Osservatorio ACCREDIA
1 2024

Norme tecniche e valutazione della conformità accreditata per lo sviluppo dei sistemi di Intelligenza Artificiale

In collaborazione con:
cini consorzio interuniversitario nazionale per l'informatica

ACCREDIA
L'ENTE ITALIANO DI ACCREDITAMENTO

Abstract

L'era digitale ha inaugurato un periodo di rapidi progressi tecnologici, con l'Intelligenza Artificiale (IA) al centro di questa trasformazione, capace di rivoluzionare vari settori della società. Questo documento esplora l'evoluzione della regolamentazione dell'IA, focalizzandosi su aspetti etici, norme tecniche e casi studio concreti che dimostrano l'applicazione pratica di tali regolamentazioni.

Il documento si apre con un'analisi dell'etica dell'IA, evidenziando come le metodologie di valutazione etica si siano evolute da semplici audit interni a modelli complessi come l'AI Impact Assessment. Si discute l'importanza di anticipare e mitigare i rischi etici per instaurare fiducia tra gli utenti e garantire che le innovazioni in IA siano responsabili e trasparenti. Viene illustrata l'evoluzione delle iniziative normative a livello internazionale, con particolare attenzione alle linee guida elaborate da organizzazioni come l'OECD (OCSE - Organizzazione per la Cooperazione e lo Sviluppo Economico) che promuovono la trasparenza, la giustizia e il rispetto dei diritti umani nell'impiego dell'IA.

Un punto focale del documento è l'AI Act dell'Unione europea, proposto nel 2021, che rappresenta uno dei più completi tentativi di legislazione specifica per l'IA. L'AI Act classifica i sistemi di IA in base al rischio e impone requisiti normativi corrispondenti, bilanciando l'innovazione con la necessità di proteggere i diritti fondamentali. La normazione tecnica, supportata da organizzazioni come il CEN (Comitato Europeo di Normazione), il CENELEC (Comitato Europeo di Normazione Elettrotecnica) e l'ISO (Organizzazione Internazionale per la Normazione) gioca un ruolo cruciale, fornendo un quadro per la valutazione e la certificazione della conformità tecnica e di rischio per i sistemi di IA. Le norme sviluppate da CEN e CENELEC sono fondamentali per garantire la presunzione di conformità ai requisiti dell'AI Act, coprendo vari aspetti, tra cui la gestione del rischio, la qualità dei dati, la trasparenza, la robustezza e la sicurezza dei sistemi di IA. Seguendo queste norme tecniche, le organizzazioni possono assicurare che i loro sistemi di IA siano conformi alle normative europee, semplificando il processo di verifica e certificazione.

Il documento prosegue con la descrizione di due PoC (Proof of Concept) in ambito biomedicale, esaminando la *detection* del melanoma e la stratificazione dei pazienti con sclerosi multipla. Questi esempi mostrano come l'IA possa supportare le decisioni cliniche seguendo una procedura che copre raccolta e preparazione dei dati, sviluppo e valutazione dei modelli, fino al rilascio. La conformità alla linea guida ISO/IEC TR 24027:2021 è essenziale per garantire che i sistemi di IA siano privi di bias e sicuri per l'uso clinico. Parallelamente, viene presentato un PoC nel settore della Pubblica Amministrazione con INAIL, che simula la conformità alla norma ISO/IEC 42001:2023 sul Quality Management per i sistemi di IA. Tale caso evidenzia l'importanza di integrare la valutazione della conformità nei cicli di sviluppo dei sistemi di IA, assicurando che le decisioni basate sull'IA rispettino elevati standard di qualità e sicurezza necessari in contesti organizzativi complessi. Infine, il documento evidenzia, nel contesto della Strategia italiana per l'Intelligenza Artificiale 2024-2026, il contributo dell'accreditamento per il raggiungimento degli obiettivi di tutela di cittadini e imprese, suggerendo per Accredia un ruolo di supporto alla Pubblica Amministrazione nelle attività di accreditamento e monitoraggio delle attività di valutazione della conformità nell'applicazione dei sistemi di IA.

Osservatorio ACCREDIA 1 2024

Norme tecniche e valutazione della conformità accreditata per lo sviluppo dei sistemi di Intelligenza Artificiale

Introduzione

L'era digitale ha portato con sé numerosi progressi tecnologici, tra cui l'Intelligenza Artificiale (IA), che sta trasformando radicalmente il tessuto della società. Mentre offre possibilità senza precedenti per l'innovazione e il miglioramento della qualità della vita, l'IA solleva anche questioni significative relative alla sicurezza, alla privacy, all'equità e all'etica. Di conseguenza, la regolazione dell'IA è emersa come un campo di interesse accademico, politico e sociale. Questo documento esplora l'evoluzione della regolazione sull'IA, concentrandosi sul Regolamento UE 2024/1689 (AI Act), lo sviluppo di regolamentazioni e framework internazionali, e le norme tecniche vigenti in materia.

La definizione di sistema di Intelligenza Artificiale

Le difficoltà nell'analisi e nella regolamentazione dei rischi e degli impatti iniziano già dalla definizione di Intelligenza Artificiale – con cui si intende il campo accademico e industriale che studia questi modelli – e dalla definizione di sistema di Intelligenza Artificiale – con cui si definisce un artefatto che include l'utilizzo di questo tipo di approcci per funzionare.
I sistemi di IA sono l'oggetto degli sforzi normativi, sia in Europa sia a livello internazionale, e la loro definizione risulta un compito complesso e sfaccettato, dovuto principalmente alla vastità e alla diversità delle tecniche e degli approcci che comprende. La difficoltà nasce dall'intersezione di vari campi disciplinari, dall'evoluzione continua delle tecnologie e dalle differenti applicazioni che abbracciano settori disparati.
Uno dei principali ostacoli è la natura dinamica e in continua evoluzione della tecnologia. Le innovazioni in IA emergono rapidamente, e continuamente vengono sviluppati nuovi modelli, che vengono implementati in sistemi sempre più complessi e diversificati. Questa rapida evoluzione rende difficile creare una definizione statica e universale di IA. Inoltre, l'IA è un campo interdisciplinare che combina informatica, matematica, neuroscienze, psicologia, ingegneria e altro ancora. Ogni disciplina porta con sé una propria comprensione di ciò che costituisce l'intelligenza e di come essa possa essere replicata artificialmente.
La vasta gamma di tecniche e approcci utilizzati nell'IA aggiunge un ulteriore livello di complessità. Il machine learning, una delle tecniche più diffuse, si basa sull'analisi di grandi quantità di dati per identificare pattern e fare previsioni. Al suo interno, troviamo sotto-discipline come l'*apprendimento supervisionato, non supervisionato e per rinforzo*, ciascuna con i propri metodi e applicazioni. Il deep learning, un sottoinsieme del machine learning, utilizza reti neurali profonde per elaborare dati complessi come immagini e audio, ma richiede grandi risorse computazionali.

La *logica fuzzy*, che permette valori intermedi tra vero e falso, è utile in contesti di incertezza, ma può essere meno precisa rispetto alla logica tradizionale. Gli *algoritmi genetici*, ispirati alla selezione naturale, cercano soluzioni ottimali attraverso processi iterativi, ma possono essere computazionalmente intensivi e non sempre garantiscono la soluzione migliore.
Ogni tecnica ha i suoi punti di forza e debolezze, adattandosi meglio a specifici problemi e applicazioni. Questa diversità tecnologica rende difficile fornire una definizione onnicomprensiva che catturi tutte le sfumature dell'IA.
Infine, la regolamentazione dell'IA, come evidenziato dall'EU AI Act, aggiunge un ulteriore strato di complessità. L'AI Act cerca di creare un quadro normativo armonizzato, classificando i sistemi di IA in base ai rischi che pongono e stabilendo requisiti di trasparenza e conformità. Tuttavia, regolamentare un campo in rapida evoluzione come l'IA è un compito arduo, poiché le normative devono essere abbastanza flessibili da adattarsi ai cambiamenti tecnologici senza soffocare l'innovazione.
All'interno del Regolamento UE la definizione di sistema di Intelligenza Artificiale è cambiata numerose volte. La prima versione dell'AI Act, il testo preliminare prodotto dalla Commissione europea del 2021, riportava una lista di sistemi di IA. Riportiamo di seguito la prima definizione nel testo del 2021:
"L'Intelligenza Artificiale, come definita all'art. 3 del Regolamento, comprende qualsiasi sistema basato su macchine che sia progettato per operare con diversi livelli di autonomia e che può mostrare adattabilità dopo il dispiegamento e che, per obiettivi espliciti o impliciti, deduce, dagli *input* che riceve, come generare *output* come previsioni, contenuti, raccomandazioni o decisioni che possono influenzare ambienti fisici o virtuali.
Tra gli approcci menzionati vi rientrano quelli:

a) di apprendimento automatico, che comprendono l'apprendimento supervisionato, l'apprendimento non supervisionato e l'apprendimento per rinforzo. Questi approcci utilizzano una vasta gamma di metodologie, tra cui il deep learning;
b) basati sulla logica e *knowledge-based*, che comprendono la rappresentazione della "conoscenza", la programmazione induttiva, le "basi di conoscenza", i modelli inferenziali e deduttivi, il ragionamento (simbolico) e i sistemi esperti.
c) statistici, a stima bayesiana e metodi di ricerca e ottimizzazione."[1]

Questo approccio è stato successivamente abbandonato perché, da più parti, sono state mosse critiche a questa definizione. In breve, la lista era considerata troppo restrittiva, in quanto non includeva alcune tecniche esistenti e non permetteva di far ricadere sotto il regolamento tecniche future. Questo limite, dato il ritmo rapidissimo con cui il mondo dell'IA evolve, ha portato ad una definizione più ad alto livello, che troviamo nella versione finale del testo:
"Sistema di IA: 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."[2]

---

[1] *"a) Machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning;*
b) Logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems;
c) Statistical approaches, Bayesian estimation, search and optimization methods."
[2] *"AI system' means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environment."*

Come si può notare, la definizione corrente permette una grande elasticità e permette, in futuro, di includere altri tipi di tecniche, a oggi non esistenti. In quest'ultima versione, l'accento è posto soprattutto sull'autonomia del sistema. Anche questa definizione ha ricevuto critiche, nello specifico quella di essere talmente ad alto livello da includere, in linea teorica, tecniche che non sono a oggi considerate Intelligenza Artificiale. Ai fini della presente ricerca, utilizzeremo pertanto la definizione di sistemi di IA elaborata nell'ambito dell'AI Act.

Perché parliamo di etica e regolamentazione dei sistemi di Intelligenza Artificiale

L'IA sta rapidamente trasformando molti aspetti della nostra società, dall'economia alla sanità, dalla sicurezza pubblica all'educazione. Tuttavia, con l'aumento dell'influenza e dell'ubiquità dei sistemi di IA, emergono anche numerosi interrogativi etici e la necessità di una regolamentazione adeguata per garantire che tali tecnologie vengano utilizzate in modo responsabile e a beneficio di tutti.

Uno dei motivi principali per cui è fondamentale discutere di etica e regolamentazione è l'impatto significativo che i sistemi di IA possono avere sui diritti umani e sulla società in generale. Gli algoritmi di IA possono influenzare decisioni critiche che riguardano la vita delle persone, come l'accesso al credito, le opportunità di lavoro, l'assistenza sanitaria e perfino le sentenze giudiziarie. Senza una regolamentazione adeguata, c'è il rischio che tali decisioni siano influenzate da bias presenti nei dati di addestramento, perpetuando e amplificando le disuguaglianze esistenti.

Ad esempio, alcuni studi hanno dimostrato che alcuni algoritmi di riconoscimento facciale hanno tassi di errore significativamente più alti per le persone di colore rispetto a quelle bianche. Questo può portare a situazioni di discriminazione ingiusta e violazioni dei diritti umani, evidenziando la necessità di norme che garantiscano l'equità e la trasparenza.

Inoltre, i sistemi di IA possono influenzare la privacy degli individui, raccogliendo e analizzando enormi quantità di dati personali. La sorveglianza massiva resa possibile dall'IA può minare la libertà individuale e creare una società di controllo, dove ogni azione è monitorata e analizzata. La regolamentazione deve quindi garantire che i diritti alla privacy e alla protezione dei dati siano rispettati, evitando l'abuso delle tecnologie di sorveglianza. L'UE, ad esempio, ha stabilito regole rigorose sull'uso della sorveglianza biometrica in tempo reale in spazi pubblici, permettendola solo in casi eccezionali e con severe garanzie per evitare abusi. Queste norme sono state create per bilanciare i benefici delle tecnologie di riconoscimento facciale con la necessità di proteggere i diritti fondamentali degli individui.

Un altro aspetto è la trasparenza e la responsabilità nei sistemi di IA. Molti algoritmi di IA sono spesso descritti come "scatole nere" perché è difficile comprendere come arrivino a determinate decisioni. Questa mancanza di trasparenza può essere problematica, specialmente quando le decisioni hanno un impatto significativo sulla vita delle persone. La regolamentazione può aiutare a stabilire requisiti per la documentazione e la spiegazione dei processi decisionali degli algoritmi, garantendo che le organizzazioni siano responsabili delle loro tecnologie. Ad esempio, l'Unione europea, con l'AI Act, ha stabilito un quadro giuridico che richieda trasparenza e accountability per i sistemi di IA ad alto rischio, assicurando che possano essere monitorati e valutati adeguatamente.

La valutazione del rischio e dell'impatto è stata posta al centro di tutte le regolamentazioni e degli approcci di governance a livello internazionale, in quanto strumento fondamentale ed efficace per garantire che i sistemi di IA vengano utilizzati in modo sicuro ed etico. Valutare i rischi associati ai sistemi di IA significa esaminare potenziali errori, bias, e conseguenze negative che potrebbero emergere dall'uso della tecnologia. Ad esempio, un algoritmo usato per il reclutamento del personale deve essere valutato per assicurarsi che non discrimini candidati in base a etnia, genere o altre caratteristiche. Inoltre, l'impatto dei sistemi di IA deve essere considerato non solo in termini tecnici, ma anche in termini sociali ed etici. Un sistema di IA che funziona perfettamente dal punto di vista tecnico potrebbe comunque avere effetti deleteri sulla società, come aumentare la disoccupazione o minare la coesione sociale. Valutare l'impatto significa considerare queste dimensioni più ampie e assicurarsi che l'IA contribuisca positivamente al benessere della società nel suo insieme. Senza questo tipo di valutazioni, esiste il rischio di perpetuare ingiustizie sistemiche e violare i diritti umani. Questa valutazione permette di identificare e mitigare i rischi prima che i sistemi vengano implementati su larga scala. Le modalità con cui la valutazione del rischio verrà effettuata nel contesto dell'EU AI Act sarà oggetto dei capitoli successivi di questo documento. Il prossimo capitolo, invece, ripercorrerà alcuni dei passaggi fondamentali che hanno portato dall'etica alla regolazione dell'IA.

1. Dall'etica al Regolamento europeo sull'Intelligenza Artificiale

1.1 L'etica dell'Intelligenza Artificiale e l'evoluzione delle metodologie di assessment

La discussione sull'etica dell'IA non è nuova, ma la sua urgenza e complessità sono cresciute in maniera esponenziale con l'avanzare delle tecnologie. Inizialmente, le preoccupazioni etiche erano focalizzate su questioni di bias e discriminazione, ad esempio nel riconoscimento facciale o nei sistemi di decisione algoritmica. Tuttavia, l'attenzione si è ampliata per includere l'impatto più ampio dell'AI sull'autonomia umana, sulla sorveglianza, sul diritto al lavoro e sulla privacy. Le metodologie di assessment etico sono diventate sempre più sofisticate. Originariamente basate su audit interni e checklist di conformità, queste metodologie si sono evolute verso modelli più integrati e olistici. Gli approcci moderni, come l'*ethics-by-design*, cercano di anticipare e mitigare i rischi etici prima che i prodotti di IA raggiungano il mercato. Questi strumenti sono vitali per instaurare fiducia tra gli utenti e per garantire che le innovazioni in IA siano responsabili e trasparenti. La necessità di un quadro normativo ha condotto all'elaborazione di varie iniziative sia a livello nazionale sia internazionale. Uno dei primi tentativi significativi di regolamentare l'IA a livello globale è stato il GDPR (General Data Protection Regulation - Regolamento UE 2016/679) che, sebbene non specificamente focalizzato sull'IA, ha stabilito importanti precedenti in termini di protezione dei dati e accountability algoritmica. Nel corso degli ultimi anni, abbiamo assistito alla nascita di specifici Framework internazionali destinati a governare l'uso dell'IA. Organizzazioni come l'OCSE e l'UNESCO hanno elaborato linee guida che enfatizzano la trasparenza, la giustizia e il rispetto per i diritti umani nell'impiego dell'IA. Queste linee guida servono come riferimento per i Paesi membri e stimolano l'adozione di normative nazionali coerenti con questi standard internazionali. Un momento decisivo nella regolamentazione dell'IA è stato l'introduzione dell'AI Act da parte dell'Unione europea. Proposto per la prima volta nel 2021, l'AI Act è uno degli esempi più comprensivi di legislazione specifica per l'IA. Classifica i sistemi IA in base al rischio che presentano e impone requisiti normativi corrispondenti.

1.2 Dall'etica dell'Intelligenza Artificiale all'EU AI Act: le "Ethics guidelines for trustworthy AI"

Prima di analizzare il Regolamento europeo è importante fornire un quadro d'insieme dell'evoluzione della riflessione accademica sulle implicazioni etico-sociali dell'IA. L'evoluzione *dell'AI ethics* ha radici accademiche, con i primi studi che risalgono agli anni '50 e '60, quando scienziati come Norbert Wiener iniziarono a esplorare le implicazioni morali dei sistemi cibernetici.

Negli anni '80 e '90, la riflessione sull'*AI ethics* ha guadagnato slancio, con accademici e filosofi che hanno iniziato a delineare principi etici per l'IA, come la prevenzione delle implicazioni etiche negative, l'equità, la trasparenza e il rispetto per l'autonomia umana.
Un punto di svolta significativo è stato il *white paper* "Ethics guidelines for trustworthy AI", pubblicato nel 2019 dalla Commissione europea. Questo documento ha delineato sette requisiti chiave per un'IA affidabile, diventando un riferimento fondamentale per le organizzazioni che sviluppano e implementano sistemi di IA e sottolineando l'importanza di un approccio etico integrato. Le "Ethics guidelines" sono state redatte dall' AI HLEG (Gruppo di Esperti di Alto Livello sull'Intelligenza Artificiale) che ancora oggi funge da faro, guidando lo sviluppo e l'implementazione dell'IA nell'Unione europea verso un percorso non solo innovativo, ma anche eticamente sostenibile.
Al centro delle linee guida c'è il concetto di "IA affidabile". La Commissione europea immagina un ecosistema di IA dove i sistemi non sono solo avanzati ma anche affidabili, etici e allineati ai valori umani. Questa si basa su tre pilastri fondamentali:

1. Legale (*lawful*): sottolinea l'importanza che i sistemi di IA rispettino le leggi e i regolamenti esistenti. Dato il carattere dinamico della tecnologia e l'evoluzione del panorama legale, garantire che i sistemi di IA rimangano conformi è di fondamentale importanza. Tale aderenza legale assicura che sviluppatori e utenti di IA operino entro i limiti stabiliti dalle Autorità regolatorie, minimizzando i rischi legali e favorendo la fiducia pubblica.
2. Etico: l'IA dovrebbe rispettare principi morali e valori sociali. Ciò significa che i sistemi di IA dovrebbero essere progettati e implementati in modi che sostengano la dignità umana, rispettino i diritti umani e promuovano il benessere sociale. Le considerazioni etiche assicurano che l'IA serva l'umanità e non danneggi involontariamente individui o gruppi.
3. Robusto: un'IA affidabile dovrebbe essere tecnicamente solida e operare in modo affidabile in condizioni diverse. Questa robustezza si estende a garantire che i sistemi di IA siano resilienti sia contro attacchi intenzionali malevoli che errori involontari. Inoltre, la robustezza sociale dell'IA, che riguarda il suo impatto più ampio sulla società, è altrettanto vitale.

Per dare istruzioni operative per l'affidabilità nei sistemi di IA, le linee guida stabiliscono che tali sistemi dovrebbero soddisfare, in particolare, sette requisiti:

1. Agenzialità umana e supervisione: al centro della visione europea dell'IA c'è l'idea che la tecnologia dovrebbe potenziare le capacità umane, non minarle. L'IA dovrebbe essere uno strumento che potenzia gli individui, migliorando il loro processo decisionale senza erodere la loro autonomia. Dovrebbero essere messi in atto meccanismi efficaci di supervisione umana per garantire che le azioni dell'IA siano allineate con le intenzioni umane.
2. Robustezza tecnica e sicurezza: man mano che i sistemi di IA diventano più integrati in settori critici come la sanità, i trasporti e la finanza, la loro affidabilità tecnica diventa fondamentale. I sistemi dovrebbero essere progettati per gestire incertezze, operare in modo sicuro e essere resilienti sia contro attacchi esterni che fallimenti interni del sistema.
3. Privacy e governance dei dati: nell'era digitale, i dati sono un asset prezioso. I sistemi di IA, intrinsecamente basati sui dati, dovrebbero dare priorità alla protezione dei dati personali. Meccanismi robusti di governance dei dati dovrebbero garantire che i dati vengano acquisiti, conservati ed elaborati in modi che rispettino la privacy individuale e si conformino alle normative sulla protezione dei dati.
4. Trasparenza: per fidarsi dell'IA, gli utenti devono capirla. I processi, gli algoritmi e i meccanismi decisionali dei sistemi di IA dovrebbero essere trasparenti.

Questa trasparenza garantisce che utenti, regolatori e il pubblico più ampio possano comprendere e fidarsi delle azioni e delle decisioni prese dall'IA.

5. Diversità, non discriminazione e equità: i sistemi dovrebbero essere progettati per essere inclusivi, soddisfacendo gruppi di utenti diversi. Inoltre, gli algoritmi di IA dovrebbero essere privi di pregiudizi, garantendo che le decisioni prese siano eque e non discriminino nessun individuo o gruppo.
6. Benessere ambientale e sociale: l'impatto più ampio dell'IA sulla società e sull'ambiente non può essere ignorato. I sistemi di IA dovrebbero essere sostenibili, minimizzando la loro impronta ambientale. Inoltre, le implicazioni sociali dell'IA, dal suo impatto sull'occupazione al suo ruolo nel plasmare il discorso pubblico, dovrebbero essere considerate, assicurando che l'IA contribuisca positivamente al progresso sociale.
7. Responsabilità: via via che i sistemi di IA esercitano un'influenza crescente su vari aspetti della società, dovrebbero essere messi in atto meccanismi per rendere responsabili sviluppatori, utenti e altre parti interessate rispetto ai risultati di questi sistemi. Questa responsabilità garantisce che in caso di errori, pregiudizi o altri problemi, ci siano vie chiare per il ricorso e l'azione correttiva.

Nel contesto delle linee guida, la Direzione Generale della Commissione europea per la Ricerca e l'Innovazione ha presentato una prospettiva sul futuro dell'industria in Europa, denominata "Industria 5.0". Questo concetto emerge come risposta alle sfide e alle opportunità in evoluzione nel panorama industriale, mirando a garantire che le industrie europee rimangano competitive, sostenibili e incentrate sull'uomo di fronte ai rapidi cambiamenti tecnologici e sociali. A differenza del paradigma dell'Industria 4.0, che si concentrava principalmente sul potenziale delle tecnologie emergenti per migliorare l'efficienza e la produttività, l'Industria 5.0 è guidata dai cambiamenti sociali e dalle realtà emergenti. Essa enfatizza il ruolo della tecnologia e dell'innovazione come componenti essenziali per la transizione a un nuovo paradigma industriale. In questa nuova visione, l'industria europea diventa più resiliente, si adatta ai cambiamenti sociali, rispetta i confini del pianeta e pone il benessere dei lavoratori industriali al centro del processo produttivo.
I temi chiave del documento possono essere riassunti nei seguenti punti:

- L'Unione europea ha costantemente promosso un approccio centrato sull'uomo nelle sue politiche. L'Industria 5.0 sottolinea ulteriormente questo aspetto sfumando le linee tra i diversi tipi di lavoratori industriali, garantendo che sia i lavoratori "blue collar" che quelli "white collar" beneficino dei progressi tecnologici.
- Il documento evidenzia l'importanza di integrare le priorità sociali e ambientali europee nell'innovazione tecnologica. Esso sostiene un approccio sistemico, enfatizzando l'interazione tra varie tecnologie. Alcune delle tecnologie abilitanti discusse includono l'interazione uomo-macchina, le tecnologie bio-ispirate, i gemelli digitali, la trasmissione e l'analisi dei dati, l'IA e le tecnologie energeticamente efficienti.
- Si sostiene un significativo passaggio dal valore per gli azionisti al valore per gli stakeholder. Ciò significa che i benefici dell'Industria 5.0 dovrebbero estendersi oltre gli interessi aziendali per abbracciare stakeholder sociali più ampi, garantendo un'interazione "win-win" tra industria e società.

Da un punto di vista implementativo, alcune metodologie sono state sviluppate per permettere una effettiva integrazione dell'etica dell'AI all'interno dei processi di design. Discutiamo nel prossimo paragrafo dei cosiddetti approcci *by-design*.

1.3 Ethics by design

L'integrazione di considerazioni etiche nello sviluppo dei sistemi di IA non può essere un'operazione di ripensamento a posteriori; deve essere intrinseca all'intero ciclo di vita del sistema. Questo approccio, noto come *ethical by design*, richiede che i principi etici siano considerati fin dalle prime fasi di progettazione, sviluppo, implementazione e utilizzo delle tecnologie di IA. Una delle metodologie chiave per implementare un design etico è il VSD (Value Sensitive Design - Design Sensibile ai Valori). Il VSD è un approccio teorico e metodologico che mira a incorporare valori umani fondamentali nel processo di design tecnologico. Questo metodo riconosce che le tecnologie non sono mai neutre e che le decisioni progettuali possono influenzare in modo significativo gli utenti e la società. Pertanto, il VSD richiede un'analisi approfondita dei valori umani coinvolti e un'attenzione costante a come questi valori possano essere integrati nelle caratteristiche tecniche del sistema.
Il VSD si afferma come una metodologia che facilita l'integrazione dei valori umani nelle creazioni tecnologiche. Originato nei primi anni '90, il VSD ha progressivamente consolidato la sua posizione come metodologia favorita per gli approcci *by-design*. Può essere riassunto nei seguenti principi:

- Privacy: in un'epoca in cui le violazioni dei dati e la sorveglianza non autorizzata sono diffuse, l'importanza della privacy è di fondamentale importante. Il VSD dà priorità alla protezione delle informazioni personali, garantendo che la tecnologia rispetti i confini degli utenti e gestisca i dati con la massima cura.
- Autonomia: ogni individuo ha il diritto di esercitare il controllo sulla propria vita, prendendo decisioni in linea con le proprie aspirazioni. Il VSD garantisce che la tecnologia rispetti e amplifichi questa autonomia, piuttosto che ridurla. Che si tratti di un algoritmo di raccomandazione o di un assistente digitale, la tecnologia dovrebbe potenziare gli utenti, non dettare le loro azioni.
- Fiducia: la fiducia è la base di qualsiasi interazione sociale. Affinché gli utenti accolgano e si affidino alla tecnologia, devono potersi fidare di essa. Il VSD enfatizza la creazione di sistemi che siano trasparenti, affidabili e coerenti, promuovendo un senso di fiducia tra gli utenti.
- Equità: mentre gli algoritmi assumono un ruolo sempre più centrale nel processo decisionale, dall'approvazione dei prestiti al reclutamento di personale, garantire l'equità diventa fondamentale. Il VSD promulga un trattamento equo, garantendo che la tecnologia sia libera da pregiudizi, siano essi espliciti o impliciti.
- Accessibilità: la tecnologia dovrebbe essere uno strumento di empowerment, accessibile a tutti, indipendentemente dalle loro capacità fisiche o cognitive. Il VSD promuove la creazione di tecnologie inclusive che si rivolgano a una base di utenti diversificata, garantendo che nessuno sia lasciato indietro.
- Benessere: oltre alla funzionalità, la tecnologia dovrebbe contribuire positivamente al benessere complessivo degli utenti. Il VSD promuove la progettazione di sistemi che migliorino la qualità della vita, riducendo lo stress, favorendo connessioni o fornendo esperienze significative.

Una delle caratteristiche distintive del VSD è la sua natura proattiva: invece di attendere che i problemi emergano per poi correre ai ripari, il VSD incoraggia la previsione. Esso spinge progettisti e sviluppatori ad anticipare le sfide potenziali, etiche o meno, e ad affrontarle direttamente durante la fase di progettazione.

1.4 L'etica dell'Intelligenza Artificiale nelle organizzazioni internazionali

Attingendo spesso alle metodologie accademiche nate negli scorsi 20 anni, per supportare lo sviluppo etico dell'IA diverse organizzazioni internazionali hanno sviluppato framework metodologici e best practice. Tra questi, il toolkit dell'OECD (OCSE - Organizzazione per la Cooperazione e lo Sviluppo Economico) per l'etica dell'IA è uno degli strumenti più completi e influenti.

Il toolkit dell'OECD fornisce una serie di linee guida e risorse per aiutare i governi e le organizzazioni a implementare pratiche etiche nell'uso dell'IA. Questo strumento si basa su cinque principi chiave: l'inclusione e la diversità, la trasparenza e la spiegabilità, la robustezza e la sicurezza, la responsabilità e la privacy, e la gestione dei dati. Ogni principio è accompagnato da raccomandazioni pratiche e casi di studio che illustrano come questi principi possono essere applicati nella pratica.

Un altro framework significativo è rappresentato dalle linee guida dell'IEEE (Institute of Electrical and Electronics Engineers) per il design etico dell'IA. Le linee guida dell'IEEE si concentrano su aspetti come il rispetto dell'autonomia umana, la non discriminazione, la responsabilità, la trasparenza e la privacy. Questi principi sono stati sviluppati attraverso un ampio processo di consultazione con esperti di vari settori, garantendo una prospettiva globale e interdisciplinare.

Oltre al toolkit dell'OECD e alle linee guida dell'IEEE, esistono numerosi altri framework internazionali che forniscono orientamenti per l'*AI ethics*. Ad esempio, il gruppo di esperti ad alto livello sull'IA della Commissione europea ha sviluppato l'ALTAI (Assessment List for Trustworthy Artificial Intelligence), uno strumento pratico che aiuta le organizzazioni a valutare e migliorare l'affidabilità dei loro sistemi di IA.

Questi toolkit e framework internazionali sono fondamentali per creare un terreno comune di comprensione e implementazione delle pratiche etiche nell'IA. Essi forniscono risorse pratiche che possono essere adattate a diversi contesti e settori, promuovendo un approccio coordinato e coerente all'etica dell'IA a livello globale.

Infine, e prima di entrare nel merito della discussione sull'AI Act, esistono alcune norme tecniche in corso di sviluppo nel CEN-CENELEC sull'etica dell'AI. Una delle più rilevanti, attualmente in fase di sviluppo, è intitolata "Competence requirements for AI ethicists professionals". Questa norma mira a stabilire un quadro di riferimento chiaro per le competenze richieste ai professionisti dell'etica dell'IA, detti "AI ethicists". La creazione di questa norma apre la strada alla definizione di schemi di certificazione per questi professionisti, garantendo che essi possiedano le conoscenze, le abilità e le responsabilità necessarie per affrontare le sfide etiche poste dall'IA. La norma europea in fase di sviluppo si allinea con i criteri dell'EQF (European Qualifications Framework) e include:

- Conoscenze: comprensione dei principi fondamentali dell'etica dell'IA, delle normative vigenti, delle teorie etiche applicabili e delle implicazioni sociali ed economiche dell'IA.
- Abilità: capacità di applicare i principi etici nello sviluppo e nell'implementazione dell'IA, valutare eticamente i sistemi di IA, gestire dilemmi etici e coinvolgere le parti interessate.
- Responsabilità e Autonomia: capacità di lavorare in contesti multidisciplinari, comunicare le valutazioni etiche a diversi stakeholder, e contribuire alla definizione di politiche e standard etici per l'IA in vari contesti organizzativi.

L'*AI ethics* rappresenta una dimensione di grande importanza per lo sviluppo di sistemi di IA che siano affidabili, trasparenti e rispettosi dei diritti umani. Uno dei pilastri fondamentali per raggiungere questi obiettivi è la normazione tecnica sulle metodologie di design etico. Standard come il VSD o norme tecniche come quelle sviluppate dall'ISO e dal CEN-CENELEC forniscono un quadro strutturato per integrare considerazioni etiche in ogni fase del ciclo di vita dei sistemi di IA. Questo approccio garantisce che le tecnologie non solo funzionino bene, ma rispettino e promuovano i valori umani fondamentali.

All'interno di questo sforzo internazionale, l'AI Act rappresenta lo sforzo normativo più significativo a livello internazionale. Il prossimo capitolo discuterà i suoi contenuti, con particolare attenzione al rapporto tra Regolamento e norme tecniche in fase di sviluppo.

2. Il Regolamento europeo sull'Intelligenza Artificiale - AI Act

2.1 Perché un Regolamento europeo sull'Intelligenza Artificiale?

Il Regolamento europeo sull'Intelligenza Artificiale (di seguito, anche "Regolamento" o "AI Act") ha l'obiettivo di stabilire un quadro normativo armonizzato e proporzionato per regolare l'impiego dell'IA all'interno dell'Unione europea. La *ratio* sottostante al Regolamento EU è fondata sull'idea che l'IA debba essere sviluppata e utilizzata in modo sicuro, etico e rispettoso dei diritti fondamentali e dei valori europei. Di conseguenza, l'atto normativo si propone di classificare i sistemi di IA in base al grado di rischio che essi comportano per la sicurezza e i diritti delle persone, oltre a istituire una serie di requisiti e obblighi per i fornitori, i distributori, gli importatori e gli utilizzatori di tali sistemi. L'AI Act rappresenta un risultato rilevante nell'evoluzione normativa dell'Unione europea, delineando un quadro giuridico volto a bilanciare la protezione dei diritti fondamentali e delle libertà individuali con la promozione dell'innovazione nel settore dell'IA.

Il percorso legislativo inerente all'AI Act è culminato con la pubblicazione del dispositivo nell'OJEU (Official Journal of European Union) il 12 luglio 2024. Tale processo ha visto il coinvolgimento attivo delle diverse Istituzioni europee: la Commissione europea ha presentato la proposta originaria nell'aprile 2021, seguita dall'adozione delle posizioni da parte del Consiglio dell'Unione europea e del Parlamento europeo negli anni successivi.

Di seguito, si riporta la cronologia del processo legislativo:

- Aprile 2021: la Commissione europea ha presentato la proposta per il Regolamento;
- Dicembre 2022: il Consiglio dell'Unione europea ha adottato la sua posizione sul testo del Regolamento;
- Dicembre 2023: è stato raggiunto un accordo politico sul testo finale dell'AI Act tra Commissione, Consiglio e Parlamento;
- Dicembre 2023: l'AI Act è stato approvato dal Parlamento europeo;
- Febbraio 2024: il Regolamento è stato approvato anche dal Consiglio dell'Unione europea;
- 13 marzo 2024: il Parlamento europeo ha definitivamente approvato il Regolamento;
- Primavera 2024: è stato ottenuto il via libera finale dal Consiglio dell'Unione europea;
- 12 luglio 2024: pubblicazione nell'OJEU.

Una volta pubblicato, la tempistica entro cui il Regolamento dispiegherà i suoi effetti è scalata nel tempo con periodi di 6, 12, 24 e 36 mesi. Ciò riflette la complessità e la portata delle norme delineate, consentendo agli attori interessati di adeguarsi gradualmente ai nuovi obblighi e requisiti.

Come già avvenuto con il GDPR, anche per l'AI Act si prevede che l'adeguamento preventivo e spontaneo rivestirà un ruolo fondamentale. La necessità di conformarsi alle disposizioni del nuovo Regolamento richiederà un attento esame delle politiche e delle pratiche aziendali, nonché un costante monitoraggio delle evoluzioni normative e tecnologiche nel campo dell'IA.
Con la pubblicazione del Regolamento, le Istituzioni europee si pongono numerosi obbiettivi; tra i principali vi sono:

- creare un mercato unico per l'IA favorendo la libera circolazione e il riconoscimento dei sistemi di IA che rispettano le norme dell'UE, promuovendo così l'integrazione e la coerenza nel mercato europeo dell'IA;
- aumentare la fiducia nei sistemi di IA, garantendo che i sistemi di IA siano affidabili, trasparenti e sviluppati secondo un principio di responsabilità, rispettando i principi etici e i diritti fondamentali delle persone;
- prevenire e mitigare i rischi rappresentati dall'IA vietando o limitando l'uso di sistemi di IA che rappresentano un rischio inaccettabile per la sicurezza, la salute, la dignità o l'autonomia delle persone. In tal senso il Regolamento si propone di proteggere gli individui e i valori democratici da possibili minacce derivanti dall'impiego non regolamentato dell'IA.
- sostenere l'innovazione e l'eccellenza nell'IA fornendo incentivi, finanziamenti e linee guida per lo sviluppo e il dispiegamento di sistemi di IA sicuri ed etici. L'obiettivo è promuovere la crescita e l'avanzamento tecnologico nell'ambito dell'IA, favorendo la cooperazione e il coordinamento tra gli Stati membri, le istituzioni e le parti interessate al fine di massimizzare i benefici dell'IA per l'intera società europea.

2.2 L'approccio al rischio del Regolamento sull'Intelligenza Artificiale e una panoramica sul contenuto regolatorio

Il Regolamento definisce all'articolo 3(1) cosa si intenda per sistema di IA. Riportiamo la definizione per completezza:
"Sistema di IA": 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."[3]
Il Regolamento ruota interamente attorno a un approccio *risk based*, prevedendo una classificazione di quei sistemi di IA considerati ad Alto Rischio.
La definizione di rischio nell'AI Act, all'articolo 3(2) enuncia:
"Rischio: la combinazione della probabilità del verificarsi di un danno e la gravità del danno stesso."[4]
Secondo l'AI Act, sono sistemi ad Alto Rischio quelli che:

- rientrano tra i sistemi disciplinati dalla normativa di armonizzazione dell'UE a cui fa riferimento l'allegato I;
- sono soggetti a una valutazione di conformità prima della loro immissione sul mercato o della loro messa in servizio, ai sensi della normativa UE.

---

[3] *"AI system' means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environment."*
[4] *Risk means the combination of the probability of an occurrence of harm and the severity of that harm.*

Inoltre, all'allegato III viene fatto un elenco di sistemi ad Alto Rischio, suddividendolo in specifici settori di incidenza, quali, a titolo meramente esemplificativo, quello della gestione delle migrazioni e controllo delle frontiere, dell'occupazione, gestione dei lavoratori e accesso al lavoro autonomo, dell'istruzione e formazione professionale, dell'accesso ai servizi pubblici e privati essenziali.
L'elenco di cui all'allegato III, in ragione dell'esponenziale crescita del livello tecnologico, è soggetto a continua revisione, al fine di evitare un disallineamento tra la normazione e la realtà tecnologica del momento.
Tale aggiornamento sarà necessario ogniqualvolta vengano introdotti nuovi sistemi di IA che, incidendo sui settori di cui all'allegato III, potrebbero costituire rischi di danno o impatto negativo:

- alla salute e alla sicurezza;
- ai diritti fondamentali in misura pari o superiore a quelli già previsti dal medesimo allegato.

Per semplicità, di seguito, si riportano le categorie di sistemi individuate all'interno del Regolamento. In particolare, i sistemi di IA possono essere suddivisi in quattro categorie di rischio: Inaccettabile, Alto, Limitato e Minimo.

- I sistemi di IA che presentano un rischio Inaccettabile, si caratterizzano per porsi in netto contrasto con quelli che sono i valori e i principi fondamentali dell'UE, come la democrazia delle Istituzioni, il rispetto della dignità umana e dello stato di diritto. Proprio perché in contrasto con i fondamenti dell'Unione, questa tipologia di sistemi è vietata o (nel caso della sorveglianza biometrica in tempo reale per motivi di sicurezza) soggetta a severe restrizioni. Tra i sistemi vietati vi sono ad esempio i sistemi di IA che manipolano il comportamento umano coartando o influenzando la volontà degli utenti o che consentono lo scoring sociale da parte delle Autorità pubbliche (per l'elenco dei sistemi vietati cfr. l'art. 5 del Regolamento).
- I sistemi di IA ad Alto Rischio (Capo III del Regolamento) si caratterizzano per poter avere un impatto sistemico, ossia significativo sui diritti fondamentali o sulla sicurezza degli utenti e dell'ambiente circostanze. Tali sistemi sono sottoposti a obblighi e requisiti (di cui si dirà) prima di poter essere immessi sul mercato o utilizzati.
  Tra le varie categorie di sistemi considerati ad Alto Rischio rientrano i sistemi di IA utilizzati per la selezione e il reclutamento del personale, l'erogazione di servizi sociali essenziali come la sanità, per l'ammissione all'istruzione, la sorveglianza biometrica a distanza (non in tempo reale), lo svolgimento di funzioni giudiziarie e di polizia, o per la gestione della sicurezza critica delle infrastrutture. È interessante notare, poi, come, nell'ambito dei modelli generali di IA, ex art. 51 comma 2 del Regolamento, sono considerati ad Alto Rischio tutti i sistemi di IA generativa dotati di potenza di calcolo superiore a 10^25 Flop.
- I sistemi di IA che presentano un rischio Limitato, pur avendo la capacità di influenzare i diritti o le volontà degli utenti, lo fanno in misura minore rispetto ai sistemi ad Alto Rischio. Tali sistemi sono sottoposti a requisiti di trasparenza, aventi la funzione di consentire agli utenti di essere consapevoli del fatto che interagiscono con un sistema di IA e di comprenderne le caratteristiche e le limitazioni.
  Ciò permetterà agli utenti di esercitare il loro diritto di scegliere se affidarsi o meno al sistema e di capire le possibili conseguenze delle loro scelte. Gli sviluppatori e gli utilizzatori di tali sistemi dovranno anche garantire che le informazioni fornite siano chiare, comprensibili e accessibili.

A titolo di mero esempio, rientrano in questa categoria i sistemi di IA utilizzati per generare o manipolare contenuti audiovisivi (si pensi ad esempio i *deepfake*), o per fornire suggerimenti personalizzati (come le *chatbot*). In sostanza, per l'utente vi è il diritto di essere cosciente di stare interagendo con un sistema di IA e non con un essere umano o, per tornare all'esempio dei deepfake, che quell'immagine è stata generata o artefatta tramite un sistema di IA.

- I sistemi di IA che presentano un rischio Minimo o nullo si caratterizzano, invece, per il fatto di non aver alcun impatto diretto sui diritti fondamentali o sulla sicurezza delle persone, oltre che per offrire ampi margini di scelta e controllo agli utenti. Laddove un sistema di IA venga categorizzato tra quelli rappresentanti un rischio minimo o nullo lo stesso è da considerarsi libero da qualsiasi obbligo normativo specifico, al fine precipuo di rafforzare e incoraggiare l'innovazione e la sperimentazione. Ciò non toglie che anche i sistemi che presentano un rischio minimo o nullo devono comunque rispettare le Leggi e i Regolamenti generali applicabili all'IA, come quelli relativi alla protezione dei dati personali, alla concorrenza, alla responsabilità civile o ai diritti dei consumatori. Rientrano in questa categoria i sistemi di IA utilizzati per scopi ludici (come i videogame) o per scopi puramente estetici (si pensi ai filtri fotografici).

Nel valutare il rischio, la Commissione tiene conto di diversi criteri, tra i quali, ad esempio, la finalità prevista dal sistema di IA, la misura in cui è usato o verrà impiegato, la portata dell'eventuale impatto negativo (danno) e il numero delle persone che potrebbero essere coinvolte, o l'eventuale previsione di legge di contromisure efficaci volte anche a prevenire o ridurre sostanzialmente i rischi.

2.3 Le regole di classificazione dei sistemi ad Alto Rischio

Per quanto attiene alla classificazione dei sistemi ad Alto Rischio, ex art. 6, par. 1 del Regolamento, è sancito che, a prescindere dal fatto che il sistema sia immesso sul mercato o messo in servizio in modo indipendente rispetto ai prodotti indicati alle lettere a) e b) di seguito, un sistema di IA è considerato ad Alto Rischio se sono soddisfatte entrambe le condizioni seguenti:

a) il sistema di IA è destinato a essere utilizzato come componente di sicurezza di un prodotto, o il sistema di IA è esso stesso un prodotto, disciplinato dalla normativa di armonizzazione dell'Unione elencata nell'allegato I;
b) il prodotto, il cui componente di sicurezza a norma della lettera a) è il sistema di IA, o il sistema di IA stesso in quanto prodotto, deve considerarsi soggetto a una valutazione della conformità da parte di terzi ai fini dell'immissione sul mercato o della messa in servizio di tale prodotto ai sensi della normativa di armonizzazione dell'Unione elencata nell'allegato I.

Oltre ai sistemi di IA ad Alto Rischio di cui all'art. 6, par. 1, ex par. 2 della medesima disposizione sono considerati ad Alto Rischio anche i sistemi di IA di cui all'allegato III (rubricato: sistemi di IA ad Alto Rischio di cui all'art. 6, par. 2). Va poi sottolineato come il par. 3 disponga una deroga al par. 2, specificando che un sistema di IA di cui all'allegato III non è da considerarsi ad Alto Rischio se non presenta un rischio significativo di danno per la salute, la sicurezza o i diritti fondamentali delle persone fisiche, anche nel senso di non influenzare materialmente il risultato del processo decisionale.

Tale deroga troverà applicazione quando è soddisfatta almeno una qualsiasi delle condizioni seguenti:

a) il sistema di IA è destinato a eseguire un compito procedurale limitato;
b) il sistema di IA è destinato a migliorare il risultato di un'attività umana precedentemente completata;
c) il sistema di IA è destinato a rilevare schemi decisionali o deviazioni da schemi decisionali precedenti e non è finalizzato a sostituire o influenzare la valutazione umana precedentemente completata senza un'adeguata revisione umana.

Fatto salvo quanto sancito in apertura del par. 2 (nel primo comma), un sistema di IA di cui all'allegato III deve sempre essere considerato ad Alto Rischio qualora esso effettui profilazione di persone fisiche. Sempre nell'art. 6 del Regolamento, è stabilito uno specifico obbligo in capo al fornitore laddove quest'ultimo non ritiene che un sistema di cui all'allegato III sia ad Alto Rischio. In particolare, in tal caso, il fornitore ne documenta la valutazione prima che tale sistema sia immesso sul mercato oppure messo in servizio. Tale fornitore è soggetto all'obbligo di registrazione di cui all'art. 49, par. 2[5]. Inoltre, su richiesta delle Autorità nazionali competenti, il fornitore è tenuto mettere a disposizione la documentazione relativa alla valutazione.
Per quanto attiene, poi, alle prospettive future, va tenuto a mente come la Commissione, entro 18 mesi dalla data di entrata in vigore del Regolamento e dopo aver consultato il Comitato europeo per l'IA, fornirà orientamenti dettagliati atti a specificare come attuare nella pratica le disposizioni ex art. 6 del Regolamento, appena descritte in conformità con quanto stabilito all'art. 96 in merito agli orientamenti della Commissione sull'attuazione del Regolamento. Questi orientamenti dovranno includere un elenco completo di esempi pratici di casi d'uso di sistemi di IA, sia ad Alto Rischio che non.
Non solo, ma rispetto a quanto disposto sin qui, ex art. 97 la Commissione ha il potere di adottare atti delegati per modificare il par. 3 dell'art. 6. Tali modifiche possono consistere nell'aggiunta di nuove condizioni o nella loro modifica, sulla base di prove concrete e affidabili che dimostrino l'esistenza di sistemi di IA inclusi nell'ambito di applicazione dell'allegato III, ma che non presentano un rischio significativo per la salute, la sicurezza o i diritti fondamentali delle persone fisiche.

