e il suo impiego effettivo, valorizzando il contesto d'uso quale fattore determinante nella produzione del rischio.

L'art. 27 dell'AI Act non si applica indistintamente a tutti i sistemi di intelligenza artificiale, ma opera con riferimento a una categoria ben delimitata, individuata sulla base del suo potenziale impatto sui diritti fondamentali.

La disposizione riguarda anzitutto i sistemi di IA qualificati ad alto rischio ovvero quelli, già descritti[24], individuati dal regolamento come suscettibili di incidere in modo significativo sui diritti, e dunque sulla vita, degli individui.

Sotto il profilo soggettivo, l'obbligo grava sugli utilizzatori di sistemi di IA ad alto rischio individuati dal regolamento nell'allegato III (con l'eccezione dei sistemi collegati ad

[24] V. *supra* par. 1.

infrastrutture critiche). Tuttavia, tale qualificazione non è, di per sé, sufficiente ad azionare la previsione di cui all'art. 27 che rileva quando i *deployers* siano organismi di diritto pubblico o enti privati che forniscano servizi pubblici, nonché nelle ipotesi di *deployers* di sistemi di IA ad alto rischio di cui all'allegato III, punto 5, lettere b) e c), ossia di sistemi destinati ad essere utilizzati per valutare l'affidabilità creditizia o il merito di credito e nel caso di assicurazioni sulla vita e assicurazioni sanitarie.

La scelta del legislatore europeo riflette la consapevolezza che i rischi per i diritti umani non derivino esclusivamente dalle caratteristiche tecniche del sistema, ma anche dalle modalità concrete di implementazione, rendendosi necessario, prevalentemente nei casi in cui i *deployers* siano organismi di diritto pubblico o enti privati che forniscono servizi pubblici, un doppio profilo di responsabilità volto ad estendersi all'intero contesto applicativo dell'IA.

Quanto al contenuto, la valutazione d'impatto richiede un'analisi strutturata e documentata degli effetti potenziali del sistema sui diritti fondamentali delle persone interessate. Essa implica, in particolare la descrizione del contesto di utilizzo e delle finalità perseguite; l'individuazione delle categorie di soggetti potenzialmente colpiti; la valutazione dei rischi specifici per i diritti fondamentali; l'indicazione delle misure previste per prevenire o mitigare tali rischi. L'art. 27, par. 2 ha infatti cura di precisare che la valutazione deve comprendere i seguenti elementi: «a) una descrizione dei processi del deployer in cui il sistema di IA ad alto rischio sarà utilizzato in linea con la sua finalità prevista; b) una descrizione del periodo di tempo entro il quale ciascun sistema di IA ad alto rischio è destinato a essere utilizzato e con che frequenza; c) le categorie di persone fisiche e gruppi verosimilmente interessati dal suo uso nel contesto specifico; d) i rischi specifici di danno che possono incidere sulle categorie di persone fisiche o sui gruppi di persone individuati a norma della lettera c), del presente paragrafo tenendo conto delle informazioni trasmesse dal fornitore a norma dell'articolo 13; e) una descrizione dell'attuazione delle misure di sorveglianza umana, secondo le istruzioni per l'uso; f) le misure da adottare qualora tali rischi si concretizzino, comprese le disposizioni relative alla governance interna e ai meccanismi di reclamo».

La FRIA presenta, dunque, una marcata dimensione procedurale e dinamica inserendosi in un processo continuo di monitoraggio e riesame, in linea con l'evoluzione del sistema e del contesto applicativo.

Dal punto di vista sistematico, l'art. 27 svolge una funzione di concretizzazione del principio di centralità dei diritti fondamentali, traducendo in obblighi operativi l'esigenza di assicurare che l'adozione di sistemi di IA sia compatibile con il quadro valoriale dell'Unione. Più specificamente, la disposizione si caratterizza per un approccio situato che perfeziona non solo le caratteristiche tecniche del sistema, ma anche le modalità e le finalità del suo impiego. Essa, inoltre, rafforza la dimensione della responsabilità, in questo caso degli utilizzatori, imponendo loro di esplicitare e giustificare le scelte adottate in relazione ai rischi individuati sulla base di un approccio che si occupa di intercettare le implicazioni concrete dell'uso dell'IA sui diritti fondamentali. In tal modo, l'art. 27 contribuisce a delineare un modello di regolamentazione multilivello, nel quale la tutela dei diritti non è affidata a un unico momento o a un solo soggetto, ma distribuita lungo l'intero ciclo di vita del sistema e tra i diversi attori coinvolti.

Come si è evidenziato, la gestione del rischio prevista dall'art. 9 è un obbligo che grava sui fornitori, strutturale e continuo; essa serve a progettare e sviluppare sistemi di IA che siano, in linea generale, sicuri e conformi, attraverso l'identificazione e la mitigazione dei rischi tipici lungo tutto il ciclo di vita del sistema. L'art. 27, invece, interviene in un momento diverso e con una funzione distinta imponendo agli utilizzatori di interrogarsi sugli effetti

prodotti da quel sistema in un determinato contesto sociale e istituzionale. Dal punto di vista giuridico, quindi, l'art. 27 non sostituisce la gestione del rischio e non la replica, ma la integra, introducendo una valutazione situata e contestuale in cui l'attenzione è rivolta non più sul sistema in sé bensì sul suo impatto concreto sui diritti fondamentali.

L'art. 9 assicura dunque una mitigazione dei rischi *by design*, laddove l'art. 27 verifica la tenuta alla prova conseguente all'uso concreto del sistema, contribuendo così a rafforzare la tutela effettiva dei diritti fondamentali. Ne deriva che il rapporto tra le due disposizioni è caratterizzato da una logica di complementarità funzionale, operando l'art. 9 sul piano della progettazione e dello sviluppo del sistema ed integrando i diritti fondamentali nella più ampia gestione del rischio; l'art. 27 interviene sul piano dell'implementazione concreta, imponendo una valutazione autonoma e approfondita delle implicazioni sugli stessi diritti nel contesto specifico di utilizzo. In questa prospettiva, l'art. 27 può essere interpretato come una specificazione qualitativa di una dimensione – quella dei diritti fondamentali – che è già presente, ma non esclusiva, nell'art. 9.

Il regolamento in parola, al contempo, prefigura una responsabilità distribuita tra i diversi attori – fornitori e utilizzatori – secondo un modello a doppio livello di tutela che mira a garantire un presidio effettivo e continuo dei rischi in forza del quale il *provider* deve progettare sistemi "*safe*" anche per i diritti; il *deployer* deve verificare che, nel contesto concreto, tali diritti non vengano compromessi.