2.4 I requisiti dei sistemi ad Alto Rischio

Proprio in ragione della connaturata rischiosità dei sistemi di IA ad Alto Rischio, è stata prevista un'ulteriore condizione ai fini della loro immissione sul mercato o messa in servizio. Invero, l'art. 9 del Regolamento prevede che debba essere strutturato, attuato, documentato e mantenuto un sistema di gestione del rischio, attraverso un costante e sistematico aggiornamento, che provveda a identificare e analizzare i rischi noti e prevedibili, a stimare e valutare i rischi potenzialmente emergenti da un uso conforme alla finalità e quelli connessi a un uso improprio; a valutare altri eventuali rischi derivanti dai dati analizzati successivamente all'immissione sul mercato e a adottare tutte le misure di gestione dei rischi in base allo stato dell'arte indicate nel Regolamento.

---

[5] *Prima di immettere sul mercato o mettere in servizio un sistema di IA che il fornitore ha concluso non essere ad Alto Rischio a norma dell'art. 6, par. 3, il fornitore o, ove applicabile, il rappresentante autorizzato si registra o registra tale sistema nella banca dati dell'UE di cui all'art. 71.*

Inoltre, sempre per quanto attiene ai sistemi di gestione del rischio, il Legislatore ha individuato dei riferimenti qualitativi ai quali devono rispondere dette misure per essere considerate appropriate.
In tal senso, è previsto che occorre:

- garantire, per quanto possibile, l'eliminazione o la riduzione delle minacce attraverso un'adeguata progettazione o fabbricazione;
- attuare adeguate misure di mitigazione e controllo per tutti quei pericoli che non sia possibile eliminare;
- fornire informazioni adeguate agli utenti, in particolar modo sulla stima e valutazione dei rischi che potenzialmente potrebbero emergere dall'uso sia conforme che improprio (purché prevedibile) del sistema di IA.

Sempre nell'ottica di eliminare o attenuare i rischi legati all'uso dei sistemi considerati ad Alto Rischio, vengono tenute in considerazione anche le conoscenze tecniche, l'istruzione, l'esperienza e la formazione che ci si può aspettare dall'utente e, in egual modo, si deve parametrare il funzionamento di detti sistemi considerando l'ambiente in cui gli stessi sono destinati ad essere utilizzati.
Ancora, in ragione dell'approccio *risk based*, il Regolamento obbliga i fornitori a sviluppare un sistema di governance dei dati, prevedendo che le informazioni impiegate per l'allenamento di modelli, l'apprendimento, la convalida e la prova, rispondano a specifici requisiti di qualità di cui si dirà diffusamente nella seconda parte del presente documento, ma a cui si accenna anche qui. In particolare, i set di dati di addestramento, convalida e prova devono essere pertinenti, rappresentativi, esenti da errori e completi. Proprio al fine di mitigare i rischi dei sistemi di IA grazie a un utilizzo attento di dati caratterizzati da un elevato livello qualitativo, nel Regolamento, in deroga all'art. 9 del GDPR, è data la possibilità da parte dei fornitori di sistemi di IA di trattare dati idonei a rivelare l'etnia, l'origine razziale, le convinzioni religiose e/o filosofiche, le opinioni politiche, i dati genetici e biometrici, l'appartenenza sindacale, lo stato di salute e l'orientamento sessuale della persona. Ma unicamente nella misura in cui il trattamento di detti dati sia essenziale a garantire un adeguato monitoraggio, il rilevamento e la correzione delle distorsioni in relazione ai sistemi di IA ad Alto Rischio. D'altro canto, pur se l'utilizzo è concesso, lo stesso potrà avvenire unicamente laddove siano garantite tutele adeguate per i diritti e le libertà fondamentali delle persone fisiche e siano attuate tutte quelle limitazioni tecniche necessarie a garantire la sicurezza e la riservatezza della vita privata.
Nelle sezioni successive del Regolamento è introdotto un set di obblighi per i fornitori i quali, fin dallo sviluppo e dalla progettazione dei sistemi d'IA, dovranno approntare il tracciamento e la registrazione automatica dei *logs*, provvedendo anche alla trasparenza delle informazioni da dare all'utente (in formato digitale o meno) e ciò con l'obiettivo di permettere a quest'ultimo di interpretare l'*output* del sistema e di utilizzarlo in modo consono.
In ragione di ciò, dovrà essere garantito dai fornitori un livello adeguato di robustezza, accuratezza e cybersicurezza per tutto il ciclo di vita del sistema, la capacità di resilienza dello stesso rispetto a errori, guasti o incongruenze che possono verificarsi durante il funzionamento del sistema o nell'ambiente in cui viene utilizzato e, in ultimo, in ordine ai tentativi di terzi non autorizzati di modificarne l'uso o le prestazioni sfruttandone le fragilità.
Visto quanto sopra, la Commissione europea ha enucleato una serie di obblighi in capo ai due protagonisti (ma non solo) del Regolamento, ovvero il fornitore e l'utilizzatore dei sistemi d'IA. I fornitori, oltre a doversi conformare con quanto è stato accennato, hanno l'obbligo di predisporre un sistema di gestione della qualità conforme a determinati requisiti meglio specificati nella Parte II del presente documento.

In sintesi, tali obblighi consistono nel:

- redigere la documentazione tecnica del loro sistema di IA ad Alto Rischio;
- conservare i *logs* generati autonomamente dal sistema;
- effettuare una procedura di valutazione sulla conformità prima di immetterlo sul mercato o in servizio e
- una volta terminata detta valutazione, in caso di esito positivo, apporre la marcatura CE, ai sensi dell'art. 48, sui sistemi rendendosi disponibili, in una logica di cooperazione, a dimostrare la conformità del sistema su richiesta di un'Autorità nazionale competente;
- in caso di non conformità, informare le Autorità nazionali competenti degli Stati membri, in cui hanno messo a disposizione o in servizio il sistema di IA, dell'esito negativo della procedura di valutazione di conformità e delle eventuali misure correttive adottate appositamente. Difatti, hanno l'obbligo di informare le Autorità qualora identifichino un incidente grave o un qualsiasi malfunzionamento.

Ad aggiungersi agli obblighi dei fornitori vi sono quelli degli importatori e dei distributori i quali, prima di immettere sul mercato tali sistemi, dovranno garantire che il loro fornitore abbia effettivamente adempiuto alla procedura di conformità, alla redazione della documentazione tecnica necessaria e alla apposizione della marcatura CE. Sempre in ragione della volontà di mitigare il rischio e di responsabilizzare i protagonisti del mercato, sia gli importatori che i distributori verranno poi considerati fornitori se, nel momento dell'immissione sul mercato o in servizio del sistema ad Alto Rischio, hanno apportato modifiche alla finalità del sistema o comunque una modifica sostanziale allo stesso; e il sistema risulti a nome loro o con il loro marchio. Come si accennava poc'anzi, tra i soggetti coinvolti nella normativa del Regolamento vi sono gli utenti stessi. Invero, il Regolamento prevede obblighi specifici per quest'ultimi, quali, a titolo di esempio, quello di un uso conforme alle istruzioni di utilizzo ed eventualmente anche di informare il fornitore o il distributore di possibili rischi che ritenessero plausibilmente derivanti da questo e, in tal caso, sospenderne l'utilizzo. L'uso conforme da parte degli utenti prevede che i dati di *input* inseriti siano sempre pertinenti con le finalità del sistema di IA.
Al fine di comprendere la natura del Regolamento e gli obblighi derivanti da esso, in considerazione delle criticità e della scarsa cooperazione che caratterizza (troppo spesso) lo sviluppo della normativa nazionale italiana a cui siamo avvezzi, è doveroso sottolineare come il Regolamento, viceversa, sia stato sviluppato e pensato secondo una logica di forte cooperazione tra i soggetti coinvolti. E questo in un'ottica di resilienza che pone al centro la costante collaborazione che trova il proprio mezzo in un flusso di informazioni senza soluzione di continuità volto alla salvaguardia della sicurezza, della salute, delle libertà e dei diritti fondamentali. Flusso e cooperazione che dovrebbero rappresentare gli strumenti essenziali volti a consentire un intervento pronto e rapido sia in caso di malfunzionamenti rischiosi che in caso di eventuali distorsioni applicative da parte degli utenti.
Poiché l'intero Regolamento ruota attorno alla cooperazione, lo stesso è redatto tenendo conto proprio di quelle logiche di trasparenza di cui s'è detto in precedenza e che emergono anche dallo stesso art. 13 dell'AI Act. In questa disposizione è sancito che sia fornito all'utente un catalogo di informazioni necessarie sia alla tutela dei diritti sia alla segnalazione di potenziali malfunzionamenti. Tra queste figurano i dati di contatto del fornitore, i rischi connessi a un utilizzo conforme o improprio e le finalità del sistema.
Il rispetto dei requisiti previsti per i sistemi ad Alto rischio dovrà essere dimostrato, come si accennava in precedenza, mediante una procedura di valutazione della conformità che, come sarà chiarito nei successivi paragrafi, potrà prevedere il coinvolgimento di un organismo di valutazione della conformità. L'organismo, per poter operare sul mercato, dovrà ottenere la qualifica di "organismo notificato".

2.5 Le Autorità notificanti e gli organismi notificati

La procedura di notifica, già prevista in vari atti normativi europei, è enucleata nella sezione IV del capo III del Regolamento. Gli organismi incaricati della valutazione della conformità devono soddisfare requisiti di indipendenza, imparzialità e competenza per poter essere designati come organismi notificati dall'Autorità competente di ciascuno Stato membro. Successivamente, sarà compito dell'Autorità notificante comunicare alla Commissione l'elenco degli organismi notificati.
L'accreditamento è un metodo per mezzo del quale è possibile dimostrare il possesso dei suddetti requisiti, pertanto gli organismi sono incoraggiati a presentare una domanda di notifica con un certificato di accreditamento allegato. In mancanza di questo, l'organismo deve fornire documentazione adeguata a dimostrare la conformità alle disposizioni dell'AI Act.
A tal proposito, si evidenzia come l'accreditamento sia già considerato un elemento qualificante, e in alcuni casi necessario, per il conseguimento della notifica in numerosi ambiti regolati dalla normativa di armonizzazione di cui all'allegato I.
A titolo esemplificativo, al 1° gennaio 2024, tutti gli organismi notificati operanti in conformità alle norme riportante nella seguente tabella (tabella 1) risultavano accreditati.

Tabella 1. Accreditamenti ai fini della notifica alla Commissione europea per alcune normative di armonizzazione

Normativa di armonizzazione dell'UE di cui all'allegato I - sezione A	Organismi accreditati e notificati
Direttiva 2014/33/UE - Ascensori	75
Direttiva 2014/34/UE - Prodotti usati in atmosfere esplosive	11
Direttiva 2006/42/CE - Macchine	33
Direttiva 2009/48/CE - Giocattoli	7
Direttiva 2014/53/UE - Apparecchiature radio	6
Direttiva 2014/68/UE - Recipienti a pressione	40
Regolamento UE 2016/424 - Apparecchi a fune	1
Regolamento UE 2016/425 - Dispositivi di protezione individuale	13
Regolamento UE 2016/426 - Apparecchi che bruciano carburanti gassosi	6

La scelta del Legislatore europeo di utilizzare al meccanismo della notifica dimostra come vi sia la volontà di ricondurre i controlli sui sistemi di IA a quelli sui prodotti già normati a livello UE. Si analizzano di seguito le disposizioni attinenti al tema.
Ex art. 28 del Regolamento vengono istituite le **Autorità di notifica**. In particolare, è previsto che ciascuno Stato membro designi o istituisca almeno un'Autorità di notifica responsabile della predisposizione e dell'esecuzione delle procedure necessarie per la valutazione, la designazione e la notifica degli organismi di valutazione della conformità e per il loro monitoraggio. Tali procedure sono sviluppate nell'ambito della collaborazione tra le Autorità di notifica di tutti gli Stati membri.

Nella medesima disposizione è sancito, inoltre, che gli Stati membri possono optare perché la valutazione e il monitoraggio siano condotti da un organismo nazionale di accreditamento in conformità al Regolamento CE 765/2008. L'Italia, ha soventemente utilizzato tale opzione, delegando ad Accredia - l'Ente Unico nazionale di accreditamento, tramite apposite convenzioni, il compito di una valutazione preliminare sugli organismi di valutazione della conformità operanti nell'ambito del nuovo quadro legislativo e il successivo monitoraggio.
Le Autorità di notifica devono essere strutturate e gestite in modo tale da evitare conflitti di interesse con gli organismi di valutazione della conformità, garantendo l'obiettività e l'imparzialità delle loro attività. Le decisioni relative alla notifica di un organismo di valutazione della conformità devono essere prese da individui competenti, diversi da coloro che hanno condotto la valutazione stessa.
Le Autorità di notifica non devono fornire né svolgere alcuna delle attività svolte dagli organismi di valutazione della conformità, né offrire servizi di consulenza su base commerciale o concorrenziale.
Inoltre, devono preservare la riservatezza delle informazioni ottenute in conformità all'art. 78 (rubricato: riservatezza, a cui si rinvia integralmente la disposizione).
Infine, sempre a riguardo delle Autorità di notifica, è stabilito che quest'ultime devono avere un numero adeguato di dipendenti competenti per garantire l'esecuzione efficace dei loro compiti. Tali dipendenti devono possedere, se necessario, competenze nei settori delle tecnologie dell'informazione, dell'IA e del diritto, compresa la tutela dei diritti fondamentali.
L'art. 29 del Regolamento è, invece, dedicato alla **domanda di notifica** presentata dagli organismi di valutazione della conformità. Sulla base di quanto disposto, gli organismi di valutazione della conformità sono tenuti a presentare una richiesta di notifica all'Autorità di notifica ex art. 28 dello Stato membro in cui sono stabilite. La richiesta deve:

- includere una descrizione delle attività di valutazione della conformità;
- contenere dei moduli di valutazione della conformità e dei tipi di sistemi di IA per cui l'organismo dichiara di essere competente;
- essere corredata di un certificato di accreditamento, se disponibile, emesso da un organismo nazionale di accreditamento che attesta la conformità dell'organismo ai requisiti dell'art. 31 del Regolamento, nel quale sono stabiliti requisiti relativi agli organismi notificati (lo si analizzerà nel prosieguo).

Per gli organismi notificati designati in base ad altre normative armonizzate dell'Unione europea, tutti i documenti e i certificati relativi a tali designazioni possono essere utilizzati per supportare la loro procedura di designazione ai sensi del Regolamento. Gli organismi notificati devono aggiornare la documentazione fornita ogniqualvolta si verifichino cambiamenti significativi, consentendo all'Autorità responsabile degli organismi notificati di monitorare e verificare il continuo rispetto dei requisiti dell'art. 31 del Regolamento. Ex art. 31 del Regolamento, è previsto che dev'essere istituito un **organismo notificato** a norma del diritto nazionale di uno Stato membro e ha personalità giuridica. Inoltre, la medesima disposizione disciplina i singoli requisiti di cui gli organismi notificati devono essere in possesso. Invero, secondo quanto disposto dal par. 1 dell'art. 30 del Regolamento, le Autorità di notifica possono notificare solo gli organismi di valutazione della conformità che siano conformi alle prescrizioni di cui all'art. 31. Rispetto ai singoli requisiti la disposizione sancisce che:

- gli organismi notificati soddisfano i requisiti organizzativi, di gestione della qualità e relativi alle risorse e ai processi necessari all'assolvimento dei loro compiti nonché i requisiti idonei di cybersicurezza;

- la struttura organizzativa, l'assegnazione delle responsabilità, le linee di riporto e il funzionamento degli organismi notificati garantiscono la fiducia nelle loro prestazioni e nei risultati delle attività di valutazione della conformità che essi effettuano;
- gli organismi notificati sono indipendenti dal fornitore di un sistema di IA ad Alto Rischio in relazione al quale svolgono attività di valutazione della conformità. Gli organismi notificati sono inoltre indipendenti da qualsiasi altro operatore avente un interesse economico nei sistemi di IA ad Alto Rischio oggetto della valutazione, nonché da eventuali concorrenti del fornitore. Ciò non preclude l'uso dei sistemi di IA ad Alto Rischio oggetto della valutazione che sono necessari per il funzionamento dell'organismo di valutazione della conformità o l'uso di tali sistemi di IA ad Alto Rischio per scopi privati;
- l'organismo di valutazione della conformità, i suoi alti dirigenti e il personale incaricato di svolgere i compiti di valutazione della conformità non intervengono direttamente nella progettazione, nello sviluppo, nella commercializzazione o nell'utilizzo di sistemi di IA ad Alto Rischio, né rappresentano i soggetti impegnati in tali attività. Essi non intraprendono alcuna attività che possa essere in conflitto con la loro indipendenza di giudizio o la loro integrità per quanto riguarda le attività di valutazione della conformità per le quali sono notificati. Ciò vale in particolare per i servizi di consulenza;
- gli organismi notificati sono organizzati e gestiti in modo da salvaguardare l'indipendenza, l'obiettività e l'imparzialità delle loro attività. Gli organismi notificati documentano e attuano una struttura e delle procedure per salvaguardare l'imparzialità e per promuovere e applicare i principi di imparzialità in tutta l'organizzazione, tra il personale e nelle attività di valutazione;
- gli organismi notificati dispongono di procedure documentate per garantire che il loro personale, i loro comitati, le affiliate, i subappaltatori e qualsiasi altra organizzazione associata o il personale di organismi esterni mantengano, conformemente all'art. 78 (rubricato: riservatezza), la riservatezza delle informazioni di cui vengono in possesso nello svolgimento delle attività di valutazione della conformità, salvo quando la normativa ne prescriva la divulgazione. Il personale degli organismi notificati è tenuto a osservare il segreto professionale riguardo a tutte le informazioni ottenute nello svolgimento dei propri compiti a norma del Regolamento, tranne che nei confronti delle Autorità di notifica dello Stato membro in cui svolge le proprie attività;
- gli organismi notificati dispongono di procedure per svolgere le attività che tengono debitamente conto delle dimensioni di un fornitore, del settore in cui opera, della sua struttura e del grado di complessità del sistema di IA interessato;
- gli organismi notificati sottoscrivono un'adeguata assicurazione di responsabilità per le loro attività di valutazione della conformità, a meno che lo Stato membro in cui sono stabiliti non si assuma tale responsabilità a norma del diritto nazionale o non sia esso stesso direttamente responsabile della valutazione della conformità;
- gli organismi notificati sono in grado di eseguire tutti i compiti assegnati loro in forza del Regolamento con il più elevato grado di integrità professionale e di competenza richiesta nel settore specifico, indipendentemente dal fatto che tali compiti siano eseguiti dagli organismi notificati stessi o per loro conto e sotto la loro responsabilità;

---

[6] Ex art. 38 del Regolamento, la Commissione garantisce che, per quanto riguarda i sistemi di IA ad Alto Rischio, siano istituiti e funzionino correttamente, in forma di gruppo settoriale di organismi notificati, un coordinamento e una cooperazione adeguati tra gli organismi notificati che partecipano alle procedure di valutazione della conformità a norma del Regolamento. Ciascuna Autorità di notifica garantisce che gli organismi da essa notificati partecipino al lavoro di un gruppo anzidetto, direttamente o mediante rappresentanti designati. La Commissione provvede allo scambio di conoscenze e migliori pratiche tra le Autorità di notifica.

- gli organismi notificati dispongono di sufficienti competenze interne per poter valutare efficacemente i compiti svolti da parti esterne per loro conto. Gli organismi notificati dispongono permanentemente di sufficiente personale amministrativo, tecnico, giuridico e scientifico dotato di esperienza e conoscenze relative ai tipi di sistemi di IA, ai dati, al calcolo dei dati pertinenti, nonché ai requisiti di cui alla sezione 2 del capo III del Regolamento;
- gli organismi notificati partecipano alle attività di coordinamento di cui all'articolo 38[6]. Inoltre, gli organismi notificati devono partecipare direttamente o essere rappresentati in seno alle organizzazioni europee di normazione o garantiscono di essere informati e di mantenersi aggiornati in merito alle norme pertinenti.

All'art. 30 dell'AI Act è stabilita la **procedura di notifica**. Nello specifico è stabilito che:

- le Autorità di notifica notificano alla Commissione e agli altri Stati membri, utilizzando lo strumento elettronico di notifica elaborato e gestito dalla Commissione (NANDO), ogni organismo di valutazione della conformità che risponda ai requisiti ex art. 31;
- la notifica anzidetta include tutti i dettagli riguardanti le attività di valutazione della conformità, il modulo o i moduli di valutazione della conformità, i tipi di sistemi di IA interessati, nonché la relativa attestazione di competenza. Qualora una notifica non sia basata su un certificato di accreditamento di cui all'art. 29, l'Autorità di notifica fornisce alla Commissione e agli altri Stati membri le prove documentali che attestino la competenza dell'organismo di valutazione della conformità nonché le misure predisposte per fare in modo che tale organismo sia monitorato periodicamente e continui a soddisfare i requisiti di cui all'art. 31 del Regolamento;
- l'organismo di valutazione della conformità interessato può eseguire le attività di un organismo notificato solo se non sono sollevate obiezioni da parte della Commissione o degli altri Stati membri entro due settimane dalla notifica da parte di un'Autorità di notifica, qualora essa includa un certificato di accreditamento di cui all'art. 29 del Regolamento, o entro due mesi dalla notifica da parte dell'Autorità di notifica, qualora essa includa le prove documentali di cui alla medesima disposizione. Se sono sollevate obiezioni, la Commissione avvia senza ritardo consultazioni con gli Stati membri pertinenti e l'organismo di valutazione della conformità. Tenutone debito conto, la Commissione decide se l'autorizzazione è giustificata. La Commissione trasmette la propria decisione allo Stato membro interessato e all'organismo di valutazione della conformità pertinente.

All'art. 32 del Regolamento viene disciplinata la **presunzione di conformità** ai requisiti relativi agli organismi notificati. In particolare, la disposizione stabilisce che se un organismo di valutazione della conformità dimostra la sua conformità ai criteri stabiliti nelle norme armonizzate pertinenti o in parti di esse pubblicate nell' OJEU allora viene considerato conforme ai requisiti dell'art. 31 del Regolamento (di cui s'è detto poc'anzi), purché dette norme armonizzate coprano tali requisiti. In altre parole, se un organismo segue le regole stabilite nell'Unione europea e dimostra la sua adesione a tali regole, si presume che sia conforme ai requisiti pertinenti.
L'art. 33 del Regolamento delinea le **responsabilità degli organismi notificati** nei confronti dei subappaltatori e delle affiliate. Sul punto viene stabilito che nel caso in cui un organismo notificato affidi specifiche attività di valutazione della conformità a un subappaltatore o a un'affiliata, deve garantire che essi soddisfino i requisiti stabiliti nell'art. 31 del Regolamento e deve informare l'Autorità di notifica di tale subappalto o affiliazione. Inoltre, è previsto che gli organismi notificati siano pienamente responsabili delle attività svolte dai subappaltatori o dalle affiliate.

Tali attività possono essere subappaltate o svolte da un'affiliata solo previo consenso del fornitore. Gli organismi notificati, inoltre, sono tenuti a rendere pubblico un elenco delle loro affiliate. Infine, è previsto che i documenti pertinenti riguardanti la valutazione delle qualifiche del subappaltatore o dell'affiliata e il lavoro svolto devono essere tenuti a disposizione dell'Autorità di notifica per un periodo di cinque anni dalla data di conclusione del contratto di subappalto.

L'art. 34 del Regolamento delinea gli **obblighi operativi degli organismi notificati** nell'ambito della verifica della conformità dei sistemi di IA ad Alto Rischio, seguendo le procedure di valutazione della conformità specificate nell'art. 43. Gli organismi notificati devono evitare di imporre oneri inutili ai fornitori durante l'esecuzione delle proprie attività e devono considerare adeguatamente le dimensioni del fornitore, il settore in cui opera, la sua struttura e il grado di complessità del sistema di IA ad alto rischio coinvolto. Questo viene fatto al fine di ridurre al minimo gli oneri amministrativi e i costi di conformità, specialmente per le microimprese e le piccole imprese, come raccomandato dalla raccomandazione 2003/361/CE. Tuttavia, gli organismi notificati devono rispettare il grado di rigore e il livello di tutela necessari per garantire la conformità del sistema di IA ad Alto Rischio ai requisiti del Regolamento. Inoltre, gli organismi notificati devono mettere a disposizione e trasmettere su richiesta tutta la documentazione pertinente, compresa quella fornita dal fornitore, all'Autorità di notifica di cui all'art. 28. Questo al fine di consentire all'Autorità di svolgere le proprie attività di valutazione, designazione, notifica e monitoraggio, nonché per agevolare la valutazione prevista nella presente sezione.

L'art. 35 del Regolamento fornisce le disposizioni in merito ai **numeri di identificazione ed elenchi di organismi notificati**. In particolare, è previsto che la Commissione assegni un numero di identificazione unico a ciascun organismo notificato, anche nel caso in cui un organismo sia notificato in conformità con più atti dell'UE. Inoltre, la Commissione è tenuta a pubblicare l'elenco degli organismi notificati in conformità con l'AI Act, comprensivo dei loro numeri di identificazione e delle attività per le quali sono stati notificati, assicurandosi che tale elenco sia costantemente aggiornato.

L'art. 36 del Regolamento disciplina i casi di **modifiche delle notifiche**, di cui si tratta a grandi linee. In sintesi, in caso di modifiche delle notifiche l'Autorità di notifica informa la Commissione e gli altri Stati membri di qualsiasi modifica rilevante alla notifica di un organismo notificato tramite lo strumento elettronico di notifica specificato nell'art. 30 del Regolamento. Le procedure di estensione della portata della notifica sono regolate dagli artt. 29 e 30 di cui si è già detto in precedenza. Per le modifiche diverse dalle estensioni della portata della notifica, si applicano le procedure indicate nei parr. da 3 a 9 dell'art. 36 (a cui si rimanda integralmente). Se un organismo notificato decide di interrompere le attività di valutazione della conformità, deve informare l'Autorità di notifica e i fornitori interessati il prima possibile e almeno un anno prima della cessazione pianificata. I certificati dell'organismo notificato possono rimanere validi per un periodo di nove mesi dopo la cessazione delle attività, a condizione che un altro organismo notificato abbia confermato per iscritto che assumerà la responsabilità per i sistemi di IA ad Alto Rischio coperti da tale certificato. Tale organismo notificato sostitutivo deve completare una valutazione completa dei sistemi di IA ad Alto Rischio coinvolti entro la fine del periodo di nove mesi, prima di rilasciare nuovi certificati per gli stessi sistemi. Se un organismo notificato cessa le sue attività, l'Autorità di notifica revoca la designazione.

Se un'Autorità di notifica ha ragioni sufficienti per ritenere che un organismo notificato non soddisfi più i requisiti dell'art. 31 del Regolamento o non adempia ai suoi obblighi, deve indagare sulla questione senza indugi e con la massima diligenza.

In tal caso, l'organismo notificato interessato deve essere informato delle obiezioni sollevate e deve essere dato loro l'opportunità di esprimere il proprio punto di vista. Se l'Autorità di notifica conclude che l'organismo notificato non soddisfa più i requisiti dell'art. 31 o non adempie ai suoi obblighi, limita, sospende o revoca la designazione, a seconda della gravità del mancato rispetto o dell'inadempimento.
La Commissione e gli altri Stati membri devono essere informati immediatamente di tali azioni. In caso di limitazione, sospensione o revoca della designazione, l'Autorità di notifica deve adottare misure adeguate per garantire che i fascicoli dell'organismo notificato interessato siano conservati e resi disponibili alle autorità di notifica negli altri Stati membri e alle autorità di vigilanza del mercato, su richiesta. Deve valutare l'impatto sui certificati rilasciati dall'organismo notificato, presentare una relazione sulle proprie constatazioni alla Commissione e agli altri Stati membri entro tre mesi dalla comunicazione delle modifiche della designazione, imporre all'organismo notificato di sospendere o ritirare i certificati rilasciati indebitamente entro un periodo ragionevole stabilito dall'Autorità, informare la Commissione e gli Stati membri dei certificati per i quali ha richiesto la sospensione o il ritiro e fornire alle Autorità nazionali competenti dello Stato membro in cui ha sede il fornitore tutte le informazioni pertinenti sui certificati per i quali ha richiesto la sospensione o il ritiro.
L'art. 37 del Regolamento stabilisce la **procedura per contestare la competenza degli organismi notificati** e per valutare il mantenimento della conformità di tali organismi ai requisiti delineati nell'art. 31 e alle relative responsabilità. In estrema sintesi, in caso di dubbi sulla competenza o sulla continuità della conformità di un organismo notificato, la Commissione è autorizzata a condurre indagini per accertare la situazione.
È importante notare che l'Autorità di notifica è tenuta a fornire alla Commissione tutte le informazioni pertinenti riguardanti la notifica o il mantenimento della competenza dell'organismo notificato interessato. Inoltre, la Commissione è incaricata di garantire che tutte le informazioni sensibili ottenute durante le indagini siano trattate in modo riservato, conformemente alle disposizioni dell'articolo 78 (inerente alla riservatezza). Se la Commissione determina che un organismo notificato non soddisfa o non soddisfa più i requisiti per la notifica, è tenuta a informare lo Stato membro che ha effettuato la notifica e a richiedere l'adozione delle misure correttive necessarie, compresa la sospensione o il ritiro della notifica.
Nel caso in cui lo Stato membro non adotti tali misure correttive, la Commissione ha il potere, mediante atto di esecuzione, di sospendere, limitare o ritirare la designazione dell'organismo notificato. È importante sottolineare che tale atto di esecuzione è soggetto alla procedura d'esame stabilita dall'art. 98, par. 2.
L'art. 38 delinea le disposizioni per il **coordinamento degli organismi notificati** nell'ambito dei sistemi di IA ad Alto Rischio. Con tale disposizione il Regolamento mira a garantire un coordinamento efficace tra gli organismi notificati e le Autorità di notifica nell'ambito della valutazione della conformità dei sistemi di IA ad Alto Rischio, promuovendo la cooperazione e lo scambio di conoscenze per garantire un'applicazione uniforme e coerente del regolamento su scala europea. Nello specifico:

- la Commissione è responsabile di garantire l'istituzione e il corretto funzionamento di un gruppo settoriale di organismi notificati, al fine di coordinare e cooperare adeguatamente tra di loro durante le procedure di valutazione della conformità stabilite dal regolamento. Il gruppo settoriale mira a facilitare una cooperazione efficace tra gli organismi notificati impegnati nella valutazione della conformità dei sistemi di IA ad Alto Rischio;

- ogni Autorità di notifica deve assicurare che gli organismi notificati da essa designati partecipino al lavoro del gruppo di coordinamento menzionato nel paragrafo precedente, direttamente o attraverso rappresentanti designati. Questa disposizione promuove un coinvolgimento attivo degli organismi notificati nel processo di coordinamento, garantendo una collaborazione stretta e efficace;
- la Commissione è incaricata di facilitare lo scambio di conoscenze e migliori pratiche tra le Autorità di notifica. Tale misura mira a promuovere la condivisione di esperienze e conoscenze tra le autorità competenti, consentendo loro di apprendere gli uni dagli altri e di migliorare le loro pratiche nell'applicazione del regolamento.

L'art. 39 del Regolamento delinea le disposizioni riguardanti gli **organismi di valutazione della conformità provenienti da Paesi terzi** e la loro possibile autorizzazione a operare come organismi notificati ai sensi del regolamento dell'Unione europea. Questa disposizione stabilisce che tali organismi, se istituiti in conformità con la legislazione di un Paese terzo con cui l'UE ha concluso un accordo, possono essere autorizzati a svolgere le attività degli organismi notificati, a patto che soddisfino i requisiti stabiliti nell'art. 31 o garantiscano un livello di conformità equivalente. La disposizione mira a consentire la partecipazione di organismi terzi nel processo di valutazione della conformità dei prodotti e dei servizi nel mercato dell'UE, purché vengano mantenuti standard di conformità paragonabili a quelli richiesti agli organismi notificati dell'UE, assicurando così un'omogeneità nella valutazione della conformità all'interno del mercato interno europeo.

2.6 Le procedure di valutazione della conformità

Gli organismi notificati saranno coinvolti, ai sensi dell'art. 43 del Regolamento, nella valutazione della conformità di alcuni sistemi di IA ad Alto Rischio. In particolare, i sistemi di IA ad Alto Rischio disciplinati dalla normativa di armonizzazione dell'UE elencata nell'allegato I, sezione A, il fornitore è tenuto a seguire la pertinente procedura di valutazione della conformità prevista da tali atti giuridici. I requisiti di cui alla sezione 2 del capo III (requisiti per i sistemi ad Alto Rischio) si applicano ai sistemi di IA ad Alto Rischio e fanno parte di tale valutazione. Si applicano anche i punti 4.3, 4.4, 4.5 e il punto 4.6, quinto comma, dell'allegato VII (di cui si dirà nel seguito).
Qualora un atto giuridico elencato nell'allegato I, sezione A, consenta al fabbricante del prodotto di sottrarsi a una valutazione della conformità da parte di terzi, purché abbia applicato tutte le norme armonizzate che contemplano tutti i requisiti pertinenti, quest'ultimo può avvalersi di tale facoltà solo se ha applicato anche le norme armonizzate o, ove applicabili, le specifiche comuni di cui all'art. 41, che contempla tutti i requisiti di cui alla sezione 2 del capo III inerente i requisiti dei sistemi di IA ad Alto Rischio. Il ricorso agli organismi notificati è inoltre previsto per i sistemi di IA afferenti alla biometrica di cui all'allegato III. Nello specifico, il fornitore sarà costretto ad applicare la procedura di valutazione della conformità di cui all'allegato VII (che prevede il coinvolgimento di un organismo notificato) nei seguenti casi:

a) non esistono le norme armonizzate e non sono disponibili le specifiche comuni;
b) il fornitore non ha applicato la norma armonizzata o ne ha applicato solo una parte;
c) esistono le specifiche comuni di cui alla lettera ma il fornitore non le ha applicate;
d) una o più norme armonizzate sono state pubblicate con una limitazione e soltanto sulla parte della norma che è oggetto di limitazione.

Al contrario, per i sistemi di IA ad Alto Rischio di cui all'allegato III, punti da 2 a 8, i fornitori sono tenuti a seguire la procedura di valutazione della conformità basata sul controllo interno di cui all'allegato VI, che non prevede il coinvolgimento di un organismo notificato.
Concluse queste premesse generali, effettuate al fine di permettere una maggior rapidità rispetto alla comprensione dei meccanismi di valutazione, si analizzano nello specifico le differenti forme di valutazioni e gli elementi che le costituiscono, differenziandole. Per quanto attiene alle valutazioni di conformità interne, queste sono condotte dal fornitore (o dal produttore, deployer o importatore, a seconda dei casi) del sistema di IA ad Alto Rischio. Nell'ambito delle valutazioni di conformità interne di cui all'allegato VI, i soggetti obbligati devono garantire la conformità dei loro sistemi tramite una procedura specifica di valutazione della conformità (allegato VI). In particolare, il fornitore e chi è sottoposto a tale obbligo:

- verifica la conformità del sistema di gestione della qualità istituito ai requisiti di cui all'art. 17;
- esamina le informazioni contenute nella documentazione tecnica al fine di valutare la conformità del sistema di IA ai pertinenti requisiti di cui al capo III, sezione 2;
- verifica, inoltre, che il processo di progettazione e sviluppo del sistema di IA e il monitoraggio successivo alla sua immissione sul mercato di cui all'art. 72 siano coerenti con la documentazione tecnica.

Tutti gli sviluppatori di sistemi di IA hanno in ogni caso la possibilità di richiedere una valutazione di conformità di terze parti se la ritengono necessaria, indipendentemente dal livello di rischio del sistema. Ciò significa che, anche nei casi in cui la valutazione interna sia obbligata, gli sviluppatori di sistemi di IA potranno chiedere (in aggiunta) una valutazione di un ente terzo secondo la procedura di cui all'allegato VII.
La procedura di valutazione della conformità che prevede il ricorso ad un organismo notificato, come anticipato, è disciplinata all'allegato VII. Tale allegato sancisce che il sistema di gestione della qualità approvato per la progettazione, lo sviluppo e la prova dei sistemi di IA a norma dell'art. 17 deve essere esaminato conformemente al punto 3 e deve essere soggetto alla vigilanza di cui al punto 5. Aggiunge, inoltre, che la documentazione tecnica del sistema di IA deve essere esaminata conformemente al punto 4.

Punto 3 (sistema di gestione della qualità)

Il sistema di gestione della qualità deve essere valutato dall'organismo notificato, che deve stabilire se soddisfa i requisiti di cui all'art. 17. La decisione deve essere notificata al fornitore o al suo rappresentante autorizzato. Tale notifica deve indicare le conclusioni della valutazione del sistema di gestione della qualità e la decisione di valutazione motivata. Il sistema di gestione della qualità approvato deve continuare a essere attuato e mantenuto dal fornitore in modo da rimanere adeguato ed efficiente.
Il fornitore deve portare all'attenzione dell'organismo notificato qualsiasi modifica prevista del sistema di gestione della qualità approvato o dell'elenco dei sistemi di IA cui si applica tale sistema. Le modifiche proposte devono essere esaminate dall'organismo notificato, che deve decidere se il sistema di gestione della qualità modificato continua a soddisfare i requisiti di cui all'art. 17 (nel Regolamento si fa riferimento più genericamente al punto 3.2 dell'allegato VII) o se è necessaria una nuova valutazione. L'organismo notificato deve notificare al fornitore la propria decisione. Tale notifica deve indicare le conclusioni dell'esame e la decisione di valutazione motivata.

Sempre nel punto 3 è previsto che la domanda presentata dal fornitore debba comprendere:
a) il nome e l'indirizzo del fornitore e, nel caso in cui la domanda sia presentata da un rappresentante autorizzato, anche il nome e l'indirizzo di quest'ultimo;
b) l'elenco dei sistemi di IA cui si applica lo stesso sistema di gestione della qualità;
c) la documentazione tecnica di ciascuno dei sistemi di IA cui si applica lo stesso sistema di gestione della qualità;
d) la documentazione relativa al sistema di gestione della qualità che deve contemplare tutti gli aspetti elencati all'articolo 17;
e) una descrizione delle procedure vigenti per garantire che il sistema di gestione della qualità rimanga adeguato ed efficace;
f) una dichiarazione scritta attestante che la stessa domanda non è stata presentata a nessun altro organismo notificato.

Punto 4 (controllo della documentazione tecnica)

Oltre alla domanda di cui al punto 3, il fornitore deve presentare una domanda a un organismo notificato di propria scelta per la valutazione della documentazione tecnica relativa al sistema di IA che il fornitore intende immettere sul mercato o mettere in servizio a cui si applica il sistema di gestione della qualità di cui al punto 3. La domanda deve comprendere:
a) il nome e l'indirizzo del fornitore;
b) una dichiarazione scritta attestante che la stessa domanda non è stata presentata a nessun altro organismo notificato;
c) la documentazione tecnica di cui all'allegato IV (a cui si rimanda integralmente).

La documentazione tecnica dev'essere esaminata dall'organismo notificato. Se del caso e nei limiti di quanto necessario per lo svolgimento dei suoi compiti, all'organismo notificato deve essere concesso pieno accesso ai set di dati di addestramento, convalida e prova utilizzati, anche, ove opportuno e fatte salve le garanzie di sicurezza, attraverso API o altri mezzi e strumenti tecnici pertinenti che consentano l'accesso remoto. Nell'esaminare la documentazione tecnica, l'organismo notificato può chiedere al fornitore di presentare elementi probatori supplementari o di eseguire ulteriori prove per consentire una corretta valutazione della conformità del sistema di IA ai requisiti di cui al capo III, sezione 2 (requisiti per i sistemi d'IA ad Alto Rischio). Qualora non sia soddisfatto delle prove effettuate dal fornitore, l'organismo notificato stesso deve effettuare prove adeguate, a seconda dei casi.
Ove necessario per valutare la conformità del sistema di IA ad Alto Rischio ai requisiti di cui al capo III, sezione 2, dopo che tutti gli altri mezzi ragionevoli per verificare la conformità sono stati esauriti e si sono rivelati insufficienti, e su richiesta motivata, anche all'organismo notificato deve essere concesso l'accesso ai modelli di addestramento e addestrati del sistema di IA, compresi i relativi parametri. Tale accesso è soggetto al vigente diritto dell'Unione in materia di protezione della proprietà intellettuale e dei segreti commerciali. La decisione dell'organismo notificato deve essere notificata al fornitore o al suo rappresentante autorizzato. Tale notifica deve indicare le conclusioni della valutazione della documentazione tecnica e la decisione di valutazione motivata.
Se il sistema di IA è conforme ai requisiti di cui al capo III, sezione 2, l'organismo notificato deve rilasciare un certificato di valutazione della documentazione tecnica dell'Unione. Tale certificato deve indicare il nome e l'indirizzo del fornitore, le conclusioni dell'esame, le eventuali condizioni

di validità e i dati necessari per identificare il sistema di IA. Il certificato e i suoi allegati devono contenere tutte le informazioni pertinenti per consentire la valutazione della conformità del sistema di IA e il controllo del sistema di IA durante l'uso, ove applicabile.
Se il sistema di IA non è conforme ai requisiti di cui al capo III, sezione 2, l'organismo notificato deve rifiutare il rilascio di un certificato di valutazione della documentazione tecnica dell'UE e deve informare in merito il richiedente, motivando dettagliatamente il suo rifiuto. Se il sistema di IA non soddisfa il requisito relativo ai dati utilizzati per l'addestramento, sarà necessario addestrare nuovamente il sistema di IA prima di presentare domanda per una nuova valutazione della conformità. In tal caso, la decisione di valutazione motivata dell'organismo notificato che rifiuta il rilascio del certificato di valutazione della documentazione tecnica dell'Unione contiene considerazioni specifiche sui dati di qualità utilizzati per addestrare il sistema di IA, in particolare sui motivi della non conformità.
Qualsiasi modifica del sistema di IA che potrebbe incidere sulla conformità ai requisiti o sulla finalità prevista dello stesso deve essere valutata dall'organismo notificato che ha rilasciato il certificato di valutazione della documentazione tecnica dell'UE. Il fornitore deve informare tale organismo notificato quando intende introdurre una delle modifiche di cui sopra o quando viene altrimenti a conoscenza del verificarsi di tali modifiche. Le modifiche previste devono essere valutate dall'organismo notificato, che deve decidere se esse rendono necessaria una nuova valutazione della conformità a norma dell'art. 43, par. 4, o se possono essere gestite tramite un supplemento del certificato di valutazione della documentazione tecnica dell'UE. In quest'ultimo caso, l'organismo notificato deve valutare le modifiche, notificare al fornitore la propria decisione e, in caso di approvazione delle modifiche, rilasciare a quest'ultimo un supplemento del certificato di valutazione della documentazione tecnica dell'UE. Per i sistemi di IA ad Alto Rischio che proseguono il loro apprendimento dopo essere stati immessi sul mercato o messi in servizio, le modifiche apportate al sistema di IA ad Alto Rischio e alle sue prestazioni che sono state predeterminate dal fornitore al momento della valutazione iniziale della conformità e fanno parte delle informazioni contenute nella documentazione tecnica di cui all'allegato IV, punto 2, lettera f), non costituiscono una modifica sostanziale.

Punto 5 (vigilanza del sistema di gestione della qualità approvato)
La finalità della vigilanza a cura dell'organismo notificato di cui al punto 3 è garantire che il fornitore sia conforme ai termini e alle condizioni del sistema di gestione della qualità approvato. Ai fini della valutazione, il fornitore deve consentire all'organismo notificato di accedere ai locali in cui hanno luogo la progettazione, lo sviluppo e le prove dei sistemi di IA. Il fornitore deve inoltre condividere con l'organismo notificato tutte le informazioni necessarie. L'organismo notificato deve eseguire audit periodici per assicurarsi che il fornitore mantenga e applichi il sistema di gestione della qualità e deve trasmettere al fornitore una relazione di audit. Nel contesto di tali audit, l'organismo notificato può effettuare prove supplementari dei sistemi di IA per i quali è stato rilasciato un certificato di valutazione della documentazione tecnica dell'UE. Sempre rispetto alla procedura di valutazione della conformità di cui all'allegato VII, ex art. 43 del Regolamento il fornitore può scegliere uno qualsiasi degli organismi notificati. Tuttavia, si osserva un'eccezione. Invero, se il sistema di IA ad Alto Rischio è destinato a essere messo in servizio dalle Autorità competenti in materia di contrasto, di immigrazione o di asilo, nonché da Istituzioni, organi o organismi dell'UE, l'Autorità di vigilanza del mercato di cui all'art. 74, parr. 8 o 9, a seconda dei casi, agisce in qualità di organismo notificato.

In definitiva, come descritto nell'allegato VII, una volta terminata con successo la valutazione di conformità, l'organismo notificato è tenuto a rilasciare un Certificato di Documentazione Tecnica UE (art. 44). Il certificato è valido per il periodo indicato che non può superare i cinque anni per i sistemi di IA disciplinati dall'allegato I e i quattro anni per i sistemi di IA disciplinati dall'allegato III. Su domanda del fornitore, la validità di un certificato può essere prorogata per ulteriori periodi, ciascuno non superiore a cinque anni per i sistemi di IA disciplinati dall'allegato I e a quattro anni per i sistemi di IA disciplinati dall'Allegato III, sulla base di una nuova valutazione secondo le procedure di valutazione della conformità applicabili. Ogni supplemento del certificato rimane valido purché sia valido il certificato cui si riferisce. Tuttavia, se l'organismo notificato ritiene che il sistema di IA in questione non soddisfi più i requisiti per i sistemi ad Alto Rischio, è tenuto a sospendere o ritirare il certificato a meno che il fornitore non adotti provvedimenti correttivi per garantire nuovamente la conformità entro un appropriato periodo temporale stabilito dall'ente notificato. Successivamente, il fornitore deve redigere la Dichiarazione di Conformità UE. Qualora l'organismo notificato scopra che il sistema ad Alto Rischio non soddisfa i suoi requisiti di conformità, deve offrire al fornitore una dettagliata spiegazione della non conformità. Il fornitore deve quindi adottare le misure correttive pertinenti per garantire la conformità; in caso contrario, deve ritirare il sistema dal mercato. Il Regolamento UE prevede un meccanismo di ricorso in quest'ultimo caso, in base all'art. 45, che conferisce al fornitore il potere di impugnare la determinazione dell'ente notificato.

Nel caso in cui la procedura di valutazione della conformità, sia essa basata sull'allegato VI o VII, abbia esito positivo, potrà essere apposta la marcatura CE, che testimonierà il rispetto della normativa di settore e dall'AI Act. La marcatura dovrà essere apposta in maniera visibile, leggibile e indelebile. Qualora ciò sia impossibile o difficilmente realizzabile a causa della natura del sistema di IA, il marchio potrà essere posizionato sull'imballaggio o sui documenti di accompagnamento. La marcatura CE, ove sia previsto il ricorso di un soggetto terzo nell'ambito della procedura di valutazione della conformità, dovrà inoltre essere seguita dal numero di identificazione dell'organismo notificato. Inoltre, vale la pena evidenziare come già il testo di compromesso definitivo del Parlamento europeo ha introdotto una nuova disposizione per garantire la protezione degli interessi delle Piccole e Medie Imprese (PMI), prevedendo che le tariffe per la conduzione di valutazioni di terze parti siano proporzionate alla dimensione e alla quota di mercato di una PMI.