Sotto un profilo sistematico, la valutazione di cui all'art. 27 presenta evidenti affinità con la valutazione d'impatto sulla protezione dei dati (anche DPIA) prevista dall'art. 35 del regolamento generale sulla protezione dei dati (anche GDPR)[25], condividendone la logica preventiva, la dimensione procedurale e l'attenzione al contesto specifico di trattamento. Tuttavia, mentre la DPIA si concentra sui rischi per la protezione dei dati personali, la valutazione d'impatto sui diritti fondamentali *ex* art. 27 assume una portata più ampia, estendendosi all'intero spettro dei diritti garantiti dall'ordinamento dell'Unione. Quest'ultima si configura, dunque, come uno strumento autonomo e di portata più ampia che riflette l'esigenza di governare i rischi dell'intelligenza artificiale secondo una prospettiva integrata e orientata ai diritti.

6. *Il bilanciamento tra diritti fondamentali nella fase di mitigazione dei rischi*

Tra le varie fasi in cui si articola la gestione del rischio, particolare rilievo assumono le tecniche di mitigazione che comprendono gli strumenti e le misure, di natura tecnica e organizzativa, diretti a prevenire, ridurre o contenere i rischi connessi all'impiego di sistemi di IA. Essi non eliminano necessariamente il rischio ma lo rendono accettabile e controllato, incidendo sia sulla probabilità del suo verificarsi sia sulla gravità delle conseguenze.

I diritti umani, che queste tecniche mirano a proteggere, presentano una dimensione naturalmente correlata e possono entrare in tensione per ragioni legittime, generando

[25] Regolamento (UE) 2016/679 del Parlamento europeo e del Consiglio del 27 aprile 2016 relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonché alla libera circolazione di tali dati e che abroga la direttiva 95/46/CE (regolamento generale sulla protezione dei dati), in *GUUE* L 119 del 4 maggio 2016, p. 1 ss. Va precisato che la possibilità di coordinare la FRIA con la DPIA e di completarla per il tramite di quest'ultima è prevista dall'art. 27, par. 4, AI Act, così come la predisposizione di un modello di questionario, anche mediante uno strumento automatizzato, a cura dell'AI Office, ai sensi dell'art. 27, par. 5.

specifiche criticità proprio nella fase di mitigazione dei rischi. Tale fenomeno, comunemente ricondotto alla nozione di interessi contrapposti, evidenzia la difficoltà di bilanciare istanze egualmente meritevoli di tutela.

Un caso paradigmatico è rappresentato dalla tensione tra utilità e dannosità nei sistemi di intelligenza artificiale generativa in cui l'adozione di misure volte a prevenire la produzione di contenuti dannosi può incidere negativamente sulla funzionalità del sistema, determinando, ad esempio, un incremento dei rifiuti di risposta alle richieste degli utenti[26]. Trasposto sul piano dei diritti fondamentali, tale dinamica riflette un potenziale conflitto tra il diritto di accesso all'informazione, qualora il sistema risulti eccessivamente restrittivo, e il principio di non discriminazione, nel caso opposto.

La mitigazione dei rischi, come si è chiarito nel paragrafo precedente, non può essere concepita come un processo meramente tecnico ma richiede, specie in casi del genere, un approccio fondato sul bilanciamento da condurre secondo criteri di ragionevolezza e volto a garantire il più elevato livello possibile di tutela dei diritti in conflitto, evitando restrizioni non necessarie o sproporzionate. Il bilanciamento tra diritti fondamentali trova la sua sede privilegiata nell'art. 27 dell'AI Act, che, attraverso la valutazione d'impatto, impone agli utilizzatori di confrontarsi con le tensioni tra diritti emergenti nel contesto concreto di utilizzo del sistema. Nell'ambito dell'art. 9, il bilanciamento assume invece una forma implicita e anticipata, riflessa nella necessità di progettare misure di mitigazione che tengano conto della influenza reciproca tra i diversi requisiti e dei potenziali effetti sui diritti fondamentali.

Questa tecnica si fonda sul presupposto che la maggior parte dei diritti fondamentali non abbia carattere assoluto, fatta eccezione per i diritti assolutamente inderogabili ed insuscettibili di eccezioni persino in caso di guerra[27], e possa, pertanto, essere oggetto di limitazioni sempre che siano giustificate da finalità legittime, quali la tutela della sicurezza o di altri diritti parimenti meritevoli di protezione. Il meccanismo, pertanto, non è strutturato sul tipo di diritto di volta in volta all'esame, ma, posto che tutti i diritti – derogabili – possono venire in considerazione, è fondato sulla distinzione tra modalità essenziali e meno essenziali di esercizio dello stesso diritto[28].

In termini operativi, il bilanciamento implica la valutazione delle diverse soluzioni alla luce di criteri consolidati nella giurisprudenza sui diritti fondamentali: il limite deve risultare dalla legge, sia necessario al perseguimento di una finalità di pubblico rilievo ed interesse e sia proporzionato al perseguimento di tale finalità. Inoltre, le limitazioni non devono restringere «le droit en cause d'une manière ou à un degré qui l'atteindraient dans sa substance

---

[26] Un esempio è rinvenibile nei contesti dell'accesso all'informazione e della partecipazione politica dove i *chatbots* di intelligenza artificiale generativa sono sempre più utilizzati, similmente ai motori di ricerca, come strumenti di informazione. Tuttavia, a causa di limiti nei dati di addestramento e della tendenza dei modelli linguistici a generare informazioni inesatte o distorte, essi non garantiscono in assoluto risposte accurate, imparziali e rappresentative. Per evitare i potenziali danni derivanti dalla diffusione di informazioni errate o faziose, molti sviluppatori impongono restrizioni sui temi trattati, in particolare sui contenuti politici ed elettorali generando una tensione tra il diritto di accesso all'informazione e il diritto alla partecipazione politica. Dovendo operare un bilanciamento, gli sviluppatori dovrebbero applicare il principio fondamentale del diritto all'informazione nel contesto democratico consentendo agli individui di accedere a informazioni politiche accurate, imparziali e pluralistiche. Sul punto si veda M. CASTELLANETA, *La libertà di stampa nel diritto internazionale ed europeo*, Bari, 2012, p. 45 ss.

[27] Si pensi ai diritti menzionati all'art. 15 CEDU.