3. La funzione della normativa tecnica nell'AI Act

3.1 La funzione delle norme nella regolamentazione UE in relazione al Regolamento sull'Intelligenza Artificiale

Per promuovere l'armonizzazione tecnica nel campo dell'IA, rendendola affidabile e uniforme nel territorio dell'UE, è stato ritenuto necessario affiancare agli obblighi regolamentari anche norme tecniche europee, al fine di coprire le principali aree tecniche coinvolte dall'AI Act. Tra i campi coinvolti, sono da includersi: i requisiti per la progettazione e lo sviluppo dei sistemi di IA definiti ad Alto Rischio, il sistema di gestione della qualità dei fornitori di IA, la valutazione di conformità e l'*auditing* dei sistemi di IA.
Per definizione una norma è una specifica tecnica, adottata da un organismo di normazione riconosciuto, non obbligatoria (Regolamento UE 1025/2012). La normazione, in generale, può avere a oggetto specifiche tecniche di prodotto o di servizio, ossia la redazione di documenti che prescrivono requisiti tecnici che un determinato prodotto, processo, servizio o sistema deve soddisfare. Pertanto le norme europee costituiscono un insieme di specifiche tecniche e/o criteri stabiliti da un organismo di normazione europeo. In generale, la normazione europea è organizzata da e per gli stakeholder, rappresentati nazionalmente tramite il Comitato Europeo di Normazione (CEN), il Comitato Europeo di Normazione Elettrotecnica (CENELEC), e la partecipazione diretta degli stakeholder attraverso l'Istituto Europeo di Normazione delle Telecomunicazioni (ETSI). L'intero processo di armonizzazione ruota attorno ai principi riconosciuti dall'Organizzazione Mondiale del Commercio (OMC) nel campo della normazione, ovvero i principi di coerenza, trasparenza, apertura, consenso, applicazione volontaria, indipendenza da interessi particolari ed efficienza. In accordo con detti principi, è importante che tutte le parti interessate rilevanti, inclusi gli Enti pubblici e le PMI, siano adeguatamente coinvolte a livello nazionale ed europeo.
L'obiettivo primario della normazione tecnica nel campo dell'IA consiste nella definizione di specifiche tecniche e/o qualitative a cui i prodotti basati sull'IA, già sul mercato o di futura introduzione, i loro processi produttivi o i servizi forniti, possono (su base volontaria) conformarsi, con il proposito di garantire, in modo armonizzato, la sicurezza e l'affidabilità dei sistemi d'IA, nonché la compatibilità e l'interoperabilità con altri prodotti o sistemi.
In linea con il Regolamento UE 1025/2012 (di seguito, anche "Regolamento sulla normazione europea"), le norme che verranno sviluppate nel campo dell'IA avranno un ruolo decisivo nel sostenere l'attuazione della nuova normativa di compliance. Invero, come indicato dalla Commissione europea nella *Implementing Decision* del 2023, trattasi di strumenti giuridici attraverso cui sarà garantito un elevato livello di protezione della sicurezza e dei diritti fondamentali per i cittadini europei in tutto il territorio dell'UE.

Oltre a ciò, laddove previsto, potranno supportare l'istituzione di condizioni di concorrenza eque e di condizioni paritarie per la progettazione e lo sviluppo dei sistemi di IA, in particolare per le PMI che sviluppano soluzioni basate sull'IA.
Ai sensi dell'art. 2, par. 1, lett. c del Regolamento, per "harmonised standard" (norme armonizzate) si intende: "una norma tecnica europea, adottata sulla base di una richiesta della Commissione ai fini dell'applicazione della legislazione dell'Unione sull'armonizzazione".
A sua volta, l'art. 3, n. 27 del Regolamento richiama questa definizione per utilizzarla ai suoi fini.

3.2 Rapporto tra Standardization Request e AI Act, soggetti coinvolti nell'attività di normazione e funzione del CEN, del CENELEC e dell'ETSI

La presente analisi si concentra sulla SR (Standardization Request) redatta dalla Commissione europea ai sensi del Regolamento. È importante notare che il presente documento analizza la richiesta di normazione della Commissione pubblicata a maggio 2023. Una nuova richiesta di normazione, aggiornata al contenuto del testo dell'AI Act, è attesa entro il 2024.
Dato questo presupposto, al fine di comprendere la relazione tra l'AI Act e la SR della Commissione è necessario fare una breve panoramica sui principali Regolamenti europei che fanno esplicito richiamo alla normazione; da individuarsi in particolare nel Regolamento Standard UE 1025/2012, nella Direttiva UE 1535/2015, nel Regolamento CE 765/2008 e nel Regolamento UE 1020/2019, nonché sugli attori coinvolti e sul processo di normazione.
Per quanto attiene al **Regolamento UE 1025/2012** sulla normazione europea, questo stabilisce i principi e i requisiti generali, istituendo un quadro giuridico atto al coordinamento delle attività di normazione nell'UE. Il principale obiettivo della norma consiste nel definire specifiche qualitative o tecniche volontarie, alle quali processi di produzione, prodotti o servizi attuali o futuri possono conformarsi. Tramite la **Direttiva UE 2015/1535** il Parlamento e la Commissione, invece, hanno determinato le procedure di notifica degli schemi di regolamentazione tecnica e delle norme UE. Il **Regolamento CE 765/2008** inserisce i requisiti di accreditamento e sorveglianza del mercato relativi alla commercializzazione dei prodotti, stabilendo i principi e i requisiti generali per l'introduzione e l'uso di norme armonizzate nel contesto della legislazione europea sui prodotti. Infine, per mezzo del **Regolamento UE 1020/2019**, avente ad oggetto la sorveglianza del mercato e la conformità dei prodotti, sono state definite le regole per garantire la conformità dei prodotti commercializzati sul mercato europeo, inclusi i requisiti relativi alla valutazione della conformità e al ricorso alle norme armonizzate.
Grazie a tali prescrizioni stato fornito il quadro giuridico e regolamentare entro cui operano gli attori coinvolti nel processo di normazione tecnica nell'UE. Tra le categorie di soggetti coinvolti vi sono:

- le ESO (Organizzazioni Europee di Normazione);
- gli NSB (Organismi Nazionali di Normazione);
- le organizzazioni europee degli stakeholder (portatori di interessi);
- i consulenti per le norme armonizzate.

Per quanto riguarda le ESO, vi sono tre responsabili della normazione nell'UE, il CEN, il CENELEC e l'ETSI. Attualmente CEN e CENELEC sono gli unici a essere direttamente coinvolti nella creazione delle norme europee inerenti ai sistemi di IA.

Cionondimeno, probabilmente l'ETSI si attiverà in futuro al fine di sviluppare ulteriori norme di cui, nel caso, non dovrà sottovalutarsi l'importanza. Pur trattandosi di Enti indipendenti dalle Istituzioni dell'UE, gli stessi possono essere incaricati direttamente dalla Commissione di sviluppare norme armonizzate destinate a essere applicate all'interno del territorio dell'UE. Oltre a poter essere incaricati direttamente dall'Istituzione europea, per tali organismi vige la facoltà di sviluppare norme ulteriori di propria iniziativa. A tal proposito, va evidenziato come tali norme non è detto ottengano lo status di norme "armonizzate" e che, in ogni caso, le norme redatte di propria iniziativa, non sarebbero pubblicate nell'OJEU. Detti Enti indipendenti, nello sviluppo delle norme anzidette, sono tenuti a coinvolgere i vari stakeholder, tra i quali vi rientrano i soggetti elencati ai numeri 2, 3 e 4 del precedente elenco e di cui si accennerà nei paragrafi che seguono. Gli NSB sono gli Enti responsabili della produzione di normativa tecnica in ciascuno degli Stati membri. Possono rappresentare il Governo, l'industria e la società civile. Le organizzazioni nazionali dei portatori d'interesse invece, rappresentano una varietà di interessi all'interno dell'UE, in particolare quelli delle PMI, dei sindacati, dell'ambiente e dei consumatori. Per quanto attiene ai consulenti per le norme armonizzate, trattasi di figure professionali assunte dalla Commissione europea, aventi come funzione quella di verificare che le norme, sviluppate dalle ESO, siano adatte alla pubblicazione da parte dell'Unione europea.
Conclusa la panoramica sui soggetti principali coinvolti nel processo di normazione, con l'intento di permettere una comprensione adeguata della relazione che intercorre tra la SR nell'ambito dell'IA e il Regolamento, è necessario descrivere per sommi capi il processo di normazione nell'Unione europea, il quale è costituito da numerosi passaggi che verranno descritti in breve e, per chiarezza, per punti.

Avvio

Il processo vede coinvolti differenti soggetti, tra cui vi sono la Commissione europea, le Autorità nazionali competenti, le organizzazioni industriali, gli Enti di ricerca e sviluppo. In questa prima fase, la Commissione europea, da cui trae origine la richiesta, una volta che ha rilevato l'esigenza di normazione, al fine di produrre benefici, quali l'armonizzazione normativa, l'interoperabilità o la sicurezza dei prodotti e dei servizi, identifica le aree d'intervento in cui considera necessario operare. Nel caso del Regolamento, la Commissione ha sviluppato una bozza di richiesta di normazione, SR per gli ESO, includendo dettagli sul campo di applicazione, sui tempi e sui requisiti legali che le norme dovrebbero specificare. Il tipo di specifica atteso dalle norme tecniche riguarda alcuni articoli dell'AI Act, generalmente quelli più tecnici. Ad esempio, l'art. 15 del Regolamento contiene indicazioni generali sulla necessità di garantire la robustezza, accuratezza e cybersicurezza dei sistemi di IA, ma non specifica in che modo. Il ruolo delle norme tecniche, in relazione al Regolamento, è quindi quello di specificare con procedure tecniche gli articoli del Regolamento, fornendo dei requisiti "di basso livello", come ad esempio le metriche per misurare l'accuratezza. Le norme che forniranno questi requisiti tecnici potranno essere armonizzate e, quindi, pubblicate nell'OJEU. Questi due processi implicano che l'insieme di requisiti della norma armonizzata siano considerati sufficienti a specificare alcuni specifici articoli del regolamento. Come descritto nel capitolo precedente, questo implica che le organizzazioni che fanno certificare la propria conformità a una norma armonizzata si presumono conformi ai requisiti definiti nel Regolamento. La bozza di SR è stata sviluppata in consultazione con gli ESO e le restanti parti interessate.

Emissione della Standardization Request

La Commissione ha pubblicato la SR in data 22 maggio 2023, rivolgendola ai Comitati Europei di Normazione (CEN, CENELEC o ETSI).

Redazione della bozza delle norme

Una volta approvata e ricevuta la SR, i Comitati Europei di Normazione incaricati si sono attivati. A seguito della consultazione dei portatori di interessi, dell'effettuazione delle ricerche tecniche necessarie, elaboreranno, entro i termini stabiliti, il "testo regolamentare" al fine di dar luce a norme conformi alle richieste espresse dalla Commissione. Le norme europee che saranno sviluppate seguiranno procedure trasparenti e partecipative per garantire la rappresentanza degli interessi di tutti gli attori coinvolti. In particolare, gli NSB forniscono un pool di esperti che si confronteranno nell'ambito di un comitato tecnico. L'obiettivo è di gestire in modo partecipato il processo di redazione delle norme. Il comitato tecnico include una selezione di soggetti interessati che svolgono la funzione di osservatori (observer). Nello specifico, per quanto attiene alla SR relativa all'AI Act, si tratterà di un comitato tecnico congiunto tra CEN e CENELEC. Il comitato tecnico formerà, poi, un gruppo di lavoro (working group) di esperti costituito dagli NSB e dagli osservatori anzidetti al fine di redigere una bozza del documento contenente l'architettura delle norme relative all'AI Act. Per completezza, è bene evidenziare che, nel gruppo di lavoro, non necessariamente vi sarà un esperto per Stato membro, invero, il comitato ha la facoltà di scegliere la composizione. Una volta raggiunto il consenso dei partecipanti sul contenuto delle norme verrà redatta la bozza.
Generalmente, quando ve ne è disponibilità, gli ESO si affidano alle norme internazionali prodotte dall'ISO (Organizzazione Internazionale per la Normazione) o dalla IEC (Commissione Elettrotecnica Internazionale). Questo con l'obiettivo di garantire la coerenza internazionale delle norme ed evitare l'introduzione da parte dei singoli Stati di norme capaci di ostacolare il commercio internazionale.

Adozione della norma

Completato il processo di redazione della bozza, la norma europea proposta verrà sottoposta a un voto tra i membri del CEN, CENELEC o ETSI. I consulenti valuteranno se le norme rispettano i requisiti contenuti nella SR e, pertanto, se sono adatti a essere pubblicati. I consulenti per le norme armonizzate possono essere coinvolti durante tutto il processo, ma le norme debbono superare definitivamente il vaglio di conformità al momento della votazione, in quanto, a seguito della consultazione, possono essere effettuate unicamente variazioni minimali. I consulenti vengono coinvolti solamente per verificare la validità dei requisiti contenuti nella norma e la corretta specificazione degli articoli del Regolamento rilevanti per quella norma tecnica. Comunque, nello specifico caso dell'AI Act, i consulenti non saranno interpellati, stante alcune dichiarazioni di DGCNECT rilasciate informalmente durante alcuni workshop pubblici. Una volta che il documento sarà approvato dalla maggioranza dei membri, laddove le norme vengano considerate conformi, verranno pubblicate nell'OJEU divenendo norme armonizzate.
L'armonizzazione: In generale, le norme potranno poi divenire obbligatorie laddove siano richiamati in atti amministrativi e documenti regolatori, quali Leggi nazionali, Regolamenti o Direttive dell'UE, ovvero possono essere utilizzati su base volontaria dalle parti interessate. Quando si tratta di norme obbligatorie, gli Stati membri dell'UE sono tenuti a incorporare le norme europee armonizzate nella loro legislazione nazionale per garantire l'omogeneità del mercato interno all'Unione. Per quanto attiene all'armonizzazione delle norme inerenti all'AI Act, gli NSB saranno responsabili dell'adozione delle norme europee a livello nazionale e dovranno, nel caso sia necessario, eliminare ogni norma che risulti in conflitto con le norme europee realizzate sulla base dei principi sanciti all'interno del Regolamento. In conclusione, per quanto attiene al rapporto tra AI Act e SR, si comprende come il Regolamento fornisca la cornice di riferimento entro la quale le norme tecniche relative a questa "famiglia" di tecnologie dovranno essere sviluppate.

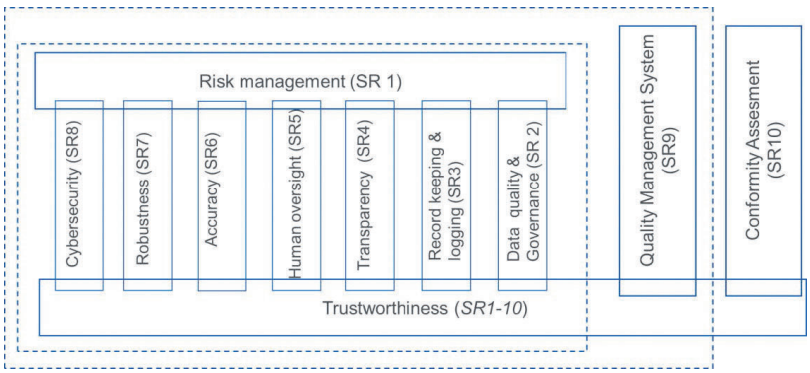
Norme, la cui redazione è promanata e vincolata alla SR emessa dalla Commissione Europea in data 22 maggio 2023, al fine di promuovere l'armonizzazione normativa e l'interoperabilità dei prodotti e servizi *Artificial Intelligence Based* sviluppati, forniti e distribuiti all'interno del mercato unico dell'Unione.

3.3 Tempistiche dei riscontri da fornire alla Commissione, periodo di validità della decisione e working programme del Comitato Tecnico Congiunto 21 di CEN e CENELEC

La SR della Commissione europea stabilisce precise tempistiche entro cui i soggetti coinvolti nel processo di normazione debbono condurre le attività previste (la SR stessa ha una validità limitata nel tempo, fino al 28 febbraio 2026, ex art. 4 della medesima). Tali scadenze, mirano a garantire un processo efficiente e tempestivo per lo sviluppo e l'implementazione delle norme tecniche in materia di IA. Il rispetto delle tempistiche non solo è fondamentale al fine di assicurare, in tempi ragionevoli, un adeguato livello di coerenza e uniformità, ma incide sulla stessa validità della Richiesta. La bozza iniziale della SR a supporto di un'IA sicura e affidabile è stata resa disponibile il 20 maggio 2022. Al terminare del mese di giugno 2022, la bozza è stata rimandata con gli emendamenti e le richieste di chiarimento. Il 5 dicembre 2022 è stata pubblicata la bozza definitiva a seguito della consultazione degli ESO e degli stakeholder. Tale pubblicazione è avvenuta tramite il sistema di notifica ai suddetti organismi. Come precedentemente indicato, il 22 maggio 2023, la Commissione ha adottato la SR C(2023)3215, accettata da CEN e CENELEC. Attualmente, il Comitato Tecnico Congiunto 21 di CEN e CENELEC sta adattando le norme partendo da quelle ISO/IEC e, laddove sia necessario, sviluppandone di nuove, come specificato nel programma di lavoro recentemente pubblicato. La scadenza per i risultati della SR C(2023)3215 è il 30 aprile 2025. Invero, a partire dal 2026, entreranno in vigore i requisiti per i sistemi di IA ad Alto Rischio, dove le norme tecniche, come vedremo nel prosieguo, svolgono un ruolo primario.
Come già indicato, il CEN e il CENELEC hanno istituito il nuovo Comitato Tecnico Congiunto CEN-CENELEC 21 Intelligenza Artificiale (CEN-CLC/JTC 21), che si muoverà sulla base delle raccomandazioni presentate nella risposta CEN-CENELEC al White Paper dell'UE sull'IA, della roadmap del Focus Group CEN-CENELEC sull'IA e della roadmap tedesca per la normazione dell'IA. Il CEN-CLC/JTC 21 è attualmente responsabile dello sviluppo e dell'adozione di norme tecniche per l'IA, dei dati correlati, oltre a dover fornire indicazioni agli altri Comitati Tecnici interessati all'IA. In particolare, il CEN-CLC/JTC 21 è stato istituito al fine di identificare e, se del caso, adottare le norme internazionali già disponibili o in fase di sviluppo da altre organizzazioni come ISO/IEC, attraverso i loro sottocomitati, come l'SC 42 dell'ISO/IEC JTC 1. L'attività del CEN-CLC/JTC 21 si sta focalizzando sulla produzione di norme tecniche volte ad affrontare le esigenze del mercato e della società dell'Unione, nonché sul sostenere la legislazione, le politiche, i principi e i valori dell'UE nell'ambito dell'IA. Il CEN e il CENELEC dopo aver accettato la SR sull' IA dalla Commissione europea e, in questo contesto, costituito il CEN-CLC/JTC 21, stanno attualmente operando al fine di sviluppare le norme che, in futuro, dovrebbero fornire ai produttori la presunzione di conformità con il Regolamento. Trattandosi di un'attività in divenire, le norme potranno subire variazioni in futuro, pertanto si considera rilevante dare una panoramica del programma di lavoro al fine di evidenziare lo stato delle norme in fase di sviluppo e le scadenze indicate dallo stesso Comitato Tecnico coinvolto nell'attività di normazione tecnica.

Di seguito, la schematizzazione grafica dell'architettura delle norme tecniche, sviluppata dal CEN-CENELEC in risposta alla SR e l'elenco analitico delle norme in fase di sviluppo.

Figura 1. Architettura del processo di normazione tecnica del CEN-CENELEC



Questa immagine riflette l'impostazione e il rapporto tra i diversi *item* della SR che verranno analizzati di seguito, e fornisce una panoramica complessiva del lavoro del CEN-CENELEC nello sviluppo di norme armonizzate. L'analisi di ogni *item* segue nelle prossime pagine.
La tabella che segue, invece, descrive l'attuale status dei diversi progetti in corso di sviluppo al CEN-CENELEC. Il loro rapporto con la SR non è possibile da valutare in maniera preliminare, in quanto soltanto l'Annex ZA di ogni progetto potrà renderlo esplicito, ovvero il documento che evidenzia il rapporto tra i requisiti della norma e il regolamento.

Tabella 2. Normativa tecnica in corso di sviluppo

Project title	Title
CEN/CLC ISO/IEC/TR 24027:2023	Information technology - Artificial intelligence (AI) - Bias in AI systems and AI aided decision making (ISO/IEC TR 24027:2021)
CEN/CLC ISO/IEC/TR 24029-1:2023	Artificial Intelligence (AI) - Assessment of the robustness of neural networks - Part 1: Overview (ISO/IEC TR 24029-1:2021)
EN ISO/IEC 22989:2023	Information technology - Artificial intelligence - Artificial intelligence concepts and terminology (ISO/IEC 22989:2022)
EN ISO/IEC 23053:2023	Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML) (ISO/IEC 23053:2022)
EN ISO/IEC 23894:2024	Information technology - Artificial intelligence - Guidance on risk management (ISO/IEC 23894:2023)
EN ISO/IEC 8183:2024	Information technology - Artificial intelligence - Data life cycle framework (ISO/IEC 8183:2023)

Project reference	Work item	Title
EN ISO/IEC 22989:2023/prA1	JT021031	Information technology — Artificial intelligence — Artificial intelligence concepts and terminology — Amendment 1
EN ISO/IEC 23053:2023/prA1	JT021032	Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML) — Amendment 1
FprCEN/CLC ISO/IEC/TS 12791	JT021013	Information technology - Artificial intelligence - Treatment of unwanted bias in classification and regression machine learning tasks (ISO/IEC DTS 12791:2023)
FprCEN/CLC/TR 18115	JT021007	Data governance and quality for AI within the European context
prCEN/CLC/TR 17894	JT021001	Artificial Intelligence Conformity Assessment
prCEN/CLC/TR XXX	JT021026	Impact assessment in the context of the EU Fundamental Rights
prCEN/CLC/TR XXX	JT021009	AI Risks - Check List for AI Risks Management
prCEN/CLC/TR XXX	JT021010	Environmentally sustainable Artificial Intelligence
prCEN/CLC/TR XXXX	JT021002	Artificial Intelligence - Overview of AI tasks and functionalities related to natural language processing
prEN ISO/IEC 12792	JT021022	Information technology - Artificial intelligence - Transparency taxonomy of AI systems (ISO/IEC DIS 12792:2024)
prEN ISO/IEC 23282	JT021012	Artificial Intelligence - Evaluation methods for accurate natural language processing systems
prEN ISO/IEC 24029-2	JT021015	Artificial intelligence (AI) - Assessment of the robustness of neural networks - Part 2: Methodology for the use of formal methods
prEN ISO/IEC 25059	JT021014	Software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - Quality model for AI systems (ISO/IEC 25059:2023)
prEN ISO/IEC 42001	JT021011	Information technology - Artificial intelligence - Management system
prEN XXX	JT021008	AI trustworthiness framework
prEN XXX	JT021025	AI tasks and evaluation methods of computer vision systems
prEN XXX	JT021024	AI Risk Management
prEN XXXXX	JT021006	AI-enhanced nudging
Preliminary	JT021030	Contributions towards ISO/IEC 27090
Preliminary	JT021029	Technical solutions to address AI specific vulnerabilities
Preliminary	JT021028	Reference architecture of knowledge engineering based on ISO/IEC 5392

Nella tabella seguente vengono analizzati gli item della *Standardization Request* mettendoli in relazione con i requisiti dell'AI Act e richiamando le norme internazionali pertinenti.

Tabella 3. Analisi di confronto della SR con i requisiti dell'AI Act

Item Standardization Request 1
Richiesta formulata nella SR Sistemi di gestione del rischio per i sistemi di IA Riferimenti dell'AI Act Il par. 2 dell'Allegato II della SR, stabilisce che la normazione europea deve determinare le specifiche dei processi di gestione del rischio per i sistemi di IA. Il par. 1 dell'art. 9 del Regolamento si sovrappone in gran parte con la descrizione dei sistemi di gestione del rischio fornita dal par. 2 della SR (*processo iterativo che continui durante l'intero ciclo di vita del sistema di IA e deve mirare a prevenire o ridurre al minimo i rischi per la salute, la sicurezza o i diritti fondamentali dei cittadini dell'Unione*), tranne per il fatto di aggiungere un ulteriore requisito. Invero, al par. 1 dell'art. 9 del Regolamento viene infatti specificato che i sistemi di gestione del rischio afferenti sistemi d' IA ad Alto Rischio debbono essere costituiti tenendo conto del fatto che, gli stessi, vanno sottoposti a revisioni e aggiornamenti sistematici e regolari. Secondo quanto disposto dal par. 2 dell'Allegato II della SR pubblicata dalla Commissione europea, rispetto a sistemi di IA che siano parte dei componenti di sicurezza di un prodotto o che costituiscono loro stessi un prodotto a sé stante, le specifiche inerenti alla gestione del rischio di detti software devono essere redatte in modo tale da poter essere integrate con sistemi di gestione del rischio già esistenti, aventi il fine di soddisfare i requisiti relativi ai sistemi di gestione del rischio elencati nell'allegato II, Sezione A, dell'AI Act. Oltre a ciò, la Commissione ha stabilito che le specifiche in merito alla gestione del rischio dei sistemi d'IA vanno redatte con modalità tali da consentirne l'utilizzo da parte degli operatori interessati e delle autorità di vigilanza del mercato. Infine, il Regolamento dedica l'intero Capo III ai sistemi d'IA ad Alto Rischio per i quali si prevede l'obbligo di stabilire, implementare, documentare e mantenere un sistema di gestione del rischio capace di soddisfare i requisiti stabiliti nella Sezione 2 del Capo III, tenendo conto degli scopi cui sono destinati e dello stato dell'arte generalmente riconosciuto in materia di IA e tecnologie correlate. Pertanto, va rilevato come i sistemi di gestione del rischio di cui all'art. 9 del Regolamento debbono, tra le cose, garantire la conformità a tali requisiti. I processi di gestione del rischio, la definizione di rischio all'interno del sistema di gestione del rischio e i requisiti di conformità previsti nel Regolamento. Ex art. 9, par. 2 del Regolamento, i processi dei sistemi di gestione del rischio inerenti all'IA ad Alto Rischio debbono comprende le seguenti fasi: (a) identificazione e analisi dei rischi noti e di quelli ragionevolmente prevedibili che il sistema di IA ad alto rischio può comportare per la salute, la sicurezza o i diritti fondamentali quando il sistema di IA ad alto rischio è utilizzato conformemente alla sua destinazione; (b) la stima e la valutazione dei rischi che possono emergere quando il sistema di IA ad alto rischio

è utilizzato conformemente alla sua destinazione e in condizioni di uso improprio ragionevolmente prevedibili[7];

(c) la valutazione di rischi ulteriori che potrebbero emergere, sulla base dell'analisi dei dati raccolti dal sistema di monitoraggio post-vendita di cui alla Sezione I del Capo IX del Regolamento;

(d) l'adozione di misure di gestione del rischio adeguate e mirate per affrontare i rischi individuati ai sensi della lettera a).

Al par. 3 dell'art. 9 del Regolamento, al fine di definire cosa rientri nei rischi indicati nella disposizione, si specifica che per rischi si intendono solo quelli che possono essere ragionevolmente ridotti o eliminati tramite lo sviluppo o la progettazione del sistema di IA (ad Alto Rischio), ovvero attraverso la predisposizione di informazioni tecniche adeguate.

Ex art. 9, par. 4 del Regolamento, le misure di gestione del rischio di cui al par. 2, lett. d), devono tenere in considerazione gli effetti e le possibili interazioni derivanti dall'applicazione combinata dei requisiti di cui alla Sezione 2 del Capo III, al fine di ridurre al minimo i rischi, raggiungendo, al contempo, un adeguato equilibrio nell'attuazione delle misure per soddisfare tali requisiti.

In aggiunta, il Regolamento stabilisce all'art. 9 par. 5 che le misure di gestione del rischio di cui all'elenco indicato al par. 2, lettera d), devono essere tali da far ritenere accettabile il rischio residuo associato a ciascun pericolo, nonché il rischio residuo complessivo dei sistemi di IA ad Alto Rischio. Inoltre, sempre al par. 5, è sancito che nell'individuare le misure di gestione del rischio appropriate deve essere garantita:

(a) l'eliminazione o la riduzione dei rischi identificati e valutati ai sensi del par. 2, per quanto tecnicamente possibile, attraverso un'adeguata progettazione e sviluppo del sistema di IA ad alto rischio;

(b) se del caso, l'attuazione di adeguate misure di mitigazione e di controllo dei rischi che non possono essere eliminati;

(c) la predisposizione delle informazioni richieste ai sensi dell'art. 13 del Regolamento e, se necessario a mitigare i rischi connessi al sistema d'IA, un'adeguata formazione dei fornitori.

Infine, con l'obiettivo di eliminare o ridurre i rischi connessi all'uso del sistema di IA ad Alto Rischio, si deve tenere in debita considerazione non solo la conoscenza tecnica, l'esperienza, l'istruzione e il *training* che ci si può aspettare da chi lo impiega, ma anche il presumibile contesto in cui il sistema è destinato a essere utilizzato.

In aggiunta a quanto indicato sin qui, il par. 6 dell'art 9 del Regolamento si sofferma sui test a cui il sistema di IA dev'essere sottoposto. In particolare, i sistemi di IA ad Alto Rischio devono essere sottoposti a test aventi lo scopo di individuare le misure di gestione del rischio più appropriate e pertinenti a garantire che gli stessi:

funzionino coerentemente allo scopo per cui sono stati progettati;

siano conformi ai requisiti stabiliti nella Sezione 2 del Capo III del Regolamento.

Le procedure di *testing* possono includere test in condizioni reali in conformità a quanto disposto all'art. 60 del Regolamento.

Per quanto attiene all'arco temporale entro cui i test dei sistemi di IA ad Alto Rischio devono essere effettuati, va evidenziato come gli stessi potranno essere eseguiti in qualsiasi momento del processo di sviluppo.

---

[7] Ex art. 3 del Regolamento, per uso improprio ragionevolmente prevedibile deve intendersi l'uso di un sistema di IA in un modo non conforme alla sua finalità prevista, ma che può derivare da un comportamento umano o da un'interazione con altri sistemi, ivi compresi altri sistemi di IA, ragionevolmente prevedibile.

Tuttavia, ex par. 8 dell'art. 9 del Regolamento, in ogni caso, debbono essere ultimati prima della loro immissione sul mercato o messa in servizio[8]. Infine, preme rilevare come detti test da una parte devono essere effettuati in base a metriche e soglie probabilistiche definite in precedenza, dall'altra, oltre a ciò devono essere adeguate agli scopi per i quali il sistema di IA ad Alto Rischio è stato sviluppato.

Per quanto attiene al rapporto tra gli obblighi regolamentari inerenti ai sistemi di gestione del rischio previsti per i sistemi di IA ad Alto Rischio ex Capo III, Sezione II del Regolamento e le norme armonizzate alla luce della SR, l'art. 40 del Regolamento (norme armonizzate e deliverables di normazione) stabilisce che: i sistemi di IA ad Alto Rischio e i Modelli di IA per Finalità Generali che risultino conformi alle norme armonizzate, o alle parti a cui lo specifico sistema d'IA è sottoposto, i cui riferimenti sono stati pubblicati nell'OJEU a norma del regolamento (UE) n. 1025/2012, si presumono conformi ai requisiti di cui alla Sezione 2 Capo III o, se applicabili, agli obblighi di cui al Capo V (avente ad oggetto gli obblighi di trasparenza per i fornitori e deployer di specifici sistemi d'IA), Sezioni II e III del Regolamento, nella misura in cui dette norme coprono tali requisiti e/o obblighi.

Norme internazionali pertinenti

- (FprCEN/CLC ISO/IEC/TS 12791 - *Work Item Number*: JT0210131), inerente al trattamento dei *bias* indesiderati nelle attività di apprendimento automatico di classificazione e regressione
  Technical Specification: ISO/IEC TS 12791;
- (prEN ISO/IEC 8183 - *Work Item Number*: JT021011), inerente la gestione del rischio dell'Intelligenza Artificiale - framework del ciclo vita dei dati
  Norma: ISO/IEC 8183:2023;
- prEN XXX - *Work Item Number*: JT021024) riguardante la gestione del rischio dell'IA
  Norma ISO/IEC 8183:2023.

All'interno della norma pubblicata EN ISO/IEC 23894:2024 - (WI=JT021016), viene fatto riferimento alla norma internazionale: ISO/IEC 23894:2023. Per quanto attiene, invece, i documenti rilevanti sul tema della gestione del rischio, dalla descrizione appare si debba far riferimento a: prCEN/CLC/TR XXX - *Work Item Number*: JT021009 – *checklist* per la gestione dei rischi legati all'IA; prEN ISO/IEC 8183 - *Work Item Number*: JT021011 – gestione del rischio dell'Intelligenza Artificiale - framework del ciclo vita dei dati.

Item Standardization Request 2

Richiesta formulata nella SR

Governance e qualità dei data set utilizzati per sviluppare sistemi di IA

Riferimenti dell'AI Act

Nell'Allegato 2, par. 2.2. della SR, la Commissione ha descritto il contenuto minimo della normativa tecnica riguardante la governance e la *data quality*, stabilendo come, dette norme, debbano includere quantomeno:

[8] *Ex art. 3 del Regolamento, per messa in servizio deve intendersi: la fornitura di un sistema di IA direttamente al deployer per il primo uso o per uso proprio nell'Unione per la finalità prevista.*

(a) le specifiche tecniche per un'adeguata *governance* dei dati e per l'implementazione di procedure di gestione degli stessi, da applicarsi agli sviluppatori/fornitori di sistemi d'IA;
(b) le specifiche sugli aspetti qualitativi dei *dataset* utilizzati per addestrare, validare e testare i sistemi di IA.

La Commissione prende in considerazione quali sono le esigenze di normazione rispetto ai requisiti e il trattamento di dati particolari per il monitoraggio delle discriminazioni e alle specifiche inerenti agli aspetti qualitativi dei *data set*.

Per quanto attiene alle specifiche di cui al punto a) par. 2.2 il Comitato, nello sviluppare le norme tecniche, deve fare particolare attenzione alla fase di generazione e raccolta dei dati, alle operazioni di preparazione degli stessi, nonché alle scelte di progettazione e alle procedure per individuare e affrontare i bias, il potenziale di discriminazione per approssimazione (*and addressing biases and potential for proxy discrimination*) o qualsiasi altra carenza pertinente nei dati;

Rispetto alle specifiche descritte al punto b) par. 2.2 queste dovranno, tra le cose, tenere in debito conto la rappresentatività, la rilevanza, la completezza e la correttezza dei dati utilizzati (ho aggiunto io dei dati utilizzati, ma sono andato a ragionamento, non è specificato).

Come è già stato indicato, il Regolamento dedica l'intero Capo III ai sistemi d'IA ad Alto Rischio per i quali, alla Sezione II, detta specifici requisiti in merito alla *governance* e la qualità dei dataset. In particolare, all'art. 10 del Regolamento (*rubricato data and data governance*) è stabilito che i sistemi di IA ad Alto Rischio che si avvalgono di tecniche che prevedono l'addestramento dei modelli di IA con i dati, sono sviluppati sulla base di data set di addestramento, di convalida e test che soddisfano i criteri di qualità di cui ai parr. da 2 a 5. In detti paragrafi è specificato che:

ex art. 10, par. 2, i *dataset* di addestramento, di convalida e di *test* devono essere soggetti a pratiche di *governance* e di gestione (*management*) dei dati adeguate allo scopo del sistema di IA ad Alto Rischio. Tali pratiche devono avere ad oggetto:

a) scelte di design pertinenti;
b) processi di raccolta dei dati e l'origine dei dati e, nel caso siano utilizzati dati personali, lo scopo originario della raccolta dei dati;
c) operazioni di preparazione al trattamento dei dati, quali l'annotazione, l'etichettatura, la pulizia (cleaning), l'aggiornamento, l'arricchimento e l'aggregazione;
d) la formulazione di ipotesi, con specifico riferimento alle informazioni che i dati dovrebbero misurare e rappresentare;
e) la valutazione della disponibilità, della quantità e dell'adeguatezza dei *data set* necessari;
f) la verifica in vista di possibili distorsioni che potrebbero incidere sulla salute e sulla sicurezza delle persone, avere un impatto negativo sui diritti fondamentali o portare a discriminazioni vietate dal diritto dell'Unione, in particolare quando i dati in uscita influenzano i dati in entrata per operazioni future;
g) misure appropriate per individuare, prevenire e attenuare i possibili pregiudizi individuati in base alla lettera f);
h) l'identificazione delle lacune o delle carenze di dati pertinenti che impediscono la conformità al presente regolamento e le modalità con cui tali lacune e carenze possono essere affrontate.

ex art. 10, par. 5, nella misura in cui sia strettamente necessario, al fine di garantire l'individuazione e la correzione dei pregiudizi in relazione ai sistemi di IA ad Alto Rischio, ai sensi del paragrafo 2, lettere f) e g), del presente articolo gli sviluppatori/fornitori di tali sistemi possono, eccezionalmente, trattare categorie particolari di dati personali, fatte salve, in ogni caso, le opportune garanzie per i diritti e le libertà fondamentali delle persone fisiche.

Oltre alle disposizioni di cui al Regolamento (UE) 2016/679, alla Direttiva (UE) 2016/680 e al Regolamento (UE) 2018/1725, affinché tale trattamento possa avvenire il Regolamento stabilisce che:

a) l'individuazione e la correzione dei pregiudizi non possano essere efficacemente effettuati mediante il trattamento di altri dati, compresi i dati sintetici o anonimizzati;
b) le categorie particolari di dati personali sono soggette a limitazioni tecniche sul riutilizzo dei dati personali e a misure di sicurezza e di tutela della privacy all'avanguardia, compresa la pseudonimizzazione;
c) le categorie particolari di dati personali sono soggette a misure volte a garantire che i dati personali trattati siano sicuri, protetti, soggetti a garanzie adeguate, tra le quali debbono comprendersi dei controlli rigorosi e una documentazione dell'accesso, per evitare abusi e garantire che solo le persone autorizzate con obblighi di riservatezza consoni abbiano accesso a tali dati personali;
d) i dati rientranti nella categoria di dati personali particolari non devono essere trasmessi, trasferiti o altrimenti accessibili ad altre parti;
e) i dati personali particolari debbono essere cancellati una volta che la distorsione sia stata corretta o che i dati personali abbiano raggiunto la fine del periodo di conservazione, a seconda di quale delle due condizioni si verifichi per prima;
f) le registrazioni delle attività di trattamento ai sensi dei regolamenti (UE) 2016/679 e (UE) 2018/1725 e della direttiva (UE) 2016/680 includono i motivi per cui il trattamento di categorie particolari di dati personali è strettamente necessario per individuare e correggere i pregiudizi;
g) i dati personali particolari devono essere cancellati una volta effettuato e concluso il loro trattamento, che deve essere espletato con le limitazioni anzidette.

Ai parr. 3 e 4 dell'art. 10 del Regolamento sono indicati ulteriori requisiti in merito alla qualità dei data set. Nello specifico è previsto che i *dataset* utilizzati per l'addestramento, la convalida e i test devono essere pertinenti, sufficientemente rappresentativi e, per quanto possibile, privi di errori, nonché completi in vista dello scopo per il quale è stato progettato il sistema d'IA.

Inoltre i suddetti data set devono avere le proprietà statistiche appropriate, anche, se del caso, per quanto attiene alle persone o ai gruppi di persone in relazione ai quali si intende utilizzare il sistema di IA ad Alto Rischio. Sul punto, poi, la Commissione evidenzia come tali caratteristiche dei *data set* possono essere soddisfatte o a livello dei singoli data set, o a livello dell'insieme dei data set raggruppati. Ciò significa che tali proprietà, laddove siano presenti nell'insieme dei data set raggruppati sarà sufficiente, non richiedendosi, contemporaneamente, che dette proprietà statistiche siano presenti anche nei singoli *data set* separati.

Infine, al par. 4 dell'art. 10 del Regolamento è stabilito che i data set devono tenere conto, nella misura richiesta dalla finalità del prodotto, delle caratteristiche o degli elementi peculiari dello specifico contesto geografico, contestuale, comportamentale o funzionale in cui il sistema di IA ad Alto Rischio è destinato a essere utilizzato.

Per quanto attiene al rapporto tra gli obblighi regolamentari inerenti alla *governance* e alla qualità dei *data set* previsti per i sistemi di IA ad Alto Rischio ex Capo III, Sezione II del Regolamento e le norme armonizzate alla luce della SR, l'art. 40 par. 1 del Regolamento (norme armonizzate e *deliverables* di normazione) stabilisce che: i sistemi di IA ad Alto Rischio e i Modelli di IA per Finalità Generali che risultino conformi alle norme armonizzate, o alle parti a cui lo specifico sistema d'IA è sottoposto, i cui riferimenti sono stati pubblicati nell'OJEU a norma del

regolamento (UE) n. 1025/2012, si presumono conformi ai requisiti di cui alla Sezione 2 Capo III o, se applicabili, agli obblighi di cui al Capo V (avente ad oggetto gli obblighi di trasparenza per i fornitori e *deployer* di specifici sistemi d'IA), Sezioni II e III del Regolamento, nella misura in cui dette norme coprono tali requisiti e/o obblighi.
In aggiunta, ex art. 42, par. 1 del Regolamento, i sistemi di IA ad Alto Rischio che sono stati addestrati e sottoposti a prova con dati che rispecchiano il contesto geografico, comportamentale, *contestuale* o funzionale specifico all'interno del quale sono destinati a essere usati si presumono conformi ai *pertinenti requisiti* di cui all'articolo 10, paragrafo 4.

Norme internazionali pertinenti

- FprCEN/CLC ISO/IEC/TS 12791 - *Work Item Number*: JT021013 1), inerente al trattamento dei bias indesiderati nelle attività di apprendimento automatico di classificazione e regressione Technical Specification: ISO/IEC TS 12791;
- prEN ISO/IEC 25059 - *Work Item Number*: JT021014 – ingegneria del *software* - requisiti di qualità e valutazione dei sistemi e del software (SQuaRE) - modello di qualità per i sistemi di IA Norma: ISO/IEC 25059:2023;
- Linea guida pubblicata CEN/CLC ISO/IEC/TR 24027:2023 - tecnologia dell'informazione - IA - Bias nei sistemi di IA e processi decisionali assistiti da IA (WI=JT021017);
- Linea guida ISO/IEC TR 24027:2021.

Per quanto attiene, invece, ai documenti rilevanti sul tema della qualità e della governance dei dati, nonché dei temi annessi quali i bias nei sistemi di IA, dalla descrizione, sembra doversi far riferimento a: prCEN/CLC/TR 17894 - *Work Item Number*: JT021001 - valutazione di conformità dell'IA.

Item Standardization Request 3

Richiesta formulata nella SR
Il tracciamento e il monitoraggio delle attività dei sistemi di IA

Riferimenti dell'AI Act
Ex Allegato 2, par. 2.3. della SR, la Commissione ha descritto il contenuto minimo delle norme inerenti al tracciamento e monitoraggio delle attività dei sistemi d'IA. Stabilendo come dette norme debbano includere quantomeno le specifiche per la registrazione automatica degli eventi per i sistemi di IA. Ciò, al fine di consentire la tracciabilità di tali sistemi durante il loro ciclo di vita e il monitoraggio delle loro operazioni anche nella fase post-vendita da parte dei *deployer*.
Le disposizioni inerenti il tracciamento e il monitoraggio delle attività dei sistemi di IA contenute nel Regolamento.
All'interno del Regolamento, gli obblighi in merito al tracciamento e al monitoraggio delle attività dei sistemi di IA sono descritti in 3 articoli inseriti in capi e sezioni differenti, ed in particolare:

- nell'art. 12 (Capo III, Sezione II) avente ad oggetto la tenuta dei registri;
- nell'art. 19 (Capo III, Sezione III) inerente ai registri generati automaticamente;
- nell'art. 73 (Capo IX, Sezione II) in merito alla segnalazione di incidenti gravi.

Procedendo con ordine, per quanto attiene alla tenuta dei registri, all'art. 12 del Regolamento, sono contenute le disposizioni attinenti alla registrazione degli eventi per i sistemi di IA ad Alto Rischio. Sistemi il cui *design*, secondo il dettato normativo, deve garantire la registrazione automatica degli eventi nel corso del loro ciclo-vita, tracciando situazioni di rischio, agevolando il monitoraggio post-vendita e controllando il funzionamento del sistema. In particolare, ai parr. 1 e 2 dell'art. 12 del Regolamento, è stabilito come i sistemi ad Alto Rischio debbano:

- ❖ essere sviluppati in modo tale da consentire la registrazione automatica degli eventi ("log") nel corso dell'intero ciclo vita;
- ❖ permettere un livello di tracciabilità del funzionamento adeguato alla finalità prevista del sistema;
- ❖ consentire di registrare gli eventi rilevanti per:

a) individuare le situazioni che possono far sì che il sistema di IA ad alto rischio presenti un rischio ai sensi dell'articolo 79, paragrafo 1[10], o una modifica sostanziale
b) facilitare il monitoraggio successivo all'immissione sul mercato di cui all'articolo 72 [11];
c) monitorare il funzionamento dei sistemi di cui all'articolo 26, paragrafo 5[12].

Al par. 3 della medesima disposizione è indicato come per i sistemi di IA ad Alto Rischio di cui al punto 1, lettera a), dell'Allegato III (sistemi di identificazione biometrica a distanza) le capacità di registrazione devono prevedere almeno:

a) la registrazione del periodo di ogni utilizzo del sistema (data e ora di inizio e data e ora di fine di ogni utilizzo);
b) il data base di riferimento rispetto al quale il sistema ha controllato i dati di *input*;
c) i dati di *input* per i quali la ricerca ha portato a una corrispondenza;
d) l'identificazione delle persone fisiche coinvolte nella verifica dei risultati, di cui all'articolo 14, paragrafo 5[13].

All'art. 19 del Capo III, Sezione III (rubricata: obblighi dei fornitori e degli installatori di sistemi di IA ad Alto Rischio e di altre parti), del Regolamento vengono disciplinati gli obblighi di conservazione da parte dei fornitori/sviluppatori del sistema d'IA ad Alto Rischio dei registri

---

[10] Per sistemi di IA che presentano un rischio la norma fa riferimento a un "prodotto che presenta un rischio", come definito all'articolo 3, punto 19, del regolamento (UE) 2019/1020, nella misura in cui presentano rischi per la salute o la sicurezza o per i diritti fondamentali delle persone. In merito alla definizione di cui all'art. 3 punto 19 del regolamento (UE) 2019/1020.

[11] L'art. 72 si esprime in merito al monitoraggio post-commercializzazione da parte dei fornitori/sviluppatori e al piano di monitoraggio per i sistemi di Intelligenza Artificiale ad Alto Rischio. In sintesi, i fornitori/sviluppatori devono stabilire e documentare un sistema di monitoraggio proporzionato alla natura delle tecnologie di IA sviluppate e ai rischi rappresentati nello specifico dal sistema ad Alto Rischio commercializzato. In sostanza, il sistema deve raccogliere, documentare e analizzare attivamente e sistematicamente i dati pertinenti sulle prestazioni dei sistemi di IA per garantire la conformità continua ai requisiti stabiliti. Se necessario e pertinente, il monitoraggio deve includere un'analisi dell'interazione con altri sistemi di IA.

[12] In sintesi, l'art. 26 par. 6 stabilisce che i distributori di sistemi di IA ad alto rischio devono monitorare l'operatività degli stessi seguendo le istruzioni per l'uso. Se rilevano rischi significativi o incidenti gravi, devono informare tempestivamente il fornitore, l'importatore o il distributore e le autorità di sorveglianza del mercato, e sospendere l'uso del sistema se necessario. Questo obbligo non si applica ai dati operativi sensibili delle autorità di contrasto. Le istituzioni finanziarie possono soddisfare l'obbligo di monitoraggio rispettando le norme di governance interna previste dalla normativa dell'Unione sui servizi finanziari.

[12] Per i sistemi di IA ad Alto Rischio sistemi di identificazione biometrica a distanza, le misure di cui al paragrafo 3 dell'art. 14 del Regolamento sono tali da garantire che nessuna azione o decisione sia presa dall'installatore sulla base dell'identificazione risultante dal sistema, a meno che tale identificazione non sia stata verificata e confermata separatamente da almeno due persone fisiche dotate delle necessarie competenze, formazione e autorità. L'obbligo di verifica separata da parte di almeno due persone fisiche non si applica ai sistemi di IA ad Alto Rischio utilizzati a fini di applicazione della legge, della migrazione, del controllo delle frontiere o asilo, qualora il diritto dell'Unione o il diritto nazionale ritenga sproporzionata l'applicazione di tale obbligo.

generati automaticamente. In particolare, viene sancito che:

- I fornitori/sviluppatori di sistemi di IA ad Alto Rischio devono conservare i registri menzionati nell'articolo 12, par. 1, generati automaticamente dai loro sistemi di IA, nella misura in cui tali registri sono sotto il loro controllo. Inoltre, senza recare pregiudizio alle normative dell'Unione o nazionali applicabili, i registri devono essere conservati per un periodo appropriato allo scopo previsto del sistema di IA ad Alto Rischio (almeno sei mesi), a meno che non sia diversamente previsto dalla normativa europea o nazionale di volta in volta applicabile, facendo particolare riferimento alle norme inerenti alla protezione dei dati personali;
- I fornitori/sviluppatori rientranti tra le istituzioni finanziarie soggette a specifici requisiti inerenti alla loro *governance* interna, sulla base di accordi di settore o di processi sviluppati in ragione della legge sui servizi finanziari dell'Unione, devono mantenere i registri generati automaticamente dai loro sistemi di IA ad Alto Rischio come parte della documentazione conservata ai sensi della legge sui servizi finanziari applicabile nel caso concreto.