[28] Sulla tecnica del bilanciamento v. G. CARELLA, *Diritto di famiglia islamico, conflitti di civilizzazione e Convenzione europea dei diritti dell'uomo*, in G. CARELLA (a cura di), *La Convenzione europea dei diritti dell'uomo e il diritto internazionale privato*,Torino, 2009, p. 67 ss.

même»[29]. L'applicazione di questi parametri può condurre all'individuazione di una o più soluzioni compatibili con il quadro dei diritti fondamentali.

Le misure di mitigazione come il bilanciamento rilevano in particolare con riguardo alla protezione della privacy la cui violazione potrebbe derivare dall'utilizzo di informazioni personali nei *datasets* di addestramento, dall'impiego dei modelli in contesti ad alta criticità e dalla possibile raccolta o divulgazione di dati in assenza di adeguate garanzie[30]. La protezione dei dati, peraltro, riveste la natura di diritto abilitante, in quanto condizione per l'effettivo esercizio di altri diritti fondamentali e la cui compromissione, dunque, non si esaurisce in una lesione circoscritta potendo incidere ad esempio su sicurezza personale e libertà individuali[31]. Così, le misure di mitigazione nel campo della protezione dei dati contribuiscono a ridurre anche rischi più ampi per i diritti umani nel loro complesso.

La tutela dei dati personali si articola attraverso un insieme integrato di misure organizzative e tecnico-operative, tra cui politiche interne di *governance* dei dati, meccanismi di controllo, minimizzazione dei dati, tecniche di pulizia dei *datasets*, *privacy* differenziale e modelli di elaborazione decentralizzata quali l'*edge AI* e il *federated learning* che consentono di limitare la circolazione dei dati personali in base ad un approccio decentrato[32].

Nell'ambito dei sistemi di intelligenza artificiale generativa, può affiorare una tensione tra l'esigenza di garantire un elevato livello di tutela dei dati personali, attraverso gli strumenti descritti, e la necessità di monitorare l'utilizzo dei sistemi al fine di prevenire abusi o impieghi impropri. L'adozione della tecnica del bilanciamento consente di individuare soluzioni che, nel rispetto dei criteri di necessità e proporzionalità, risultino maggiormente compatibili con la tutela complessiva dei diritti fondamentali. Nel campo dell'IA, questa tecnica può essere messa a punto attraverso strumenti quali la modulazione graduata delle restrizioni, l'introduzione di forme di supervisione umana nei casi più sensibili, nonché l'adozione di meccanismi di trasparenza e contestabilità, secondo il noto test elaborato dalla giurisprudenza della Corte di giustizia, che richiede la verifica dell'idoneità, della necessità e della proporzionalità in senso stretto della misura di volta in volta in rilievo[33].

---

[29] Corte europea dei diritti dell'uomo, sentenza del 18 dicembre 1987, 11329/85, *F.* v. *Switzerland*, plenary, par. 32. Il principio della preservazione della sostanza del diritto è stato affermato a partire dalla decisione del 23 luglio 1968, 1474/62 e altri, *Case "relating to certain aspects of the laws on the use of languages in education in Belgium" v. Belgium* (merits), plenary, par. 5.

[30] In argomento M. VEALE, R. BINNS, J. AUSLOOS, *When Data Protection by Design and Data Subject Rights Clash*, in *International Data Privacy Law*, 2018, p. 105 ss.

[31] Corte europea dei diritti dell'uomo, sentenze del 4 dicembre 2008, 30562/04 e 30566/04, *S. and Marper v. the United Kingdom*, GC, par. 67; del 27 giugno 2017, 931/13, *SatakunnanMarkkinapörssiOy and SatamediaOy v. Finland*, GC, par. 133 s.

[32] Cfr. C. TRONCOSO e altri, *Decentralized Privacy-Preserving Proximity Tracing*, in *IEEE Data Engineering Bulletin*, 2020; D. ANNUNZIATA, F. GIAMPAOLO, F. PICCIALLI, E. PREZIOSO, *Federated Learning ed Edge AI: innovazioni per un'AI sicura e decentralizzata*, in *Agenda Digitale* (online), 17 giugno 2025.

[33] V., in tal senso, Corte di giustizia UE, sentenze dell'8 luglio 2010, causa C-343/09, *Afton Chemical*, EU:C:2010:419, punto 45; del 9 novembre 2010, Grande Sezione, causa C-92/09, *Volker und Markus Schecke* e *Eifert*, ECLI:EU:C:2010:662, punto 74; del 23 ottobre 2012, Grande Sezione, cause riunite C-581/10 e C-629/10, *Nelson* e *a.*, ECLI:EU:C:2012:657, punto 71; del 22 gennaio 2013, Grande Sezione, causa C-283/11, *Sky Österreich*, ECLI:EU:C:2013:28, punto 50; del 17 ottobre 2013, causa C-101/12, *Schaible*, ECLI:EU:C:2013:661, punto 29; dell'8 aprile 2014, Grande Sezione, cause riunite C-293/12 e C-594/12, *Digital Rights Ireland* e *Seitlinger* e *a.*, ECLI:EU:C:2014:238, punto 46.
Il requisito dell'idoneità impone che le misure adottate siano effettivamente capaci di prevenire o ridurre i rischi per i diritti fondamentali derivanti dall'impiego del sistema di intelligenza artificiale; il criterio della necessità richiede che, tra le diverse opzioni disponibili, sia prescelta quella meno restrittiva per i diritti coinvolti, purché ugualmente efficace nel conseguire l'obiettivo di tutela. Infine, la proporzionalità in senso stretto implica una

7. *Tecniche di mitigazione del rischio e approccio* rights-based *nei sistemi di IA*

Nel paragrafo precedente è emerso che le tensioni tra diritti fondamentali non restino confinate su un piano meramente teorico ma rilevino direttamente sulla progettazione e sul funzionamento dei sistemi di IA. Le metodologie, sempre di mitigazione del rischio, di allineamento (anche *alignment*) rappresentano uno degli strumenti attraverso cui le operazioni di bilanciamento vengono tradotte in soluzioni tecniche.

L'allineamento dei modelli costituisce un insieme di metodi volto a orientare il comportamento dei sistemi di intelligenza artificiale generativa affinché i relativi *output* risultino coerenti con valori, norme e vincoli giuridici pertinenti, in particolare con il rispetto dei diritti fondamentali. Tali tecniche intervengono per ridurre la probabilità che il modello generi contenuti dannosi o inappropriati in quanto discriminatori o fuorvianti, nonché le "allucinazioni" ovvero risposte plausibili ma false o prive di fondamento fattuale[34]. Si tratta, evidentemente, di strumenti di prevenzione del rischio. Dal punto di vista giuridico, gli *output* dei sistemi generativi possono incidere su una pluralità di diritti fondamentali, tra cui la dignità umana, il principio di non discriminazione, la sicurezza e il diritto a un'informazione corretta.

Esse si collocano, sul piano normativo, nell'ambito dell'art. 9 dell'AI Act, in quanto strumenti tecnici di mitigazione del rischio integrati nel processo di progettazione e sviluppo del sistema. Tuttavia, la loro concreta configurazione ed efficacia risultano strettamente connesse alla valutazione d'impatto sui diritti fondamentali di cui all'art. 27, che consente di individuare i rischi specifici nel contesto d'uso e di orientare, conseguentemente, le misure di mitigazione. In tale prospettiva, l'*alignment* rappresenta il punto di raccordo tra la gestione del rischio *by design* e la valutazione situata degli impatti sui diritti umani.

Gran parte degli impatti sui diritti umani derivano da *bias* e inesattezze nei *datasets*; pertanto, le tecniche di allineamento che mirano a ridurre tali problemi contribuiscono anche a mitigare i relativi impatti sui diritti fondamentali. Tuttavia, queste metodologie non sono esenti da critiche[35] e, pur rappresentando strumenti di mitigazione del rischio, non

valutazione comparativa tra i benefici derivanti dalla misura adottata e il sacrificio imposto agli altri diritti o interessi, al fine di evitare restrizioni eccessive o sproporzionate.

[34] Con riguardo al profilo tecnico-operativo le principali tecniche di allineamento dei modelli di intelligenza artificiale generativa annoverano il *Reinforcement Learning from Human Feedback* (anche RLHF), il *Reinforcement Learning from AI Feedback* (RLAIF) e i sistemi *rule-basedrewards* (RBR).
Il RLHF consiste nell'addestrare il modello utilizzando valutazioni fornite da esseri umani che orientano il sistema verso risposte considerate più appropriate, accurate o conformi a determinati standard etici e normativi. Il RLAIF rappresenta una variante in cui il *feedback* è generato da altri sistemi di IA, progettati per valutare la qualità e la sicurezza degli *output*. I *rule-based rewards*, infine, si fondano sull'introduzione di regole esplicite che guidano il comportamento del modello, premiando o penalizzando determinate tipologie di risposta sulla base di criteri predefiniti. Si veda G. MASI, *RLHF, addestrare l'IA con feedback umano: una guida completa*, in *Agenda Digitale* (online), 24 giugno 2025.

[35] In particolare, il RLHF può riflettere *bias* e preferenze soggettive dei valutatori umani, con il rischio di incorporare visioni culturali o normative non universalmente condivise; analogamente, il RLAIF può amplificare errori o distorsioni già presenti nei sistemi di IA utilizzati per fornire il *feedback*. Anche gli approcci basati su regole presentano limiti in quanto difficilmente riescono a riprodurre la complessità e la variabilità dei

garantiscono di per sé una piena conformità ai diritti fondamentali, rendendo necessario un controllo continuo e un'integrazione con ulteriori misure di gestione. La loro efficacia aumenta quando sono applicate in modo mirato, sulla base dei rischi individuati attraverso una valutazione d'impatto sui diritti fondamentali. In tal modo, anziché limitarsi a rendere il modello genericamente più sicuro, esse consentono di intervenire su specifici impatti negativi, quali discriminazione, disinformazione o incitamento all'odio. La valutazione d'impatto fornisce, dunque, una mappatura dei rischi in grado di orientare l'allineamento del modello in modo più preciso ed efficace.

Il comportamento del sistema può dunque essere guidato non solo da criteri tecnici, ma anche da principi normativi espliciti, incorporati nei processi di addestramento e valutazione, quali la dignità, la non discriminazione e la libertà di espressione. Un esempio significativo è rappresentato dall'approccio adottato da *Anthropic* nello sviluppo del modello Claude, in cui la Dichiarazione universale dei diritti umani del 1948 è stata utilizzata come una delle fonti per definire i principi guida del sistema, secondo un processo RLAIF in cui il modello viene migliorato utilizzando *feedback* generati da un'altra IA piuttosto che da esseri umani[36]. In termini giuridici, questo modello rappresenta un tentativo di tradurre standard normativi in vincoli tecnici, integrando la logica della valutazione d'impatto con strumenti operativi di mitigazione. Tuttavia, esso solleva anche questioni rilevanti, tra cui la selezione e l'interpretazione dei principi utilizzati, nonché il rischio di una loro applicazione semplificata o parziale all'interno di sistemi complessi.

La progettazione centrata sull'uomo che il regolamento (UE) 2024/1689 pone a fondamento della disciplina[37] consiste nel non causare danni agli utenti, applicando principi come *privacy-by-design* e *safety-by-design* ovvero integrando misure di mitigazione fin dalle prime fasi dello sviluppo del modello. Una progettazione simile può essere estesa allo *human rights-by-design* facendovi confluire la mitigazione di impatti noti sui diritti umani. Questa inclusione permette di identificare tensioni tra diritti e a trovare soluzioni prima che siano incorporate nel prodotto, realizzando un tipo di approccio che verrà meglio chiarito nella parte conclusiva.

L'impostazione tratteggiata è tuttavia subordinata alla chiarezza nell'identificazione dei rischi e delle misure di mitigazione. Quando i potenziali impatti sui diritti umani sono identificati e descritti in modo preciso, i fornitori e gli utilizzatori saranno in grado di selezionare in modo più consapevole le misure di mitigazione appropriate. In assenza di trasparenza, le decisioni potrebbero essere vaghe, incomplete o orientate esclusivamente a criteri tecnici o economici, e volte soprattutto ad agire sulle conseguenze delle violazioni più che sulle cause.

Essa inoltre svolge una funzione essenziale in termini di responsabilità. Rendere espliciti i rischi e le misure adottate consente a soggetti esterni – quali autorità di vigilanza,

---

contesti reali. Sul punto si rinvia a G. BOELLA, *L'allineamento dei modelli può prevenire i danni intenzionali dell'IA?*, in *MagIA*(online), 4 dicembre 2023; G. MOSCA, *Analisi approfondita dell'RLHF: il rapporto tra AI e feedback umano*, in *AI4Business* (online), 14 febbraio 2024; M. PEPICIELLO, *Limiti e problemi aperti nel Reinforcement Learning from Human Feedback*, in *AgorAI* (online), 20 ottobre 2024.