Rispetto, invece, agli obblighi di segnalazione di incidenti gravi e malfunzionamenti, l'art. 73 del Regolamento stabilisce il dovere per i fornitori/sviluppatori di sistemi di IA di segnalare immediatamente alle autorità di sorveglianza del mercato dell'Unione qualsiasi incidente grave causato da sistemi di IA ad Alto Rischio che abbiano immesso sul mercato[14].
In sintesi, l'art. 73 stabilisce che il fornitore e il *deployer* (se applicabile) di sistemi di IA ad Alto Rischio sono tenuti a segnalare eventuali incidenti gravi alle autorità di vigilanza del mercato degli Stati membri in cui si sono verificati tali incidenti. La segnalazione dovrà avvenire immediatamente dopo che il fornitore ha stabilito un nesso causale tra il sistema di IA e l'incidente grave o sussiste la ragionevole probabilità di tale nesso, entro un massimo di 15 giorni dalla conoscenza dell'incidente. In caso di infrazione diffusa o incidente grave, la segnalazione deve avvenire entro due giorni dalla conoscenza dell'incidente. Se si verifica il decesso di una persona, la segnalazione deve essere immediata o entro massimo 10 giorni dalla conoscenza dell'incidente. Il fornitore deve condurre indagini necessarie sull'incidente e sul sistema di IA interessato, cooperando con le autorità competenti e seguendo le procedure stabilite. L'autorità di vigilanza del mercato informa le autorità nazionali competenti e la Commissione fornisce orientamenti entro 12 mesi dall'entrata in vigore del regolamento. Le autorità di vigilanza adottano misure appropriate entro sette giorni dalla ricezione della notifica di incidente.
Per quanto concerne il rapporto tra gli obblighi regolamentari inerenti al tracciamento e monitoraggio delle attività previsti per i sistemi di IA ad Alto Rischio ex Capo III, Sezione II del Regolamento e le norme armonizzate alla luce della SR, l'art. 40 par. 1 del Regolamento (norme armonizzati e deliverables di normazione) stabilisce che: i sistemi di IA ad Alto Rischio e i Modelli di IA per Finalità Generali che risultino conformi alle norme armonizzate, o alle parti a cui lo specifico sistema d'IA è sottoposto, i cui riferimenti sono stati pubblicati nell'OJEU a norma del regolamento (UE) n. 1025/2012, si presumono conformi ai requisiti di cui alla Sezione 2 Capo III o, se applicabili, agli obblighi di cui al Capo V (avente ad oggetto gli obblighi di trasparenza per i fornitori e deployer di specifici sistemi d'IA), Sezioni II e III del Regolamento, nella misura in cui dette norme coprono tali requisiti e/o obblighi.

[14] *Ex art. 3 del Regolamento, per messa a disposizione sul mercato deve intendersi la fornitura di un sistema di IA o di un modello di IA per finalità generali per la distribuzione o l'uso sul mercato dell'Unione nel corso di un'attività commerciale, a titolo oneroso o gratuito.*

Norme internazionali pertinenti

Per quanto attiene alla valutazione della presenza di *norme* e di documenti pertinenti al tema del tracciamento e monitoraggio dei sistemi di IA, si suggerisce di fare riferimento ai documenti del *working programme* e alla normativa tecnica pubblicata riguardante la gestione del rischio.

Item Standardization Request 4

Richiesta formulata nella SR

Gli obblighi di trasparenza e le informazioni per gli utenti di sistemi di IA

Riferimenti dell'AI Act

Nell'Allegato 2, par. 2.4. della SR, la Commissione ha descritto il contenuto minimo delle norme riguardanti la trasparenza e le informazioni rivolte agli utenti dei sistemi di IA. Stabilendo come, dette norme, debbano fornire specifiche relative alle:

- soluzioni di progettazione e sviluppo che garantiscano la trasparenza del funzionamento del sistema di IA al fine di consentire agli utenti di comprendere i risultati del sistema e di utilizzarli in modo appropriato; e
- istruzioni per l'uso che accompagnano i sistemi di IA, comprese le istruzioni sulle capacità e i limiti del sistema e sulle misure di manutenzione e cura, tenendo in particolare considerazione:

i. la necessità di identificare e distinguere adeguatamente le informazioni pertinenti e comprensibili per diverse categorie di utenti professionali e non professionali;

ii. senza pregiudizio del punto i., la necessità di garantire che le informazioni identificate siano sufficienti per consentire agli utenti di interpretare l'output del sistema e utilizzarlo in modo appropriato in modo da mitigare i rischi.

All'interno del Regolamento, gli obblighi in merito alla trasparenza e alle informazioni rivolte agli utenti dei sistemi di IA sono descritti in 3 disposizioni inserite in capi e sezioni differenti, ed in particolare:

- nell'art. 13 (Capo III, Sezione II), avente ad oggetto la trasparenza e le informazioni da fornire ai *deployer* inerenti ai sistemi ad Alto Rischio;
- nell'art. 50 (Capo IV), riguardante gli obblighi di trasparenza per i fornitori/sviluppatori e i *deployer* di determinati sistemi d' IA;
- nell'art. 53 (Capo V, Sezione II), inerente agli obblighi per i fornitori di sistemi d' IA classificabili tra i *general-purpose model*.

Nel dettaglio, all'art. 13 del Regolamento, si stabilisce che i sistemi di IA ad Alto Rischio devono essere progettati e sviluppati in modo trasparente per consentire ai *deployer* di interpretare l'*output* e utilizzarlo correttamente.

---

[15] *Ex art, 3 del regolamento, per fornitore/sviluppatore, deve intendersi una persona fisica o giuridica, un'autorità pubblica, un'agenzia o un altro organismo che sviluppa un sistema di IA o un modello di IA per finalità generali o che fa sviluppare un sistema di IA o un modello di IA per finalità generali e immette tale sistema o modello sul mercato o mette in servizio il sistema di IA con il proprio nome o marchio, a titolo oneroso o gratuito.*

[16] *Ex art. 3 del Regolamento, per deployer: una persona fisica o giuridica, un'autorità pubblica, un'agenzia o un altro organismo che utilizza un sistema di IA sotto la propria autorità, tranne nel caso in cui il sistema di IA sia utilizzato nel corso di un'attività personale non professionale.*

La disposizione sancisce, inoltre, l'obbligo del fornitore/sviluppatore[15] di fornire istruzioni per l'uso che siano complete, accessibili e che, inoltre, includano informazioni chiare sulle caratteristiche, le prestazioni e i limiti del sistema. Le istruzioni che accompagnano il sistema d' IA ad Alto Rischio devono anche comprendere informazioni sul monitoraggio umano, sulle risorse computazionali necessarie e sulla manutenzione del sistema. In particolare, ai parr. 1 e 2 dell'art. 13 del Regolamento, è stabilito come i sistemi ad Alto Rischio debbano essere:

- progettati e sviluppati in modo tale da garantire che il loro funzionamento sia sufficientemente trasparente per consentire ai fornitori di interpretare l'*output* del sistema e utilizzarlo in modo appropriato al fine di permettere (ottenere) il rispetto degli obblighi del fornitore e del *deployer*[16], stabiliti nella Sezione III del Capo II;
- accompagnati da istruzioni per l'uso in un formato digitale appropriato o in altro modo, che includano informazioni concise, complete, corrette e chiare, pertinenti, accessibili e comprensibili per i deployer.

Al par. 3 della medesima disposizione è indicato come per i sistemi di IA ad Alto Rischio:
Le istruzioni per l'uso devono contenere almeno le seguenti informazioni:

- l'identità e i dettagli di contatto del fornitore e, se applicabile, del suo rappresentante autorizzato;
- le caratteristiche, le capacità e i limiti di prestazione del sistema di IA ad Alto Rischio, tra cui:

i. il suo scopo previsto;
ii. il livello di accuratezza, compresi i suoi indicatori, la robustezza e la cybersicurezza di cui all'articolo 15 (di cui si dirà nel capitolo relativo alle norme inerenti alla cybersicurezza), contro cui il sistema di IA ad Alto Rischio è stato testato e convalidato e che possono essere previsti, nonché eventuali circostanze conosciute e prevedibili che possono avere un impatto su quel livello previsto di accuratezza, robustezza e cybersicurezza;
iii. eventuali circostanze conosciute o prevedibili, legate all'uso del sistema di IA ad Alto Rischio in conformità con il suo scopo previsto o in condizioni di abuso ragionevolmente prevedibili, che possono comportare rischi per la salute e la sicurezza o per i diritti fondamentali di cui all'articolo 9, par. 2) (di cui si è già detto nel cap. 2 del presente documento);
iv. quando applicabile, le capacità tecniche e le caratteristiche del sistema di IA ad Alto Rischio per fornire informazioni pertinenti a spiegare il suo *output*;
v. quando appropriato, le prestazioni riguardanti persone specifiche o gruppi di persone su cui il sistema è destinato ad essere utilizzato.
vi. se del caso, le specifiche dei dati di input o qualsiasi altra informazione pertinente in termini di set di dati di addestramento, convalida e test utilizzati, tenendo conto dello scopo previsto del sistema di IA ad alto rischio;
vii. se del caso, informazioni che consentano agli operatori di interpretare i risultati del sistema di IA ad alto rischio e di utilizzarli in modo appropriato.

All'art. 50 del Regolamento, vengono disciplinati gli obblighi di trasparenza per i fornitori e i *deployer* di determinati sistemi d'IA. In particolare, i fornitori devono assicurare che i sistemi di IA progettati per interagire direttamente con le persone forniscano informazioni chiare agli individui interessati, informandoli sulla presenza e l'interazione con un sistema di IA, a meno che ciò non sia ovvio per una persona ragionevolmente informata, attenta e consapevole, tenendo conto delle circostanze e del contesto di utilizzo.

Detto obbligo non si applica ai sistemi di IA autorizzati per finalità legali come l'accertamento, la prevenzione, l'indagine o il perseguimento di reati, a condizione che siano adeguatamente protetti i diritti e le libertà dei terzi, a meno che tali sistemi non siano disponibili al pubblico per segnalare un reato.

I fornitori di sistemi di IA, inclusi quelli a scopo generale, che generano contenuti audio, immagine, video o testuali sintetici, devono garantire che gli *output* del sistema siano chiaramente identificati come generati o manipolati artificialmente in un formato leggibile meccanicamente. Devono assicurarsi che le soluzioni tecniche siano efficaci, interoperabili, robuste e affidabili, nella misura in cui ciò sia tecnicamente possibile, tenendo conto delle peculiarità e dei limiti dei diversi tipi di contenuti, dei costi di implementazione e dello stato dell'arte generalmente riconosciuto, come indicato nelle norme tecniche pertinenti. Questo obbligo non si applica se i sistemi di IA svolgono una funzione di assistenza per l'editing standard o non modificano in modo sostanziale i dati di input forniti dall'utente o la loro semantica, o se sono autorizzati dalla legge per finalità di accertamento, prevenzione, indagine o perseguimento di reati. I responsabili dell'implementazione di un sistema di riconoscimento delle emozioni o di un sistema di categorizzazione biometrica devono informare le persone coinvolte sul funzionamento del sistema e devono trattare i dati personali in conformità alle normative UE pertinenti, a seconda dei casi. Questo obbligo non si applica ai sistemi di IA utilizzati per la categorizzazione biometrica e il riconoscimento delle emozioni autorizzati per accertare, prevenire o indagare reati, a condizione che siano protetti adeguatamente i diritti e le libertà dei terzi e in conformità con il diritto dell'Unione.

I responsabili dell'implementazione di un sistema di IA che genera o manipola immagini, audio o video per creare un c.d. "deep fake" devono informare chiaramente che il contenuto è stato generato o manipolato artificialmente. Tuttavia, questa responsabilità non si applica se l'uso del "deep fake" è autorizzato dalla legge per finalità di accertamento, prevenzione, indagine o perseguimento di reati. Nel caso in cui il contenuto rientri in un'opera o programma artisticamente, creativamente, satiricamente o fittiziamente analoghi, gli obblighi di trasparenza si limitano a rendere noto in modo adeguato l'esistenza di tali contenuti generati o manipolati, senza impedire l'esposizione o il godimento dell'opera stessa.

Per quanto riguarda un sistema di IA che genera o manipola testo pubblicato per informare il pubblico su questioni di interesse pubblico, i responsabili dell'implementazione devono dichiarare chiaramente che il testo è stato generato o manipolato artificialmente. Tuttavia, questo obbligo non si applica se l'uso è autorizzato dalla legge per accertare, prevenire, indagare o perseguire reati, o se il contenuto generato dall'IA è stato sottoposto a revisione umana o controllo editoriale e una persona fisica o giuridica ha la responsabilità editoriale della pubblicazione del contenuto.

Le informazioni fornite ai soggetti interessati di cui all'art. 50 del Regolamento, devono essere chiare e distinguibili al più tardi al momento della prima interazione o esposizione. Queste informazioni devono essere conformi ai requisiti di accessibilità pertinenti. È importante notare che quanto riportato nell'art. 50 lascia impregiudicati sia i requisiti e gli obblighi stabiliti nel capo III del Regolamento inerente ai sistemi ad Alto Rischio, sia gli altri obblighi di trasparenza stabiliti dal diritto dell'Unione o da quello nazionale per coloro che implementano i sistemi di IA.

Vale la pena ricordare che ex. art 507 l'Ufficio AI incoraggia e facilita l'elaborazione di codici di condotta a livello dell'Unione per agevolare l'effettiva attuazione degli obblighi relativi all'individuazione e all'etichettatura di contenuti generati o manipolati artificialmente.

La Commissione può adottare atti di esecuzione per approvare tali codici di condotta secondo la procedura di cui all'articolo 56 (6).
Se ritiene che il codice non sia adeguato, la Commissione può adottare un atto di esecuzione che specifichi norme comuni per l'attuazione di tali obblighi secondo la procedura di esame di cui all'articolo 98, paragrafo 2.
All'art. 53, del Capo V, Sezione II, del Regolamento, vengono disciplinati gli obblighi informativi e di trasparenza per i fornitori/*deployer* di sistemi d'IA classificabili tra i *general-purpose model*. In particolare, i fornitori/sviluppatori di detti sistemi devono:
a) redigere e mantenere aggiornata la documentazione tecnica del sistema (modello), compresi i processi di formazione e di test e i risultati della valutazione. Questa documentazione deve contenere, almeno, gli elementi stabiliti nell'Allegato XI del Regolamento (rubricato: informazioni da presentare al momento della registrazione di sistemi di IA ad Alto Rischio) al fine di fornirlo, su richiesta, all'Ufficio designato e alle autorità competenti nazionali. Secondo quanto disposto nell'Allegato XI – il quale descrive le informazioni a riguardo dei test in condizioni reali da registrare in conformità all'articolo 60 (rubricato test di sistemi di IA ad Alto Rischio in condizioni reali, al di fuori delle *sandbox* normative sull'IA) – gli elementi che devono essere contenuti nella documentazione tecnica sono i seguenti:

- numero univoco di identificazione a livello dell'Unione per i test in condizioni reali;
- nome e dettagli di contatto del fornitore o del potenziale fornitore e degli utenti coinvolti nei *test* in condizioni reali;
- una breve descrizione del sistema di IA, lo scopo dello stesso e le altre informazioni necessarie per l'identificazione del sistema;
- una descrizione riassuntiva delle principali caratteristiche del piano per i test in condizioni reali;
- informazioni sulla sospensione o la terminazione dei test in condizioni reali;

b) redigere e mantenere aggiornate le informazioni e la documentazione relativa da fornire ai fornitori/sviluppatori di sistemi di IA che intendono integrare nei loro *general-purpose model*. Senza che ciò rechi pregiudizio alla necessità di rispettare e proteggere i diritti di proprietà intellettuale e le informazioni aziendali riservate o i segreti commerciali conformemente alla legislazione dell'Unione e nazionale. Nello specifico le informazioni e la documentazione devono:

- consentire ai fornitori di sistemi di IA di comprendere adeguatamente le capacità e i limiti del *general-purpose model* e di adempiere ai loro obblighi ai sensi del presente regolamento;
- contenere, almeno, gli elementi stabiliti nell'Allegato XII , nel quale è specificato che: la documentazione tecnica per i fornitori di *general-purpose model* destinati a fornitori *downstream* che intendono integrare il modello nel loro sistema di IA dovrà includere almeno:
  - una descrizione generale del *general-purpose model*, che comprenda:

[18] *Tale disposizione, in sintesi, sancisce che gli Stati membri possono stabilire eccezioni o limitazioni per l'estrazione di testo e dati da opere accessibili legalmente. Tali estrazioni possono essere conservate per il tempo necessario. Tuttavia, questa eccezione si applica solo se i detentori dei diritti non hanno espressamente escluso l'uso automatizzato delle opere disponibili pubblicamente online.*

a) le attività che il modello è destinato a svolgere e il tipo e la natura dei sistemi di IA in cui può essere integrato;
b) le politiche di utilizzo accettabili applicabili;
c) la data di rilascio e i metodi di distribuzione;
d) come il modello interagisce o può essere utilizzato per interagire con *hardware* o *software* che non sono parte del modello stesso, se applicabile;
e) le versioni del *software* rilevanti relative all'uso del *general-purpose model*, se applicabile;
f) l'architettura e il numero di parametri;
g) modalità (ad esempio, testo, immagine) e formato degli input e degli output;
h) la licenza per il modello;
    • una descrizione degli elementi del modello e del processo per il suo sviluppo, compresi/e:
aa) i mezzi tecnici (ad esempio, istruzioni per l'uso, infrastruttura, strumenti) necessari affinché il *general-purpose model* possa essere integrato nei sistemi di IA;
bb) le modalità (ad esempio, testo, immagine, ecc.) e formato degli *input* e degli *output* e la loro dimensione massima (ad esempio, lunghezza della finestra di contesto, ecc.);
cc) le informazioni sui dati utilizzati per l'addestramento, il test e la convalida, se applicabile, compreso il tipo e la provenienza dei dati e le metodologie di cura;
c) adottare misure idonee a conformarsi alla legge sul diritto d'autore dell'Unione, e, in particolare, a identificare e conformarsi, anche attraverso tecnologie all'avanguardia, a una riserva di diritti espressa ai sensi dell'articolo 4, paragrafo 3, della direttiva (UE) 2019/790 (rubricato: eccezioni o limitazioni ai fini dell'estrazione di testo e di dati)[18];
d) redigere e rendere pubblicamente disponibile un riassunto sufficientemente dettagliato sul contenuto utilizzato per la formazione del modello di IA a uso generale, secondo un modello fornito dall'Ufficio AI.

Per quanto attiene al rapporto tra gli obblighi di trasparenza e le informazioni per gli utenti dei sistemi di IA previsti per i sistemi di IA ad Alto Rischio ex Capo III, Sezione II del Regolamento e le norme armonizzate alla luce della SR, l'art. 40 par. 1 del Regolamento (norme armonizzati e *deliverables* di normazione) stabilisce che: i sistemi di IA ad Alto Rischio e i Modelli di IA per Finalità Generali che risultino conformi alle norme armonizzate, o alle parti a cui lo specifico sistema d'IA è sottoposto, i cui riferimenti sono stati pubblicati nell'OJEU a norma del regolamento (UE) n. 1025/2012, si presumono conformi ai requisiti di cui alla Sezione 2 Capo III o, se applicabili, agli obblighi di cui al Capo V (avente ad oggetto gli obblighi di trasparenza per i fornitori e deployer di specifici sistemi d'IA), Sezioni II e III del Regolamento, nella misura in cui dette norme coprono tali requisiti e/o obblighi.

Norme internazionali pertinenti

Per quanto attiene alla valutazione della presenza di norme e di documenti pertinenti al tema degli obblighi di trasparenza e delle informazioni per gli utenti dei sistemi di IA, non sono stati riscontrati documenti o normativa tecnica specifica elaborata o richiamata all'interno del *working programme* elaborato dal comitato tecnico competente.

Item Standardization Request 5

Richiesta formulata nella SR
I requisiti di supervisione umana nei sistemi d'IA

Riferimenti dell'AI Act
Nell'Allegato 2, par. 2.5. della SR, la Commissione ha descritto il contenuto minimo della normativa tecnica sul tema, stabilendo come, gli stessi, debbano specificare misure e procedure per la supervisione umana dei sistemi di IA che siano:

- individuate e, se è tecnicamente possibile, integrate nel sistema di IA dal fornitore prima che quest'ultimo venga immesso sul mercato o divenga operativo;
- individuate dal fornitore prima di immettere il sistema di IA sul mercato o di metterlo in servizio e che siano adeguate a essere attuate dall'utente.

Tali procedure devono includere misure che consentano agli utenti di comprendere, monitorare, interpretare, valutare e influenzare gli aspetti rilevanti del funzionamento del sistema di IA. Tali norme devono, inoltre, stabilire (se ciò è necessario) misure di supervisione adeguate, tenendo conto dello specifico scopo per il quale il sistema di IA è stato creato.
In aggiunta, per quanto attiene i sistemi di IA destinati all'identificazione biometrica da remoto delle persone, le misure di supervisione umana devono, tra le cose, prevedere la possibilità che non venga effettuata alcuna azione o presa alcuna decisione dall'utente sulla base dell'identificazione risultante dal sistema, a meno che questa non sia stata verificata e confermata separatamente da almeno due persone fisiche.
All'interno dell'AI Act, gli obblighi inerenti alla supervisione umana nei sistemi d'IA sono descritti nell'art. 15, Capo III, Sezione II, del Regolamento. In particolare, al par. 2 dell'art. 14 del Regolamento è sancita una disposizione di ordine generale, per mezzo della quale è specificato come la supervisione umana debba mirare a prevenire o minimizzare i rischi per la salute, la sicurezza o per i diritti fondamentali che possono emergere quando viene utilizzato un sistema di IA ad Alto Rischio. Tale prevenzione/minimizzazione dei rischi è doverosa sia quando l'utilizzo è in conformità con lo scopo per il quale il sistema d'IA è stato creato, sia qualora lo stesso sia utilizzato impropriamente, purché i rischi che s'intende prevenire/minimizzare siano ragionevolmente prevedibili e gli stessi persistano nonostante l'applicazione degli altri requisiti stabiliti nella medesima Sezione del regolamento (Sezione II).

Oltre a quanto disposto all'interno del par. 2, l'art. 14 del Regolamento stabilisce:

- al par. 1, che i sistemi di IA ad Alto Rischio devono essere progettati e sviluppati in modo tale da poter essere efficacemente supervisionati da persone fisiche durante il periodo in cui sono in uso e ciò anche per mezzo di interfacce uomo-macchina appropriate;
- al par. 3, che le procedure di sorveglianza siano commisurate ai rischi, al livello di autonomia e al contesto di utilizzo del sistema di IA ad Alto Rischio. Inoltre, allo stesso paragrafo viene aggiunto che, dette procedure, debbono essere garantite da:

[19] *La disposizione fa riferimento ai sistemi di identificazione biometrica a distanza, tra i quali non sono compresi i sistemi di Intelligenza Artificiale destinati a essere utilizzati per la verifica biometrica il cui unico scopo è confermare che una determinata persona fisica è quella che dichiara di essere.*

- misure identificate e integrate, quando tecnicamente fattibile, nel sistema di IA ad Alto Rischio dal fornitore prima che venga messo sul mercato o messo in servizio;
- e/o misure identificate dal fornitore prima di mettere il sistema di IA ad Alto Rischio sul mercato o di renderlo operativo (metterlo in servizio), che siano adatte ad essere implementate dal *deployer* (la congiunzione e sta ad indicare che, nel caso sia necessario, vanno adottate entrambe le misure).

Al par. 4 del medesimo articolo del Regolamento è specificato che ai fini dell'attuazione dei paragrafi 1, 2 e 3, il sistema di IA ad Alto Rischio è fornito al *deployer* in modo tale che le persone fisiche alle quali è affidata la sorveglianza umana abbiano la concreta possibilità, ove opportuno e proporzionato, di:

- comprendere correttamente le capacità e i limiti pertinenti del sistema di IA ad Alto Rischio ed essere in grado di monitorarne debitamente il funzionamento, anche al fine di individuare e affrontare anomalie, disfunzioni e prestazioni inattese;
- rimanere consapevoli del fatto che può portare ad un eccesso di affidamento rispetto all'*output* prodotto da un sistema di IA ad Alto Rischio ("distorsione dell'automazione"), in particolare in relazione ai sistemi di IA ad Alto Rischio utilizzati per fornire informazioni o raccomandazioni per le decisioni che devono essere prese da persone fisiche;
- interpretare correttamente l'*output* del sistema di IA ad Alto Rischio, tenendo conto ad esempio degli strumenti e dei metodi di interpretazione disponibili;
- decidere, in qualsiasi situazione particolare, di non usare il sistema di IA ad Alto Rischio o altrimenti di ignorare, annullare o rovesciare l'*output* del sistema di IA ad Alto Rischio;
- intervenire sul funzionamento del sistema di IA ad Alto Rischio o interrompere il funzionamento del sistema mediante un interruttore di "arresto" o una procedura analoga che consenta al sistema di arrestarsi in condizioni di sicurezza.

Infine, al par. 5, art. 14 del Regolamento, viene aggiunto che, per i sistemi di IA ad Alto Rischio di cui all'allegato III, punto 1, lettera a)[19] del Regolamento, le misure di cui al paragrafo 3 del presente articolo debbono essere architettate in modo tale da garantire che il *deployer* non compia azioni o adotti decisioni sulla base dell'identificazione risultante dal sistema, a meno che l'identificazione non sia stata verificata e confermata separatamente da almeno due persone fisiche dotate della necessaria competenza, formazione e autorità.
Il requisito di una verifica separata da parte di almeno due persone fisiche non si applica ai sistemi di IA ad Alto Rischio utilizzati a fini di contrasto, migrazione, controllo delle frontiere o asilo, qualora il diritto dell'Unione o nazionale ritenga sproporzionata l'applicazione di tale requisito.
Per quanto attiene al rapporto tra gli obblighi inerenti alla supervisione umana nei sistemi di IA ad Alto Rischio sui sistemi di IA previsti per i sistemi di IA ad Alto Rischio ex Capo III, Sezione II del Regolamento e le norme armonizzate alla luce della SR, l'art. 40 par. 1 del Regolamento (norme armonizzate e *deliverables* di normazione) stabilisce che: i sistemi di IA ad Alto Rischio e i Modelli di IA per Finalità Generali che risultino conformi alle norme armonizzate, o alle parti a cui lo specifico sistema d'IA è sottoposto, i cui riferimenti sono stati pubblicati nell'OJEU del regolamento (UE) n. 1025/2012, si presumono conformi ai requisiti di cui alla Sezione 2 Capo III o, se applicabili, agli obblighi di cui al Capo V (avente

ad oggetto gli obblighi di trasparenza per i fornitori e *deployer* di specifici sistemi d'IA), Sezioni II e III del Regolamento, nella misura in cui dette norme coprono tali requisiti e/o obblighi.

Norme internazionali pertinenti
Per quanto riguarda alla valutazione della presenza di norme e di documenti pertinenti al tema degli obblighi di trasparenza e delle informazioni per gli utenti dei sistemi di IA, non sono stati riscontrati documenti o norme specifiche elaborate o richiamate all'interno del *working programme* pubblicato dal Comitato Tecnico Competente. Ad ogni modo, per ragioni di contiguità tematica, si suggerisce di prendere in considerazione i documenti del *working programme* e la normativa tecnica pubblicata riguardante la gestione del rischio.

Item Standardization Request 6

Richiesta formulata nella SR
L'accuratezza nei sistemi d'IA

Riferimenti dell'AI Act
Nell'Allegato 2, par. 2.6. della SR, la Commissione ha descritto il contenuto minimo della normativa tecnica sul tema, stabilendo come, ai fini dello sviluppo degli stessi per "accuratezza" deve intendersi la capacità del sistema di IA di svolgere il compito per cui è stato progettato. Aggiunge, inoltre, che il concetto di accuratezza non deve intendersi in senso restrittivo e che, pertanto, non deve essere confuso con la definizione più stretta di accuratezza statistica, che altro non è se non una delle varie metriche possibili per valutare le prestazioni dei sistemi di IA.
La Commissione prosegue, specificando come tali norme debbano stabilire specifiche capaci di garantire un adeguato livello di accuratezza dei sistemi di IA e, allo stesso tempo, di consentire ai fornitori/sviluppatori di dichiarare le metriche e i livelli di accuratezza necessari a valutare il sistema. Infine, aggiunge che la normazione europea inerente all'accuratezza deve occuparsi di stabilire, se giustificato dalla necessità, un insieme di strumenti e metriche appropriati e rilevanti per misurare l'accuratezza rispetto a livelli opportunamente definiti.
Nel perimetro del Regolamento, gli obblighi inerenti all'accuratezza (concetto da intendersi così come descritto al par. 7.1 del presente documento) dei sistemi d'IA sono descritti nell'art. 15, Capo III, Sezione II. Quest'articolo disciplina congiuntamente sia i parametri inerenti all'accuratezza, sia quelli riguardanti la robustezza e la cybersicurezza, sancendo come i sistemi di IA ad Alto Rischio devono essere progettati e sviluppati in modo tale da conseguire un adeguato livello di accuratezza, robustezza e cybersicurezza e da operare in modo coerente con tali aspetti durante tutto il loro ciclo di vita. Nella medesima disposizione è specificato che al fine di affrontare gli aspetti tecnici relativi alle modalità di misurazione degli adeguati livelli di accuratezza e robustezza di cui al paragrafo 1 e altre metriche di prestazione pertinenti, la Commissione, in cooperazione con i portatori di interessi e le organizzazioni pertinenti, quali le autorità di metrologia e di analisi comparativa, incoraggia, se del caso, lo sviluppo di parametri di riferimento e metodologie di misurazione.

Per quanto attiene specificamente all'accuratezza, al par. 3 dell'art. 15 del Regolamento è stabilito come i livelli di accuratezza e le pertinenti metriche di accuratezza dei sistemi di IA ad Alto Rischio debbono essere dichiarate nelle istruzioni per l'uso che accompagnano il sistema. Per quanto concerne il rapporto tra gli obblighi inerenti all'accuratezza previsti per i sistemi di IA ad Alto Rischio ex Capo III, Sezione II del Regolamento e le norme armonizzate alla luce della SR, l'art. 40 par. 1 del Regolamento (norme armonizzate e deliverables di normazione) stabilisce che: i sistemi di IA ad Alto Rischio e i Modelli di IA per Finalità Generali che risultino conformi alla normativa armonizzata, o alle parti a cui lo specifico sistema d'IA è sottoposto, i cui riferimenti sono stati pubblicati nell'OJEU a norma del regolamento (UE) n. 1025/2012, si presumono conformi ai requisiti di cui alla Sezione 2 Capo III o, se applicabili, agli obblighi di cui al Capo V (avente ad oggetto gli obblighi di trasparenza per i fornitori e *deployer* di specifici sistemi d'IA), Sezioni II e III del Regolamento, nella misura in cui dette norme coprono tali requisiti e/o obblighi.

Norme internazionali pertinenti
Per quanto attiene alla valutazione della presenza di normativa tecnica e di documenti pertinenti al tema degli obblighi di trasparenza e delle informazioni per gli utenti dei sistemi di IA, attualmente, non risulta esservi un documento specifico all'interno del *working programme*. Dalla descrizione dei singoli *deliverable* (documenti) appare presumibile che il tema sia trattato in riferimento ad altre macro-tematiche quali, a titolo di mero esempio, la gestione del rischio, i bias dei sistemi d'IA, i requisiti di qualità e valutazione dei sistemi e del software ecc.

Item Standardization Request 7

Richiesta formulata nella SR
Le specifiche di robustezza nei sistemi d'IA

Riferimenti dell'AI Act
Nell'Allegato 2, par. 2.7. della SR, la Commissione ha descritto il contenuto minimo della normativa tecnica sul tema, stabilendo come, quest'ultimi debbano stabilire le specifiche per la robustezza dei sistemi di IA. Specificando come tali specifiche debbano riguardare la considerazione di varie fonti di errori, difetti e incongruenze, nonché la comprensione di come i sistemi di IA interagiscono con l'ambiente in cui vengono utilizzati.
Nel perimetro dell'AI Act, gli obblighi inerenti alla solidità dei sistemi d'IA sono descritti nell'art. 15, Capo III, Sezione II. Quest'articolo, oltre alla solidità (robustezza), disciplina congiuntamente anche i parametri dell'accuratezza e della cybersicurezza. Nello specifico, sancisce che i sistemi di IA ad Alto Rischio devono essere progettati e sviluppati in modo tale da conseguire un adeguato livello di accuratezza, robustezza e cybersicurezza e da operare in modo coerente con tali aspetti durante tutto il loro ciclo di vita.
Nella medesima disposizione è specificato che al fine di affrontare gli aspetti tecnici relativi alle modalità di misurazione degli adeguati livelli di accuratezza e robustezza di cui al paragrafo 1 e altre metriche di prestazione pertinenti, la Commissione, in cooperazione con i portatori di interessi e le organizzazioni pertinenti, quali le autorità di metrologia e di

analisi comparativa, incoraggia, se del caso, lo sviluppo di parametri di riferimento e metodologie di misurazione.
Per quanto concerne specificamente la robustezza, al par. 5 dell'art. 15 del Regolamento è stabilito come i sistemi di IA ad Alto Rischio devono essere il più resilienti possibile per quanto riguarda errori, guasti o incongruenze che possono verificarsi all'interno del sistema o nell'ambiente in cui esso opera; soprattutto nel caso detti sistemi interagiscano con persone fisiche o altri sistemi. A tale riguardo, pertanto, debbono essere adottate adeguate misure tecniche e organizzative.
Sempre all'art. 15 del Regolamento si stabilisce come la robustezza dei sistemi di IA ad Alto Rischio possa essere conseguita mediante soluzioni tecniche di ridondanza, che possono includere piani di backup o fail-safe[20]. Infine, è ulteriormente specificato che i sistemi di IA ad Alto Rischio che proseguono il loro apprendimento dopo essere stati immessi sul mercato o dopo che sono stati messi in servizio, devono essere sviluppati in modo tale da eliminare o ridurre il più possibile il rischio di *output* potenzialmente distorti che influenzano gli input per operazioni future (*feedback loops*) e garantire che tali feedback loops siano oggetto di adeguate misure di attenuazione.
Per quanto riguarda il rapporto tra gli obblighi inerenti alla robustezza previsti per i sistemi di IA ad Alto Rischio ex Capo III, Sezione II del Regolamento e le norme armonizzate alla luce della SR, l'art. 40 par. 1 del Regolamento (norme armonizzate e deliverables di normazione) stabilisce che: i sistemi di IA ad Alto Rischio e i Modelli di IA per Finalità Generali che risultino conformi alla normativa armonizzata, o alle parti a cui lo specifico sistema d'IA è sottoposto, i cui riferimenti sono stati pubblicati nell'OJEU a norma del regolamento (UE) n. 1025/2012, si presumono conformi ai requisiti di cui alla Sezione 2 Capo III o, se applicabili, agli obblighi di cui al Capo V (avente ad oggetto gli obblighi di trasparenza per i fornitori e *deployer* di specifici sistemi d'IA), Sezioni II e III del Regolamento, nella misura in cui dette norme coprono tali requisiti e/o obblighi.

Norme internazionali pertinenti

- Linea guida CEN/CLC ISO/IEC/TR 24029-1:2023 - IA - Valutazione della robustezza delle reti neurali - (WI=JT021018)
  Linea guida ISO/IEC TR 24029-1:2021.

Per quanto attiene, invece, ai documenti rilevanti sul tema, sembra potersi fare riferimento a: Soluzioni tecniche per affrontare le vulnerabilità specifiche dell'IA (WI=JT021029). Documento preliminare che si occupa delle soluzioni tecniche per affrontare le vulnerabilità specifiche dell'IA. Dalla descrizione dei singoli *deliverable* (documenti) appare presumibile che il tema sia trattato in riferimento ad altre macro-tematiche quali, a titolo di mero esempio, la gestione del rischio, i bias dei sistemi d'IA, i requisiti di qualità e valutazione dei sistemi e del software, ecc. (par. 2.4, 3.4. parte II del presente documento).

[20] Un fail-safe in un sistema di Intelligenza Artificiale è una misura progettata per garantire che il sistema si comporti in modo "sicuro" anche in presenza di guasti o errori. In particolare, l'obiettivo principale di un fail-safe è garantire che il sistema di IA mantenga la sicurezza e la stabilità anche in situazioni impreviste o critiche, riducendo al minimo i rischi per gli utenti e l'ambiente circostante.

Item Standardization Request 8

Richiesta formulata nella SR
I requisiti in tema di cybersicurezza nei sistemi di IA

Riferimenti dell'AI Act
Nell'Allegato 2, par. 2.8. della SR, la Commissione ha descritto il contenuto minimo della normativa tecnica sul tema, stabilendo come debbano fornire specifiche organizzative e tecniche adeguate per garantire che i sistemi di IA siano resilienti rispetto ai tentativi di alterarne il funzionamento, l'utilizzo, il comportamento, le prestazioni o comprometterne le proprietà di sicurezza da parte di terze parti malintenzionate e mosse dalla volontà di sfruttare le vulnerabilità dei sistemi di IA.
In ragione di ciò, le soluzioni organizzative e tecniche dovranno includere, se appropriato, misure volte a prevenire e controllare gli attacchi informatici per mezzo dei quali viene tentato di manipolare risorse specifiche dell'IA, come set di dati di addestramento, i modelli, o che cercano di sfruttare vulnerabilità delle risorse digitali di un sistema di IA o nell'infrastruttura ICT sottostante.
Queste soluzioni tecniche saranno adeguate alle circostanze e ai rischi rilevanti.
Infine, la Commissione specifica come queste norme europee debbano tenere conto delle disposizioni essenziali per i prodotti con componenti digitali come elencato nelle Sezioni 1 e 2 dell'Allegato I relativo alla proposta del Regolamento, contenente l'elenco delle normative armonizzate dell'Unione Europea a cui l'*Artificial Intelligence Act* fa riferimento.
Nel perimetro dell'AI Act, gli obblighi inerenti alla cybersicurezza dei sistemi d'IA sono descritti nell'art. 15, Capo III, Sezione II.
Quest'articolo, oltre alla cybersicurezza, disciplina congiuntamente i parametri di accuratezza e robustezza, sancendo come i sistemi di IA ad Alto Rischio devono essere progettati e sviluppati in modo tale da conseguire un adeguato livello di accuratezza, robustezza e cybersicurezza e da operare in modo coerente con tali aspetti durante tutto il loro ciclo di vita.
In particolare, al par. 5 del medesimo articolo il Regolamento stabilisce che I sistemi di IA ad Alto Rischio debbano essere resilienti ai tentativi di terzi non autorizzati di modificarne l'uso, gli *output* o le prestazioni sfruttando le vulnerabilità del sistema.
Aggiunge, inoltre, che le soluzioni tecniche volte a garantire la cybersicurezza dei sistemi di IA ad Alto Rischio devono essere adeguate alle circostanze e ai rischi pertinenti.
Infine, la disposizione precisa che le soluzioni tecniche finalizzate ad affrontare le vulnerabilità specifiche dell'IA devono includere, ove opportuno, misure volte a prevenire, accertare, rispondere, risolvere e controllare gli attacchi per mezzo dei quali i terzi "aggressori" cercano di:

- manipolare il set di dati di addestramento (data poisoning - "avvelenamento dei dati") o i componenti pre-addestrati utilizzati nell'addestramento (model poisoning - "avvelenamento dei modelli");
- utilizzare *input* affinché il modello di IA commetta un errore (*adversarial examples* - "esempi antagonistici", o *model evasion*, - "evasione dal modello");
- apprestare attacchi alla riservatezza o di sfruttare per scopi nocivi i difetti del modello.

Per quanto riguarda il rapporto tra gli obblighi inerenti alla cybersicurezza previsti per i sistemi di IA ad Alto Rischio ex Capo III, Sezione II del Regolamento e le norme armonizzate alla luce della SR, l'art. 40 par. 1 del Regolamento (norme armonizzate e deliverables di normazione) stabilisce che: i sistemi di IA ad Alto Rischio e i Modelli di IA per Finalità Generali che risultino conformi alle norme armonizzate, o alle parti a cui lo specifico sistema d'IA è sottoposto, i cui riferimenti sono stati pubblicati nell'OJEU a norma del regolamento (UE) n. 1025/2012, si presumono conformi ai requisiti di cui alla Sezione 2 Capo III o, se applicabili, agli obblighi di cui al Capo V (avente ad oggetto gli obblighi di trasparenza per i fornitori e *deployer* di specifici sistemi d'IA), Sezioni II e III del Regolamento, nella misura in cui dette norme coprono tali requisiti e/o obblighi.
Inoltre, all'art. 42, par. 2 del Regolamento è previsto che I sistemi di IA ad Alto Rischio che sono stati certificati o per i quali è stata rilasciata una dichiarazione di conformità nell'ambito di un sistema di cybersicurezza a norma del regolamento (UE) 2019/881 e i cui riferimenti sono stati pubblicati nell'OJEU si presumono conformi ai requisiti di cybersicurezza di cui all'articolo 15 del Regolamento, nella misura in cui tali requisiti siano contemplati nel certificato di cybersicurezza o nella dichiarazione di conformità o in parti di essi.

Norme internazionali pertinenti
Soluzioni tecniche per affrontare le vulnerabilità specifiche dell'IA (WI=JT021029).
Dalla descrizione dei singoli *deliverable* (documenti) appare presumibile che il tema sia trattato in riferimento ad altre macro-tematiche, tra le quali, le più attinenti sembrano essere la gestione del rischio e i requisiti di qualità e valutazione dei sistemi e del software (par. 2.4, 3.4. parte II del presente documento).

Item Standardization Request 9

Richiesta formulata nella SR
La gestione del monitoraggio della qualità nei sistemi di IA incluso il monitoraggio post commercializzazione

Riferimenti dell'AI Act
Nell'Allegato 2, par. 2.9. della SR, la Commissione ha descritto il contenuto minimo degli *standard* sul tema, stabilendo come debbano determinare le specifiche per un sistema di gestione della qualità da implementare all'interno delle organizzazioni da parte dei fornitori/sviluppatori di sistemi di IA. Tali sistemi di gestione della qualità devono garantire, tra le altre cose, la costante conformità di un sistema d'IA con gli aspetti descritti in tutti i capitoli precedenti del presente documento (dal 2, inerente alla gestione del rischio, al 9, avente ad oggetto la cybersicurezza).
La Commissione, aggiunge che dev'essere data particolare considerazione all'implementazione di misure del sistema di gestione della qualità da parte delle piccole e medie imprese e organizzazioni.
In ragione di ciò, le specifiche contenute negli *standard* dovranno essere redatte in modo che gli aspetti del sistema di gestione della qualità relativi al sistema di IA possano essere integrati

nell'intero sistema di gestione del fornitore/sviluppatore, in particolare con i sistemi di gestione della qualità esistenti stabiliti per soddisfare i requisiti dei sistemi di gestione della qualità contenuti nella legislazione armonizzata dell'Unione elencata nell'Allegato II, Sezione A della proposta di legge sull'IA, contenente parte dell'elenco delle normative armonizzate dell'Unione Europea a cui l'*Artificial Intelligence Act* fa riferimento.
Nel perimetro dell'AI Act, i requisiti di conformità in merito alla gestione del monitoraggio della qualità nei sistemi d'IA e al monitoraggio post-vendita sono descritti in 2 articoli inseriti in capi e sezioni differenti, ed in particolare:

- nell'art. 17 (Capo III, Sezione III) avente ad oggetto gli obblighi dei fornitori/sviluppatori e *deployer* rispetto ai sistemi di gestione della qualità;
- nell'art. 72 (Capo IX, Sezione I) inerente al monitoraggio post-vendita, alla condivisione delle informazioni e alla sorveglianza del mercato.

Per quanto attiene agli obblighi dei fornitori/sviluppatori e *deployer* rispetto ai sistemi di gestione della qualità, nell'art. 17 del Regolamento, si stabilisce che i fornitori di sistemi AI ad Alto Rischio devono istituire un sistema di gestione della qualità che assicuri la conformità al Regolamento. In sintesi, viene sancito che detto sistema di gestione deve essere sistematico, ordinato ed estrinsecarsi sotto forma procedure e istruzioni, includendo aspetti quali: la strategia per la conformità normativa, le tecniche di progettazione, lo sviluppo, il collaudo, la validazione/gestione dei dati e dei rischi, il monitoraggio post-vendita, la segnalazione degli incidenti, la comunicazione con le autorità competenti nonché la conservazione della documentazione e la responsabilità.
Tutto ciò con la precisazione che l'attuazione di tali aspetti deve essere proporzionata alla dimensione dell'organizzazione del fornitore, anche se, va specificato, il Regolamento prevede che I fornitori rispettino, in ogni caso, il grado di rigore e il livello di protezione necessari per garantire la conformità dei loro sistemi di IA ad Alto Rischio al presente regolamento.
Vediamo la disposizione nel dettaglio. il par. 1 dell'art. 17 del Regolamento a proposito del monitoraggio della qualità, stabilisce come tale sistema è documentato in modo sistematico e ordinato sotto forma di politiche, procedure e istruzioni scritte e che deve comprendere almeno:

a) una strategia per la conformità normativa, compresa la conformità alle procedure di valutazione della conformità e alle procedure per la gestione delle modifiche dei sistemi di IA ad Alto Rischio;

---

[21] *Si riporta unicamente l'elenco dell'oggetto della normazione per le singole normative cfr. All. I Sezione A per le normative di riferimento):*
- Macchine;
- Sicurezza dei giocattoli;
- Imbarcazioni da diporto e natanti personali;
- Ascensori e componenti di sicurezza per ascensori;
- Dispositivi e sistemi di protezione per "atmosfere" potenzialmente esplosive;
- Apparecchi radio;
- Attrezzature a pressione;
- Impianti a fune;
- Dispositivi di protezione individuale;
- Apparecchi a gas;
- Dispositivi medici;
- Dispositivi medici diagnostici in vitro.

b) le tecniche, le procedure e gli interventi sistematici da utilizzare per la progettazione, il controllo della progettazione e la verifica della progettazione del sistema di IA ad Alto Rischio;
c) le tecniche, le procedure e gli interventi sistematici da utilizzare per lo sviluppo e per il controllo e la garanzia della qualità del sistema di IA ad Alto Rischio;
d) le procedure di esame, prova e convalida da effettuare prima, durante e dopo lo sviluppo del sistema di IA ad alto rischio e la frequenza con cui devono essere effettuate;
e) le specifiche tecniche, comprese le norme, da applicare e, qualora le pertinenti norme armonizzate non siano applicate integralmente, o non includano tutti i requisiti pertinenti di cui alla Sezione 2, i mezzi da usare per garantire che il sistema di IA ad alto rischio sia conforme a tali requisiti;
f) i sistemi e le procedure per la gestione dei dati, compresa l'acquisizione, la raccolta, l'analisi, l'etichettatura, l'archiviazione, la filtrazione, l'estrazione, l'aggregazione, la conservazione dei dati e qualsiasi altra operazione riguardante i dati effettuata prima e ai fini dell'immissione sul mercato o della messa in servizio di sistemi di IA ad alto rischio;
g) il sistema di gestione dei rischi di cui all'articolo 9
h) la predisposizione, l'attuazione e la manutenzione di un sistema di monitoraggio successivo all'immissione sul mercato a norma dell'articolo 72 (che analizzeremo nel prosieguo);
i) le procedure relative alla segnalazione di un incidente grave a norma dell'articolo 73
j) la gestione della comunicazione con le autorità nazionali competenti e le autorità pertinenti, comprese quelle che forniscono o sostengono l'accesso ai dati, gli organismi notificati, altri operatori, clienti o altre parti interessate;
k) i sistemi e le procedure per la conservazione delle registrazioni e di tutte le informazioni e la documentazione pertinenti;
l) la gestione delle risorse, comprese le misure relative alla sicurezza dell'approvvigionamento;
m) un quadro di responsabilità che definisca le responsabilità della dirigenza e di altro personale per quanto riguarda tutti gli aspetti elencati nel presente paragrafo.

Infine, al par.3 della medesima disposizione è specificato che i fornitori di sistemi di IA ad Alto Rischio, soggetti agli obblighi relativi ai sistemi di gestione della qualità o che si trovino a ricoprire una funzione equivalente secondo il diritto settoriale dell'Unione, possono includere gli aspetti elencati nel par. 1 come parte dei sistemi di gestione della qualità in conformità a tale legge.
Per quanto attiene, invece, l'art. 72, la disposizione si esprime in merito al monitoraggio post-commercializzazione da parte dei fornitori/sviluppatori e al piano di monitoraggio per i sistemi di IA ad Alto Rischio. In sintesi, i fornitori/sviluppatori devono stabilire e documentare un sistema di monitoraggio proporzionato alla natura delle tecnologie di IA sviluppate e ai rischi rappresentati nello specifico dal sistema ad Alto Rischio commercializzato. In sostanza, il sistema deve raccogliere, documentare e analizzare attivamente e sistematicamente i dati pertinenti sulle prestazioni dei sistemi di IA per garantire la conformità continua ai requisiti stabiliti. Se necessario e pertinente, il monitoraggio deve includere un'analisi dell'interazione con altri sistemi di IA. Gli obblighi specificamente previsti dalla disposizione sono stabiliti:

- nel par. 1 secondo il quale i fornitori devono istituire e documentare un sistema di monitoraggio successivo all'immissione sul mercato che sia proporzionato alla natura delle tecnologie di IA e ai rischi del sistema di IA ad Alto Rischio;

- nel par. 2 in cui viene stabilito come Il sistema di monitoraggio successivo all'immissione sul mercato deve essere in grado di raccogliere, documentare e analizzare attivamente e sistematicamente i dati pertinenti che possono essere forniti dai *deployer* o che possono essere raccolti tramite altre fonti sulle prestazioni dei sistemi di IA ad Alto Rischio per tutta la durata del loro ciclo di vita e consente al fornitore di valutare la costante conformità dei sistemi di IA ai requisiti di cui al capo III, Sezione 2. Se risulti necessario, il monitoraggio successivo all'immissione sul mercato dovrà includere un'analisi dell'interazione con altri sistemi di IA. Tale obbligo non riguarda i dati operativi sensibili dei *deployer* che sono autorità di contrasto.;
- nel par. 3, in cui è previsto che il sistema di monitoraggio successivo all'immissione sul mercato deve poggiare su un piano di monitoraggio successivo all'immissione sul mercato. Il piano di monitoraggio successivo all'immissione sul mercato fa parte della documentazione tecnica richiamata all'allegato IV del Regolamento. Rispetto a quanto stabilito nel presente paragrafo, si tenga presente come la Commissione adotterà un atto di esecuzione che stabilisce disposizioni dettagliate in cui definirà un modello per il piano di monitoraggio successivo all'immissione sul mercato e un elenco di elementi da includere nel piano entro 18 mesi dall' entrata in vigore del Regolamento.

Infine, è rilevante considerare come, nell'art. 72 del Regolamento è previsto che per i sistemi di IA ad Alto Rischio disciplinati dalla normativa di armonizzazione dell'Unione elencata nell'Allegato I, sezione A[21], qualora tale normativa preveda già un sistema e un piano di monitoraggio successivo all'immissione sul mercato, al fine di garantire la coerenza, evitare duplicazioni e ridurre al minimo gli oneri aggiuntivi, i fornitori possono scegliere di integrare, se del caso, i necessari elementi di cui ai parr. 1, 2 e 3 utilizzando il modello di cui al paragrafo 3 nei sistemi e nei piani già esistenti in virtù di tale normativa, a condizione che consegua un livello di protezione equivalente. Quanto appena descritto si applica anche ai sistemi di IA ad Alto Rischio di cui all'allegato III, punto 5, immessi sul mercato o messi in servizio da istituti finanziari soggetti a requisiti in materia di governance, dispositivi o processi interni stabiliti a norma del diritto dell'Unione in materia di servizi finanziari. Per quanto attiene specificamente ai sistemi di gestione della qualità e alle valutazioni di conformità ne parleremo diffusamente nel prosieguo relativamente al decimo item della SR. Per quanto riguarda il rapporto tra gli obblighi inerenti alla gestione del monitoraggio della qualità nei sistemi d'IA ex Capo III, Sezione III del Regolamento e le norme armonizzate alla luce della SR, l'art. 40 par. 1 del Regolamento (norme armonizzati e deliverables di normazione) stabilisce che: i sistemi di IA ad Alto Rischio e i Modelli di IA per Finalità Generali che risultino conformi alle norme armonizzate, o alle parti a cui lo specifico sistema d'IA è sottoposto, i cui riferimenti sono stati pubblicati nell'OJEU a norma del regolamento (UE) n. 1025/2012, si presumono conformi ai requisiti di cui alla Sezione 2 Capo III o, se applicabili, agli obblighi di cui al Capo V (avente ad oggetto gli obblighi di trasparenza per i fornitori e *deployer* di specifici sistemi d'IA), Sezioni II e III del Regolamento, nella misura in cui dette norme coprono tali requisiti e/o obblighi.

Norme internazionali pertinenti

Dalla descrizione dei singoli *deliverable* (documenti) appare presumibile che il tema sia trattato in riferimento ad altre macro-tematiche, tra le quali, le più attinenti sembrano essere la gestione del rischio, i requisiti di qualità e valutazione dei sistemi e del software e la valutazione della conformità per i sistemi d'IA (par. 2.4, 3.4 e 11.4 parte II del presente documento).

Item Standardization Request 10

Richiesta formulata nella SR
Valutazione della conformità per i sistemi d'IA

Riferimenti dell'AI Act
Nell'Allegato 2, par. 2.10. della SR, la Commissione ha descritto il contenuto minimo della normativa tecnica sul tema, stabilendo come debbano fornire procedure e processi per le attività di valutazione della conformità relative ai sistemi di IA e ai sistemi di gestione della qualità dei fornitori di IA, inclusi criteri per valutare la competenza delle persone incaricate di tali attività. Inoltre, deve considerare sia gli scenari in cui la valutazione della conformità è effettuata dal fornitore stesso che con il coinvolgimento di un'organizzazione terza esterna professionale.
Ai sensi del Regolamento UE sull'IA, un requisito cruciale per i sistemi d'IA ad Alto Rischio è che i loro fornitori effettuino valutazioni di conformità ex ante (precedenti alla commercializzazione).
Le valutazioni di conformità sono un requisito legale primario previsto dall'articolo 43 del Regolamento, che ha lo scopo di garantire la responsabilità nello sviluppo e nella distribuzione di Sistemi di IA ad Alto Rischio. Il Regolamento definisce le valutazioni di conformità come un mezzo per dimostrare se i requisiti stabiliti nel Capo III, Sezione 2 della legislazione relativa a un sistema di IA ad Alto Rischio siano stati soddisfatti. Sul punto è fondamentale tenere in considerazione i seguenti punti chiave in merito alle valutazioni di conformità:

- gli sviluppatori di sistemi d'IA ad Alto Rischio devono eseguire le valutazioni di conformità prima di mettere il sistema sul mercato o usarlo per la prima volta nell'UE;
- possono essere condotte internamente dal fornitore o attraverso un'entità terza notificata, a seconda della presenza e dell'uso di norme armonizzate. Sul punto, va sottolineato come esista un'eccezione per i sistemi ad Alto Rischio progettati per l'identificazione biometrica remota o per trarre inferenze sulle caratteristiche personali basate su dati biometrici, che richiede il coinvolgimento di un terzo nel processo di valutazione della conformità;
- dovrà essere condotta una nuova valutazione di conformità in caso di modifiche significative a un sistema ad Alto Rischio. Tali modifiche possono essere innescate da cambiamenti rilevanti che influenzano la conformità del sistema ai requisiti per i sistemi di IA ad Alto Rischio o quando modificano lo scopo previsto per il sistema;

Le valutazioni di conformità devono essere tipicamente effettuate dai fornitori dei sistemi ad Alto Rischio. Tuttavia, ci sono due eccezioni in cui la responsabilità di condurre una valutazione di conformità rientra nella competenza:

- del produttore del prodotto, se il sistema in questione viene introdotto insieme al prodotto sotto il suo nome o è correlato a prodotti elencati nell'Allegato II Sezione A del Regolamento;
- dei *deployer* o importatori, se introducono il sistema d'IA ad Alto Rischio sotto il loro nome o marchio, modificano il suo scopo previsto determinato dal fornitore o apportano modifiche sostanziali.

Per quanto riguarda il rapporto tra gli obblighi in merito alla valutazione della conformità nei sistemi d'IA e le norme armonizzate alla luce della SR, l'art. 40 par. 1 del Regolamento (norme armonizzate e *deliverables* di normazione) stabilisce che: i sistemi di IA ad Alto Rischio e i Modelli di IA per Finalità Generali che risultino conformi agli standard armonizzati, o alle parti a cui lo specifico sistema d'IA è sottoposto, i cui riferimenti sono stati pubblicati nell'OJEU a norma del regolamento (UE) n. 1025/2012, si presumono conformi ai requisiti di cui alla Sezione 2 Capo III o, se applicabili, agli obblighi di cui al Capo V (avente ad oggetto gli obblighi di trasparenza per i fornitori e *deployer* di specifici sistemi d'IA), Sezioni II e III del Regolamento, nella misura in cui dette norme coprono tali requisiti e/o obblighi.

Norme internazionali pertinenti

Tra i documenti parte del *working programme* rilevanti sul tema, da quanto descritto sembra potersi far riferimento ai:

- prCEN/CLC/TR 17894 - *Work Item Number*: JT021001 - valutazione di conformità dell'IA;
- prEN ISO/IEC 23282 - *Work Item Number*: JT021012 - metodi di valutazione per sistemi di elaborazione del linguaggio naturale;
- prEN ISO/IEC 25059 - *Work Item Number*: JT021014 - ingegneria del software - requisiti di qualità e valutazione dei sistemi e del software (SQuaRE) - modello di qualità per i sistemi di IA;
- prEN XXX - *Work Item Number*: JT021025 - task di IA e metodi di valutazione dei sistemi di visione artificiale.

Inoltre, valutato che il presente capitolo si caratterizza per un tema ampio, si suggerisce di far riferimento a tutte le Sezioni dei Capitoli precedenti aventi ad oggetto le norme individuate dal Comitato Tecnico Congiunto e i documenti del *working programme* rilevanti.

4. I Framework internazionali

4.1 Il Risk Management Framework del NIST

Il 26 gennaio 2023, il NIST (National Institute of Standards and Technology) del Dipartimento del Commercio, ovvero un'Agenzia del Governo[22] degli Stati Uniti ha approvato la sua prima versione del "documento di orientamento per l'uso volontario da parte delle organizzazioni che progettano, sviluppano, implementano o utilizzano sistemi di IA per aiutare a gestire i rischi delle tecnologie". Tale documento, denominato RMF (Risk Management Framework), da un lato, si prefigge lo scopo di offrire linee guida per la gestione sicura e controllata dei sistemi di IA, dall'altro vuole monitorare i rischi e contenere i possibili effetti negativi di tali tecnologie sulle libertà civili e sui diritti dei cittadini. Tale documento è stato accompagnato da un secondo documento, il *playbook*, che "fornisce azioni suggerite per raggiungere i risultati stabiliti nel Core dell'AI RMF (Risk Management Framework). I suggerimenti sono allineati a ciascuna sottocategoria delle quattro funzioni dell'AI RMF (Governare, Mappare, Misurare, Gestire)".

Genesi evolutiva del RMF

Il primo passaggio evolutivo nella costituzione del RMF è stata l'approvazione del National AI Initiative Act del 2020, diventata legge il 1° gennaio 2021. Tale atto prevede "un programma coordinato in tutto il Governo federale per accelerare la ricerca e l'applicazione dell'IA per la prosperità economica e la sicurezza nazionale". Il secondo passaggio è stata la richiesta da parte del Congresso destinata al NIST di sviluppare il quadro applicativo del National AI Initiative Act. La bozza più recente di tale quadro risale ad agosto 2022. Lo sviluppo dell'RMF da parte del NIST è nato nel quadro della collaborazione tra settore pubblico e privato nello sforzo più ampio di aderire non solo ai parametri richiesti dal National AI Initiative Act del 2020, ma anche a quelli delle raccomandazioni della National Security Commission on Artificial Intelligence e dal Plan for Federal Engagement in Developing Technical Standards and Related Tools.
Il 30 marzo 2023 invece è stato creato il Trustworthy and Responsible AI Resource Center (AIRC), un Ente di ricerca con lo scopo di "supportare tutti gli attori dell'IA nello sviluppo e nella diffusione di tecnologie di IA affidabili e responsabili". Tale Ente supporta il RMF e lo rende operativo permettendo "un'esperienza interattiva, basata sui ruoli, che fornisca l'accesso a un'ampia gamma di risorse rilevanti per l'IA".

---

[22]https://www.agendadigitale.eu/sicurezza/nist-ai-risk-management-framework-rmf-come-garantire-laffidabilita-dellia-secondo-gli-usa/

L'approccio risk based e l'anatomia del rischio

Il Quadro adottato dal RMF del NIST prevede un approccio basato sul rischio per "identificare, valutare e mitigare i rischi associati ai sistemi di IA". Tale approccio prevede cinque fasi principali: la fase di "preparazione, di valutazione del rischio, di mitigazione del rischio, di comunicazione del rischio e di monitoraggio continuo". La fase di preparazione ha lo scopo di definire gli obiettivi, di identificare i requisiti e di stabilire il contesto operativo dentro il quale i sistemi di IA devono essere applicati.

La fase di valutazione implica, invece, un processo di identificazione delle potenziali minacce, di valutazione delle possibili vulnerabilità del sistema e di stima delle possibili conseguenze. Le valutazioni stabilite durante tale fase costituiscono la base programmatica per sviluppare le diverse strategie di mitigazione del rischio, tra cui il monitoraggio dei sistemi di IA in tempo reale, l'implementazione di sistemi di sicurezza e di controllo e l'addestramento degli utenti ai fini di un uso consapevole.

La fase di comunicazione del rischio implica poi la trasmissione di informazioni alle parti interessate circa l'esistenza dei rischi e delle loro implicazioni. Tale processo rappresenta l'elemento fondamentale per promuovere la fiducia nel loro utilizzo e la comprensione adeguata circa i potenziali rischi. Per assicurare un controllo continuo il RMF garantisce un monitoraggio continuo nella valutazione dei sistemi di IA e nel loro aggiornamento. La gestione del rischio definita dal RMF vuole innanzitutto rispondere al carattere mutevole ed in continua evoluzione dei sistemi di IA e in tal senso si prefigge un'architettura dinamica ed iterativa. La gestione del rischio all'interno del RMF viene intesa quale misura per bilanciare e prevenire i possibili impatti negativi dei sistemi di IA e per offrire, al contempo gli strumenti necessari per ottimizzare le opportunità. Il dispositivo del NIST descrive il rischio come "la misura composita della probabilità che un evento si verifichi e dell'entità o del grado delle conseguenze dell'evento corrispondente". Tale definizione ricomprende l'adattamento delle disposizioni della norma ISO 31000:2018 e della OMB Circular A-130:2016. In particolare, la ISO 31000:2018 afferma che gli impatti e le possibili conseguenze dei sistemi dei rischi possono essere "positivi, negativi o entrambi e possono tradursi in opportunità o minacce" e che "la gestione del rischio si riferisce alle attività coordinate per dirigere e controllare un'organizzazione rispetto al rischio". La quantificazione del possibile impatto negativo invece è data dalle disposizioni della OMB Circular A-130:2016. Stante le disposizioni di tale circolare, nel processo di valutazione di un potenziale effetto negativo, il livello di rischio è determinato da "1) l'impatto negativo, o l'entità del danno, che si verificherebbe se la circostanza o l'evento si verificasse e 2) la probabilità che si verifichi". L'impatto del danno viene descritto e misurato anche in base al suo raggio di azione che può concernere "individui, gruppi, comunità, organizzazioni, società, ambiente e pianeta". Nella sezione 1.1 del Risk Management Framework si insiste sul fatto che, a differenza di altri processi di gestione del rischio, il quadro di gestione di riferimento non considera solamente gli impatti negativi ma si sforza egualmente di identificare "le opportunità per massimizzare gli impatti positivi". Una gestione del rischio volta alla duplice finalità della minimizzazione del rischio e della massimizzazione delle opportunità positive vuole portare alla creazione di sistemi di IA più liberi di generare potenziale positivo e più affidabili in termini di sicurezza. Secondo le disposizioni dell'atto "La gestione del rischio può consentire agli sviluppatori e agli utenti dell'IA di comprendere gli impatti e di tenere conto dei limiti e delle incertezze insiti nei loro modelli e sistemi, il che, a sua volta, può migliorare le prestazioni e l'affidabilità complessive del sistema e la probabilità che le tecnologie dell'IA vengano utilizzate in modo vantaggioso".

Definizione di sistemi di IA e carattere volontario delle disposizioni

Il Risk Management Framework elaborato dal NIST adotta una definizione di sistema di IA in linea con la definizione dell'OECD. Più precisamente il sistema di IA viene definito come "un sistema ingegnerizzato o basato su macchine in grado, per un determinato insieme di obiettivi, di generare output come previsioni,

raccomandazioni o decisioni che influenzano ambienti reali o virtuali. I sistemi di IA sono progettati per operare con vari livelli di autonomia". Le disposizioni previste dal RMF si distinguono rispetto ad altri atti quali l'AI Act per il loro carattere volontario e privo di quadro sanzionatorio. Tale caratteristica è frutto di una scelta precisa da parte del legislatore statunitense, ovvero la necessità di rispondere in primo luogo alle esigenze di flessibilità e di adattamento normativo delle società e delle organizzazioni, a prescindere dalla loro struttura normativa e delle loro dimensioni. In linea con tale scopo l'atto prevede l'utilizzo di risorse specifiche per supportare un continuo processo di aggiornamento e di allineamento rispetto agli standard e rispetto ai feedback della comunità AI.

Struttura

L'architettura del RMF si struttura in due parti distinte, ciascuna delle quali è suddivisa a sua volta in sottosezioni. La prima parte, nello specifico, è divisa in quattro sottosezioni. La prima è volta alla descrizione e all'inquadramento dei diversi tipi di rischio e analizza in particolare la misura del rischio (Risk Measurement), la tolleranza al rischio (Risk Tolerance) e la priorità attribuita al rischio (Risk Prioritization). La seconda tratta dell'audience, mentre la terza delle caratteristiche necessarie per garantire il carattere affidabile o trustworthy dei sistemi di IA. In tale sottosezione viene posta particolare attenzione alle seguenti caratteristiche e requisiti: "validità, affidabilità, *safety*, *security*, resilienza, *accountability*, *explainability*, trasparenza, interpretabilità, privacy ed equità". La quarta ed ultima sottosezione fornisce invece gli strumenti necessari per misurare l'efficacia dell'implementazione delle misure descritte nell'atto. La seconda parte del dispositivo è improntata sul Core ovvero sul nucleo centrale e portante del Risk Management Framework. Tale nucleo è costituito da quattro sottosezioni che offrono linee guida pratiche affinché le aziende e le organizzazioni possano rispondere ai potenziali rischi dati dall'utilizzo dei sistemi di IA. Nello specifico le sottosezioni o funzioni sono denominate: *govern, map, measure e manage*. Mentre la prima funzione, ovvero quella di govern, può essere applicata indistintamente a tutto il ciclo della vita del sistema di IA, le altre possono essere implementate solamente a contesti specifici e a fasi della vita prestabilite dei sistemi di IA.

Le sfide per la gestione del rischio delle IA

Il Risk Management Framework dispone, nella sezione dedicata al rischio, di un elenco dettagliato di sfide da prendere in considerazione nel processo di gestione del rischio, al fine di creare un sistema di IA affidabile.

Misura del rischio

Stante le disposizioni dell'atto sono "i rischi o i fallimenti dell'IA non ben definiti o adeguatamente compresi, difficili da misurare quantitativamente o qualitativamente". Alcuni dei problemi di misurazione del rischio espressamente disciplinati sono i seguenti:

- "Rischi legati al software, all'hardware e ai dati di parti terze";
- "Tracciamento dei rischi emergenti";
- "Disponibilità di metriche affidabili";
- "Rischio da stabilire nelle diverse fasi del ciclo di vita dell'IA";
- "Rischio in contesti reali";
- "Inscrutabilità";
- "Human Baseline", ovvero il grado di partecipazione umana all'interno del processo decisionale.

Tolleranza al rischio

Il Risk Management Framework ha lo scopo di essere utilizzato per definire il rischio e la sua priorità; tuttavia, non prevede la valutazione o la misurazione della tolleranza al rischio.

Per tolleranza al rischio l'atto fa espressamente riferimento al grado di "disponibilità dell'organizzazione o dell'attore dell'IA a sopportare il rischio per raggiungere i propri obiettivi". Tale definizione è stata adattata alle disposizioni della ISO GUIDE 73:2009. La ISO specifica che" La tolleranza al rischio e il livello di rischio accettabile per le organizzazioni o la società sono altamente contestuali e specifici per ogni applicazione e caso d'uso". Tale disposizione sottolinea il fatto che il margine di tolleranza al rischio può essere influenzato da fattori esogeni quali norme prestabilite e linee politiche. Tale influenza predispone il livello di tolleranza al rischio ad un notevole margine di variazione e di fluttuazione nello spazio e nel tempo. Per tali considerazioni, il Framework si prefigge di essere flessibile e di "integrare le pratiche di rischio esistenti", in linea con regolamenti, leggi e norme applicabili. Premettendo che le organizzazioni hanno obbligo di conformità rispetto alle linee guida e alle disposizioni normative esistenti, l'atto specifica che in caso di mancanza di linee guida consolidate, starà alle organizzazioni stesse definire un margine di tolleranza al rischio "ragionevole". Definita tale tolleranza, le organizzazioni potranno successivamente applicare il Risk Management Framework e utilizzarlo per la gestione del rischio e per il processo di valutazione e monitoraggio.

Priorità del rischio

Secondo le disposizioni dell'AI RMF, sono le organizzazioni a dovere stabilire quali sono i rischi più elevati in capo a determinati sistemi di IA, a doverne misurare gli impatti e i possibili effetti, attribuendo di conseguenza il livello di priorità a ciascun rischio. Perché un rischio sia inaccettabile devono esservi le seguenti condizioni: i potenziali effetti negativi del sistema devono essere "imminenti" e "significativi", i danni potenzialmente causati devono presentare potenziali "rischi catastrofici" o quantomeno essere "effettivamente gravi". Qualora invece l'immissione sul mercato e la diffusione dei sistemi di IA siano considerati a basso rischio, tenendo in considerazione anche lo specifico ambito di applicazione, l'organizzazione o la società in questione può attribuire liberamente al rischio un livello di priorità più basso. Il Risk Management Framework considera anche quale elemento di classificazione della priorità del rischio, la "destinazione d'uso" del sistema, ovvero se il sistema sia destinato ad interagire direttamente con l'essere umano o meno. I sistemi di IA che vengono addestrati sulla base di un cospicuo "insieme di dati sensibili o protetti, come le informazioni di identificazione personale, o in cui i risultati dei sistemi di IA hanno un impatto diretto o indiretto sugli esseri umani" richiedono in tal senso un livello di priorità e di attenzione maggiori rispetto ai sistemi di IA progettati per interagire esclusivamente con "sistemi computazionali e addestrati su insiemi di dati non sensibili" (quali ad esempio i dati raccolti dall'ambiente fisico). Tali sistemi sono considerati potenzialmente meno rischiosi e possono quindi essere trattati secondo un livello di priorità più basso. Il Framework specifica, tuttavia, che anche in questi casi occorrono una valutazione ed un monitoraggio continui, al fine di poter attribuire al sistema in questione un livello di priorità che si adatti anche al contesto e all'abito di riferimento.

Il *Core*

Il *Core* dell'AI RMF costituisce il nucleo centrale dell'atto e ha lo scopo di fornire "azioni e risultati che rendano possibili il dialogo e la comprensione delle attività per gestire i rischi dell'IA e sviluppare sistemi di IA affidabili". Il *Core* è composto da quattro funzioni distinte: *Govern* o *governance, Map* o tracciamento, *Measure* o misura e *Manage*, o gestione. Ciascuna delle funzioni è a sua volta divisa in ulteriori categorie e sottocategorie. Tali sezioni si caratterizzano per la presenza di azioni e risultati da raggiungere. Il nucleo costitutivo dell'AI RMF è stato progettato al fine di enucleare punti di vista eterogenei e prospettive multidisciplinari quali i feedback degli attori esterni all'organizzazione.

La funzione di governance

Come è stato precedentemente anticipato, la funzione di governance si applica a ogni fase del ciclo vitale del sistema di IA e influenza direttamente l'intero processo di gestione del rischio di IA. Gli aspetti di governance che nello specifico riguardano i parametri di valutazione e di conformità, possono essere applicati ed integrati ad ogni fase processuale delle altre funzioni disciplinate dal Core. Lo scopo della funzione governance è quello di creare e di promuovere una "cultura della gestione del rischio" per rendere tale processo efficace ed efficiente oltre che per "delineare processi, documenti e schemi organizzativi che anticipino, identifichino e gestiscano i rischi che un sistema può comportare, anche per gli utenti e per altri soggetti della società". Le altre finalità della funzione sono elencate come segue:

- "delineare processi, documenti e schemi organizzativi che anticipano, identificano e gestiscono i rischi che un sistema può comportare, anche per gli utenti e per altri soggetti della società, e le procedure per raggiungere tali risultati";
- "incorporare processi di valutazione degli impatti potenziali";
- "fornire una struttura con cui le funzioni di gestione del rischio di IA possono allinearsi ai principi, alle politiche e alle strategie dell'organizzazione;"
- "collegare gli aspetti tecnici della progettazione e dello sviluppo di sistemi di IA ai valori e ai principi dell'organizzazione";
- "consentire di sviluppare pratiche e competenze organizzative per le persone coinvolte nell'acquisizione, nella formazione, nell'implementazione e nel monitoraggio di tali sistemi";
- "affrontare l'intero ciclo di vita del prodotto e i processi associati, comprese le questioni legali e di altro tipo relative all'uso di sistemi e dati software o hardware di terzi".

Il tracciamento

Come viene precisato nell'introduzione della funzione *Map*, qui di seguito tradotta come tracciamento, la previsione accurata dei possibili impatti dei sistemi di IA può essere difficoltosa. Il fatto che determinate parti interessate e responsabili di una parte del processo del sistema IA, non dispongano del controllo o della piena visibilità delle altre parti del processo rende la stima e la misurazione del rischio un'operazione delicata e complessa. L'"interdipendenza" esistente tra le varie fasi processuali della vita dei sistemi di IA e tra gli attori che vi prendono parte, rende la classificazione e il monitoraggio dei rischi potenziali incerta e poco affidabile. Un esempio in tal senso che viene preso in considerazione è "la decisione iniziale nell'identificazione degli scopi e degli obiettivi di un sistema di IA che può alterarne il comportamento e le capacità, e le dinamiche del contesto di impiego che possono modellare gli impatti delle decisioni del sistema di IA". Tale esempio lascia presagire il grado di incidenza che le condizioni e le decisioni di talune attività esercitano sulle attività restanti del processo o sulle fasi della vita dei sistemi di IA. La funzione *Map* si prefigge di raccogliere informazioni volte alla prevenzione dei rischi e di informare e di contribuire alla creazione di sistemi di IA più affidabili, attraverso le seguenti modalità:

- "la migliore capacità di comprensione dei contesti";
- "la verifica delle ipotesi sul contesto d'uso;"
- "il riconoscimento della non funzionalità dei sistemi quando questi non sono funzionali all'interno o al di fuori del loro contesto di utilizzo";
- "la disamina del contesto in cui si trovano";
- "l'identificazione degli usi positivi e vantaggiosi dei sistemi di IA esistenti";
- "migliorare la comprensione dei limiti dei processi di IA e ML";
- "identificare i vincoli nelle applicazioni reali che possono portare a impatti negativi";
- "identificare gli impatti negativi noti e prevedibili relativi all'uso previsto dei sistemi di IA";
- "anticipare i rischi dell'uso dei sistemi di IA al di là dell'uso previsto".

Una volta ottenuto il quadro di informazioni offerto dalla funzione *Map*, i fruitori del Framework dispongono di conoscenze contestuali sufficienti per poter fare una valutazione decisionale completa circa la possibilità di "progettare, sviluppare o implementare un sistema di IA". I fruitori devono obbligatoriamente continuare ad utilizzare la funzione MAP in caso di evoluzione o trasformazione di elementi quali "il contesto, le capacità, i rischi, i benefici e i potenziali impatti" che possono mutare nel tempo.

Misura

Secondo le disposizioni dell'atto, la funzione misura "impiega strumenti, tecniche e metodologie quantitative, qualitative o miste per analizzare, valutare, confrontare e monitorare il rischio associato all'IA e i relativi impatti". Per effettuare tale operazione si avvale delle informazioni raccolte durante la funzione *Map* e offre gli elementi necessari affinché possa operare la funzione *Manage*. La funzione misura prevede il "monitoraggio delle metriche relative alle caratteristiche di affidabilità, all'impatto sociale e alle configurazioni uomo-IA". I protocolli ideati e sviluppati in seno alla funzione prevedono l'utilizzo obbligatorio di metodologie di test del software e di valutazione delle prestazioni, insieme a relative misure di incertezza, confronti con parametri di riferimento delle prestazioni e relazioni e documentazioni formalizzate dei risultati. Secondo le disposizioni dell'atto i processi di revisione indipendente sono inoltre atti a migliorare l'efficacia dei test e a diminuire sensibilmente i conflitti di interesse nonché i possibili bias interni.

Gestione

La funzione gestione è volta invece all'allocazione delle risorse di rischio e alla costituzione di un "trattamento" del rischio. Tale trattamento è costituito da piani di recupero, di risposta e di comunicazione circa eventuali eventi o incidenti. Le informazioni ottenute dalle funzioni Governance e Tracciamento vengono utilizzate in questa specifica funzione al fine di garantire la massima riduzione di possibili impatti negativi o di guasti al sistema.

4.2 Il Framework normativo UK

Contesto di riferimento

Il 29 marzo 2023 il Dipartimento per la Scienza, per l'Innovazione e per la Tecnologia ha pubblicato il libro bianco "AI Regulation: A Pro-Innovation Approach", che rappresenta la proposta di regolamentazione dell'IA da parte del Governo britannico. Così come altri testi normativi precedentemente esaminati il testo in questione analizza le sfide e i potenziali rischi dati dall'utilizzo dei sistemi di IA, senza tuttavia tralasciare gli aspetti potenzialmente positivi dati da un utilizzo etico, consapevole e proporzionato dei sistemi di IA. Il quadro normativo si applica a tutto il Regno Unito e la sua applicazione è oggetto di interazione e scambio con le restanti normative esistenti nella comunità internazionale quali ad esempio il Data Protection Act del 2018 e l'Equality Act del 2010.

Obiettivi

Stante le disposizioni dell'atto, gli obiettivi del quadro normativo sono i seguenti:

- "Promuovere la crescita e la prosperità agevolando l'innovazione responsabile e riducendo l'incertezza normativa, al fine di incoraggiare gli investimenti nei sistemi di IA e sostenerne l'adozione in tutta l'economia, creando posti di lavoro e rendendoli più efficienti";
- "Aumentare il grado di fiducia pubblico rispetto ai sistemi di IA, affrontando i rischi e proteggendo i nostri valori fondamentali";

- "Rafforzare la posizione del Regno Unito in qualità di leader globale delle IA. Lo sviluppo delle tecnologie IA può affrontare alcune delle sfide globali più urgenti, dal cambiamento climatico alle future pandemie. È inoltre sempre più riconosciuto a livello internazionale che i sistemi di IA richiedono nuove risposte normative per essere in grado di veicolare un'innovazione responsabile".

Il libro bianco adotta i principi disciplinati nel "Piano per la regolamentazione digitale", o "Plan for Digital Regulation" ovvero un approccio "proporzionato", attento al bilanciamento tra la valutazione del rischio e la "realizzazione di benefici e opportunità".

Definizione di sistema di Intelligenza Artificiale

Per stabilire una definizione di sistema di IA, l'atto in questione parte da due premesse. La prima costituisce la constatazione che, allo stato attuale, non esiste una definizione univoca di sistema di IA che trovi consenso in seno alla comunità internazionale. La seconda premessa fa riferimento al carattere volutamente adattabile e flessibile del testo normativo in questione ed espone il rischio potenziale nell'adottare una definizione rigida ed immutabile. Premesse tali considerazioni l'AI Regulation offre una definizione basata su due elementi caratteristici di ogni sistema di IA che necessitano di una specifica tutela normativa. Tali elementi sono:

1. L'adattabilità dell'IA, elemento che può rendere difficile spiegare l'intento o la logica dei risultati del sistema": durante la fase di "addestramento" i sistemi di IA sono portati alla deduzione di connessioni e all'elaborazione di schemi. Tale operazione può rendere complessa l'attribuzione dell'autore dell'operato in quanto le deduzioni logiche finali spesso non risultano essere distinguibili da quelle umane.
2. L'autonomia dell'IA, elemento che può rendere difficile l'attribuzione della responsabilità rispetto ai risultati": per autonomia in questo caso si fa espressamente riferimento al processo decisionale adottato nell'introduzione dei sistemi di IA che talvolta manca di supervisione o di controllo continuo da parte dell'agente umano.

La coesistenza dell'adattabilità e dell'autonomia rende difficile non solo l'accertamento dei risultati prodotti, ma anche la comprensione o l'individuazione dello schema logico adottato. La definizione proposta dal libro bianco, basata sui due elementi sopracitati, cerca di tenere conto di tali difficoltà e ad offrire al contempo la flessibilità necessaria per includere tecnologie future e per allinearsi a future definizioni normative di carattere internazionale.
Il quadro di riferimento è *context specific*. Tale caratteristica esclude una categorizzazione aprioristica delle tecnologie e dei livelli di rischio e parametra invece i sistemi di IA in base ai risultati dati in ambiti di applicazione specifici. Un esempio riportato all'interno dell'AI Regulation" è quello delle infrastrutture critiche. Secondo il Legislatore britannico classificare tutti i sistemi di IA applicati nelle infrastrutture critiche quali sistemi ad alto rischio, è un'operazione non proporzionata e non efficace. L'AI Regulation spiega che l'utilizzo di taluni sistemi di IA nelle infrastrutture critiche quali ad esempio l'"identificazione di graffi superficiali su macchinari" deve considerarsi quale attività dal rischio relativamente basso. L'approccio *context specific* consente alle Autorità di regolamentazione di effettuare un'operazione di bilanciamento tra i rischi derivanti dall'utilizzo dei sistemi di IA e i costi che "derivano dalla perdita di determinate opportunità".

Un'altra caratteristica importante dell'AI Regulation è l'attinenza dell'atto a cinque principi fondamentali:

1. "Sicurezza e robustezza";
2. "Trasparenza e spiegabilità adeguate";
3. "Equità";
4. "Responsabilità e governance";
5. "Contestabilità e possibilità di fare ricorso".

Tali principi si adattano alle linee guida dell'OCSE (Organizzazione per la Cooperazione e lo Sviluppo Economico) e sono inoltre stati oggetto di revisione in seguito alla pubblicazione del "documento programmatico sulla regolamentazione dell'IA". In particolare sono stati ampliati ulteriormente i concetti di governance e di robustezza ed è stata inclusa una riflessione specifica sulla privacy in seguito ad un confronto avvenuto con Il Centro per l'etica e l'innovazione dei dati (CDEI).

Implementazione dei principi

Per quanto riguarda l'applicazione pratica di tali principi e delle disposizioni generali dell'atto, l' AI Regulation attribuisce un ruolo fondamentale alla normativa tecnica. Al fine di garantire un'implementazione chiara ed efficace di tali principi, il Governo britannico ha redatto il "Roadmap to an effective AI assurance ecosystem", un documento che identifica nello specifico le aree di priorità in cui è necessario intervenire, definendo i ruoli e le responsabilità da assegnare a ciascun attore in seno all'ecosistema dei sistemi IA. Insieme al "Roadmap" è stato creato inoltre il "UK AI Standards Hub" al fine di incoraggiare l'utilizzo di normativa tecnica. Per aiutare gli innovatori a capire inoltre come applicare tecniche di valutazione e di controllo, il 7 giugno 2023 è stato redatto dal Governo britannico il Portfolio per l'implementazione delle tecniche di controllo qualità. Tali tecniche sono in linea con i principi prescritti dalle disposizioni dell'OECD e ricomprendono tecniche di Risk management, assessment e auditing.

Alcune tecniche previste sono le seguenti:

- "Impact assessment: Utilizzata per prevedere l'effetto di un sistema su ambiente, uguaglianza, diritti umani, protezione dei dati o altri risultati."
- "Impact evaluation : Simili alle tecniche di impact assessment (o valutazioni d'impatto), ma condotta dopo l'implementazione di un sistema in modo retrospettivo."
- "Bias audit o verifica dei pregiudizi: Valutazione degli input e degli output dei sistemi algoritmici per determinare l'eventuale presenza di pregiudizi ingiusti nei dati di input, nel risultato di una decisione o di una classificazione effettuata dal sistema."
- "Audit di conformità: Esame dell'aderenza di un'azienda alle politiche e alle procedure interne, alle normative esterne o ai requisiti legali. Tra le tipologie specializzate di audit di conformità vi sono gli audit di sistema e di processo e le ispezioni normative."
- "Tecniche di certificazione: Processi in cui un organismo indipendente attesta che un prodotto, un servizio, un'organizzazione o un individuo sono stati testati e soddisfatti rispetto a standard oggettivi di qualità o prestazioni."
- "Tecniche di Valutazione della conformità: Forniscono la garanzia che un prodotto, un servizio o un sistema fornito soddisfi le aspettative specificate o dichiarate, prima dell'immissione sul mercato. La valutazione della conformità comprende attività quali test, ispezioni e certificazioni."

4.3 Il Framework della Cina

Lo sviluppo della tecnologia legata ai sistemi di IA costituisce una priorità assoluta all'interno dell'agenda politica cinese[23] che, entro il 2030, punta a rendere la Cina il leader mondiale nel settore delle Intelligenze Artificiali.

A tal proposito la Risoluzione storica del Partito comunista cinese dichiara espressamente che l'"*intelligentazione* del Paese rappresenta il «traguardo storico nella lotta secolare del Partito»".[24] L'importanza politica dello sviluppo tecnologico e in particolar modo il ruolo centrale del cloud, dei big data e dei sistemi di IA è stata ribadita dallo stesso segretario generale del Partito comunista cinese Xi Jinping durante la XXXIV sessione di studio collettivo dell'Ufficio politico del XIX Comitato centrale del Partito, tenutosi il 10 ottobre 2021.[25] Il Governo cinese aveva espresso per la prima volta l'intenzione di investire nel settore delle intelligenze artificiali già nel 2012,[26] durante il XVIII Congresso nazionale del Partito comunista cinese, esprimendo la necessità di puntare sulla capacità di innovazione e di sviluppo scientifico delle IA di nuova generazione. Nel 2015 poi è stato il Consiglio di Stato cinese ad adottare lo "Schema d'azione per la promozione dello sviluppo dei big data", per poi adottare nuovamente nel 2017 il "Piano di sviluppo dell'IA di nuova generazione". Nell'anno seguente, il Ministero dell'Istruzione ha pubblicato infine il "Piano d'azione per l'innovazione dell'IA nelle Università". I piani strategici della Cina in ambito AI hanno tuttavia iniziato a destare la curiosità della comunità internazionale nel 2015, data in cui il Governo cinese ha iniziato ad emanare diversi documenti tecnici volti all'utilizzo dei sistemi di IA in diversi ambiti e settori.[27] Al 2015 risale in particolare il piano decennale "Made in China 2025", ovvero il piano programmatico contenente le linee guida per rendere la Cina il leader mondiale della produzione high-tech, inclusa l'AI. Nel 2017 è stato adottato invece il piano sistematico di strategia e di sviluppo dedicato specificatamente ai sistemi AI, trattasi del "New Generation Artificial Intelligence Development Plan"[28]. Tale documento getta le basi programmatiche per tutti gli sviluppi tecnologici e normativi in merito ai sistemi di IA. Gli organi principali atti all'organizzazione e all'implementazione del "New Generation Artificial Intelligence Development Plan" sono il Comitato consultivo per la strategia sull'IA, istituito nel novembre del 2017 e il Ministero della Scienza e della Tecnologia. Nonostante il ruolo di tale organi pubblici sia centrale nell'organizzazione e nello sviluppo del Development Plan, il Governo cinese ha previsto dei meccanismi a latere per incentivare e incoraggiare anche l'intervento del settore privato. Il Ministero della Scienza e della Tecnologia in particolare ha creato un gruppo di ricerca denominato "AI National Team" composto da aziende tech e volto alla ricerca di specifiche applicazioni dell'IA All'interno di tale gruppo rientrano i gruppi tech più influenti a livello nazionale ed internazionale.

Iniziative degne di nota sono state affidate anche a città e province al fine di delineare le proprie strategie locali di IA.

---

[23] Dal reportage https://en.people.cn/n3/2022/0210/c90000-9955688.html è possibile constatare che solo nel 2021 il Governo cinese ha investito 2,79 trilioni di yuan nella ricerca e nello sviluppo dei sistemi di IA. La cifra equivale a 2.79 punti del PIL.

[24] La Risoluzione storica è consultabile al seguente indirizzo:
htpp://www.gov.cn/zhengce/2021-11/16/content_5651269.htm

[25] Il discorso è disponibile al seguente link:
https://baijiahao.baidu.com/s?id=1722038734476920998&wfr=spider&for=pc

[26] https://dirittocinese.com/2022/10/03/disciplina-dellintelligenza-artificiale-e-intelligentizzazione-della-giustizia-in-cina/

[27] Tra questi documenti figura, ad esempio, il progetto Internet plus (互联网+), un progetto presentato dal primo ministro Li Keqiang durante l'inaugurazione della sessione annuale dell'Assemblea nazionale del popolo, il 5 marzo 2015, volto ad integrare cloud computing, big data ed internet mobile alle industrie tradizionali al fine di promuovere ed incrementare l'economia nazionale, fonte: G. Negro, Internet plus: un progetto strategico per lo sviluppo tecnologico, in "Orizzonte Cina", Vol. 7 (2016), p. 13

[28] H. Roberts, J. Cowls, J. Morley, M. Taddeo, V. Wang, L. Floridi, The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation, in "AI e Society", vol. 36 (2021).

La quarta sessione della XIII Assemblea Nazionale del Popolo ha poi votato e approvato nel 2021 le risoluzioni sul "Quattordicesimo Piano quinquennale per lo sviluppo nazionale, sociale ed economico e lo schema degli obiettivi per il 2035". Il 1° gennaio dell'anno successivo è invece entrata in vigore la Legge sul progresso scientifico e tecnologico. L'art. 38 di tale Legge attribuisce allo stato il "dovere di promuovere l'applicazione dei risultati della ricerca e dell'innovazione tecnologica". Per quanto riguarda i principi che governano la produzione e la distribuzione dei sistemi di IA invece il Comitato di esperti per la Governance nazionale ha pubblicato nel giugno del 2019 i "Principi di governance per l'IA di nuova generazione: per lo sviluppo di un'IA responsabile"[29]. All'interno del testo sono disciplinati i principi dell'affidabilità e dell'armonia, della Giustizia e dell'equità, della condivisione e dell'inclusività, della privacy e del rispetto e della controllabilità e della Sicurezza. Secondo tali disposizioni, i sistemi di IA devono essere affidabili, controllabili, tracciabili e spiegabili. Il legislatore è intervenuto nel 2021 per emanare la Legge sulla sicurezza dei dati.

La bozza di Regolamento sull'Intelligenza Artificiale generativa

Nell'aprile del 2023 è stata infine pubblicata dall'autorità "Cyberspace Administration of China" la bozza preliminare di Regolamento sull'Intelligenza Artificiale generativa, denominata "Measures for the Management of Generative Artificial Intelligence Service". Tale bozza si ascrive all'ambizione del Governo cinese di diventare leader mondiale nel settore delle AI entro il 2030, un'ambizione già delineata nel "New generation of Artificial Intelligence Development Plan" sopra citato. La bozza rispetta le disposizioni del "Codice etico di nuova generazione dell'IA" insieme alle disposizioni dei "Principi di governance dell'IA di nuova generazione", dove sono trattati i principi volti a sviluppare i sistemi di IA in condizioni di equità, sicurezza e affidabilità. Il 13 luglio 2023[30] alcune delle disposizioni presenti nella bozza sono state riviste e modificate dalla Commissione nazionale per lo sviluppo e la riforma. Il risultato è una bozza di Regolamento composta da 24 articoli suddivisi in cinque sezioni la cui entrata in vigore è prevista entro il mese di agosto 2023. La nuova bozza di Regolamento annovera tra le sue fonti di riferimento la "Personal Information Protection Law of the People's Republic of China", ovvero la Legge cinese sulla protezione delle informazioni personali, la Cybersecurity Law of the People's Republic of China, ovvero la Legge nazionale sulla sicurezza delle reti e la "Data Security Law of the People's Republic of China", ovvero la Legge sulla sicurezza dei dati.

La suddivisione del Regolamento e i rispettivi articoli

La bozza è suddivisa in cinque distinte sezioni. La prima tratta delle disposizioni generali, la seconda dello sviluppo tecnologico e della governance, la terza degli standard di servizio, la quarta della Supervisione, dell'ispezione e delle responsabilità legali e la quinta delle disposizioni complementari.

Sezione 1: Disposizioni generali

La sezione che concerne le disposizioni generali ricopre i primi quattro articoli della bozza e tratta lo scopo del Regolamento, il suo ambito di applicazione, la ratio del Regolamento e i requisiti principali che i sistemi di IA generativa devono rispettare. Lo scopo del Regolamento è quello di "promuovere il sano sviluppo e l'applicazione standardizzata dell'IA generativa, di salvaguardare la sicurezza nazionale e gli interessi pubblici sociali, e proteggere i diritti e gli interessi legittimi dei cittadini, delle persone giuridiche e di altre organizzazioni." (art. 1), allo scopo di "incoraggiare l'innovazione e lo sviluppo dell'IA generativa e attua una supervisione dei servizi di IA generativa che sia inclusiva e prudente, oltre che differenziata e gerarchica" (art. 3).

---

[29] Il testo è disponibile al link https://baijiahao.baidu.com/s?id=1636567627966397830&wfr=spider&for=pc
Per la traduzione inglese si veda L. LASKAI, G. WEBSTER, Translation: Chinese expert group offers 'governance principles' for 'responsible AI', in New America, 2019.
[30] https://www.wired.it/article/intelligenza-artificiale-pentagono-difesa-stati-uniti

L'art. 2 disciplina invece l'ambito di applicazione del Regolamento che si ascrive "alla ricerca, allo sviluppo e all'utilizzo di prodotti di IA generativa per fornire servizi al pubblico nel territorio della Repubblica popolare cinese". La bozza di Regolamento si applica quindi a qualsiasi applicazione tecnologica in grado di produrre "testi, immagini, suoni, video, codici e altri contenuti basati su algoritmi". Come è stato fatto notare dalla dottrina, si rileva nel testo una generale tendenza al controllo operativo sull'implementazione dei sistemi di IA generativa, al fine di evitare, probabilmente, che il funzionamento tecnico di tali sistemi comprometta la sicurezza nazionale del Paese e dell'ordine pubblico. In tal senso l'art. 4 dispone che la "fornitura e l'utilizzo dei servizi di IA generativa devono rispettare la morale e l'etica sociale e attenersi ai valori fondamentali del socialismo" oltre che essere conformi "alle leggi e ai regolamenti amministrativi". Occorre notare che i parametri della morale e dell'etica, disciplinati dall'art. 4, costituiscono requisiti di carattere generale, la cui interpretazione si presta a margini di discrezionalità particolarmente ampi, anche in sede applicativa e processuale.