[36] V. Claude's Constitution, 2023-2026, sviluppata da Anthropic per guidare il comportamento e la sicurezza del suo modello linguistico Claude. La Costituzione di Claude, costituita da un elenco di principi ispirati a fonti come la Dichiarazione universale dei diritti umani, nasce come parte dell'approccio denominato «constitutional AI», introdotto da Anthropic per sostituire o comunque integrare la supervisione umana diretta nel processo di addestramento. Y. BAI e altri, *Constitutional AI: Harmlessness from AI Feedback*, https://arxiv.org/pdf/2212.08073 (online), 15 dicembre 2022.

[37] V. considerando 1, 6, 27 e 176, nonché art. 1, par. 1 del regolamento (UE) 2024/1689.

società civile o utenti – di valutare l'operato dei fornitori, verificando se questi abbiano effettivamente adempiuto i propri obblighi di prevenzione e mitigazione. In questo modo, la trasparenza diventa uno strumento di *accountability*, poiché espone le scelte effettuate a scrutinio pubblico e istituzionale[38].

8. *La gestione del rischio e il paradigma non compensatorio dei diritti umani*

Alla luce del modello delineato dall'AI Act, appare necessario confrontare l'impostazione adottata con alcuni standard in tema di gestione del rischio e diritti umani, al fine di verificarne coerenza e adeguatezza. Assumono particolare rilievo i *Guiding Principles on Business and Human Rights* (anche UNGPs) con la relativa guida interpretativa[39], che differenziano la mitigazione dell'impatto negativo sui diritti umani dalla mitigazione dei rischi per i diritti umani sancendo che «[t]he mitigation of adverse human rights impact refers to actions taken to reduce itsextent, with anyresidual impact thenrequiringremediation. The mitigation of human rights risks refers to actions taken to reduce the likelihood of a certain adverse impact occurring»[40]. Nell'individuazione delle misure appropriate, richiamano il criterio relativo alla natura e all'intensità del legame tra l'attore economico e l'impatto negativo (imputazione) e quello concernente il grado di influenza esercitabile su tale impatto (capacità di influenza)[41]. L'interazione tra questi principi consente di delineare un modello di

---

38 V. *infra* par. 9.

39 Il Consiglio per i diritti umani delle Nazioni Unite ha adottato i Principi Guida nella risoluzione 17/4 del 16 giugno 2011. Sui Principi si rinvia a M. FASCIGLIONE, *Per uno studio dei Principi Guida ONU su imprese e diritti umani*, Roma, 2020, p. 35 ss., che ne ha curato anche la traduzione a p. 4 ss. Il secondo pilastro dei Principi Guida, dedicato alla responsabilità delle imprese di rispettare i diritti umani, è meglio chiarito in una guida interpretativa ufficiale del 2012, *The Corporate Responsibility to Respect Human Rights: An Interpretive Guide*, redatta dall'United Nations Office of the High Commissioner for Human Rights (OHCHR).
Gli standard internazionali di *governance* dell'IA comprendono, tra l'altro, la risoluzione n. 78/265 del 21 marzo 2024, approvata mediante *consensus*, dall'Assemblea generale delle Nazioni Unite, *Seizing the Opportunities of Safe, Secure and Trustworthy Artificial Intelligence Systems for Sustainable Development*; la raccomandazione C/MIN(2019)3/FINAL del 22 maggio 2019 dell'OECD, modificata il 3 maggio 2024, *Recommendation of the Council on Artificial Intelligence*; gli *International GuidingPrinciples on Artificial Intelligence* e il *Voluntary Code of Conduct for AI Developers*, adottati dal G7 il 30 ottobre 2023; la *Recommendation on the Ethics of Artificial Intelligence*, adottata dall'UNESCO il 23 novembre 2021, nonché la citata Convenzione quadro del Consiglio d'Europa del 2024 sull'intelligenza artificiale e i diritti umani, la democrazia e lo Stato di diritto.

40 In particolare, il Principio 15 sottolinea che le imprese dovrebbero dotarsi di politiche e processi appropriati alla loro dimensione e alle circostanze, inclusi «[a] human rights due diligence process to identify, prevent, mitigate and account for how they address their impacts on human rights» (lett. b). In tema di *due diligence* sui diritti umani, v. J. BONNITCHA, R. MCCORQUODALE, *The Concept of "Due Diligence" in the UN Guiding Principles on Business and Human Rights*, in *European Journal of International Law*, 2017, p. 899 ss.; J.G. RUGGIE, J.F. SHERMAN, *The Concept of "Due Diligence" in the UN Guiding Principles on Business and Human Rights: A Reply to Jonathan Bonnitcha and Robert McCorquodale*, *ivi*, p. 921 ss.; M. FASCIGLIONE, *Impresa e diritti umani nel diritto internazionale*, Torino, 2024, p. 142 ss.

41 Nella guida interpretativa OHCHR si legge che «[l]everage is an advantage that gives power to influence. In the context of the Guiding Principles, it refers to the ability of a business enterprise to effect change in the wrongful practices of another party that is causing or contributing to an adverse human rights impact». V. anche il Principio 19, lett. b), UNGPs: «Appropriate action will vary according to: (i) Whether the business enterprise causes or contributes to an adverse impact,or whether it is involved solely because the impact is directly linked to itsoperations, products or services by a business relationship;(ii) The extent of its leverage in addressing the adverse impact». M. FASCIGLIONE, *Per uno studio dei Principi Guida ONU*, cit., p. 46 ss.

due diligence fondato sul rischio e sulla responsabilità graduata, nel quale le misure di mitigazione risultano proporzionate sia al ruolo dell'impresa nella produzione dell'impatto sia alla sua capacità effettiva di incidervi. La mitigazione non si configura come un insieme statico di obblighi bensì come un processo continuo di valutazione e intervento, volto a garantire una tutela effettiva dei diritti umani lungo l'intera catena del valore[42].

Un ulteriore criterio ribadito dai predetti Principi, secondo cui gli impatti negativi sui diritti umani non sono suscettibili di compensazione attraverso benefici, contribuisce a definire la natura della mitigazione. Tale criterio esclude una logica compensativa, tipica di approcci utilitaristici, in base alla quale gli effetti negativi su alcuni individui potrebbero essere giustificati dai vantaggi complessivi prodotti dall'attività economica. I diritti umani assumono, invece, carattere non negoziabile, in quanto fondati su standard minimi inderogabili[43].

Ne consegue che, nell'ambito dei processi di *human rights due diligence*, l'impresa non può limitarsi a dimostrare che un'attività produca benefici complessivi superiori ai danni, ma deve prevenire e mitigare gli impatti negativi in quanto tali. La mitigazione non può essere sostituita da forme di compensazione *ex post* né giustificata attraverso logiche di *trade-off* tra diritti e interessi economici. È importante rimarcare, invero, le differenze tra mitigazione e riparazione[44] poiché, rispettivamente, l'una attiene alla prevenzione e alla riduzione degli impatti, l'altra interviene successivamente e non può legittimare retroattivamente la violazione.

La non tollerabilità dell'impatto residuo sui diritti umani impone una riconsiderazione della stessa logica preventiva dovendosi escludere che il rischio residuo possa essere accettato sulla base di criteri meramente probabilistici o di costo, richiedendosi, piuttosto, che l'azione preventiva sia commisurata alla natura qualitativa dei diritti coinvolti. Essi non possono essere ridotti a soglie astratte o a criteri meccanici di valutazione, risultando indispensabile invece un approccio sostanziale e contestuale qual è quello adottato con la tecnica del bilanciamento[45]. Ciò impone di individuare un equilibrio tra l'esigenza di evitare l'accettazione dell'impatto residuo già in fase preventiva e la necessità di non imporre standard irrealistici, quali la totale eliminazione del rischio ovvero tra l'esigenza di garantire

---

[42] In questo senso, anche, il Principio 17, lett. c),UNGPs afferma pure che la *due diligence* in materia di dirittiumani «[s]hould be ongoing, recognizing that the human rights risks may change over time as the business enterprise's operations and operating context evolve».

[43] Rileva in particolare il Principio 11 UNGPs, in cui si evidenzia che «[b]usiness enterprises should respect human rights. This means that they should avoid infringing on the human rights of others and should address adverse human rights impacts with which they are involved», nonché la relative interpretazione OHCHR secondo cui «[i]t is also important to note in this context that there is no equivalent of a carbon off-set for harm caused to human rights: a failure to respect human rights in one area cannot be cancelled out by a benefit provided in another».

[44] Nell'ambito dell'IA, la riparazione comprende il risarcimento economico; la *restituito* ovvero riportare la situazione allo stato precedente (ad esempio cancellazione dei dati, correzione della decisione); la cessazione della violazione, dunque interrompere il comportamento lesivo; le garanzie di non ripetizione (ad esempio modifiche strutturali del sistema, cambiamento delle procedure). Nella guida interpretativa OHCHR si chiarisce che: «[r]emediation and remedy refer to both the processes of providing remedy for an adverse human rights impact and the substantive outcomes that can counteract, or make good, the adverse impact. These outcomes may take a range of forms, such as apologies, restitution, rehabilitation, financial or non-financial compensation, and punitive sanctions (whether criminal or administrative, such as fines), as well as the prevention of harm through, for example, injunctions or guarantees of non-repetition».

[45] V. *supra* par. 6.

un elevato livello di tutela dei diritti umani e il riconoscimento dei limiti concreti dei processi di gestione del rischio.

La prevenzione non va interpretata come un obbligo di risultato – ossia come la completa eliminazione di ogni possibile rischio – bensì come un obbligo di condotta qualificata, fondato sul processo di *due diligence*. Non è, dunque, ammissibile una compensazione tra prevenzione e riparazione, né una strategia preventiva che si limiti a ridurre la gravità del danno lasciandone invariata la probabilità.

Le misure preventive mostrano difatti la loro diversa potenziale efficacia posto che alcune di esse si limitano a ridurre il rischio, mentre altre sono idonee a eliminarlo; analogamente, talune intervengono sulle cause profonde degli impatti sui diritti umani, mentre altre si limitano a trattarne gli effetti. Conseguentemente misure meramente superficiali non possono compensare l'omessa adozione di interventi più incisivi e strutturali, laddove questi siano plausibili, dovendosi sempre privilegiare soluzioni che incidano sulle cause profonde degli impatti.

Quando l'eliminazione totale del rischio non è concretamente realizzabile, la prevenzione deve essere intesa come riduzione della causa del rischio da cui deriva che le misure preventive non possono limitarsi ad intervenire sugli effetti dovendosi bensì indirizzare anche verso l'identificazione e il trattamento di ciò che ha cagionato il rischio. Pur non essendo tutte le misure facilmente riconducibili a questa logica, un adeguato sistema di gestione del rischio deve valorizzare gli interventi volti ad affrontare i fattori strutturali che generano gli impatti negativi. Non accettare anticipatamente l'impatto residuo significa dimostrare che le opzioni di intervento sulle cause del rischio sono state effettivamente considerate e integrate nel processo decisionale.

9. *Rischio residuo e diritti umani nell'AI Act. Struttura e limiti del modello di prevenzione*

Le considerazioni svolte sin ora permettono di distinguere tra concezione "economica" e "giuridico-fondamentale" del rischio. Mentre i modelli aziendalistici ammettono forme di bilanciamento tra costi e benefici (*risk offsetting*), l'AI Act si colloca in una prospettiva differente, fondata sulla prevenzione e sulla mitigazione degli impatti sui diritti fondamentali accogliendo un modello di gestione del rischio orientato alla verifica ed alla responsabilità. La non tollerabilità dell'impatto residuo sui diritti umani trova un significativo riscontro nell'impianto dell'AI Act, in particolare negli articoli 9 e 27, che, pur non escludendo in via assoluta la persistenza di rischi residui, delineano un modello di gestione del rischio fondato su un obbligo di condotta qualificata e non meramente su un criterio di accettabilità del rischio.

L'art. 9, relativo ai sistemi di IA ad alto rischio, impone un processo continuo di identificazione, valutazione e mitigazione dei rischi lungo l'intero ciclo di vita del sistema richiedendo che le misure adottate siano idonee a ridurre tali rischi a un livello accettabile. La nozione di rischio "accettabile" non può tuttavia essere interpretata in senso puramente tecnico o probabilistico, come avviene nei modelli tradizionali di *risk management*, ma deve essere letta alla luce della natura dei diritti fondamentali, che non ammettono logiche compensative.