Sezione 2: Sviluppo tecnologico e governance

La seconda sezione del Regolamento è composta dagli artt. 5-8 e tratta della governance dei sistemi di IA generativa. Gli artt. 5 e 6 in particolare delineano i tratti principali della governance come segue: "incoraggiare l'applicazione innovativa della tecnologia di IA generativa in vari settori e campi, generare contenuti positivi, sani ed edificanti di alta qualità, esplorare e ottimizzare gli scenari applicativi e costruire un ecosistema applicativo" (art. 5); e, ancora, "incoraggiare l'innovazione indipendente delle tecnologie di base come gli algoritmi di IA generativa, i framework, i chip e le piattaforme software di supporto, effettuare scambi e cooperazioni internazionali su base paritaria e reciprocamente vantaggiosa e partecipare alla formulazione di regole internazionali relative all'IA generativa" (art. 6).

L'art. 7 attribuisce invece la responsabilità generale delle eventuali conseguenze, date dai sistemi di IA generativa, in capo ai fornitori. L'articolo attribuisce inoltre, in capo ai fornitori, l'obbligo di effettuare una valutazione dei rischi potenziali ed effettivi di sicurezza preventiva all'immissione sul mercato dei sistemi di IA generativa. Tale valutazione dovrà inoltre essere a sua volta valutata ed autorizzata attraverso il rilascio di una licenza amministrativa.

Sezione 3: Standard di servizio

La terza sezione disciplina invece gli standard di servizio e comprende gli artt. 9-15. L'art. 8 della seconda sezione attribuisce in capo ai fornitori l'obbligo di creare "regole di etichettatura chiare, specifiche e operative". A tale disposizione si affiancano gli artt. 7 e 13 secondo i quali i fornitori dei sistemi di IA generativa sono responsabili della rimozione e della cancellazione dei dati personali, nei casi in cui la permanenza arrechi danno ai soggetti interessati.

Stante le disposizioni dell'art. 9 invece "i fornitori devono assumersi la responsabilità legale di essere produttori di contenuti informativi di rete e adempiere agli obblighi relativi alla sicurezza delle informazioni di rete". Se si tratta di informazioni personali, devono assumersi le responsabilità connesse con il trattamento delle informazioni personali ai sensi della legge e adempiere ai propri obblighi di protezione delle informazioni personali".

All'interno della terza sezione sono trattati anche i diritti individuali; in particolar modo, l'art. 15, dispone che debbano essere pubblicizzate "le procedure di trattamento e le tempistiche di feedback, accettando e gestendo tempestivamente i reclami e le segnalazioni del pubblico e fornendo un feedback sui risultati delle loro azioni".

Sezione 4: Supervisione, ispezione e responsabilità legali

La sezione quattro tratta della supervisione, delle ispezioni e delle responsabilità legali e comprende gli artt. 16-21. In tale sezione l'art. 19 attribuisce ai dipartimenti competenti il compito di condurre e supervisionare l'ispezione dei servizi di IA generativa in base ai loro compiti. Dispone inoltre che "il fornitore collabora in conformità con la legge, spiegando la fonte, la scala, il tipo, le regole di etichettatura, il meccanismo dell'algoritmo, ecc. dei dati di addestramento come richiesto, e fornendo il supporto e l'assistenza tecnica e di dati necessari. L'art. 20 tratta invece delle possibili sanzioni disponendo che, in caso di violazione delle disposizioni, saranno erogate sanzioni dalle Autorità competenti in conformità alle leggi sulla sicurezza informatica, sulla sicurezza dei dati, sulla protezione delle informazioni personali e sul progresso scientifico e tecnologico. Qualora si tratti di una violazione della gestione della sicurezza pubblica o di un reato, l'articolo dispone che saranno imposte sanzioni amministrative e penali in conformità con la legge.

Sezione 5: Disposizioni complementari

La sezione 5 tratta delle disposizioni complementari e comprende gli artt. 22-24. Nello specifico l'art. 22 tratta delle definizioni della tecnologia di IA generativa, dei fornitori di tali servizi e degli utenti. In particolare, all'interno della tecnologia di IA generativa vengono ricompresi tutti i modelli e le tecnologie correlate "che hanno la capacità di generare contenuti come testo, immagini, audio e video". L'art. 24 afferma invece che l'entrata in vigore delle disposizioni è stata il 15 agosto 2023.

Conclusioni

Il Regolamento cinese sui sistemi di IA generativa rientra all'interno del piano programmatico più vasto e ambizioso che porterebbe la Cina a diventare leader mondiale nei sistemi di IA entro il 2030. La volontà di regolamentare l'utilizzo e lo sviluppo delle IA generative si ascrive ad una modello di governance centralizzato, fondato sul controllo statale delle infrastrutture tecnologiche, in particolare le infrastrutture che si occupano di "realizzare un costante monitoraggio del flusso comunicativo veicolato online a presidio di interessi nazionali che giustificano la prioritaria salvaguardia della sicurezza del Paese".

4.4 L'OCSE (Organizzazione per la Cooperazione Economica Europea)

L'OCDE (Organisation de Coopération et de Développement Èconomiques) o OECD (Organisation for Economic Co-operation and Development) è un'organizzazione internazionale, istituita il 14 dicembre 1960 ed entrata in vigore il 30 settembre 1961 sostituendo la precedente OECE (Organizzazione per la Cooperazione Economica Europea), a sua volta istituita nel 1948 all'interno del Piano Marshall volto alla ricostruzione dell'economia europea dopo la seconda guerra mondiale. Fondata da nove Paesi tra cui l'Italia, l'OCSE conta oggi tra i suoi membri 36 paesi. Lo scopo principale dell'organizzazione è la creazione di politiche volte al miglioramento del benessere sociale ed economico dei cittadini. Le politiche finalizzate allo scopo sono "la realizzazione dei più alti livelli di crescita economica e di occupazione sostenibile" insieme all'integrazione dei mercati, al fine di" favorire la concorrenza e la competitività, mantenendo al contempo la stabilità economico-finanziaria". L'organizzazione rappresenta un luogo di confronto e di scambio sulle diverse esperienze politiche a livello internazionale dove poter trovare risposte e soluzioni a problemi di natura economico-sociale comuni ai diversi Stati membri.

Gli strumenti legali dell'OCSE nell'ambito dei sistemi di Intelligenza Artificiale

Per quanto riguarda i sistemi di IA, l'OCSE ha contribuito a trattare la materia attraverso la pubblicazione di diversi documenti, tra cui raccomandazioni, pareri, principi e metriche. Saranno qui di seguito trattati quelli più rilevanti.

AI Principles

In primo luogo, nel 2019, sono stati pubblicati gli "AI Principles" ovvero un insieme di raccomandazioni e pareri elaborati da 50 esperti e poi sottoscritti dai Governi degli Stati membri, oltre che da sei governi aggiuntivi, per creare sistemi di IA affidabili e sicuri. I principi sono stati adottati nel maggio dello stesso anno e dispongono quanto segue:

1. "L'IA dovrebbe portare benefici alle persone e al pianeta, favorendo la crescita inclusiva, lo sviluppo sostenibile e il benessere."
2. "I sistemi di IA dovrebbero essere progettati in modo da rispettare lo stato di diritto, i diritti umani, i valori democratici e la diversità, e dovrebbero includere adeguate salvaguardie – ad esempio, consentendo l'intervento umano ove necessario – per garantire una società equa e giusta."
3. "I sistemi di IA devono essere caratterizzati da trasparenza e divulgazione responsabile, per garantire che le persone capiscano quando sono coinvolte e possano contestarne i risultati."
4. "I sistemi di IA devono funzionare in modo robusto, sicuro e protetto per tutto il loro ciclo di vita e i rischi potenziali devono essere costantemente valutati e gestiti."
5. "Le organizzazioni e gli individui che sviluppano, implementano o gestiscono i sistemi di IA devono essere ritenuti responsabili del loro corretto funzionamento in linea con i principi di cui sopra."

OECD Framework

Nel 2022 invece è stato redatto un documento rilevante denominato "OECD Framework for the Classification of AI Systems: a tool for effective AI policies". Tale framework è nato con l'obiettivo di classificare i diversi sistemi di IA in base al loro impatto e ai rischi potenzialmente causati.

Il Framework distingue i modelli di IA classificando diverse dimensioni applicative:

- Persone e pianeta;
- Contesto economico nell'implementazione dei sistemi di IA;
- Data e input;
- Modelli di IA;
- Compiti e output.

Persone e pianeta

Questo sottogruppo comprende gli utenti, sia individuali che categorie di individui, che interagiscono con il sistema di IA con scopi interattivi o per applicazioni specifiche in contesti oggettivi. Ad esempio, settori come il sistema giudiziario o la selezione del personale possono essere inclusi in questa categoria. In tali settori, dove la responsabilità è cruciale, l'intervento umano nell'uso dei sistemi di IA è essenziale. Questi sistemi possono automatizzare compiti e aumentare la produttività umana.

Contesto economico nell'implementazione dei sistemi di IA

Il concetto di contesto economico si riferisce all'ambiente economico e settoriale in cui viene attuato un sistema di IA. In linea generale, si valuta l'applicazione dell'IA e si fornisce una descrizione dell'ambito organizzativo e funzionale per il quale il sistema di IA è stato sviluppato. Inoltre, si specifica il settore in cui l'IA viene implementata, ad esempio nel campo della sanità o delle finanze. Questo implica la chiara definizione delle funzioni, del modello di business, degli aspetti critici e della maturità tecnologica legati all'implementazione dell'IA.

Ogni settore industriale identificato rappresenta un contesto unico, con implicazioni diverse in termini di struttura e regolamentazione per i sistemi di IA. Anche se gli stessi tipi di sistemi IA possono essere impiegati in più settori per svolgere compiti simili in diverse aree funzionali, le conseguenze e gli orientamenti da seguire possono variare.

Ad esempio, un algoritmo predittivo utilizzato per ottimizzare la logistica potrebbe avere effetti diversi se applicato al settore delle assunzioni aziendali. Gli attori chiave in questa dimensione sono gli operatori di sistema che pianificano progettano, gestiscono e monitorano i sistemi di IA. Il Framework integra l'approccio al rischio proposto dalle regole e azioni stabilite dalla Commissione europea nell'aprile 2021.

Dati e input

Questa categoria riguarda l'elaborazione dei dati provenienti da insiemi di regole, sia di origine umana che artificiale (algoritmi). I dati e gli input sono razionalizzati dalla IA in base alla provenienza, al metodo di raccolta, alla struttura, al formato alle proprietà e alle caratteristiche tecniche. Attualmente, l'espansione e lo sviluppo dell'IA sono favoriti dalla maturità delle reti neurali e dalla disponibilità di grandi quantità di dati insieme alla potenza di calcolo.

Modelli di IA

Gli elementi che costituiscono un modello di IA sono le caratteristiche che consentono la sua classificazione in un particolare sistema o modello. La classificazione e la costruzione di processi di inferenza sono cruciali per l'assegnazione delle politiche di utilizzo. Proprietà chiave possono variare tra diversi modelli di IA, includendo la trasparenza, la comprensibilità e le implicazioni sulla protezione dei diritti umani, come la privacy.

Compiti e Output

Questa categoria approfondisce i compiti eseguiti dall'IA, come la classificazione dei dati, il rilevamento di pattern o anomalie, le previsioni, la personalizzazione, il supporto all'interazione uomo-tecnologia e l'ottimizzazione dei processi.

Ciclo vitale dei sistemi di IA

Il ciclo di vita dei sistemi di IA è importante per la loro classificazione in modelli specifici. I principi chiave evidenziati dall'OCSE riguardo al ciclo di vita includono la valutazione dei momenti di:

- Pianificazione e progettazione del processo, raccolta ed elaborazione dei dati e costruzione del modello.
- Verifica e convalida dei dati.
- Dispiegamento dei dati.
- Funzionamento e monitoraggio dei dati.

Il Framework dell'OCSE adotta questo modello come una struttura complementare per comprendere le caratteristiche tecniche fondamentali di un sistema di IA. La categorizzazione dei sistemi di IA è legata alle diverse fasi del ciclo di vita dell'IA al fine di individuare i vari attori principali associati a ciascuna dimensione, creando così una connessione con le responsabilità e la gestione del rischio. Il Framework incorpora l'approccio al rischio delineato nell'AI Act con l'obiettivo di assicurare fiducia, efficienza ed affidabilità nei confronti dei sistemi di IA.

Strumenti e metriche per l'affidabilità

L'OCSE ha sviluppato una serie di criteri e strumenti per l'IA che consentono alle organizzazioni di valutare e migliorare le proprie pratiche di governance dell'IA. Nel corso del 2023, è stata rilasciata un'edizione aggiornata di questi criteri e strumenti, che incorpora miglioramenti e nuove caratteristiche. Il documento distingue due elenchi: gli strumenti e le metriche. Entrambi comprendono varie tecniche e approcci suddivisi per ambito. Ad esempio, i criteri di ricerca all'interno della sezione degli strumenti possono essere tecnici, procedurali o educativi. Un esempio di strumento è l'ECPAIS, che rappresenta il "Programma di Certificazione Etica per Sistemi Autonomi e Intelligenti", sviluppato dall'organizzazione IEEE. Lo scopo di questo strumento è formulare specifiche per i processi di certificazione e di etichettatura, in modo da promuovere la trasparenza, la responsabilità e la riduzione dei pregiudizi algoritmici negli AIS (Sistemi Autonomi e Intelligenti). Un'altra metrica, ad esempio, è il FIP (Fréchet Inception Distance), che rappresenta uno strumento utilizzato per valutare la qualità delle immagini generate da un modello generativo, come una GAN (Rete Generativa Avversaria). A differenza dell'IS (Indice di Inizio Precedente), che considera esclusivamente la distribuzione delle immagini generate, la FID (Distanza di Inizio Frechet) confronta la distribuzione delle immagini generate con quella di un insieme di immagini reali noto come *ground truth*.

Le metriche ricomprendono le seguenti indicazioni:

- Indice di Preparazione per l'IA: un insieme di strumenti che valuta il livello di prontezza di una nazione all'adozione e all'utilizzo di sistemi di IA. Questa valutazione considera fattori sociali e tecnologici, tra cui ricerca e sviluppo infrastrutture tecniche, adozione da parte delle imprese, strategie governative e accettazione sociale. Nel 2021, gli Stati Uniti si sono posizionati in cima all'Indice di Preparazione per l'IA dell'OCSE con un punteggio totale di 0,93. Al secondo e terzo posto si sono collocati rispettivamente la Corea del Sud (0,88) e Singapore (0,87).
- Quadro Etico dell'IA: un insieme di linee guida volte a garantire che lo sviluppo e l'implementazione dei sistemi di IA rispettino valori etici norme sociali e diritti umani. Questo quadro si basa su cinque principi fondamentali: Valori centrati sull'umanità, Equità, Trasparenza, Solidità, Sicurezza e Responsabilità. L'Unione europea ha sviluppato il proprio quadro etico per l'IA, il quale abbraccia la maggior parte di questi principi, come responsabilità, trasparenza e IA orientata all'umanità.
- Set di Strumenti per la governance dell'IA: un insieme di buone pratiche e suggerimenti rivolti a Governi e altre parti interessate per promuovere una governance dell'IA responsabile ed efficace. Questo set di strumenti si compone di sei componenti principali: Strategie Nazionali per l'IA, Valutazione dell'Impatto, Quadro di Governance, Normative e Standard, Sviluppo delle Capacità e Cooperazione Internazionale.

4.5 Confronto e analisi comparativa dei Framework di regolazione dell'Intelligenza Artificiale

Un'analisi comparativa dei Framework di regolazione dell'IA adottati da Regno Unito, Cina, Stati Uniti e OCSE permette di esplorare come queste potenze mondiali e organizzazioni internazionali affrontino i rischi connessi all'IA, promuovano l'innovazione e influenzino gli interessi globali.
I principi e gli approcci normativi variano significativamente tra questi attori. Il Regno Unito, ad esempio, favorisce un ambiente normativo orientato all'innovazione che minimizza le barriere regolatorie, incoraggiando così lo sviluppo tecnologico. Al contrario, la Cina adotta un approccio più centralizzato, esercitando un controllo governativo esteso su tutto il ciclo di vita dell'IA, dalla ricerca allo sviluppo fino all'implementazione, enfatizzando la sicurezza e la sovranità nazionale. Negli Stati Uniti, il National Institute of Standards and Technology (NIST) ha introdotto l'"AI Risk Management Framework", un modello che si basa sulla gestione del rischio e incoraggia l'adozione volontaria di pratiche responsabili. Questo approccio cerca di bilanciare la flessibilità con la necessità di mantenere standard elevati di sicurezza e affidabilità. D'altra parte, l'OCSE non stabilisce Leggi ma fornisce principi e raccomandazioni destinati a orientare i Governi verso Regolamenti che supportino uno sviluppo sostenibile e inclusivo dell'IA.
L'approccio regolativo adottato può avere impatti significativi sull'innovazione. Nel Regno Unito, la flessibilità del quadro normativo è vista come un catalizzatore per l'innovazione, permettendo a startup e PMI di esplorare e sviluppare nuove tecnologie senza l'onere di regolamenti oppressivi. Questo clima di "libertà regolata" è pensato per attrarre talenti e investimenti, rendendo il Regno Unito un hub per le nuove tecnologie di IA. Il modello cinese, pur essendo efficace nell'accelerare l'innovazione all'interno dei suoi confini, può limitare la collaborazione internazionale e l'apertura a causa delle sue politiche di controllo rigoroso. Questo può sfociare in una sorta di isolamento tecnologico, per il quale l'IA cinese rischia di svilupparsi in modo divergente rispetto alle tendenze globali. Negli USA, l'accento posto sul rispetto volontario delle linee guida federali consente alle aziende di adattarsi rapidamente ai cambiamenti tecnologici. Questo può però portare a una regolamentazione ineguale, con alcune aziende che potrebbero non impegnarsi a fondo nella gestione dei rischi.
L'OCSE cerca di mediare tra queste due estremità, promuovendo principi che incoraggino l'innovazione e allo stesso tempo salvaguardino i diritti e la sicurezza pubblica. Queste differenze nei principi regolatori riflettono divergenti filosofie politiche e impostazioni economiche, e influenzano direttamente la capacità di ciascun sistema di promuovere o frenare l'innovazione nel campo dell'IA. Una cooperazione e un dialogo internazionale efficaci sono quindi essenziali per armonizzare questi approcci, garantendo che l'IA possa svilupparsi in maniera che i benefici siano diffusi nella comunità globale, rispettando al contempo norme etiche e di sicurezza condivise.

Gestione del rischio e misure implementate

Nell'ambito della regolamentazione dell'IA, la gestione dei rischi e l'implementazione di salvaguardie efficaci sono fondamentali per assicurare che la tecnologia sia sviluppata e utilizzata in modo sicuro ed etico. L'analisi dei diversi approcci rivela variazioni significative nella priorità data ai rischi e nelle misure di mitigazione proposte. Il Regno Unito si distingue per il suo approccio flessibile e orientato all'innovazione, ma questa flessibilità porta con sé sfide specifiche nella gestione dei rischi. Il Framework britannico tende a delegare una notevole responsabilità alle aziende private per l'autovalutazione dei rischi e l'implementazione delle misure di mitigazione.

Questo approccio è visibile nel Libro Bianco sull'IA, che enfatizza la responsabilità delle aziende nel garantire che i loro sistemi di IA siano sicuri e affidabili. Tuttavia, la mancanza di prescrizioni rigide può portare a discrepanze nel modo in cui i rischi vengono gestiti, a seconda delle capacità e delle risorse delle singole aziende. Per compensare, il Governo britannico incoraggia la collaborazione tra il settore pubblico e quello privato e sostiene iniziative come il Centre for Data Ethics and Innovation, che si propone di sviluppare una governance dell'IA responsabile e orientata all'etica.
In contrasto con l'approccio britannico, la Cina adotta un modello di regolamentazione altamente centralizzato. Il Governo cinese impone rigorosi controlli su tutti gli aspetti della ricerca, dello sviluppo e dell'implementazione dell'IA. Questi controlli sono intesi a mitigare i rischi per la sicurezza e l'ordine pubblico, oltre a proteggere i diritti dei cittadini. Le recenti regolamentazioni sull'IA generativa e i principi di governance delineati per l'IA riflettono un impegno verso sistemi di IA sicuri, controllabili e trasparenti. Tuttavia, questo controllo statale potrebbe anche limitare l'innovazione e la collaborazione internazionale, e c'è il rischio che tali misure di controllo possano essere usate per scopi di sorveglianza e repressione piuttosto che per la protezione dei diritti individuali. Gli Stati Uniti, attraverso il RMF (Risk Management Framework) del NIST, hanno adottato un approccio basato sulla gestione dei rischi che incoraggia le organizzazioni a seguire volontariamente le linee guida per la sicurezza e l'affidabilità dei sistemi di IA. Questo approccio mira a promuovere l'innovazione consentendo al contempo una certa flessibilità nell'applicazione delle norme. Il Framework enfatizza l'importanza della trasparenza, della responsabilità e della resilienza dei sistemi di IA. Tuttavia, la natura volontaria di queste linee guida può portare a un'adozione incoerente, con alcune aziende che potrebbero non adottare pienamente le pratiche raccomandate.
Infine l'OCSE, tramite i suoi principi sull'IA, ha cercato di stabilire un terreno comune per la gestione dei rischi associati all'IA a livello internazionale. I principi dell'OCSE sono stati adottati da numerosi Paesi e mirano a promuovere un IA che rispetti i diritti umani e i valori democratici, inclusi la sicurezza e la privacy. L'organizzazione sottolinea la necessità di trasparenza e contestabilità dei sistemi di IA, richiedendo che le decisioni prese dagli algoritmi siano comprensibili e contestabili dagli utenti. Sebbene questi principi forniscono una solida base per la sicurezza e l'etica nell'IA, la loro efficacia dipende dalla volontà e dalla capacità dei singoli Paesi di implementare regolamentazioni che li rispettino. L'armonizzazione delle politiche di IA a livello globale è un'ambizione dell'OCSE, che promuove un approccio coordinato per affrontare rischi transnazionali come la manipolazione delle elezioni e la diffusione di disinformazione. Tuttavia, la diversità di approcci tra Regno Unito, Cina e Stati Uniti, che riflette variazioni nelle priorità politiche e nei contesti economici, complica la creazione di un framework unificato. Di fronte a queste differenze normative, è essenziale una maggiore cooperazione internazionale e la creazione di un Framework globale che possa essere adottato e adattato localmente. Questo dovrebbe includere standard condivisi per la trasparenza, responsabilità e gestione del rischio, arricchito da meccanismi per la condivisione delle migliori pratiche e delle lezioni apprese. Un tale sforzo congiunto potrebbe facilitare una governance più efficace dell'IA a livello internazionale.

Il rapporto tra la regolamentazione EU e i Framework internazionali

L'approccio dell'Unione europea alla regolamentazione dell'IA è cristallizzato nell'AI Act. Quando comparato con i Framework internazionali degli Stati Uniti, Regno Unito, Cina e gli orientamenti dell'OCSE, l'AI Act si distingue per il suo approccio dettagliato, che cerca di bilanciare l'innovazione tecnologica con rigorose salvaguardie per i diritti dei cittadini e la sicurezza.

Uno degli aspetti chiave del AI Act è la sua classificazione dei sistemi AI basata sui rischi, dove la valutazione dell'Alto Rischio non è delegata alle aziende, ma alcuni casi d'uso vengono esplicitamente individuati nel testo del Regolamento, come discusso nei capitoli precedenti.
Diversamente dagli approcci prevalentemente volontari come quello statunitense, che si basano sull'autoregolamentazione guidata da principi del Risk Management Framework del NIST, l'UE impone requisiti legali specifici, e sanzioni di forte impatto nel caso in cui questi requisiti non vengano rispettati. Questo sistema di classificazione rischia di essere più restrittivo, ma fornisce chiarezza legale e protezione all'utente, tentando di prevenire i danni prima che possano verificarsi.
In confronto con il Regno Unito, che adotta un modello pro-innovazione con Regolamenti relativamente flessibili per promuovere lo sviluppo tecnologico, l'AI Act è più prescrittivo.
Invece, guardando alla Cina, l'approccio centralizzato e il controllo governativo sull'IA contrastano con la tendenza dell'UE di incoraggiare una vasta partecipazione delle parti interessate, inclusi cittadini e gruppi della società civile, nella formazione della politica sull'IA. Il modello cinese, che privilegia la sicurezza e il controllo statale, potrebbe essere più efficace nel coordinare e implementare rapidamente le politiche, ma è anche critico per la potenziale limitazione delle libertà individuali e per le implicazioni etiche che possono divergere dalle norme europee.
Infine, rispetto agli orientamenti dell'OCSE, che promuovono principi non vincolanti per guidare lo sviluppo responsabile dell'IA tra i paesi membri, l'AI Act fornisce un quadro legislativo concreto che potrebbe servire da modello per una regolamentazione internazionale. L'OCSE si concentra più sulla promozione del dialogo e della cooperazione internazionale, mentre l'UE sta attivamente stabilendo norme che potrebbero influenzare gli standard globali in tema di IA.

5. Proof of Concept

I PoC (Proof of Concept) sono considerati sempre più importanti per verificare le modalità di attuazione della normativa tecnica e per guidare l'implementazione di nuovi dispositivi normativi, come l'AI Act. Questi non solo dimostrano la fattibilità di soluzioni innovative, ma fungono anche da ponte tra la teoria normativa e la pratica applicativa, offrendo una piattaforma per testare, in ambiente controllato, l'efficacia e la conformità dei sistemi di IA rispetto agli standard. Per Enti come Accredia, i PoC assumono una significativa rilevanza in quanto strumenti essenziali per comprendere i processi di accreditamento, ispezione e certificazione di tali tecnologie. Questi processi sono fondamentali per garantire che i sistemi di IA non solo aderiscano alle norme di qualità e sicurezza stabilite dalle norme tecniche, ma siano anche implementati in modo da rispettare i rigorosi requisiti dell'AI Act europeo. L'adozione di linee guida e norme tecniche, come la ISO/IEC TR 24027:2021 e la ISO/IEC 42001:2023 nel contesto dei PoC, serve a illustrare concretamente come i sistemi di IA possano essere progettati e valutati per assicurare che i bias siano minimizzati e che la gestione della qualità sia mantenuta a livelli ottimali. Questa verifica attraverso i PoC è essenziale per il processo di accreditamento, in quanto fornisce le prove necessarie che i sistemi rispettino i criteri qualitativi richiesti per l'approvazione e la certificazione finale. Inoltre, i PoC facilitano l'elaborazione di linee guida dettagliate per le ispezioni, essendo questi capaci di identificare specifiche aree di rischio e di efficacia che gli ispettori possono poi monitorare e valutare. Attraverso i PoC, gli Enti di accreditamento possono sviluppare procedure di ispezione più mirate ed efficaci, che si traducono in processi di certificazione più affidabili e trasparenti. L'implementazione dell'AI Act richiede una comprensione approfondita di come le normative possano essere applicate nella pratica, e i PoC offrono una visione anticipata di potenziali sfide e soluzioni. Sono strumenti vitali per valutare la conformità dei sistemi di IA, non solo rispetto alla normativa tecnica esistente, ma anche per anticipare come queste tecnologie possano adattarsi ai nuovi requisiti legislativi.

5.1 I PoC in ambito medico: la gestione dei bias in sistemi di IA per la *detection* del melanoma e per la stratificazione dei pazienti con sclerosi multipla

Nel contesto biomedicale, i PoC illustrano l'integrazione dell'IA nella pratica clinica, mettendo in luce come questa tecnologia possa rivoluzionare la gestione delle malattie. Ci concentreremo, in particolare, sulla *detection* del melanoma e sulla stratificazione dei pazienti con sclerosi multipla. Questi casi di studio non solo dimostrano la capacità dell'IA di supportare le decisioni cliniche, ma sottolineano anche l'importanza di una procedura che comprende la raccolta e preparazione dei dati, lo sviluppo e la valutazione dei modelli, e infine il rilascio del sistema.

La conformità a normative specifiche, come la ISO/IEC TR 24027:2021, è essenziale per assicurare che i sistemi di IA presentino un bias accettabile per l'uso clinico. Questa linea guida pubblicata dall'ISO è particolarmente rilevante perché mira a eliminare potenziali pregiudizi nei sistemi di IA, garantendo che le diagnosi e i trattamenti proposti non siano solo accurati, ma anche equi e imparziali. La verifica di conformità secondo la ISO/IEC TR 24027:2021 è un passaggio cruciale, che assicura che i sistemi di IA rispettino i principi di affidabilità, supportando così l'implementazione del recentemente introdotto AI Act. L'AI Act richiede che i sistemi di IA siano trasparenti, etici, e che operino senza discriminazioni, integrando requisiti normativi stringenti per prevenire rischi ai diritti fondamentali e alla sicurezza delle persone. I PoC nel settore biomedicale rappresentano quindi non solo un esercizio di conformità tecnica, ma anche un allineamento ai valori e alle norme europee sull'uso etico dell'IA. Parallelamente, nel settore della Pubblica Amministrazione, il caso di studio con l'INAIL (Istituto Nazionale per l'Assicurazione contro gli Infortuni sul Lavoro) evidenzia l'importanza della norma ISO/IEC 42001:2023 sul Quality Management per i sistemi di IA. Questa norma si concentra sull'implementazione di sistemi di gestione della qualità, fondamentali per garantire che le decisioni prese dai sistemi di IA siano consistenti, riproducibili e affidabili. In un ambiente complesso come quello della Pubblica Amministrazione, dove le decisioni possono avere impatti ampi e significativi sulla vita dei cittadini, l'adeguamento a tali norme assicura che ogni processo decisionale automatizzato sia sottoposto a controlli rigorosi prima di essere implementato. Questo approccio è particolarmente pertinente nell'ottica dell'AI Act, che sottolinea l'importanza di una gestione adeguata del rischio e della qualità nell'IA. La norma ISO/IEC 42001:2023, quindi, non solo risponde a esigenze tecniche, ma si allinea anche agli obiettivi dell'AI Act per promuovere un utilizzo responsabile dell'IA, basato su elevati standard di qualità. La necessità di aderire a norme tecniche rigorose e riconosciute globalmente è un presupposto non solo per la conformità normativa, ma anche per il rafforzamento della posizione competitiva a livello internazionale nel settore dell'IA.

5.1.1 La norma ISO/IEC TR 24027:2021: sommario dei contenuti

La linea guida ISO/IEC TR 24027:2021 si concentra sulla gestione dei pregiudizi (bias) nei sistemi di IA, evidenziando come questi possano influenzare negativamente i processi decisionali assistiti da IA. Nella linea guida dell'ISO viene stabilito un quadro concettuale per identificare, comprendere e affrontare i pregiudizi in modo da minimizzare gli impatti negativi e promuovere l'equità nelle decisioni automatizzate.

Definizione e tipologie di pregiudizio

Il pregiudizio è descritto come una differenza sistematica nel trattamento di persone, gruppi o oggetti. Questi possono essere di natura sociale o statistica e possono manifestarsi o essere amplificati a vari livelli del sistema, sia a livello di input (dati), sia nel processo decisionale (algoritmi), sia nel comportamento del sistema. Inoltre, la ISO/IEC TR 24027:2021 distingue tra:

- Pregiudizio desiderato: necessario per il funzionamento corretto del sistema di IA secondo i suoi obiettivi specifici.
- Pregiudizi indesiderati: emergono involontariamente e non sono allineati con gli obiettivi del sistema, portando a decisioni ingiuste o discriminatorie.

Gestione e impatto dei pregiudizi

È cruciale riconoscere che mentre alcuni pregiudizi sono essenziali per il funzionamento dei sistemi di IA, i pregiudizi indesiderati possono essere introdotti accidentalmente durante lo sviluppo o l'uso del sistema.

Questi possono compromettere l'equità, conducendo a decisioni che discriminano o danneggiano certi gruppi o individui. La norma enfatizza l'importanza di trattare i pregiudizi indesiderati attraverso pratiche metodiche di raccolta e analisi dei dati e la continua valutazione delle performance del sistema IA.

Equità e correlazione con il pregiudizio

L'equità, secondo la linea guida ISO, è un trattamento o un risultato che rispetta le norme stabilite e non è influenzato da favoritismi o discriminazioni ingiuste. Questo concetto è distinto, ma strettamente correlato, al pregiudizio, poiché i pregiudizi sono uno dei molti fattori che possono compromettere l'equità di un sistema. Gli sviluppatori sono invitati a considerare come i loro sistemi trattano diversi gruppi di utenti e ad implementare controlli per mitigare gli impatti negativi dei pregiudizi indesiderati.

Fonti di pregiudizio nei sistemi di IA

La norma identifica varie fonti di pregiudizio nei sistemi di IA, tra cui:

- Pregiudizi introdotti dalla progettazione: risultanti da carenze strutturali nel design del sistema.
- Pregiudizi derivanti da pregiudizi cognitivi umani: influenze umane che possono alterare il comportamento del sistema.
- Pregiudizi nei set di dati: pregiudizi esistenti nei dati usati per addestrare i modelli di IA, che possono essere appresi o amplificati dal sistema.

Ciclo di vita dello sviluppo di IA e pregiudizi

La gestione dei pregiudizi è considerata un processo che si estende per tutto il ciclo di vita del sistema IA, dalla raccolta dei dati e dalla progettazione del modello fino all'implementazione e al monitoraggio post-lancio. Ogni fase presenta opportunità uniche per introdurre o mitigare i pregiudizi:

- Durante la raccolta dei dati: assicurare che i dati non riflettano pregiudizi storici o sociali non desiderati.
- Nella progettazione del modello: scegliere architetture e algoritmi che minimizzino il rischio di amplificare pregiudizi esistenti.
- Nel testing e nella validazione: utilizzare set di dati di test diversificati per valutare come il sistema gestisce vari scenari e gruppi di utenti.
- Nel monitoraggio continuo: osservare il comportamento del sistema nell'ambiente reale per identificare e correggere nuovi pregiudizi che possono emergere nel tempo.

La ISO/IEC TR 24027:2021 mette in luce diverse strategie efficaci per la mitigazione dei pregiudizi nei sistemi di IA, sottolineando la necessità di un approccio sistematico e multidimensionale per affrontare queste sfide. Le strategie raccomandate includono l'adeguamento dei set di dati, l'applicazione di tecniche di apprendimento avanzate e algoritmi progettati per promuovere l'equità, nonché il monitoraggio continuo dei sistemi per individuare e correggere i pregiudizi che possono emergere nel tempo. Uno degli aspetti fondamentali della mitigazione dei pregiudizi è l'adeguamento dei set di dati utilizzati per addestrare i sistemi di IA. Questo processo include la revisione e la modifica dei dati per garantire che riflettano una rappresentazione equa e bilanciata di tutte le parti interessate in base agli obiettivi del sistema. L'adeguamento può comportare la rimozione di bias storici o culturali impliciti nei dati, l'incremento della rappresentatività di minoranze o gruppi sottorappresentati, e la correzione di eventuali distorsioni nei dati di input. Ad esempio, potrebbe essere necessario sovracampionare i dati provenienti da gruppi minoritari o utilizzare tecniche di apprendimento per bilanciare le classi di output. La norma consiglia l'uso di tecniche di apprendimento e algoritmi specificamente progettati per ridurre i pregiudizi e promuovere l'equità.

Questi algoritmi sono spesso sviluppati con meccanismi incorporati che cercano attivamente di identificare e compensare i pregiudizi durante il processo di apprendimento. Ciò può includere l'uso di tecniche di regolarizzazione per minimizzare il pregiudizio nei parametri del modello, o l'implementazione di algoritmi che modificano i pesi attribuiti a diversi segmenti di dati per assicurare un trattamento equo tra i gruppi.
Un'altra strategia cruciale è il monitoraggio continuo dei sistemi di IA dopo la loro implementazione. Questo permette agli sviluppatori di osservare come i sistemi si comportano in ambienti reali e di identificare nuovi pregiudizi o problemi di equità che possono emergere con l'evoluzione dei contesti sociali e tecnologici. Il monitoraggio continuo implica l'analisi regolare delle decisioni prese dal sistema, l'esame delle metriche di performance disaggregate per diversi gruppi demografici e la verifica della conformità con i principi etici e normativi vigenti. La mitigazione efficace dei pregiudizi nei sistemi di IA richiede un impegno continuo e collaborativo tra diverse parti interessate. Gli sviluppatori di sistemi, gli eticisti, gli utenti finali e altre parti rilevanti devono lavorare insieme per garantire che i sistemi non solo siano tecnicamente efficaci, ma anche socialmente responsabili. Questa collaborazione può includere la condivisione di best practices, la formazione continua sulle questioni etiche legate all'IA e il coinvolgimento attivo delle comunità colpite nelle fasi di progettazione e revisione dei sistemi.

5.1.2 Proposta di protocollo per la verifica della conformità di sistemi basati sull'Intelligenza Artificiale nel settore biomedicale

Questa sezione presenta una prima ipotesi di protocollo per la verifica di conformità alla ISO/IEC TR 24027:2021, analizzata nel paragrafo precedente, specificamente pensato per sistemi basati su metodi di IA applicati al settore biomedicale.
Dopo una rapida introduzione di un ciclo di sviluppo classico di un sistema di IA in ambito biomedicale e la descrizione delle caratteristiche associate alla prontezza commerciale di un software, il paragrafo presenta un protocollo di verifica della conformità per la rilevazione ed il controllo dei bias basato sulla linea guida ISO/IEC TR 24027:2021. Successivamente vengono presentati i due sistemi di IA in ambito biomedicale che saranno utilizzati come caso di studio per la validazione del processo di verifica proposto; in ultimo si riportano le conclusioni del protocollo.
Il cuore del capitolo è costituito dalle due sezioni centrali, nelle quali viene definito e illustrato uno schema di valutazione della conformità che verrà poi applicato a due casi di studio ben distinti: un sistema per la *detection* del melanoma, orientato all'uso da parte di un qualsiasi soggetto (e.g., paziente); un sistema a supporto della stratificazione della progressione di malattia di pazienti affetti da sclerosi multipla, orientato all'uso da parte di un medico specializzato. In entrambi i casi sarà simulata una valutazione di conformità di un sistema di IA alla ISO/IEC TR 24027:2021, sulla base della procedura definita in questo documento.

Ciclo di sviluppo di un sistema di IA in ambito biomedicale

I sistemi di IA che maggiormente vengono utilizzati al fine di supportare il processo decisionale in medicina sono modelli che consentono di:

- rilevare le malattie;
- fornire diagnosi e prognosi;
- creare strumenti di valutazione del rischio.

Indipendentemente dal fine ultimo, la creazione di questi modelli segue uno schema di progettazione e realizzazione che può essere formalizzato in una pipeline costituita da quattro passi:
1. Raccolta e preparazione dei dati (data collection & data preparation);
2. Sviluppo del modello (model development);
3. Valutazione del modello (model evaluation);
4. Rilascio del modello (model deployment).

È importante notare che ognuna di queste quattro fasi è delicata, nei termini dei bias (o pregiudizi) che possono emergere durante l'intero processo. Nel contesto dell'IA e dell'apprendimento automatico, i bias si riferiscono a errori sistematici nei dati (fase 1) o nel processo algoritmico (fasi da 2 a 4) che portano a risultati ingiusti per alcuni gruppi. I bias possono derivare dal processo di raccolta dei dati, dal modo in cui i dati vengono elaborati e utilizzati nell'addestramento dei modelli, o dalle ipotesi e le scelte intrinseche fatte durante lo sviluppo degli algoritmi.
I bias nei sistemi di IA possono portare a pratiche discriminatorie o a un trattamento diseguale degli individui in base a caratteristiche come il sesso, l'età, l'etnia (ove applicabile) o lo status socio-economico. L'obiettivo dell'identificazione e della riduzione dei pregiudizi è garantire che i sistemi di IA operino in modo equo, etico e senza perpetuare le disuguaglianze esistenti.
Le sezioni successive dettagliano ognuno dei 4 passi sopra elencati, evidenziando le possibili sorgenti di bias, così come presentati in letteratura e nella ISO/IEC TR 24027:2021. Questo documento affronta i bias in relazione ai sistemi di IA, in particolare per quanto riguarda il processo decisionale assistito dall'IA. Vengono descritte le tecniche di misurazione e i metodi per valutare i bias, con l'obiettivo di affrontare e trattare le vulnerabilità associate.

Passo 1: Raccolta e preparazione dei dati

La creazione di un sistema di IA inizia con una fase di raccolta e preparazione dei dati. Questo processo è fondamentale, indipendentemente dal tipo di modello di IA che si intende sviluppare, poiché la qualità, la varietà e la numerosità dei dati raccolti influenzano direttamente la capacità del sistema di apprendere e di funzionare correttamente. La raccolta dei dati non si limita ad una particolare famiglia di modelli, come il *deep learning* o i *foundation models*, ma si estende a tutte le categorie di modelli di IA, con l'obiettivo di ottenere un insieme di dati che sia rappresentativo, ampio e pertinente all'ambito di applicazione. I dati possono provenire da una varietà di fonti, sia open source che proprietarie, e devono essere raccolti tenendo conto delle necessità specifiche del modello. A seconda dell'obiettivo, la quantità di dati necessaria può variare significativamente: modelli generalisti possono richiedere volumi importanti di dati, mentre per modelli più specializzati può essere sufficiente una quantità minore, focalizzata su un dominio specifico.

In tutti i casi, una volta raccolti i dati devono essere accuratamente preparati per l'addestramento, attraverso le seguenti attività:
- Categorizzazione: Definire la natura dei dati raccolti e etichettarli (nel caso di apprendimento supervisionato) secondo il loro dominio applicativo;
- Filtraggio: Rimuovere dati non pertinenti, di qualità insufficiente o sensibili;
- Eliminazione dei duplicati: Assicurarsi che i dati siano unici per prevenire distorsioni nell'addestramento;
- Normalizzazione: La preparazione può includere anche la normalizzazione o standardizzazione dei dati, per garantire che siano in un formato coerente e ottimale per l'addestramento dei modelli.

Il risultato di questa fase è la creazione di una "Base Data Pile": un insieme di dati puliti e organizzati, pronti per essere utilizzati nell'addestramento del modello. Questa collezione di dati è essenziale per la costruzione di sistemi di IA affidabili e performanti e può essere versionata ed etichettata per garantire tracciabilità e conformità ai principi di governance dei sistemi di IA, assicurando così trasparenza e affidabilità nel processo di sviluppo.
In questa fase 1 esiste un rischio significativo di introdurre bias che possono influenzare negativamente l'intero processo di sviluppo del sistema di IA. I bias possono manifestarsi in diverse forme:

- Bias nella selezione dei dati: Se i dati raccolti non rappresentano equamente tutte le variabili o i gruppi interessati, il modello potrebbe apprendere in modo distorto. Ad esempio, se un modello di diagnosi medica viene addestrato su dati prevalentemente raccolti da una popolazione specifica (e.g., caucasica), potrebbe non essere altrettanto accurato o efficace per altre popolazioni;
- Bias di etichettatura: Nel caso di apprendimento supervisionato, durante la fase di categorizzazione e etichettatura i bias possono insinuarsi attraverso etichette imprecise o parziali, influenzando come il modello percepisce e classifica i dati. Se i dati vengono etichettati in modo soggettivo o basato su presupposti errati, questo può portare a un apprendimento sbilanciato del modello;
- Bias attraverso la pulizia dei dati: Operazioni come il filtraggio e l'eliminazione dei duplicati, se non eseguite con attenzione, possono rimuovere dati importanti o variabilità necessaria per un addestramento completo del modello. La decisione su quali dati considerare non pertinenti o di bassa qualità può essere soggettiva e portare a una perdita di informazioni cruciali;
- Bias derivanti da dati non rappresentativi: Anche la selezione di dati open source o proprietari può introdurre bias se questi dataset non sono completamente rappresentativi del dominio di applicazione o se riflettono pregiudizi esistenti nella società o nella raccolta dei dati.

Per mitigare questi bias, è fondamentale adottare approcci consapevoli e metodici durante la raccolta e la preparazione dei dati. Ciò include la verifica della rappresentatività e della diversità dei dati, l'utilizzo di tecniche di etichettatura oggettive, la pulizia dei dati con criteri chiari e trasparenti, e l'adozione di pratiche di revisione e valutazione critica dei dati raccolti. Inoltre, l'implementazione di tecniche di debiasing e l'analisi di sensibilità possono aiutare a identificare e ridurre i bias prima che influenzino le fasi successive dello sviluppo del modello di IA. La consapevolezza e l'attenzione a questi aspetti nella fase iniziale sono cruciali per sviluppare sistemi di IA che siano equi, affidabili e privi di discriminazioni.

Passo 2: Sviluppo del modello

La fase di sviluppo si focalizza sull'addestramento del modello utilizzando il "Base Data Pile" allestito nella fase iniziale. Questo passaggio è fondamentale per generare un modello capace di interpretare i dati e fornire output accurati secondo le necessità. La procedura di addestramento è un'attività che richiede impegno significativo, con una variabilità notevole in termini di risorse e tempo necessari, dipendenti dall'ampiezza e dalla complessità del modello selezionato. Il processo inizia con la scelta della tipologia e dell'architettura del modello, decisione che può variare ampiamente a seconda delle specifiche esigenze e degli obiettivi del progetto. La gamma di modelli disponibili include diverse architetture, ciascuna con i propri punti di forza e limitazioni, che vanno valutati attentamente per identificare l'opzione più adatta allo scopo. In questa fase è cruciale una buona interazione tra l'esperto di dominio (ad esempio, il medico) e il progettista del sistema di IA. Un aspetto cruciale di questa fase è l'attenzione all'errore di generalizzazione, ovvero la capacità del modello di performare bene non solo sui dati su cui è stato addestrato (training set), ma anche su nuovi dati mai visti prima (test set). Per minimizzare questo errore e garantire l'affidabilità del modello, è essenziale organizzare i dati in tre distinti insiemi:

- un training set per l'addestramento;
- un validation set per ottimizzare i parametri del modello e valutarne le prestazioni durante l'addestramento;
- un test set per testare la generalizzazione del modello su dati non utilizzati nel processo di addestramento.

I tre set sono egualmente importanti e quindi, avendo a disposizione una base di dati "infinita", la situazione ideale sarebbe quella in cui i set sono molto ampi. Nel caso reale tuttavia, il set di dati a disposizione è limitato. La scelta delle proporzioni di divisione (e.g., 60/20/20) può influenzare l'intero processo di addestramento. Per limitare l'impatto di tale scelta possono essere messe in atto tecniche quali la stratificazione, la cross-validation e l'addestramento ripetuto. La gestione della complessità computazionale rappresenta un altro punto di attenzione, dato che l'addestramento di modelli complessi può richiedere risorse significative in termini di elaborazione e tempo. È quindi fondamentale bilanciare la complessità del modello con le risorse disponibili, cercando soluzioni che ottimizzino l'efficienza senza compromettere la qualità delle prestazioni.
La fase di addestramento, pur essendo onerosa, è assolutamente necessaria per lo sviluppo di un modello efficace. Una volta che il modello è stato addestrato e ha dimostrato di poter generalizzare bene da nuovi dati, le fasi successive di test, valutazione e implementazione possono procedere con maggiore facilità, avendo stabilito una solida base su cui costruire. Anche in questa fase esistono diversi modi in cui possono essere introdotti bias che influenzano l'intero processo di sviluppo di un sistema di IA. Questi bias possono compromettere la capacità del modello di generalizzare correttamente su nuovi dati e portare a decisioni o previsioni ingiuste o discriminatorie.
Di seguito, vengono esplorati alcuni dei principali fattori che contribuiscono all'introduzione di bias durante la fase di addestramento:

- Selezione del training set: se il training set non è rappresentativo della popolazione (i.e., dei possibili input) che il modello incontrerà nel mondo reale, esiste un alto rischio che il modello sviluppi bias. Ad esempio, se un modello viene addestrato principalmente su dati relativi a un certo gruppo demografico, potrebbe non performare bene o potrebbe essere ingiusto nei confronti di altri gruppi non adeguatamente rappresentati nei dati di addestramento.
- Overfitting e underfitting: l'overfitting si verifica quando un modello apprende troppo bene i dettagli e il rumore specifici del training set, a scapito della sua capacità di generalizzare su nuovi dati. Al contrario, l'underfitting si verifica quando il modello non è in grado di apprendere adeguatamente la struttura sottostante dei dati di addestramento. Entrambi questi fenomeni possono portare a un errore di generalizzazione e possono essere influenzati da bias nei dati di addestramento.
- Valutazione durante l'addestramento: la scelta dei criteri e delle metriche utilizzate per valutare le prestazioni del modello durante l'addestramento può anch'essa introdurre bias. Se le metriche scelte non riflettono equamente l'importanza di vari aspetti del compito di previsione o decisione, il modello potrebbe essere ottimizzato in modo da favorire certi risultati a scapito di altri, introducendo bias nelle sue prestazioni.
- Bias nei validation e test set: analogamente alla selezione del training set, se i validation e test set non sono rappresentativi o contengono bias preesistenti, la valutazione della capacità di generalizzazione del modello può essere distorta. Questo può portare a una percezione inaccurata delle prestazioni del modello su dati reali e diversificati.