L'obbligo di gestione del rischio *ex* art. 9 si traduce nella necessità di adottare misure che non si limitino a ridurre gli effetti del rischio ma che siano dirette, ove possibile, a

intervenire sulle sue cause, privilegiando soluzioni strutturali rispetto a interventi meramente correttivi. Ne deriva che la persistenza di un rischio residuo non può essere giustificata in via automatica sulla base di considerazioni di costo o fattibilità tecnica, ma presuppone la dimostrazione che tutte le misure ragionevolmente praticabili siano state adottate.

Tale impostazione è rafforzata dall'art. 27, che introduce la valutazione d'impatto sui diritti fondamentali in capo ai *deployers*, sancendo un'analisi contestuale degli effetti del sistema nel concreto ambito di utilizzo. L'attenzione si sposta così dal sistema in astratto al suo impatto situato, rafforzando una concezione sostanziale della prevenzione. L'art. 27, infatti, richiede di individuare e valutare i rischi specifici per i diritti fondamentali e di adottare misure adeguate a prevenirli o mitigarli, senza prevedere alcuna forma di giustificazione fondata su un bilanciamento utilitaristico tra danni e benefici.

Il coordinamento tra le due disposizioni consente, pertanto, di delineare un modello integrato nel quale la gestione del rischio *by design* prevista dall'art. 9 è sottoposta a una verifica *in concreto* attraverso la valutazione d'impatto di cui all'art. 27. In questo modello, la prevenzione assume una centralità rafforzata e si configura come un obbligo di diligenza qualificata che esclude la possibilità di considerare l'impatto residuo come automaticamente tollerabile né *ex ante* né *ex post*.

L'AI Act, pur evitando l'(impossibile) eliminazione di ogni tipologia di rischio, richiede agli operatori un impegno sostanziale e dimostrabile nella prevenzione degli impatti sui diritti fondamentali, coerente con la loro natura di esigere il rispetto di standard minimi inderogabili e con un approccio alla gestione del rischio che privilegia l'intervento sulle cause rispetto alla mera attenuazione degli effetti.

La riduzione delle cause del rischio implica tipologie di interventi sulle determinanti strutturali del rischio stesso, di cui si è già detto[46], quali la qualità dei dati, la definizione degli obiettivi di addestramento, l'architettura del modello e le finalità d'uso, assumendo rilievo tecniche come il *debiasing* dei *datasets*, l'introduzione di vincoli nel processo di apprendimento e l'integrazione di criteri di tutela dei diritti già nella fase di progettazione[47]. Essa inoltre può essere indirettamente rafforzata attraverso il potenziamento dei meccanismi *a posteriori*, ad esempio mediante regimi di responsabilità più stringenti capaci di orientare le scelte preventive delle imprese. Tali misure si distinguono da interventi meramente correttivi sugli *output*, in quanto mirano a prevenire il rischio alla sua origine.

Il rifiuto dell'approccio del *business case* – secondo cui il rispetto dei diritti umani rappresenterebbe una scelta discrezionale dell'impresa, subordinata a valutazioni di convenienza economica o reputazionale – emerge con chiarezza nel più recente sviluppo del diritto dell'Unione europea in materia di tecnologie digitali e sostenibilità. In tale contesto, l'AI Act si inserisce in un più ampio mutamento paradigmatico, che segna il passaggio da modelli volontaristici di responsabilità sociale d'impresa a un sistema fondato su obblighi giuridici vincolanti[48].

---

[46] V. *supra* paragrafi 6 e 7.

[47] In dottrina, il tema del *debiasing* dei *datasets* è stato ampiamente analizzato, sia sotto il profilo tecnico sia giuridico, evidenziandosi come le distorsioni nei dati costituiscano una delle principali fonti di discriminazione algoritmica. V., *ex multis*, S. BAROCAS, M. HARDT, A. NARAYANAN, *Fairness and Machine Learning. Limitations and Opportunities*, Cambridge-London, 2023, p. 221 ss.; M. MOLINARI, *Debiasing AI: l'uso dell'intelligenza artificiale per un futuro etico e inclusivo*, in *https://hrintelligence.it/debiasing-ai-intelligenza-artificiale-per-futuro-etico-e-inclusivo/* (online), 20 dicembre 2023.

[48] Sul punto G. CARELLA, *La responsabilità civile dell'impresa transnazionale per violazioni ambientali e di diritti umani: il contributo della proposta di direttiva sulla* due diligence *societaria a fini di sostenibilità*, in *Freedom, Security & Justice: European Legal Studies*, 2022, p. 10 ss.

Questo mutamento trova un fondamento teorico nei citati UN Guiding Principles, i quali, pur nella loro natura di *soft law*, hanno introdotto il concetto di *human rights due diligence* come standard di condotta per le imprese. Tuttavia, proprio il carattere non vincolante degli UNGPs ha favorito, nella prassi, una loro interpretazione in chiave di *business case*, limitandone l'effettività. È tale limite che il legislatore europeo ha inteso superare con la direttiva (UE) 2024/1760 del Parlamento europeo e del Consiglio, del 13 giugno 2024, relativa al dovere di diligenza delle imprese ai fini della sostenibilità (*Corporate Sustainability Due Diligence Directive*, CSDDD), la quale positivizza il modello della *due diligence*. La direttiva impone infatti alle imprese un insieme organico di obblighi giuridici, che trasformano la tutela dei diritti umani in un elemento strutturale della *governance* aziendale[49].

In particolare, l'art. 5 prevede l'integrazione del dovere di diligenza nelle politiche e nei sistemi di gestione del rischio, rendendo la considerazione degli impatti sui diritti umani parte integrante delle strategie societarie. L'art. 8 introduce un obbligo continuo di individuazione e valutazione degli impatti negativi effettivi e potenziali, esteso all'intera catena del valore (*value chain*), superando così approcci limitati alla sola sfera di controllo diretto dell'impresa. Gli articoli 10 e 11 delineano poi obblighi di natura operativa in forza dei quali le imprese devono, rispettivamente, prevenire o mitigare i rischi potenziali e porre fine agli impatti negativi già verificatisi. Di particolare rilievo è l'art. 12, che sancisce un obbligo di riparazione, rafforzando la dimensione rimediale e introducendo un nesso più diretto tra violazione e responsabilità. Il tratto decisivo che sottrae la materia alla logica della volontarietà non risiede, tuttavia, in uno specifico apparato rimediale o sanzionatorio – la cui intensità è stata anzi ridimensionata dalla successiva riforma, che ha tra l'altro soppresso il regime armonizzato di responsabilità civile dell'Unione – bensì nella positivizzazione del dovere di diligenza quale obbligo giuridico vincolante, sottratto alla discrezionalità dell'impresa e funzionalmente orientato alla prevenzione e alla mitigazione degli impatti negativi.

La dottrina ha evidenziato come solo una evoluzione di questo tipo possa segnare il passaggio verso una *hard law corporate accountability* in cui la *due diligence* non rappresenti uno standard di buone prassi bensì un obbligo di condotta giuridicamente esigibile la cui violazione può incidere sulla responsabilità dell'impresa[50]. In un quadro del genere la tutela dei diritti fondamentali si configura come limite strutturale all'innovazione e al mercato, non negoziabile rispetto a logiche di convenienza economica. Il definitivo superamento dell'approccio *business case* si realizza attraverso la positivizzazione della *due diligence* e l'introduzione di obblighi che vincolano le imprese a prevenire, mitigare e riparare gli impatti negativi, inclusi quelli derivanti dall'uso di sistemi di intelligenza artificiale *ex* AI Act.

---

[49] Occorre tuttavia dare conto del fatto che la CSDDD è stata nel frattempo ridefinita dalla stagione di semplificazione (c.d. pacchetto Omnibus I) ovvero dalla direttiva (UE) 2025/794, c.d. Stop the Clock, in vigore dal 17 aprile 2025, in *GUUE* L 2025/794 del 16 aprile 2025, p. 1 ss., e, soprattutto, dalla direttiva (UE) 2026/470, in *GUUE* L 2026/470 del 26 febbraio 2026, in vigore dal 18 marzo 2026, la quale ha ristretto in modo sensibile l'ambito soggettivo (soglie elevate a 5.000 dipendenti e 1,5 miliardi di euro di fatturato), ha soppresso il regime armonizzato di responsabilità civile dell'Unione previsto all'art. 29, par. 1 – rimettendone la disciplina al diritto nazionale – e ha abrogato l'obbligo di adottare un piano di transizione climatica. Il termine di recepimento delle modifiche è fissato al 26 luglio 2028 e l'applicazione, unificata per tutte le imprese soggette, al 26 luglio 2029.

[50] R. MCCORQUODALE, *Pluralism, Global Law and Human Rights: Strengthening Corporate Accountability for Human Rights Violations*, in *Global Costitutionalism*, 2013, p. 287 ss.; K. BUHMANN, Business and Human Rights, in A. RASCHE, M. MORSING, J. MOON, A. KOURULA (eds.), *Corporate Sustainability: Managing Responsible Business in a Globalised World*, Cambridge, 2023, p. 435 ss.

Il quadro emergente non è tuttavia esente da critiche. Oltre a quelle già formulate[51], va evidenziato che l'approccio *business case*, sebbene formalmente abbandonato, tende a riemergere in forma indiretta proprio attraverso l'adozione del modello *risk-based* che accomuna gli strumenti normativi in parola. La selezione e la priorità degli interventi, infatti, sono inevitabilmente influenzate da valutazioni di probabilità e gravità del rischio, ma anche da considerazioni di costo, esposizione reputazionale e sostenibilità economica delle misure adottate. La tutela dei diritti fondamentali rischia così di essere interpretata attraverso logiche di efficienza aziendale.

Questa preoccupazione è emersa chiaramente nel discorso del 30 gennaio 2025 pronunciato dal Commissario per i diritti umani del Consiglio d'Europa «Human rights oversight of artificial intelligence»[52], in cui si mette in luce come la regolamentazione dell'IA sia fortemente condizionata dal settore privato sia perché le grandi imprese sviluppano le tecnologie sia perché influenzano anche il linguaggio e le categorie con cui il fenomeno viene regolato. In particolare, il Commissario evidenzia come esse contribuiscano a definire cosa si intende per rischio, quali rischi siano rilevanti e quali interessi meritino protezione prioritaria. Se, dunque, il concetto di rischio è in parte "costruito" dagli stessi soggetti regolati, si crea un conflitto proprio con l'obiettivo di tutela dei diritti fondamentali.

Occorre riflettere poi sulla indeterminatezza strutturale degli obblighi di *due diligence*. Le definizioni di "misure appropriate", "impatti negativi potenziali" e "catena del valore" sono volutamente elastiche al fine di garantire adattabilità a contesti diversi. Tuttavia, questa flessibilità comporta un significativo sacrificio in termini di certezza del diritto e di prevedibilità delle conseguenze giuridiche, rendendo complesso sia per le imprese individuare con precisione il contenuto dei propri obblighi, sia per le autorità di vigilanza accertarne eventuali violazioni. Il rischio concreto è quello di una deriva verso forme di *formal compliance*, in cui l'adempimento si esaurisce nella predisposizione di procedure e documentazione senza incidere effettivamente sulla prevenzione degli impatti negativi.

La progressiva proceduralizzazione della tutela dei diritti umani, attraverso sistemi di *due diligence* sempre più articolati, può condurre a una riduzione della sostanza a favore della forma, trasformando la protezione dei diritti in un insieme di adempimenti standardizzati e verificabili *ex ante*, piuttosto che in un'effettiva capacità di incidere sugli impatti reali delle attività d'impresa. In tale prospettiva, la prevenzione potrebbe essere ridotta, nella prassi, a una mera attività di riduzione del rischio, con la conseguenza di legittimare implicitamente una forma di accettazione anticipata dell'impatto residuo.

---

[51] In particolare, v. *supra* par. 4. Si veda anche F. BORGIA, *Il modello europeo*, cit., p. 1182 ss. Un'ulteriore conferma di tale tendenza può trarsi dallo stesso Digital Omnibus on AI, approvato dal Parlamento europeo il 16 giugno 2026 e in attesa di adozione formale del Consiglio, il quale, in dichiarata ottica di competitività e semplificazione, posticipa l'applicazione degli obblighi sui sistemi ad alto rischio (al 2 dicembre 2027 e al 2 agosto 2028) e introduce margini di flessibilità nella classificazione e negli obblighi di registrazione. Si tratta di misure che, pur non incidendo sull'architettura *risk-based* né sui contenuti degli articoli 9 e 27, mostrano come la pressione del settore regolato continui a orientare la definizione e la tempistica delle garanzie.

[52] Human rights oversight of artificial intelligence, https://www.coe.int/en/web/commissioner/-/human-rights-oversight-of-artificial-intelligence#:~:text=The%20AI%20Act%20requires%20so,to%20human%20rights%20test%20technology, 30 gennaio 2025.