Per mitigare questi potenziali bias, è essenziale adottare pratiche di addestramento consapevoli e inclusive, come la cura nella selezione di dati rappresentativi per tutti i set (training, validation, test), l'utilizzo di tecniche per prevenire l'overfitting (come la regolarizzazione), e la scelta di metriche di valutazione che considerino equità e inclusività. Inoltre, l'analisi e la correzione dei bias nei dati prima dell'addestramento e la valutazione continua delle prestazioni del modello su diversi gruppi demografici possono contribuire a ridurre l'impatto dei bias sulle decisioni prese dai sistemi di IA.

Passo 3: Valutazione del modello

Questa fase si attiva al completamento dell'addestramento, con l'obiettivo di valutare l'efficacia del modello attraverso il confronto con benchmark predefiniti, al fine di determinare quanto accuratamente il modello possa operare su dati nuovi, ovvero dati non inclusi nel training set. In questa fase, il modello viene sottoposto a test di inferenza, ovvero la produzione di output basati su dati precedentemente non analizzati, per valutare le sue capacità predittive. Questa operazione di inferenza, sebbene richieda meno risorse computazionali rispetto all'addestramento, può comunque necessitare di risorse notevoli a causa della sua complessità computazionale. Per quantificare le prestazioni del modello, si utilizzano dataset di benchmark – idealmente riconosciuti dalla comunità scientifica e accessibili open source – che permettano di confrontare i risultati ottenuti con quelli di altri modelli e di valutare in maniera oggettiva la qualità delle sue prestazioni. L'analisi su questi dataset di test consente di generare un profilo dettagliato del modello, che include informazioni sulla sua formazione (grazie al versionamento dei dati effettuato nella Fase 1) e sui risultati ottenuti nei test benchmark.

Fino a questo punto del processo, la guida dello sviluppo è tipicamente nelle mani di uno specialista in IA. Completata la fase di benchmarking, le prestazioni del modello possono essere ulteriormente verificate e migliorate con il contributo di un esperto del dominio di applicazione, il quale, pur non essendo necessariamente versato in IA, apporta un valore aggiunto critico nella valutazione delle capacità del modello. L'esperto di dominio gioca un ruolo chiave nel mettere alla prova il modello, offrendo spunti e suggerimenti per ottimizzarne le prestazioni.

Questo può includere la fornitura di ulteriori dati di test specifici al dominio, che possono essere utilizzati per affinare e migliorare l'accuratezza del modello. Il processo di perfezionamento e ottimizzazione del modello, condotto con l'assistenza dell'esperto di dominio, è conosciuto come "tuning". Attraverso il tuning, si mira a raffinare il modello per massimizzarne l'efficacia e assicurare che le sue prestazioni siano ottimali nell'ambito di applicazione specifico, garantendo così un sistema di IA non solo tecnicamente avanzato ma anche perfettamente allineato alle necessità pratiche del settore di riferimento.

In questa fase 3 emerge un'opportunità per l'identificazione e la mitigazione dei bias.

Difatti, la valutazione delle prestazioni attraverso dataset di benchmark e test può rivelare se e come i bias, eventualmente introdotti nelle fasi precedenti del processo di sviluppo, influenzino i risultati del modello in scenari reali:

- Rilevamento dei bias tramite benchmarking: utilizzando dataset di benchmark diversificati e rappresentativi di varie popolazioni e scenari, è possibile identificare bias di performance che potrebbero non essere stati evidenti durante l'addestramento. La comparazione dei risultati del modello con standard riconosciuti permette di evidenziare discrepanze e distorsioni nelle sue capacità predittive.
- Analisi delle prestazioni per gruppi demografici: effettuare analisi disaggregate delle prestazioni del modello per diversi gruppi demografici o categorie può rivelare bias specifici. Questo approccio consente di identificare se il modello performa in modo equo su tutti i gruppi o se vi sono disparità che necessitano di interventi correttivi.

- Feedback dell'esperto di dominio: l'intervento di esperti del dominio di applicazione può fornire insight preziosi non solo per ottimizzare le prestazioni ma anche per rilevare bias specifici del settore che potrebbero non essere stati considerati dai data scientists. Il loro contributo può guidare l'identificazione di situazioni in cui il modello potrebbe manifestare comportamenti ingiusti o inappropriati.

Per mitigare i bias identificati in questa fase, è cruciale adottare strategie di tuning del modello che non si limitino alla mera ottimizzazione delle prestazioni complessive, ma che includano anche interventi specifici per ridurre o eliminare le disparità rilevate. Questo può comportare l'aggiustamento dei parametri del modello, l'introduzione di tecniche di debiasing, o l'espansione del dataset di addestramento con dati supplementari che migliorino la rappresentatività e la diversità. Inoltre, la trasparenza nel reporting delle prestazioni del modello, includendo dettagli sulle metriche di valutazione utilizzate e sulle eventuali disparità rilevate, contribuisce alla costruzione di sistemi di IA più equi ed etici. Attraverso un processo di valutazione consapevole e inclusivo, è possibile assicurare che i modelli di IA sviluppati non solo siano tecnicamente validi ma anche giusti e imparziali, rispecchiando un impegno verso l'equità e l'etica nell'IA.

Passo 4: Rilascio del modello

La Fase 4 del ciclo di sviluppo di un sistema di IA riguarda il rilascio e l'implementazione del modello per l'uso effettivo. Questa fase cruciale richiede una pianificazione accurata per assicurare che il modello sia accessibile e utilizzabile nel modo più efficace possibile dall'utenza finale, che sia essa composta da aziende (B2B) o consumatori finali (B2C). In questa ottica, e tenendo sempre in considerazione la gestione dei bias, i punti salienti da considerare sono:

- Ottimizzazione dell'esperienza utente: è fondamentale identificare e comprendere le necessità specifiche degli utenti finali per sviluppare un'interfaccia utente o una *wrapper application* che faciliti l'interazione con il modello. Questo approccio può variare significativamente a seconda del contesto di utilizzo; ad esempio, un'applicazione destinata al supporto decisionale medico richiederà un'interfaccia intuitiva per i professionisti sanitari, mentre una destinata al grande pubblico dovrà essere accessibile anche a utenti senza competenze tecniche specifiche.
- Modalità di distribuzione: La scelta tra una soluzione basata su cloud o un'applicazione eseguibile localmente dipende da numerosi fattori, inclusi la necessità di connettività, i requisiti di privacy e sicurezza dei dati, e la facilità di accesso per l'utente. Ogni opzione presenta vantaggi e sfide che devono essere valutati attentamente per massimizzare l'efficacia del modello una volta implementato.
- Gestione continua e iterazione: dopo il rilascio, è vitale instaurare un processo di feedback continuo che permetta di monitorare le prestazioni del modello e di identificare opportunità di miglioramento. Questo approccio assicura che il modello rimanga efficace nel tempo, adattandosi a eventuali cambiamenti nei dati o nelle esigenze degli utenti.
- Considerazioni sui bias: durante la fase di rilascio, è importante considerare come i bias identificati nelle fasi precedenti possano influenzare l'utilizzo del modello nel mondo reale. La trasparenza riguardo ai limiti del modello e alle potenziali distorsioni è fondamentale, specialmente in applicazioni critiche come nel settore sanitario, dove le decisioni influenzate da un modello possono avere conseguenze significative sulla vita delle persone.
- Mitigazione dei bias: è essenziale implementare meccanismi per mitigare l'impatto dei bias, che possono includere l'aggiornamento periodico del modello con nuovi dati per assicurare che rimanga rappresentativo e imparziale, così come l'offerta di formazione agli utenti sulle migliori pratiche per interpretare e utilizzare gli output del modello in modo etico ed equo.

- Valutazione dell'impatto: una valutazione dell'impatto del modello sul pubblico e sui gruppi specifici può aiutare a identificare problemi non previsti di bias o discriminazione, consentendo l'adozione tempestiva di misure correttive.

In sintesi, la Fase 4 non solo segna il punto in cui il modello diventa operativo, ma inaugura anche un periodo di monitoraggio e adattamento continuo, durante il quale gli sviluppatori possono raffinare e migliorare il modello in risposta al feedback degli utenti e all'evoluzione delle esigenze. Garantire l'equità, la trasparenza e l'accessibilità del modello durante questa fase è cruciale per il successo a lungo termine di qualsiasi applicazione di IA. In tale contesto è da sottolineare anche l'intervento dell'accreditamento nell'integrare il processo di valutazione della conformità ai requisiti cogenti all'interno del ciclo di sviluppo del modello di IA. La valutazione della conformità per i dispositivi medici e in particolare per quelli contenenti sistemi di IA richiede imparzialità, competenza e coerente funzionamento dell'organizzazione preposta a tale attività. La valutazione della conformità deve essere posizionata strategicamente prima della fase di rilascio o implementazione (Fase 4), per assicurare che il modello sia conforme ai requisiti normativi e alle aspettative di qualità prima di essere accessibile agli utenti finali. È bene tuttavia differenziare tra un prodotto destinato a essere immesso sul mercato e uno di natura più strettamente di ricerca (prodotto scientifico). Mentre per il primo è indispensabile una valutazione che comprenda aspetti come la sicurezza del paziente, la conformità normativa e l'efficacia clinica, per i prodotti scientifici o di ricerca, la valutazione della conformità normativa non è obbligatoria e l'enfasi potrebbe essere maggiormente posta sulla validità metodologica, l'innovazione e la capacità di contribuire alla base di conoscenza medica. In entrambi i casi, è fondamentale che il processo di valutazione della conformità sia concepito per promuovere l'innovazione e il progresso tecnologico, mantenendo al contempo standard rigorosi di sicurezza, efficacia e equità. Considerare la valutazione della conformità quale elemento fondamentale e continuativo del ciclo di sviluppo dell'IA in medicina non solo aumenta la fiducia degli utenti e dei professionisti del settore ma assicura anche che le soluzioni di IA siano sviluppate e implementate in modo responsabile e sostenibile.

Prodotto Commerciale vs. Prodotto di Ricerca

Per definire un prodotto come "commerciale" si può utilizzare la scala del TRL (Technology Readiness Level). La scala TRL è utilizzata per definire la maturità di una tecnologia e può aiutare a comprendere le varie sfide che devono essere superate nelle diverse fasi di sviluppo di una tecnologia.
Sviluppati dalla NASA negli anni '70, i TRL sono stati adottati e adattati a un'ampia gamma di industrie per i rispettivi settori. Sebbene esistano scale aggiuntive rispetto alla classificazione TRL, per esempio scale per giudicare le soluzioni in base alla loro prontezza commerciale (CRL), al livello di prestazioni (TPL), alla producibilità (MRL), alla capacità di integrarsi in sistemi più ampi (IRL), la scala TRL è maggiormente diffusa e pertanto in questo documento si farà riferimento ad essa.

- TRL 1-3: ricerca iniziale. Indipendentemente dalla scala, ai livelli più bassi (da 1 a 3), una tecnologia in via di sviluppo è ben lontana dall'essere pronta per il mercato. A questi livelli di sviluppo, l'incertezza abbonda e può essere difficile prevedere tutte le sfide e la portata dei benefici della tecnologia. Lo sviluppo, a questi TRL, è spesso condotto da istituzioni accademiche che lavorano su scale temporali medio-lunghe e prevede studi di fattibilità e ricerche di proof of concept (PoC).
- TRL 4-6: sviluppo tecnologico. L'area TRL compresa tra 4 e 6 comprende le tecnologie che sono state dimostrate a livello concettuale, ma che devono ancora essere convalidate e dimostrate in ambienti più rappresentativi dal punto di vista industriale o commerciale. La prototipazione e la dimostrazione della fattibilità costituiscono una parte fondamentale di queste fasi.

Questa area del TRL viene spesso definita la "Valle della Morte" a causa delle numerose tecnologie che spesso si bloccano a questo livello, di solito a causa della mancanza di investimenti e di diritti di proprietà sulle stesse da parte del mondo accademico e dell'industria, che di solito si concentrano sulle estremità opposte della scala.

- TRL 7-9: commercializzazione. Le tecnologie a TRL più elevato (da 7 a 9) sono sul punto di diventare un prodotto commerciale o sono già sul mercato. In questi stadi più elevati si conoscono meglio le tecnologie, ma di solito anche i costi e i rischi sono più elevati.

Lo sviluppo a questi livelli comporta in genere la verifica e la qualificazione di un sottosistema o di un sistema completo oltre ad altre attività finalizzate all'immissione sul mercato di un prodotto completamente maturo. Le tecnologie a TRL più elevati sono spesso di proprietà dell'industria e da essa sviluppate. Da quanto discusso sopra, si evince come un prodotto possa definirsi commerciale se è inquadrabile nell'area TRL che va dal livello 7 al livello 9. La figura 2 riporta uno schema realizzato da KPMG che permette di avere una maggiore comprensione della scala TRL rispetto ad un possibile mercato.

Figura 2. Valutazione di un TRL

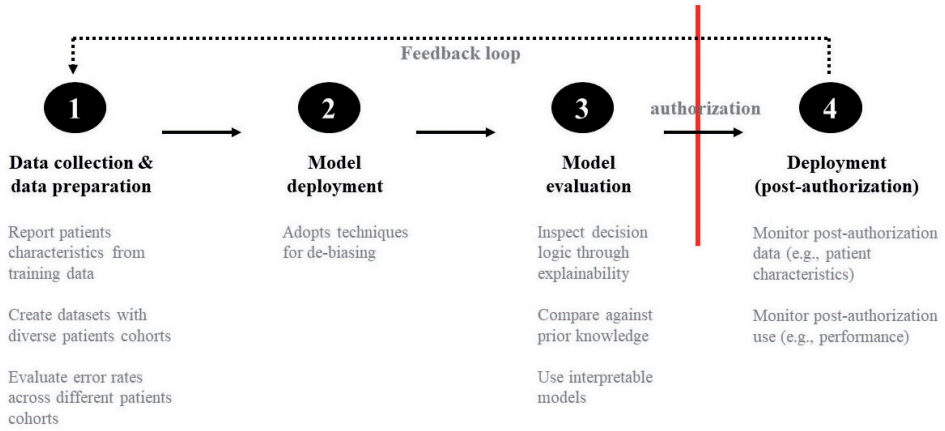
Revenue TRL	Revenue Definition	TRL	*H2020	**EARTO	KPMG Definition and Description
0	**Idea**	0			Idea formulation
1	**Basic Research**	1	Basic Principles Observed	Basic Principles Observed	Basic principles translated from scientific research
2	**Technology Formulation**	2	Technology Concept Formulated	Technology Concept Formulated	A technology concept is formulated in terms of how the basic principles can be applied
3	**Applied Research**	3	Experimental Proof of concept	First assessment of feasibility of the concept and technologies	First assessment of concept through actual research assessing technical and market feasibility
4,5	**Small Scale Prototype / Large Scale Prototype**	4	Validation of integrated prototype in a laboratory	Validation of integrated prototype in a laboratory	Early feasibility tested in laboratory by integrating basic technological components
6	**Prototype System**	5	Testing of prototype in a user environment	Testing of prototype in a user environment	System, actual use and manufacturing tested and validated in user environment
7	**Demonstration System**	6	Technology demonstrated in relevant environment	Per-production of the product, including testing in a user environment	Full integration of product and manufacturing technology now established in pilot line/plant
8	**First kind of commercial S ystem**	7	Low scale pilot production, producing actual commercial products	Low scale pilot production demonstrated	Low scale pilot production demonstrated. Product launched to early markets
9	**Full commercial application**	8	System Completed and qualified	Manufacturing fully tested, validated and qualified	Manufacturing of system has been fully completed. Product is launched to majority markets
		9	Market expansion, incremental changes in product create new versions	Production and product fully operational and competitive	Product fully operational and competitive. Production and manufacturing optimized through continuous incremental innovations

Fonte: KPMG, Assessment of a TRL
https://assets.kpmg.com/content/dam/kpmg/ie/pdf/2019/11/ie-assessment-of-a-trl.pdf (al 6 febbraio 2024)

In particolare, la scala TRL viene affiancata dalle sue interpretazioni nell'ambito dei progetti Europei Horizon 2020 (H2020) e alle indicazioni della European Association of Research and Technology Organisations (EARTO). Nel momento in cui un modello viene invece sviluppato da un gruppo di ricerca per fini accademici non è strettamente necessario (sebbene raccomandabile) che ne sia verificata la conformità rispetto a norme tecniche, poiché lo stesso, difficilmente sarà commercializzato così come viene descritto nella pubblicazione scientifica in cui appare. Il modello pubblicato, se di valore, sarà stato validato su benchmark pubblici e di esso si conoscerà l'architettura di base. Inoltre, sarà stato addestrato su dati anch'essi pubblici. In caso contrario, il modello sarà di scarso valore accademico. Un prodotto scientifico (o di ricerca) avrà tipicamente un TRL massimo pari a 6. Con specifico riferimento alla creazione di un modello di IA per la medicina, esso sarà stato creato passando per le fasi 1, 2 e 3 del ciclo di sviluppo presentato in precedenza.

Collocazione della valutazione della conformità nel ciclo di sviluppo di un sistema di IA in ambito biomedicale. Con specifico riferimento al ciclo di sviluppo di un sistema di IA in ambito biomedicale è possibile collocare l'intervento di un organismo notificato a valle della fase di valutazione del modello (fase 3) e prima della fase di rilascio del modello (fase 4). La figura 3 fornisce uno schema da cui si evince con il tratto rosso la possibile collocazione della fase di valutazione della conformità.

Figura 3. In rosso viene mostrato dove interviene la fase di verifica di conformità nel ciclo di sviluppo di un sistema predittivo per la medicina



Fonte: Vokinger, K.N., Feuerriegel, S. & Kesselheim, A.S. Mitigating bias in machine learning for medicine. Commun Med 1, 25 (2021). https://doi.org/10.1038/s43856-021-00028-w

L'importanza di un processo di valutazione della conformità per i sistemi basati su IA nel settore medico si evidenzia considerando le potenziali conseguenze di un rilascio prematuro di tali tecnologie senza un'adeguata valutazione dei rischi. Un sistema di IA non rispondente ai requisiti cogenti può portare a complicazioni significative, non solo sotto il profilo della sicurezza dei pazienti ma anche relativamente alla privacy e alla gestione dei dati. Una volta che un modello è entrato nel mercato o è stato adottato in contesti clinici, intervenire per correggere difetti o rimuovere il sistema può essere estremamente difficile, costoso e dannoso anche per gli utilizzatori e i pazienti.

Inoltre per i dispositivi medici e medico-diagnostici in vitro è obbligatoria la certificazione secondo i Regolamenti UE 2017/745 e 2017/746, prima dell'immissione sul mercato dei prodotti, con l'intervento di un Organismo Notificato, a seconda della classe di rischio del dispositivo medico stesso.
In tale contesto è da sottolineare anche l'intervento dell'accreditamento: uno strumento importante per garantire la conformità dei sistemi basati sull'IA. Esso attesta la competenza, l'imparzialità e l'affidabilità degli organismi di valutazione della conformità, offrendo alle imprese, e quindi ai consumatori, una garanzia sui processi di verifica e certificazione. Gli organismi notificati, che devono rispettare requisiti di indipendenza, competenza e imparzialità possono dimostrare la propria conformità ai requisiti grazie ad un certificato di accreditamento[31] da allegare alla domanda di notifica. Uno scenario è quello sollevato dal caso di ChatGPT in Italia, che può essere utilizzato per illustrare i rischi associati al lancio di tecnologie di IA senza un adeguato processo di valutazione di conformità alla normativa tecnica. Difatti, la decisione del GPDP (Garante per la Protezione dei Dati Personali) di bloccare l'accesso a ChatGPT a causa della mancanza di filtri per la verifica dell'età e dell'incertezza riguardo alla gestione dei dati solleva questioni cruciali anche per i sistemi di IA in ambito biomedicale, in quanto in contesti medici, la verifica dell'utente e la protezione dei minori sono di fondamentale importanza, e un sistema di IA che non implementa adeguate misure di controllo dell'accesso può esporre i pazienti a rischi non necessari e compromettere la sicurezza dei dati sensibili. La mancanza di trasparenza su come i dati sono raccolti, elaborati e conservati da un sistema di IA può avere implicazioni dirette sulla privacy dei pazienti e sulla conformità alle normative sulla protezione dei dati. In ambito medico, dove i dati hanno un alto grado di sensibilità, è imperativo che i sistemi di IA siano progettati per garantire la massima protezione delle informazioni sanitarie dei pazienti. Questi scenari sottolineano l'importanza di integrare processi di valutazione della conformità che valutino in modo comprensivo i sistemi di IA prima del loro rilascio, soprattutto quando destinati all'uso in contesti critici come quello medico. Tra tutti, cosi come evidenziato dalla linea guida ISO/IEC TR 24027:2021, nel contesto di applicazioni biomedicali, particolare attenzione va posta alla verifica di conformità dei sistemi di IA rispetto alla presenza di bias. In quest'ottica, la lezione appresa dal caso di ChatGPT evidenzia la necessità di un approccio proattivo nella regolamentazione e verifica dei sistemi di IA, particolarmente in settori ad alto impatto sociale e personale come la medicina. L'adozione di processi preventivi di verifica di conformità a norme tecniche non solo protegge gli utenti e i pazienti ma contribuisce anche a costruire un ecosistema di IA più sicuro, etico e responsabile.

Bias nei sistemi di IA

Secondo la linea guida ISO/IEC TR 24027:2021, le fonti di distorsione (bias) indesiderata nei sistemi di IA sono classificate a grandi linee in tre gruppi principali, che illustrano come queste distorsioni interagiscano e abbiano un impatto sullo sviluppo e sul funzionamento dei sistemi di IA:

1. Pregiudizi cognitivi umani: comprende i pregiudizi che hanno origine dai processi di pensiero e dalle percezioni umane. I pregiudizi cognitivi umani possono influenzare i sistemi di IA in diversi modi, compreso il modo in cui i dati vengono raccolti, interpretati e utilizzati nel processo decisionale. Questi pregiudizi possono penetrare nei sistemi di IA attraverso decisioni soggettive prese nel processo di progettazione o attraverso i dati scelti per l'addestramento dei modelli di IA.
2. Data bias: le distorsioni dei dati si riferiscono alle distorsioni presenti nel set di dati utilizzato per addestrare i sistemi di IA.

---

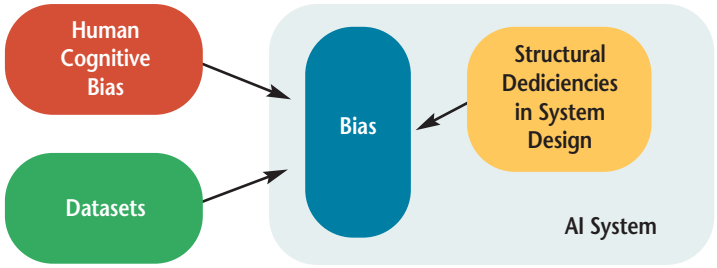
[31] Se la domanda di notifica non si basa su un certificato di accreditamento, l'autorità di notifica deve fornire alla Commissione e agli altri Stati (che avranno maggiore tempo per un'eventuale obiezione) prove documentali per dimostrare il rispetto dei requisiti.

Ciò può verificarsi a causa di pregiudizi sociali incorporati nelle fonti di dati, che portano a problemi come il campionamento non rappresentativo. Ad esempio, i pregiudizi sociali che si riflettono nel linguaggio possono essere perpetrati e persino amplificati dalle tecnologie di IA, come i modelli di incorporazione delle parole, dando luogo a sistemi di IA che ereditano e propagano tali pregiudizi.

3. Pregiudizi introdotti da decisioni ingegneristiche: questa fonte di pregiudizi riguarda le scelte fatte durante la progettazione e lo sviluppo dei sistemi di IA. Le decisioni ingegneristiche, che vanno dalla selezione degli algoritmi alle metodologie di elaborazione dei dati e di valutazione dei modelli, possono inavvertitamente introdurre pregiudizi nei sistemi di IA. Queste decisioni, influenzate dalle prospettive soggettive degli ingegneri o dai limiti nella comprensione delle diverse esigenze, possono plasmare il sistema di IA in modo tale da privilegiare alcuni gruppi rispetto ad altri.

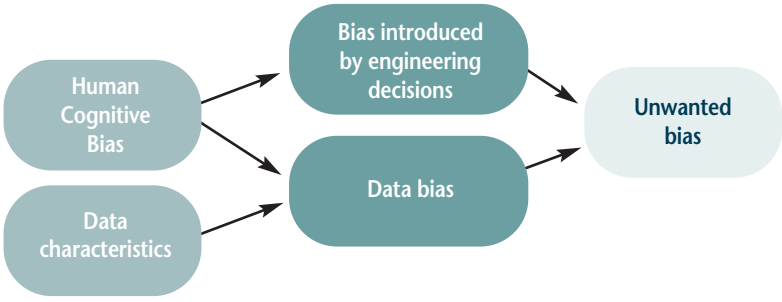
La figura 4 mostra uno schema delle possibili fonti di distorsione in un sistema di IA.

Figura 4. Possibili fonti di distorsione in un sistema di IA



Il documento evidenzia inoltre le complesse relazioni tra queste fonti di pregiudizio, indicando che i pregiudizi in un'area possono influenzare ed esacerbare i pregiudizi in un'altra. Ad esempio, i pregiudizi cognitivi umani possono portare a decisioni ingegneristiche distorte o alla selezione di dati distorti, che a loro volta influiscono sull'equità e sull'efficacia dei sistemi di IA. Riconoscere e affrontare queste fonti interconnesse di pregiudizi è fondamentale per sviluppare sistemi di IA che siano giusti, trasparenti ed equi.

Figura 5. Relazione tra gruppi di bias di alto livello (sorgente: ISO/IEC TR 24027:2021)



Fonte: ISO/IEC TR 24027:2021

La norma suggerisce di valutare i pregiudizi nei sistemi di IA concentrandosi sulla consapevolezza dei possibili bias, compresi quelli statistici e sociali, che possono determinare un comportamento scorretto del sistema. Il processo di valutazione prevede l'analisi dei risultati del sistema utilizzando una o più metriche di equità per scoprire le tracce di bias. Queste metriche mirano a catturare varie nozioni di equità e includono, tra le altre cose, la valutazione delle differenze tra i valori medi osservati e i valori reali. Il documento evidenzia che le metriche tradizionali della distorsione statistica possono essere insufficienti per rilevare comportamenti ingiusti o discriminatori, portando allo sviluppo di metriche che mirano a catturare diversi aspetti dell'equità. Tali metriche sono spesso discusse nella letteratura sulla "equità algoritmica" e sono progettate per confrontare diversi tipi di tassi di errore tra gruppi di persone. Il processo di valutazione non ha una corrispondenza diretta uno-a-uno tra la nozione generale di bias o fairness e le metriche statistiche di bias o fairness. La sfida principale consiste nel determinare le metriche più appropriate per ogni contesto. Il documento osserva che la maggior parte del lavoro sulle metriche di equità si è concentrato sull'equità dei sistemi di IA basati sulla classificazione o sulla regressione rispetto agli attributi demografici. Gli approcci per valutare la parzialità e l'equità includono la suddivisione dei dati in set di dati di addestramento, di convalida e di test, suddividendoli ulteriormente in base alle caratteristiche rilevanti per individuare eventuali parzialità. Le caratteristiche possono essere considerate in modo indipendente o intersecante, con gli obiettivi di equità e le caratteristiche demografiche rilevanti che vengono determinate esplicitamente prima del test. Le metriche di equità predeterminate vengono quindi calcolate per ciascun gruppo e i confronti tra i gruppi vengono effettuati per valutare se un sistema di classificazione può essere considerato sufficientemente equo o imparziale in base alle misure basate sulle metriche tra i gruppi che rientrano in un margine di differenza accettabile ("delta").
Uno degli obiettivi principali della verifica di conformità di un sistema di IA dovrebbe essere quello di appurare quanto e in che maniera sia stato preso in considerazione il problema del bias. Si vuole quindi progettare un processo di verifica che miri proprio ad affrontare queste questioni, assicurando che ogni sistema di IA sia sottoposto a un'analisi approfondita per identificare (e potenzialmente suggerire strategie di mitigazione) i bias in tutte le sue fasi di sviluppo, sulla base delle indicazioni presenti nella linea guida ISO/IEC TR 24027:2021.

Schema di verifica di conformità (ISO/IEC TR 24027:2021) per un sistema di IA

La ISO/IEC TR 24027:2021, fornendo linee guida per identificare e trattare i bias, può essere utilizzata in schemi di valutazione di conformità e le indicazioni, in essa contenute, essere usate come quadro chiaro e metodico di elementi che gli organismi di valutazione della conformità considerano per appurare la presenza di equità e l'assenza di pregiudizi nei sistemi di IA e nella presa di decisioni assistita dall'IA. L'obiettivo dello schema di valutazione della conformità è valutare che i sistemi di IA siano progettati e implementati in modo da minimizzare il rischio di bias, evitando che il sistema risulti discriminatorio, e che lo sviluppo dei modelli di IA sia stato eseguito con trasparenza e responsabilità. Attraverso un processo di verifica strutturato e basato su metriche specifiche, è possibile valutare sistematicamente i sistemi di IA rispetto a requisiti di equità e giustizia, rendendoli affidabili dal punto di vista non solo tecnico, ma anche sociale. Sulla base dell'analisi dettagliata delle sezioni 7 e 8 del documento ISO/IEC TR 24027:2021 e del quadro di riferimento del processo di verifica derivato, la tabella 4 delinea i controlli che dovrebbero essere eseguiti per la verifica di conformità dei sistemi di IA. La tabella è stata progettata per allinearsi alle sezioni della norma tecnica di riferimento, fungendo da bussola per navigare attraverso i critici processi di individuazione, valutazione, mitigazione e controllo dei bias nei sistemi di IA.

Essa non solo cataloga i controlli indispensabili ma associa ciascuno di essi alla specifica sezione della ISO/IEC TR 24027:2021, fornendo ulteriori dettagli sull'output atteso per ogni verifica effettuata. Questa metodologia può rappresentare un utile strumento per gli sviluppatori di sistemi di IA e, al contempo, fornire un quadro completo e dettagliato delle azioni necessarie per affrontare in modo efficace ogni potenziale distorsione, assicurando che ogni passo sia intrapreso con chiarezza e precisione.

Tabella 4. Controlli da effettuarsi sulla base delle indicazioni fornite dalla ISO/IEC TR 24027:2021

Fase/Aspetto	Analisi da eseguire	Sezioni di riferimento	Metrica o concetto descritto nel documento
Identificazione dei possibili pregiudizi (bias)	- Identificare le potenziali fonti di distorsione nei dati e negli algoritmi. - Categorizzare i pregiudizi (ad esempio: dati, algoritmici).	- 3.2: Elenco possibili bias - 5.2: Overview bias - 6: Dettaglio operativo bias	Discutere nel dettaglio le varie fonti e tipi di distorsione, se presenti (ad esempio, distorsione dei dati, distorsione algoritmica).
Quantificazione di bias ed equità (fairness)	-Valutare l'equità utilizzando metriche appropriate. - Valutare le disparità tra i diversi gruppi demografici.	- 5.2: Overview fairness - 7: Quantificazione di bias e fairness	Impiegare metriche di equità, come l'uguaglianza delle probabilità, la parità demografica e l'uguaglianza predittiva.
Tecniche di attenuazione dei pregiudizi	- Applicare la pre-elaborazione, l'in-elaborazione e la post-elaborazione per ridurre le distorsioni. - Utilizzare un design che tenga in considerazione i principi di equità.	- 8: Tecniche di attenuazione	Utilizzare tecniche di mitigazione che includono la pre-elaborazione dei dati, la modifica dell'algoritmo e gli aggiustamenti post-hoc.
Convalida e verifica	- Eseguire test indipendenti utilizzando set di dati di riserva. - Eseguire test di validità e test degli utenti. - Eseguire test di validità e test degli utenti.	- 8.4: Tecniche di validazione	Uso di set di dati di convalida esterni, misurando le performance indipendenti e la capacità di generalizzazione tramite matrici di confusione e altre metriche di prestazione.

Ne consegue che il documento, o fascicolo tecnico, che accompagna il software sottoposto a valutazione da parte dell'organismo notificato dovrebbe non solo fissare e documentare i criteri previsti, ma anche attestare la conformità a tali criteri. Dall'altra parte l'organismo di valutazione della conformità dovrebbe valutare come il fabbricante ha implementato la tabella e l'ha utilizzata per guidare l'intero processo di revisione e miglioramento del sistema. In particolare, quindi, la richiesta di verifica di conformità dovrebbe prevedere un documento contenente:

- Una dettagliata descrizione delle fasi di progettazione, sviluppo e implementazione, inclusa una descrizione completa del dataset e dei processi di stima degli errori e tecniche utilizzate per lo stesso (e.g., hold-out, cross-validation, ecc.).
- Un'analisi completa dei controlli effettuati rispetto alla presenza di bias (ossia, la Tabella 1), evidenziando anche le professionalità coinvolte nel processo per comprendere e incorporare diverse prospettive di equità.
- Una relazione sulla mitigazione dei pregiudizi e sulle valutazioni di equità.

- Report sulle strategie di monitoraggio continuo (se previste), compresa la valutazione dinamica dell'equità e della parzialità.

Partendo dalla tabella 4, che serve come fondamento per gli sviluppatori di sistemi IA nella loro missione di identificare, analizzare, mitigare e verificare i bias, possiamo estendere il concetto per creare una checklist operativa per gli operatori dell'organismo di valutazione della conformità impegnati nel processo di verifica. Questa trasformazione, da una guida di riferimento a uno strumento pratico di verifica, culmina, nella tabella 5, in una checklist dettagliata che facilita la compilazione dei risultati specifici per ogni controllo da parte degli operatori. Attraverso questo strumento, gli operatori possono sistematicamente valutare se un sistema di IA rispetti i criteri prestabiliti di imparzialità ed equità.
La tabella 5, quindi, rappresenta uno strumento cruciale nel processo di verifica di conformità, permettendo agli operatori di annotare con precisione e valutare l'adeguatezza dei sistemi di IA. Essa non solo traccia un percorso chiaro per la registrazione dei risultati di ogni verifica, ma definisce anche le metriche e le metodologie specifiche da utilizzare per una valutazione approfondita dei pregiudizi presenti.

Tabella 5. Schema di controllo da parte dell'organismo di certificazione

Human cognitive biases (sez. 6.2)

Bias	Metodo di verifica	Azione intrapresa	Note	Giudizio di conformità
Automation				
Group Attribution				
Implicit				
Confirmation				
In-Group				
Out-Group homogeneity				
Societal				
Rule-Based				
Requirement				

Data biases (sez. 6.3)

Data selection (sec. 6.3.2 & 6.3.4.)				
Data labelling (sec. 6.3.3 & 6.3.5)				
Data processing (sec. 6.3.6)				
Simpson's paradox (sec. 6.3.7)				
Data aggregation (sec. 6.3.8)				

Engineering biases (sez. 6.3.9 & 6.4)

Bias	Metodo di verifica	Azione intrapresa	Note	Giudizio di conformità
Distributed training (sec. 6.3.9)				
Features (sec. 6.4.2)				
Algorithm (sec. 6.4.3)				
Hyperparameter (sec. 6.4.4)				
Informativeness (sec. 6.4.5)				
Model bias (sec. 6.4.6)				
Model expressiveness (sec. 6.4.7.2)				

La tabella riporta l'intero set di bias evidenziati nello standard di riferimento, organizzati per tipologia. Per ognuno di essi, l'operatore può indicare nella colonna "esito" se il documento fornito dagli sviluppatori ha affrontato quel particolare bias e, nel caso di esito conforme, quali sono le azioni intraprese. In particolare, le tre colonne assumono il seguente significato:

- Metodo di verifica: se il bias è stato studiato (e.g., bias individuato, bias non presente, bias ignorato) e tramite quale metrica (tra quelle riportate nella sezione 7 della ISO/IEC TR 24027:2021). Questo campo dovrebbe essere compilato in maniera sintetica ma precisa, per permettere una rapida individuazione delle problematiche.
- Azione intrapresa: se un bias è stato individuato, riportare le eventuali azioni correttive o di mitigazione intraprese (e.g., nessuna azione, correzione del dataset, ecc.) e tramite quale tecnica (tra quelle previste nella sezione 8 della ISO/IEC TR 24027:2021). Questo campo dovrebbe essere compilato in maniera sintetica ma precisa, per permettere una rapida individuazione delle problematiche.
- Note: eventuali annotazioni aggiuntive che permettano di meglio comprendere l'esito del controllo (e.g., bias non presente data la natura del task) e/o per motivare l'uso di metriche diverse da quelle presenti nella sezione 7 della ISO/IEC TR 24027:2021;
- Giudizio di conformità: conforme se la gestione di quel bias (se presente) è in linea con la ISO/IEC TR 24027:2021.

La riga finale riporta il numero di check superati con successo. L'operatore può associare alla tabella una serie di informazioni in testo libero per meglio dettagliare gli aspetti più cruciali emersi durante la verifica. Questo approccio garantisce un processo sistematico e trasparente per la certificazione dei sistemi di IA rispetto agli standard di parzialità e assenza di bias.

Use case

Il processo di verifica di conformità proposto sarà applicato ai due casi di studio precedentemente introdotti. Questi serviranno non solo a dimostrare l'efficacia del nostro processo di verifica ma anche a fornire insights preziosi su come i principi di equità e trasparenza possono essere integrati nelle diverse fasi di sviluppo di un sistema di AI. Di seguito, presentiamo i due casi di studio proposti, evidenziando le quattro fasi di sviluppo per ognuno e come la metodologia proposta, ispirata dalla linea guida ISO/IEC TR 24027:2021, possa essere impiegata efficacemente in una verifica di conformità.

A valle di ciò, si applicherà il processo proposto per valutare come il bias sia stato affrontato in ogni fase, specialmente nell'equità di diagnosi tra diversi gruppi demografici, provando (ove necessario) a commentare possibili mitigazioni. Si noti che in entrambi i casi non sarà riportata la fase 4, in quanto il processo di verifica di conformità è simulato prima della fase di messa in esercizio.

Descrizione dello Use Case 1: *detection* del melanoma

Il melanoma è la forma più mortale di cancro della pelle. La diagnosi precoce delle lesioni maligne è fondamentale per ridurre la mortalità. L'uso di tecniche di DL (Deep Learning) sulle immagini dermoscopiche può aiutare a tenere traccia del cambiamento nel tempo dell'aspetto della lesione, che è un fattore importante per il rilevamento precoce di lesioni maligne.

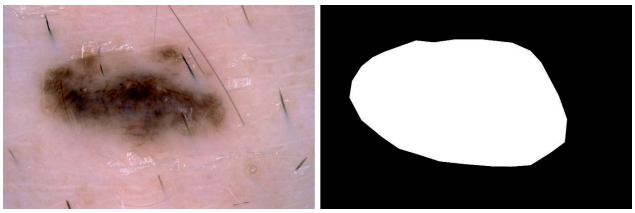
Lo Use Case 1 si concentra sull'analisi di un sistema di IA con una architettura DL denominata Attention Squeeze U-Net per la segmentazione dell'area della lesione cutanea progettata specificatamente per dispositivi embedded (per esempio, smartphone). Lo scopo del sistema di IA è quello di aumentare l'empowerment del paziente attraverso l'adozione di algoritmi di DL che possano essere eseguiti localmente su smartphone o dispositivi embedded a basso costo.

I metodi DL possono essere applicati per affrontare tre compiti principali:

1. Segmentazione dell'area della lesione.
2. Identificazione degli attributi della lesione.
3. Classificazione della malattia.

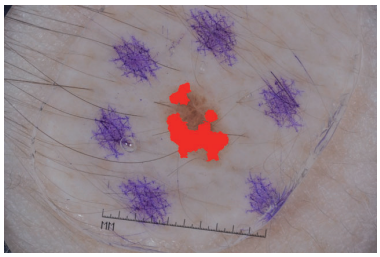
L'obiettivo della segmentazione dell'area della lesione è creare una maschera binaria da un'immagine dermoscopica che fornisca una precisa separazione tra l'area lesionata e quella sana circostante (figura 6).

Figura 6. Segmentazione dell'area della lesione. Sinistra: immagine dermoscopica in ingresso. Destra: maschera binaria in output, dove pixel bianchi appartengono all'area della lesione e i pixel neri gli sono estranei (sorgente: ISIC 2017)



Fonte: ISIC 2017

Figura 7. L'obiettivo nel processo di identificazione degli attributi consiste nel localizzare criteri dermoscopici clinici nell'immagine in input (i globuli sono evidenziati in rosso)



Nella classificazione, le immagini in input sono etichettate in base a diverse classi diagnostiche. Al di là della tipica categorizzazione in benigni e melanomi, è possibile raggruppare le immagini dermoscopiche in più di due classi. Ciò fornisce una migliore discriminazione tra melanoma, altri tipi di cancro della pelle che sono meno aggressivi del melanoma e lesioni benigne. Per esempio, è possibile usare una classificazione basata su sette classi, tra cui melanoma, nevo melanocitico, carcinoma basocellulare, cheratosi attinica, cheratosi benigna, dermatofibroma e lesione vascolare. In questo Use Case 1 si prenderà in esame la segmentazione dell'area della lesione utilizzando un metodo convoluzionale profondo pixel-wise in grado di girare su dispositivi embedded. Mentre qualsiasi interpretazione medica della segmentazione delle lesioni cutanee può essere eseguita solo da parte di esperti, l'anamnesi della lesione può essere creata dal paziente in modo indipendente.

Fase 1: Raccolta e preparazione dei dati

Il "Base Data Pile" è costituito da immagini dermoscopiche e dalle corrispondenti annotazioni di ground-truth provenienti dal dataset ISIC 2017[32]. In particolare, come set di addestramento si utilizzano i seguenti dati proveniente da ISIC 2017:

- Tutte le 2000 immagini dermoscopiche della cartella dei dati di addestramento in formato JPEG.
- Le corrispondenti 2000 immagini di maschere binarie in formato PNG della cartella dei dati di addestramento.

Poiché ISIC 2017 contiene immagini con diverse dimensioni, tutte le immagini sono state ridimensionate al formato 384×512 pixel.

I dati ISIC 2017 presentano alcune criticità, in particolare:

- Un numero considerevole di immagini contiene artefatti come bolle d'aria o di olio, peli del corpo e cerotti colorati.
- L'etichettatura delle lesioni cutanee non segue uno schema predefinito, poiché le annotazioni potrebbero essere state fatte da esperti diversi o con l'ausilio di algoritmi semi-automatizzati.

Fase 2: Sviluppo del modello

Il sistema utilizzato per la rilevazione dei melanomi si basa su reti neurali convoluzionali (CNN) addestrate su un ampio dataset di immagini di melanomi e lesioni cutanee. Le reti neurali convoluzionali sono state scelte perché sono in grado di riconoscere automaticamente le caratteristiche delle immagini, senza la necessità di una programmazione esplicita. Inoltre, le CNN sono in grado di apprendere in modo autonomo le caratteristiche delle immagini, migliorando in tale maniera la precisione del modello. Il sistema elabora le immagini mediche per identificare caratteristiche indicative di melanomi, come la forma, la dimensione, il colore e la simmetria. Inoltre, il sistema utilizza tecniche di segmentazione per isolare le lesioni cutanee dalle aree circostanti. Una volta che le caratteristiche delle immagini sono state identificate, il sistema prende decisioni diagnostiche basate su tali caratteristiche.

Il modello utilizzato per la segmentazione dell'area delle lesioni, in particolare, si chiama Attention Squeeze U-Net e si basa sulle seguenti architetture:

- U-Net;
- Squeeze U-Net;
- Attention U-Net.

---

[32] *Codella, N.C.F., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., Halpern, A.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC) (2018)..*

Architettura U-Net. È un modello encoder-decoder sviluppato per applicazioni mediche e biomedicali. La sua architettura simmetrica, che ricorda la forma di una "U", lo rende particolarmente adatto per la segmentazione di immagini. Gli approcci DL creano nello specifico una mappa delle caratteristiche di un'immagine e la convertono in un vettore, poi utilizzato per la classificazione. Nella segmentazione delle immagini, i metodi DL convertono la mappa delle caratteristiche dell'immagine in un vettore, ma creano anche un'immagine maschera dal vettore creato. A causa della perdita di informazioni nella fase di codifica, la conversione del vettore delle caratteristiche in un'immagine può generare distorsioni. L'idea di U-Net è quella di memorizzare informazioni sulla trasformazione applicata in ogni fase di codifica per utilizzarle nella fase di decodifica, facilitando così la generazione dell'immagine maschera dal vettore delle caratteristiche, preservando la sua integrità strutturale. Tuttavia, nella sua implementazione Keras, U-Net ha più di 30 milioni di parametri addestrabili, un numero considerevole quando si tratta della limitata potenza computazionale e memoria di un dispositivo embedded. La necessità di calcolare milioni di parametri rallenta il processo di inferenza e può generare errori legati alle risorse esaurite.
Architettura Squeeze U-Net. Per ridurre la dimensione del modello, sono state proposte modifiche di U-Net. Squeeze U-Net, in particolare, è un modello efficiente in termini di memoria ed energia ispirato a U-Net, dove i livelli di down- e upsampling sono sostituiti dai moduli fire. Un modulo fire, introdotto in SqueezeNet, utilizza convoluzioni fire point-wise insieme ad uno stadio di inception, che vengono poi concatenati per formare l'output. Il modello Squeeze U-Net, in questa maniera, necessita solo di 2,5 milioni di parametri, ossia più di dieci volte meno di U-Net.
Architettura Attention U-Net. Squeeze U-Net è riuscito a ridurre il numero di parametri da apprendere dai 30 milioni di U-Net a 2,5 milioni. Il meccanismo di concatenazione in Squeeze U-Net può tuttavia essere un fattore limitante quando si tratta di immagini mediche, poiché le caratteristiche di alto livello e le caratteristiche di basso livello sono concatenate tra loro, rischiando di perdere molte informazioni utili. Per risolvere il problema, è possibile introdurre un blocco di attenzione all'interno del blocco di upsampling. Il meccanismo di attenzione è integrato, in particolare, nelle connessioni skip. Come precedentemente accennato, l'idea dietro U-Net è di far guidare le caratteristiche del percorso di contrazione alle caratteristiche del percorso di espansione concatenandole. Applicare un blocco di attenzione prima della concatenazione permette alla rete di capire quali caratteristiche della connessione di skip sono più rilevanti e quindi di ponderare in maniera più efficace. Moltiplicando la connessione di skip e la distribuzione di attenzione, la rete può quindi concentrarsi su una parte particolare dell'input, anziché introdurre la singola caratteristica.
Per aumentare i dati in ingresso al sistema è stata utilizzata una tecnica di data *augmentation* che permette di ottenere nuovi campioni di addestramento a partire dal "Base Data Pile" originale. In particolare, sono state utilizzate tre trasformazioni per ciascuna immagine originale:

- capovolgimento verticale (vertical flipping);
- ribaltamento orizzontale (horizontal flipping);
- entrambi.

La procedura di aumento ha permesso di portare il numero di campioni di addestramento a 8.000 immagini.

Fase 3: Valutazione del modello

Per poter quantitativamente valutare il modello è necessario confrontare:

- Il set di previsioni, costituito dalle maschere binarie di segmentazione generate dal modello addestrato.

- Il set di maschere “ground-truth” create da esperti dermatologi, che rappresenta il nostro obiettivo.

Minore sarà la differenza tra il set di previsioni e il set di ground-truth, maggiore sarà la bontà del modello. Il confronto quantitativo può essere effettuato in termini di:

- positivi “veri” (TP), cioè pixel etichettati come lesione che effettivamente sono appartenenti alla lesione;
- positivi “falsi” (FP), cioè pixel etichettati come lesione che invece non sono appartenenti alla lesione;
- negativi “veri” (TN), cioè pixel etichettati come non lesione che effettivamente sono non appartenenti alla lesione;
- negativi “falsi” (FN), cioè pixel etichettati come non lesione che invece sono appartenenti alla lesione.

Come misura della differenza tra il set di previsioni e il set di ground-truth si sono usate:

- Accuracy (pixel-wise);
- Coefficiente di Dice;
- Indice di somiglianza di Jaccard.

L'accuratezza (a livello di pixel) è la percentuale di pixel nell'immagine di predizione che sono etichettati correttamente e può essere definita come:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Sebbene l'accuratezza sia facile da calcolare e comprendere, non è utile quando le due classi “lesione” e “background” sono estremamente sbilanciate, cioè quando una classe domina l'immagine e l'altra copre solo una piccola parte di essa. Poiché le immagini dermoscopiche sono tipicamente sbilanciate, possono essere usate metriche migliori per affrontare il problema della *detection* della lesione cutanea. Una metrica che può essere applicata in presenza di sbilanciamento delle classi è il coefficiente di somiglianza di Dice (detto anche semplicemente Dice).
Dice misura l'accordo tra due insiemi divisi per la media delle loro dimensioni. In modo formale, tenendo conto dei valori di TP, FP e FN, Dice può essere scritto come:

$$Dice = \frac{TP + TP}{(FP + TP) + (TP + FN)} = \frac{2TP}{2TP + FP + FN}$$

Nel caso della segmentazione di un'immagine, un coefficiente Dice più alto indica che la predizione corrisponde al ground-truth più delle predizioni che producono coefficienti Dice più bassi. Il punteggio Dice riflette sia l'accordo sulle dimensioni che sulla localizzazione ed è quindi più in linea con la qualità percettiva rispetto all'accuratezza pixel-wise. Un’altra metrica che può essere applicata in presenza di sbilanciamento delle classi è l'indice di somiglianza di Jaccard (JSI), che misura la sovrapposizione di due insiemi. L'indice di Jaccard è pari a 0 se i due insiemi sono disgiunti, cioè non hanno membri in comune, mentre è pari a 1 se sono identici. In altre parole, l’obiettivo di un sistema di AI è quello di avvicinarci il più possibile a un JSI pari a 1.

Anche il JSI può essere espresso in termini di TP, FP e FN nel modo seguente:

$$JSI = \frac{TP}{TP + FP + FN}$$

Esiste una versione modificata di JSI denominata Threshold Jaccard (Indice di Jaccard a soglia). Si tratta di una variante di JSI che ha lo scopo di penalizzare i risultati in cui la percentuale di errori FP e FN è superiore a una certa soglia. Per il compito di segmentazione della lesione cutanea, tipicamente l'indice di Jaccard a soglia viene calcolato secondo la regola:

$$Threshold\ Jaccard = \begin{cases} 0, JSI < 0,65 \\ JSI, altrimenti \end{cases}$$

Il valore soglia pari a 0,65 dipende dal dominio di applicazione ed è calcolato empiricamente. La scelta del Threshold Jaccard al posto del JSI si basa sull'osservazione che quest'ultimo non riflette accuratamente il numero di immagini in cui la segmentazione automatizzata fallisce di molto, cioè JSI è eccessivamente ottimistico.
Forniamo ora la valutazione quantitativa del modello di lesion segmentation basato su architettura Attention Squeeze U-Net. Al fine di valutare le prestazioni del nostro approccio, consideriamo il benchmark di test ISIC 2017. La scelta di ISIC 2017 è dovuta alla disponibilità di un ampio set di test annotato (600 immagini) e open source. Nei nostri esperimenti, utilizziamo tutte le 600 immagini dermoscopiche JPEG dalla cartella dei dati di prova e le corrispondenti 600 immagini di maschere binarie in formato PNG dalla cartella dei dati di test ground-truth. Per avere anche una comparazione tra pari, abbiamo confrontato la nostra rete Attention Squeeze U-Net con altre tre reti, cioè U-Net, Attention U-Net e Squeeze U-Net. Per tutte le reti, abbiamo effettuato un addestramento di 100 epoche.
I risultati riportati nella figura 8 mostrano anche che il modello con architettura Attention Squeeze U-Net è in grado di ottenere i migliori risultati (evidenziati in grassetto).

Figura 8. Risultati delle reti sul set di test ISIC 2017 (600 immagini)

Network	Dice	Threshold Jaccard
U-Net	0.8965	0.7591
Attention U-Net	0.8766	0.7043
Squeeze U-Net	0.8987	0.7597
Attention Squeeze U-Net	**0.9035**	**0.7758**

Tabella 6: Use Case 1 "Melanoma" - Schema di controllo da parte dell'organismo di certificazione

Human cognitive biases (sez. 6.2)

Bias	Metodo di verifica	Azione Intrapresa	Note	Giudizio di conformità
Automation	Controllo tramite dermatologo esperto	Tutti i risultati dell'algoritmo sono stati analizzati da un esperto dermatologo con 20 anni di esperienza.	Sono stati evidenziate le debolezze del modello AI per classe di lesione.	Conforme
Group Attribution	Questo aspetto non è stato verificato esplicitamente	E' stato utilizzato il dataset pubblico ISIC 2017 largamente utilizzato nella comunità scientifica.	Il dataset ISIC 2017 potrebbe portare ad bias di tipo "Group Attribution" avendo solo immagini legate ad un certo fototipo.	Non conforme
Implicit	Questo aspetto è stato contrastato utilizzando un secondo dataset per il test	E' stato utilizzato il dataset pubblico ISIC 2017 largamente utilizzato nella comunità scientifica per il training e il dataset PH2 per il test.	Il dataset ISIC 2017 potrebbe portare ad bias di tipo "Implicit" se nel dataset sono presenti più immagini provenienti da un unico paziente o annotate da un'unica persona.	Non conforme
Confirmation	Questo aspetto non è stato contrastato	Per ovviare al confirmation bias sarebbe necessaria l'inclusione di un sufficiente numero di lesioni benigne "facili" da individuare. Questa azione non è stata intrapresa in questo studio.	Il dataset ISIC 2017 presenta una sovrarappresentazione di lesioni benigne difficili da diagnosticare. Andrebbe esteso con un numero elevato di lesioni benigne.	Non conforme
In-Group	Questo aspetto non è stato verificato	Per ovviare all'in-group bias sarebbe necessaria l'inclusione di un sufficiente numero di lesioni provenienti da pazienti con fototipi vari. Questa azione non è stata intrapresa in questo studio.	Le immagini ISIC vengono ottenute principalmente negli Stati Uniti, in Europa e in Australia, pertanto è presente un racial bias legato a lesioni cutanee su pelle chiara.	Non conforme
Out-Group homogeneity	Questo aspetto non è stato verificato	Per ovviare all'out-group bias sarebbe necessaria l'inclusione di un sufficiente numero di lesioni provenienti da pazienti con fototipi vari. Questa azione non è stata intrapresa in questo studio.	Le immagini ISIC vengono ottenute principalmente negli Stati Uniti, in Europa e in Australia, pertanto è presente un racial bias legato a lesioni cutanee su pelle chiara.	Non conforme
Societal	Questo aspetto non è stato verificato	Il pregiudizio e la discriminazione sociale si riflettono nei dataset contenenti dati storici a causa di possibili decisioni "unfair" precedentemente prese dagli esseri umani che hanno collezionato i dati. Va quindi creato un dataset storicamente bilanciato. Questa azione non è stata intrapresa in questo studio.	Il dataset ISIC 2017 rappresenta con un numero elevato di sample le tonalità della pelle più chiare, tuttavia esso non contiene tutte le tonalità della pelle.	Non conforme
Rule-Based	Questo aspetto è stato preso in considerazione coinvolgendo esperti dermatologi nella verifica dei risultati.	È stata realizzata una analisi dell'errore per classe di lesione. Tale analisi ha permesso di evidenziare come un parametro critico per limitare i falsi negativi sia l'utilizzo di immagini con un buon contrasto. Immagini con basso contrasto tendono a produrre un errore nel sistema.	La diagnosi basata sulla sola immagine può essere fortemente migliorata aggiungendo informazioni specifiche correlate come sede anatomica della lesione, sesso, età, fototipo (che potrebbero essere ricavati da un'immagine presa da un contro-sito sano laterale) e altre informazioni anamnestiche.	Conforme
Requirement	Questo aspetto non è stato verificato	Il sistema dovrebbe essere testato su pazienti aventi differenti tonalità della pelle. Questa azione non è stata intrapresa in questo studio.	Il dataset PH2 è stato usato per i test. Esso rappresenta con un numero elevato di sample le tonalità della pelle più chiare, tuttavia esso non contiene tutte le tonalità della pelle. Per tale motivo il sistema di *detection* potrebbe fallire con pazienti aventi fototipi diversi da Fitzpatrick II e III.	Non conforme

Data biases (sez. 6.3)

Bias	Metodo di verifica	Azione Intrapresa	Note	Giudizio di conformità
Data selection (sec. 6.3.2 & 6.3.4.)	Controllo della distribuzione dei campioni rispetto alla popolazione	Non sono state intraprese azioni mitiganti.	Il dataset utilizzato soffre di data selection bias.	Non conforme
Data labelling (sec. 6.3.3 & 6.3.5)	Controllo del metodo di etichettatura del dataset	Non sono state intraprese azioni mitiganti.	Il dataset utilizzato soffre di data labelling bias poiché sono stati usati metodi di labelling automatico su alcuni campioni.	Non conforme
Data processing (sec. 6.3.6)	Alcuni campioni contengono artefatti come bolle d'aria/olio, peli del corpo e cerotti colorati	I campioni sono stati etichettati in modo da evidenziare la presenza di artefatti.	Il sistema di *detection* considera lo sfondo (il background) nelle maschere come una classe, trattando quindi la rilevazione del bordo delle lesioni come un problema di classificazione multi-classe.	Conforme
Simpson's paradox (sec. 6.3.7)	Sono state condotte due valutazioni: 1) considerando tutto il dataset e 2) dividendo i risultati per classe di lesione	Sono stati misurati quantitativamente e qualitativamente sia i risultati sull'intero dataset che sui dati di test divisi per classe.	E' stato possibile identificare classi di lesioni particolarmente problematiche, poiché sia quantitativamente che qualitativamente esse hanno livelli di errore più alti rispetto alla media.	Conforme
Data aggregation (sec. 6.3.8)	La provenienza dei dati per i dataset di training e test è stata verificata.	Non sono state intraprese azioni mitiganti.	Il dataset utilizzato soffre di data aggregation bias.	Non conforme

Engineering biases (sez. 6.3.9 & 6.4)

Bias	Metodo di verifica	Azione Intrapresa	Note	Giudizio di conformità
Distributed training (sec. 6.3.9)	Non è stato utilizzato un metodo di training distribuito in questo sistema.	Nessuna azione necessaria.	Il training del sistema è avvenuto utilizzando un approccio centralizzato.	Conforme
Features (sec. 6.4.2)	Alcune caratteristiche delle immagini sono state prese in considerazione, come la presenza di bolle d'aria/olio, peli del corpo e cerotti colorati	Il problema della rilevazione del bordo delle lesioni viene modellato come un problema di classificazione multi-classe.	Il sistema tratta in modo differenziato le features della pelle rispetto alle features dei peli cutanei e degli elementi esterni, come cerotti e bolle di aria o olio eventualmente presenti nell'immagine dermoscopica.	Conforme
Algorithm (sec. 6.4.3)	Il sistema è stato testato su un dataset pubblico ben noto in letteratura	Il sistema è stato quantitativamente comparato con altri sistemi esistenti in letteratura, sfruttando in particolare i risultati di una competizione pubblica organizzata su dati ISIC.	Il sistema è risultato essere in linea con i risultati allo stato dell'arte della sua pubblicazione (ottobre 2022).	Conforme
Hyperparameter (sec. 6.4.4)	Il sistema è stato progettato per utilizzare un numero ridotto di parametri	Il sistema è basato sull'architettura Attention Squeeze U-Net che utilizza circa 2,6 milioni di parametri.	Il sistema ottiene performce comparabili con reti che utilizzano un numero di parametri anche 10 volte superiore.	Conforme
Informativeness (sec. 6.4.5)	Il sistema soffre di group bias legato alla mancanza di dati relativi ad alcune tonalità di pelle.	Nessuna azione è stata intrapresa per mitigare questo bias.	Il sistema potrebbe essere ri-addestrato su un insieme di immagini dermoscopiche maggiormente rappresentativo della popolazione mondiale.	Non conforme
Model bias (sec. 6.4.6)	Questo aspetto non è stato verificato.	Nessuna azione è stata intrapresa per mitigare questo bias.	Andrebbero eseguiti test comparativi mirati su una classe di dati che sia sotto-rappresentata nel dataset di training.	Non conforme
Model expressiveness (sec. 6.4.7.2)	Una delle peculiarità del sistema risiede nel poter essere eseguito su dispositivi mobili, per tale motivo il numero di parametri è un aspetto cruciale.	Il modello proposto è stato confrontato con altri modelli disponibili in letteratura aventi un numero di parametri simile e maggiore.	Il modello proposto ha performance simili o migliori a tutti gli altri modelli, anche con molti più parametri, presi in considerazione.	Conforme

Descrizione dello Use Case 2: stratificazione dei pazienti con sclerosi multipla

Nella SM (Sclerosi Multipla) esiste un divario riconosciuto tra la ricerca sulla RM (Risonanza Magnetica) e la necessità di stratificare i pazienti in base alla gravità della malattia e al rischio di progressione clinica. In effetti, la disponibilità di biomarcatori di neuroimmagine utili dal punto di vista clinico è fondamentale per identificare tempestivamente i pazienti con una prognosi potenzialmente peggiore e per sintonizzare la gestione clinica in base alle caratteristiche individuali. Il ML offre strumenti preziosi per la stratificazione dei pazienti con SM. Oltre ai modelli prognostici, che mirano esplicitamente a prevedere gli esiti longitudinali sulla base delle caratteristiche della risonanza magnetica, esistono diverse tecniche di ML non supervisionate per modellare la progressione della malattia basandosi esclusivamente sui cambiamenti dei biomarcatori oggettivi.

Il caso in oggetto riguarda la stratificazione dei pazienti affetti da SM in base alla gravità della malattia e al rischio di progressione clinica, al fine di identificare sottogruppi di pazienti con SM che richiedono diverse strategie di gestione e terapia. Questo tipo di analisi è fondamentale per identificare tempestivamente pazienti con prognosi potenzialmente peggiore e adattare la gestione clinica in base alle caratteristiche individuali. Tuttavia, quando si utilizzano metodi di stratificazione basati sull'IA può essere difficile valutare il ruolo dei fattori confondenti non correlati alla malattia, che possono influenzare la gravità della malattia e il rischio di progressione predetti. Questi fattori possono includere l'età del paziente, il sesso, la presenza di altre condizioni mediche, lo stile di vita, la dieta, l'ambiente, la genetica, ecc. Inoltre, il modello di IA utilizzato per la stratificazione può essere influenzato dalla qualità e dalla quantità dei dati utilizzati per l'addestramento. Se i dati non sono rappresentativi della popolazione di pazienti, o se sono stati raccolti in modo non uniforme o incompleto, il modello di IA potrebbe produrre risultati inaccurati.

Fase 1: Raccolta e preparazione dei dati

La metodologia adottata per la raccolta e la preparazione dei dati nell'ambito dello studio sulla stratificazione dei pazienti con sclerosi multipla attraverso l'apprendimento automatico non supervisionato si basa su un approccio strutturato e meticoloso. Questo processo può essere descritto attraverso le seguenti fasi principali, integrate con ulteriori dettagli relativi al dataset utilizzato:

- Selezione dei Partecipanti e Fonti dei Dati: sono stati inclusi pazienti affetti da SM a decorso recidivante-remittente, oltre a controlli sani (HC) e una popolazione esterna di pazienti con SM per il calcolo degli z-score e la selezione delle caratteristiche MRI. I dati sono stati raccolti retrospettivamente dal database radiologico e di ricerca clinica dell'Università di Napoli "Federico II" e dell'AOU Federico II, partendo da ottobre 2006. Il dataset comprende 1129 sequenze MRI di soggetti di diversi sessi ed età, tutti affetti da SM con vario grado di severità.
- Valutazione Clinica: la EDSS (Expanded Disability Status Scale) è stata utilizzata per valutare la disabilità clinica entro una settimana dall'esecuzione della MRI. Inoltre, per i pazienti con valutazioni cliniche e neuropsicologiche a lungo termine disponibili, è stata effettuata una classificazione al follow-up dopo 10 ± 2 anni dalla MRI basale.
- Acquisizione e Elaborazione dei Dati MRI: Utilizzo di uno scanner 3-Tesla, con sequenze 3D T1-pesate per le analisi volumetriche e FLAIR T2-pesate per quantificare il volume totale delle lesioni demielinizzanti (TLV).

- Segmentazione e Parcellizzazione: Segmentazione automatica (e manuale, se necessario) delle lesioni e parcellizzazione della materia grigia in 116 regioni basate sull'Atlante Anatomico Automatico (AAL).
- Preparazione dei Dati per l'Analisi: i volumi delle lesioni demielinizzanti e di 116 regioni della materia grigia sono stati quindi automaticamente segmentati e espressi come punteggi z normalizzati, utilizzando un modello di deep learning (rete neurale convoluzionale 3D) per l'estrazione delle caratteristiche morfologiche. Il calcolo degli z-score è stato effettuato per i volumi delle lesioni e delle regioni di materia grigia rispetto ai controlli sani e alla popolazione esterna di pazienti con SM, invertendo i segni degli z-score quando necessario per mantenere coerenza nell'indicazione del peggioramento.

Il dataset risultante è stato diviso in un subset di traning e uno di test. In particolare, il dataset di training comprende 425 sequenze MRI raccolte durante il primo accesso del paziente, rappresentanti la condizione di baseline (con un'età media di 35,9 ± 9,9 anni, e un rapporto femmine/maschi di 301/124), mentre il dataset di test include 704 sequenze MRI da acquisizioni longitudinali (successive alla baseline), utilizzate per le visite di controllo e ulteriori valutazioni. Le caratteristiche demografiche dei soggetti arruolati sono riportate nella tabella 4.

Figura 9. Caratteristiche demografiche, cliniche e di risonanza magnetica della popolazione studiata

	MS	HC	HCMS (external site)
Number of subjects	425	148	80
Number of MRI scans	1129	148	80
Age (y)	35.9 ± 9.9	35.9 ± 13.0	40.4 ± 11.9
Female Sex*	301 (70.8)	77 (52.0)	56 (70.0)
DD (y)	12.7 ± 8.3	-	10.3 ± 7.4
EDSS**	2.5 (2.0 - 3.5)	-	2.0 (1.5 - 3.0)
TLV (mL)	10.1 ± 13.4	-	3.4 ± 5.3
WBV (mL)	1238.8 ± 127.9	1385.1 ± 147.4	1370.4 ± 153.3

**I dati sono il numero di soggetti, con le percentuali tra parentesi.*
***I dati sono mediani, con intervalli interquartili tra parentesi. SM, sclerosi multipla; HC, controlli sani; DD, durata della malattia; EDSS, Expanded Disability Status Scale; TLV, volume totale della lesione; WBV, volume del cervello intero.*

Fase 2: Sviluppo del modello

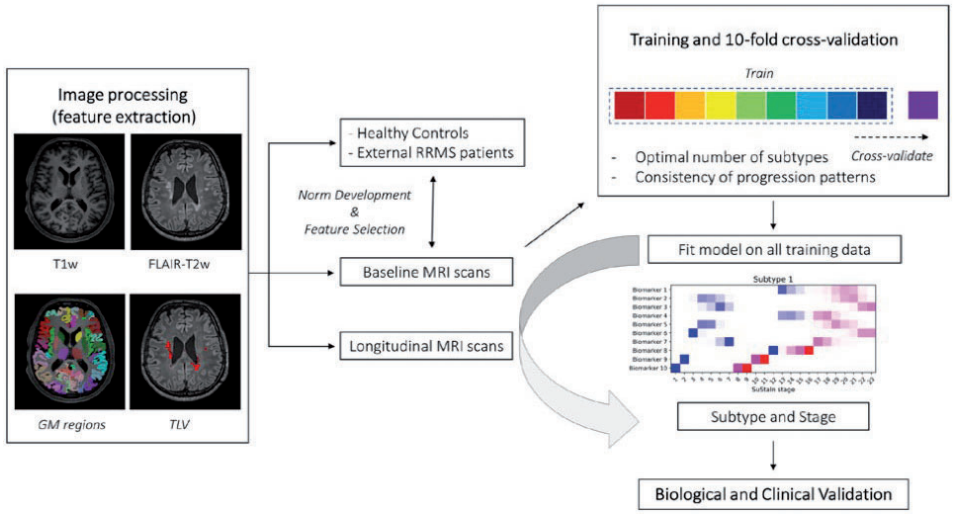
Il sistema è progettato per supportare un medico radiologo nella stratificazione dei pazienti affetti da MS in base alla gravità della malattia e al rischio di progressione clinica. In particolare, il progetto esplora l'impiego dell'apprendimento automatico non supervisionato per classificare i pazienti affetti da sclerosi multipla utilizzando dati di risonanza magnetica cerebrale. L'obiettivo è stratificare i pazienti in gruppi basati su biomarcatori MRI specifici, attraverso l'uso dell'algoritmo SuStaIn (Subtype and Staging Inference). Tale stratificazione punta a essere prognosticamente significativa e biologicamente affidabile.

La metodologia seguita può essere esemplificata come segue:

- Selezione delle caratteristiche: sono state selezionate 11 misure chiave, tra cui il carico di lesioni demielinizzanti e i volumi di regioni specifiche della materia grigia, basandosi su una differenza significativa (Cohen's f > 0,25) rispetto ai controlli sani.
- Applicazione dell'algoritmo non supervisionato: SuStaIn è stato impiegato per identificare sottogruppi di pazienti con pattern distinti di evoluzione dei biomarcatori. Due sottotipi principali sono stati identificati: "deep gray matter (DGM)-first" e "cortex-first", basati sui modelli di atrofia.

L'approccio si concentra sulla modellazione oggettiva della progressione della SM, offrendo agli specialisti uno strumento supportato da dati per la stratificazione dei pazienti in base alla gravità della malattia e al rischio di progressione clinica. La metodologia adottata permette di riflettere accuratamente sulla natura progressiva della SM, associando specifici sottotipi e stadi di malattia a differenze nella disabilità fisica e cognitiva a lungo termine. La procedura è riportata in figura 10.

Figura 10. Rappresentazione schematizzata della procedura di stratificazione utilizzata



Fase 3: Valutazione del modello

La fase di validazione e analisi del modello nel progetto è stata strutturata per assicurare che i sottotipi identificati avessero una solida base clinica e biologica. Questa fase comprende diversi elementi chiave, organizzati per garantire una comprensione approfondita dell'efficacia del modello e della sua applicabilità nel contesto reale della gestione della sclerosi multipla:

- Analisi di regressione e correlazione: queste analisi sono state impiegate per confermare la rilevanza clinica dei sottotipi identificati, verificando la loro correlazione con esiti clinici a lungo termine.

L'obiettivo era di dimostrare che i sottotipi e gli stadi definiti dal modello riflettessero effettivamente variazioni significative nella progressione della malattia e negli esiti per i pazienti.

- Stabilità dei sottotipi e progressione degli stadi: la stabilità dei sottotipi nel tempo e la progressione degli stadi sono state valutate utilizzando l'alpha di Krippendorff, una metrica che esprime la concordanza tra le valutazioni e modelli di regressione lineare multilivello, per analizzare le traiettorie di progressione della malattia all'interno dei sottogruppi identificati.
- Training e valutazione del modello: è stato usato un approccio di tipo 10-fold Cross Validation per testare l'accuratezza e la generalizzabilità del modello su diversi sottoinsiemi di dati. La cross-validation ha permesso di valutare le prestazioni del modello in scenari vari, riducendo il rischio di overfitting e assicurando che le previsioni fossero affidabili.

Da un punto di vista delle performance, data la natura del task (stratificazione dei pazienti rispetto alla progressione longitudinale della patologia) e la tipologia di apprendimento utilizzata (i.e., non supervisionata), sono state utilizzate due metriche di performance ad-hoc:

1. Coefficiente di Bhattacharyya: Questa metrica è stata adottata per misurare la somiglianza tra i modelli di progressione dei sottotipi attraverso le diverse iterazioni della cross-validation. Un alto coefficiente di Bhattacharyya indica una forte somiglianza tra i modelli, suggerendo che il modello è capace di identificare coerentemente i sottotipi di malattia.
2. Alpha di Krippendorff per la stabilità nel tempo: oltre alla sua utilizzazione nella valutazione della stabilità dei sottotipi, l'alpha di Krippendorff ha fornito una misura quantitativa della consistenza con cui i sottotipi sono stati identificati attraverso le diverse fasi temporali del dataset.

In termini numerici, la procedura utilizzata ha permesso di identificare due sottotipi identificati, il "deep gray matter (DGM)-first" (n = 238) e il "cortex-first" (n = 187), con un valore di Krippendorff's $\alpha = 0{,}806$ (indice di una consistenza elevata dei sottotipi nel tempo) e un incremento annuale dello stadio (risultati dell'analisi di regression) significativo con passo annuale ($b = 0{,}20$; $p < 0{,}001$). Questa metodologia di validazione e analisi ha permesso di confermare la validità del modello di stratificazione dei pazienti con SM, assicurando che le classificazioni basate su biomarcatori MRI fossero non solo biologicamente plausibili ma anche clinicamente significative. L'approccio multi-metrica alla validazione ha rafforzato la fiducia nelle capacità predittive del modello, evidenziando la sua potenziale applicazione nella pratica clinica per migliorare la diagnosi, il monitoraggio e la personalizzazione dei trattamenti per i pazienti con SM.

Simulazione di verifica di conformità alla ISO/IEC TR 24027:2021

Tabella 7. Schema di controllo da parte dell'organismo di certificazione per il caso sclerosi

Human cognitive biases (sez. 6.2)

Bias	Metodo di verifica	Azione Intrapresa	Note	Giudizio di conformità
Automation	Controllo tramite radiologo esperto	Il radiologo esperto ha verificato se il processo era a rischio della presenza del bias	Il sistema ha potenziali rischi, ma essendo un CAD, l'automation bias non si pone	Conforme
Group Attribution	Controllo tramite radiologo esperto	Il radiologo esperto ha verificato se il processo era a rischio della presenza del bias	Non può verificarsi data la pipeline realizzata (per paziente)	Conforme
Implicit	Controllo tramite radiologo esperto	Il radiologo esperto ha verificato se il processo era a rischio della presenza del bias	Non può verificarsi data la pipeline realizzata (marcatori clinici noti in letteratura)	Conforme
Confirmation	Controllo tramite radiologo esperto	Il radiologo esperto ha verificato se il processo era a rischio della presenza del bias	Non può verificarsi data la pipeline realizzata (indicatori clinici della patologia noti in letteratura)	Conforme
In-Group	Controllo tramite radiologo esperto	Il radiologo esperto ha verificato se il processo era a rischio della presenza del bias	Dati anonimi, rischio assente	Conforme
Out-Group homogeneity	Controllo tramite radiologo esperto	Il radiologo esperto ha verificato se il processo era a rischio della presenza del bias	Il bias potrebbe essere presente, essendo il CAD pensato per la stratificazione di soli pazienti con sclerosi. Essendo però usato da un medico esperto, non sussiste il rischio	Conforme
Societal	Questo aspetto non è stato verificato	Il pregiudizio e la discriminazione sociale si riflettono nei dataset contenenti dati storici a causa di possibili decisioni "unfair" precedentemente prese dagli esseri umani che hanno collezionato i dati. Non è quindi applicabile al caso in oggetto, essendo in questo caso tutti soggetti con patologia conclamata	Nessuna	Conforme
Rule-Based	Analisi dei risultati con metriche note in letteratura	Le performance del sistema sono state riportate in termini di Alpha di Krippendorff e di Coefficiente di Bhattacharyya, due indicatori relativi all'affidabilità clinica dei risultati	La metodologia di validazione utilizzata ha confermato la validità del modello di stratificazione dei pazienti con SM, assicurando che le classificazioni basate su biomarcatori MRI fossero non solo biologicamente plausibili ma anche clinicamente significative.	Conforme
Requirement	Questo aspetto non è stato verificato	Andrebbe verificato se caratteristiche specifiche del dataset hanno condizionato lo sviluppo del tool/pipeline. Potrebbe essere necessario testare l'intera procedura su un dataset esterno, raccolto in condizioni simili, ma da altri centri		Non conforme

Engineering biases (sez. 6.3.9 & 6.4)

Bias	Metodo di verifica	Azione Intrapresa	Note	Giudizio di conformità
Distributed training (sec. 6.3.9)	Non è stato utilizzato un metodo di training distribuito in questo sistema.	Nessuna azione necessaria	Il training del sistema è avvenuto utilizzando un approccio centralizzato.	Conforme
Features (sec. 6.4.2)	Analisi della procedura tramite esperto	Nessuna azione intrapresa	Sono state selezionate 11 features da parte di un medico esperto. La scelta non è stata motivata dettagliatamente. Andrebbe chiarito l'aspetto	Non conforme
Algorithm (sec. 6.4.3)	Analisi della procedura tramite esperto	Nessuna azione intrapresa	Il sistema è risultato è basato su un tool di ML (sustain) ben noto e riconosciuto dalla comunità	Conforme
Hyperparameter (sec. 6.4.4)	Analisi della procedura tramite esperto	Nessuna azione intrapresa	La stabilità dei risultati ottenuti in cross-validation sembra escludere questo bias	Conforme
Informativeness (sec. 6.4.5)	Analisi della procedura tramite esperto	Nessuna azione intrapresa	Il bias non si pone dato il problema in esame	Conforme
Model bias (sec. 6.4.6)	Analisi della procedura tramite esperto	Nessuna azione intrapresa	Il bias non si pone dato il problema in esame	Conforme
Model expressiveness (sec. 6.4.7.2)	Controllo non effettuato	Nessuna azione intrapresa. Poteva essere necessario confrontare i risultati ottenuti al variare delle feature scelte e/o degli iperparametri		Non conforme

Conclusioni

In base ai risultati trovati rispetto agli use case valutati, è evidente come sia complesso il processo di valutazione di conformità dei sistemi basati sull'IA e come questi sistemi possano includere bias che potrebbero pregiudicare le decisioni mediche.

In tale contesto, l'AI Act stabilisce regole chiare per gli organismi notificati che verificheranno la conformità dei sistemi basati su IA ad alto rischio (tra cui anche i dispositivi medici regolamentati anche dal Regolamento UE 2017/745 e i dispositivi medici diagnostici in vitro, regolamentati anche dal Regolamento UE 2017/746). Gli organismi notificati dovranno rispettare requisiti di indipendenza, competenza e imparzialità. Ruolo fondamentale sarà assunto, secondo quanto presente nell'AI Act, dall'ente nazionale di accreditamento di cui al Regolamento CE 765/2008.

Accredia, in qualità di Ente italiano di accreditamento, potrà dunque offrire il suo contributo nella valutazione e nel monitoraggio degli organismi di valutazione della conformità. Inoltre tali organismi, per essere notificati alla Commissione europea dovranno essere in possesso di un certificato di accreditamento o dimostrare di essere in possesso di tutti i requisiti necessari per la verifica, il riconoscimento e il monitoraggio periodico ai requisiti specifici dell'AI Act.

5.2 Il PoC su INAIL: Quality Management per i sistemi di Intelligenza Artificiale nelle organizzazioni

L'IA sta rapidamente trasformando il modo in cui le Pubbliche Amministrazioni operano, migliorando l'efficienza, la trasparenza e la qualità dei servizi offerti ai cittadini. In questo contesto, l'applicazione di norme tecniche rigorose come la ISO/IEC 42001:2023 sul Quality Management per i sistemi di IA diventa fondamentale per assicurare che tali tecnologie siano implementate in maniera sicura, etica e responsabile. Questo capitolo esplora il PoC sviluppato con INAIL come un esempio chiave di come la ISO/IEC 42001:2023 possa essere applicata nelle pubbliche amministrazioni per migliorare la gestione della qualità dei sistemi di IA. La ISO/IEC 42001:2023 fornisce un quadro di riferimento per la gestione della qualità specificamente adattato ai sistemi di IA, concentrandosi su aspetti cruciali come la trasparenza, la responsabilità, la robustezza e l'equità. Per le Pubbliche Amministrazioni, che operano in ambienti complessi e altamente regolamentati, l'adozione di questo standard è essenziale per garantire che le decisioni basate sull'IA siano coerenti, riproducibili e allineate ai più alti standard di qualità e sicurezza. Infatti, INAIL, come Ente pubblico, gestisce l'assicurazione obbligatoria contro gli infortuni sul lavoro e le malattie professionali, e ricopre un ruolo chiave in tema di salute e sicurezza sul lavoro, e l'implementazione di sistemi di IA conformi agli standard di qualità è fondamentale per svolgere efficacemente questo ruolo. Attraverso il PoC, INAIL ha potuto sperimentare e validare l'efficacia di questi sistemi in un ambiente controllato, identificando best practice e aree di miglioramento che possono essere estese ad altre pubbliche amministrazioni. Nel contesto di questo PoC, la norma ISO/IEC 42001:2023 verrà applicata sul caso di un sistema antifrode basato su sistemi di IA. La descrizione di questo sistema può essere trovata nelle sezioni successive.
L'applicazione della ISO/IEC 42001:2023 nelle Pubbliche Amministrazioni non solo migliora la qualità e l'affidabilità dei sistemi di IA, ma contribuisce anche a costruire la fiducia dei cittadini nei confronti dell'uso dell'IA nel settore pubblico. La trasparenza e la responsabilità, elementi chiave dello standard, sono particolarmente rilevanti in questo contesto, dove le decisioni prese dall'IA possono avere un impatto significativo sulla vita delle persone. Garantire che queste decisioni siano basate su dati accurati, processi trasparenti e principi etici è essenziale per mantenere la legittimità e la credibilità delle istituzioni pubbliche. Inoltre, l'adozione della ISO/IEC 42001:2023 può facilitare la conformità al recentemente introdotto AI Act. I PoC, come quello sviluppato con INAIL, giocano un ruolo cruciale in questo processo, fornendo esempi pratici di come le pubbliche amministrazioni possano adeguarsi alle nuove normative e migliorare la gestione della qualità dei loro sistemi di IA. Di seguito introduciamo i concetti fondamentali della norma 42001:2023 su Quality Management.

5.2.1 La norma ISO/IEC 42001:2023 sul Quality Management

La norma ISO/IEC 42001:2023 stabilisce un quadro per la gestione della qualità dei sistemi di IA all'interno delle organizzazioni. Fornisce i requisiti e le linee guida necessarie per stabilire, implementare, mantenere e migliorare continuamente un sistema di gestione dell'IA, assicurando che l'uso, lo sviluppo e il monitoraggio dei sistemi di IA siano eseguiti in modo responsabile e conforme alle aspettative degli stakeholder e agli obblighi normativi.
Questa norma è stata sviluppata per assistere le organizzazioni di qualsiasi dimensione e settore nell'integrazione efficace dei sistemi di IA nelle loro operazioni quotidiane, garantendo al contempo che tali tecnologie avanzate siano utilizzate in modo etico e sostenibile.

Copre aspetti critici come la leadership, la pianificazione, il supporto, l'operazione, la valutazione delle prestazioni e il miglioramento continuo, fornendo un approccio olistico alla gestione dell'IA. Il focus principale della norma è promuovere un'implementazione dell'IA che non solo rispetti gli standard tecnici e di sicurezza ma che incoraggi anche una cultura organizzativa che valorizzi l'innovazione responsabile e consapevole. Attraverso l'adozione di questa norma, le organizzazioni possono non solo migliorare la loro competitività ma anche contribuire positivamente alla società, gestendo le potenziali implicazioni etiche e sociali legate all'IA.
Nelle successive sezioni, ci concentreremo nell'analisi dei capitoli pertinenti agli obiettivi di questo documento (non riportando informazioni tecniche come ad esempio quelle contenute nel capitolo 3, riguardante i riferimenti normativi).

Capitolo 4

Il capitolo 4 "Context of the organization" della norma ISO/IEC 42001:2023 si focalizza sull'importanza di comprendere a fondo il contesto organizzativo per un'efficace gestione dei sistemi di IA. Esso stabilisce una cornice di riferimento chiave per le organizzazioni che intendono implementare un sistema di gestione per l'IA, assicurando che tali sistemi siano gestiti in modo responsabile e conforme alle esigenze e aspettative degli stakeholder.
La sottosezione 4.1 enfatizza la necessità per le organizzazioni di determinare le questioni interne ed esterne rilevanti ai fini del loro scopo e che influenzano la loro capacità di raggiungere i risultati previsti dal sistema di gestione dell'IA. Questo include considerazioni sui ruoli dell'organizzazione relativi ai sistemi di IA che sviluppa, fornisce o utilizza, e sulle implicazioni di tali ruoli nell'ambito dei sistemi di IA, includendo provider, produttori, clienti, partner e soggetti interessati. È essenziale per l'organizzazione comprendere e gestire i propri ruoli in relazione ai sistemi di IA per garantire che le pratiche di gestione siano adeguatamente allineate con le aspettative legali e di mercato.
La sottosezione 4.2 sottolinea l'importanza di identificare le parti interessate pertinenti e di comprendere le loro esigenze e aspettative rispetto al sistema di gestione dell'IA. L'organizzazione deve determinare quali di queste esigenze e aspettative verranno affrontate tramite il sistema di gestione, essenziale per garantire che le pratiche adottate siano efficaci e rispettino i requisiti delle parti interessate.
Nella sottosezione 4.3, si discute della determinazione dell'ambito del sistema di gestione dell'IA. Questo passaggio è cruciale per definire i confini e l'applicabilità del sistema di gestione, assicurando che sia adeguato al contesto e agli obiettivi specifici dell'organizzazione. L'ambito deve essere documentato e deve riflettere le questioni esterne e interne identificate, nonché i requisiti delle parti interessate.
Infine, la sottosezione 4.4 si occupa della realizzazione del sistema di gestione dell'IA, ponendo le basi per l'istituzione, l'attuazione, il mantenimento e il miglioramento continuo del sistema in conformità ai requisiti della norma. Ciò include definire e documentare i processi necessari e le loro interazioni all'interno dell'organizzazione per assicurare un approccio sistematico e strutturato alla gestione dell'IA.
Complessivamente, il capitolo 4 fornisce una guida essenziale per le organizzazioni nel comprendere e gestire il contesto in cui operano i sistemi di IA, assicurando che le strategie e le pratiche adottate siano robuste, appropriate e allineate con gli obiettivi organizzativi e le aspettative delle parti interessate. Questo è fondamentale non solo per il successo del sistema di gestione dell'IA, ma anche per la sua accettazione e legittimità nel contesto più ampio in cui l'organizzazione opera.

Capitolo 5

Il capitolo 5 "Leadership" della norma ISO/IEC 42001:2023 affronta il ruolo cruciale della leadership nell'efficace gestione dei sistemi di IA all'interno delle organizzazioni. Questa sezione sottolinea l'importanza di un impegno attivo e consapevole da parte della direzione al più alto livello, il cui ruolo è fondamentale per assicurare che il sistema di gestione dell'IA sia completamente integrato nei processi aziendali e allineato agli obiettivi strategici dell'organizzazione. La sottosezione 5.1, "Leadership and commitment", evidenzia specifici compiti che la direzione deve assumere per dimostrare il suo impegno. Tra questi, la direzione deve garantire che le politiche e gli obiettivi specifici per l'IA siano stabiliti e che siano compatibili con la direzione strategica dell'organizzazione. È anche essenziale che la direzione assicuri la disponibilità delle risorse necessarie per il sistema di gestione dell'IA, promuova l'importanza di una gestione efficace dell'IA e sostenga il miglioramento continuo del sistema. Un aspetto cruciale messo in luce è la comunicazione della rilevanza di un efficace sistema di gestione dell'IA e la conformità ai requisiti del sistema di gestione dell'IA da parte di tutti i livelli dell'organizzazione. Questo include la direzione e il supporto del personale per contribuire all'efficacia del sistema di gestione dell'IA. La leadership deve anche promuovere il miglioramento continuo e supportare altri ruoli pertinenti affinché dimostrino a loro volta leadership nelle loro aree di responsabilità. Inoltre, la sezione pone l'accento sulla creazione e la modellazione di una cultura organizzativa responsabile nell'uso, nello sviluppo e nella governance dei sistemi di IA. Questo approccio culturale è visto come una dimostrazione significativa di impegno e leadership da parte della direzione, essenziale per il successo del sistema di gestione dell'IA.
La sottosezione 5.2, "AI policy", si concentra sulla necessità per la direzione di stabilire una politica per l'IA che sia appropriata agli scopi dell'organizzazione, fornisca un quadro per la definizione degli obiettivi di IA, e includa un impegno a soddisfare i requisiti applicabili e al miglioramento continuo del sistema di gestione dell'IA. Questa politica deve essere documentata, comunicata all'interno dell'organizzazione e resa disponibile alle parti interessate, come appropriato.
Infine, la sottosezione 5.3, "Roles, responsibilities and authorities", enfatizza il ruolo della direzione nel garantire che le responsabilità e le autorità per i ruoli rilevanti siano assegnate e comunicate all'interno dell'organizzazione. Ciò include l'assegnazione della responsabilità per garantire che il sistema di gestione dell'IA sia conforme ai requisiti del documento e per riferire sulla performance del sistema di gestione dell'IA alla direzione.
In conclusione, il capitolo sulla leadership della norma ISO/IEC 42001:2023 evidenzia come la direzione al vertice debba non solo mostrare impegno, ma anche guidare attivamente l'implementazione e il miglioramento continuo dei sistemi di gestione dell'IA, creando un ambiente in cui la cultura dell'innovazione responsabile è promossa e sostenuta a tutti i livelli dell'organizzazione.

Capitolo 6

Il capitolo 6 "Planning" della norma ISO/IEC 42001:2023 è dedicato alla pianificazione delle azioni per affrontare rischi e opportunità legati all'implementazione dei sistemi di IA. Questo capitolo è cruciale per le organizzazioni che intendono gestire efficacemente i rischi associati all'IA, garantendo allo stesso tempo che il sistema di gestione dell'IA sia in grado di raggiungere i risultati previsti.
La sottosezione 6.1 "Actions to address risks and opportunities" sottolinea la necessità di un'analisi approfondita dei rischi e delle opportunità. L'organizzazione deve stabilire e mantenere criteri di rischio specifici per l'IA, che aiutino a distinguere i rischi accettabili da quelli non accettabili, facilitando così le valutazioni e il trattamento dei rischi. Questa parte della norma enfatizza l'importanza di integrare e implementare le azioni di mitigazione all'interno dei processi del sistema di gestione dell'IA, e di valutare l'efficacia di tali azioni.

È essenziale che le informazioni documentate su queste azioni siano mantenute, come parte del processo continuo di miglioramento e adattamento della gestione dei rischi legati all'IA .
La sottosezione 6.2 "AI objectives and planning to achieve them" si concentra sulla definizione degli obiettivi specifici per l'IA, che devono essere consistenti con la politica IA dell'organizzazione e misurabili, ove praticabile. Gli obiettivi devono considerare i requisiti applicabili e essere monitorati e comunicati all'interno dell'organizzazione. Questo approccio mira a garantire che gli obiettivi di IA siano chiaramente definiti, gestibili e allineati con gli obiettivi strategici più ampi dell'organizzazione.
Infine, la sottosezione 6.3 "Planning of changes" discute la necessità di pianificare i cambiamenti in modo controllato, assicurando che ogni modifica al sistema di gestione dell'IA sia gestita efficacemente per minimizzare interruzioni e potenziali impatti negativi. L'adattamento e la flessibilità sono sottolineati come componenti essenziali per il mantenimento dell'efficacia del sistema di gestione dell'IA di fronte ai cambiamenti interni ed esterni.
Complessivamente, la sezione "Planning" della norma ISO/IEC 42001:2023 stabilisce una base solida per l'identificazione, valutazione e gestione dei rischi legati all'IA, garantendo che le strategie di mitigazione siano integrate nei processi aziendali e che gli obiettivi di IA siano chiari e ben allineati con la politica e gli obiettivi strategici dell'organizzazione. Questo è fondamentale per la sostenibilità e il successo a lungo termine delle iniziative IA in un contesto aziendale.

Capitolo 7

Il capitolo 7 "Support" della norma ISO/IEC 42001:2023 è focalizzato sul fornire le risorse necessarie, competenze, sensibilizzazione, comunicazione e informazioni documentate per supportare il sistema di gestione dell'IA. Questo capitolo è essenziale per garantire che l'organizzazione abbia le risorse e le competenze necessarie per gestire efficacemente i sistemi di IA.
7.1 Risorse: l'organizzazione deve determinare e fornire le risorse necessarie per stabilire, attuare, mantenere e migliorare continuamente il sistema di gestione dell'IA. Questo include risorse umane, tecnologiche e finanziarie, assicurando che siano adeguatamente allocate per supportare tutte le attività legate all'IA.
7.2 Competenze: è fondamentale che le persone che lavorano sotto il controllo dell'organizzazione abbiano le competenze necessarie, ottenute attraverso l'educazione, la formazione o l'esperienza. L'organizzazione deve prendere misure appropriate per acquisire queste competenze e valutare l'efficacia di tali azioni. Le informazioni documentate devono essere mantenute come prova della competenza.
7.3 Consapevolezza: le persone coinvolte devono essere consapevoli della politica dell'IA, del loro contributo all'efficacia del sistema di gestione dell'IA e delle implicazioni di non conformità ai requisiti del sistema di gestione dell'IA. Questo aiuta a garantire che tutti le persone che lavorano per l'organizzazione comprendano l'importanza delle loro attività e come queste influenzino il successo generale del sistema di gestione dell'IA.
7.4 Comunicazione: l'organizzazione deve stabilire la comunicazione interna ed esterna relativa al sistema di gestione dell'IA, decidendo cosa comunicare, quando, con chi e come. Questo assicura che tutte le parti interessate siano informate adeguatamente e che la comunicazione supporti efficacemente gli obiettivi del sistema di gestione dell'IA.
7.5 Informazioni documentate: la gestione delle informazioni documentate è vitale per la tracciabilità e la verifica delle attività del sistema di gestione dell'IA. L'organizzazione deve includere sia le informazioni richieste dalla norma che quelle determinate come necessarie dall'organizzazione stessa per l'efficacia del sistema di gestione dell'IA.

Ciò include la creazione e l'aggiornamento delle informazioni documentate e il controllo di tali informazioni per garantire che siano disponibili e adeguatamente protette.
In sintesi, il capitolo 7 si concentra sul fornire le risorse e supporti necessari per il funzionamento efficace del sistema di gestione dell'IA, assicurando che l'organizzazione sia equipaggiata per gestire le sfide associate all'uso dell'IA in modo responsabile e efficace.

Capitolo 8

Il capitolo 8 "Operation" della norma ISO/IEC 42001:2023 tratta la pianificazione operativa e il controllo dei sistemi di IA all'interno delle organizzazioni. Questo capitolo è fondamentale per garantire che i processi legati ai sistemi di IA siano gestiti in modo efficace e sicuro, rispettando i criteri stabiliti e rispondendo adeguatamente ai rischi identificati.
La sottosezione 8.1, "Operational planning and control", specifica che l'organizzazione deve pianificare, implementare e controllare i processi necessari per soddisfare i requisiti e le azioni determinate nella Clausola 6, attraverso l'istituzione di criteri per i processi e il controllo degli stessi in conformità ai criteri stabiliti. Questo include l'attuazione dei controlli relativi all'operazione del sistema di gestione dell'AI, come sviluppo e utilizzo del ciclo di vita dei sistemi di IA.
Nella sezione 8.2, "AI risk assessment", si enfatizza l'importanza delle valutazioni periodiche dei rischi legati all'IA, che devono essere eseguite a intervalli pianificati o quando si propongono o si verificano cambiamenti significativi. Questo aiuta a mantenere il sistema di gestione dell'IA allineato con gli eventuali cambiamenti nel contesto operativo e tecnologico.
La sottosezione 8.3, "AI risk treatment", discute come l'organizzazione debba implementare il piano di trattamento dei rischi associati all'IA e verificare la sua efficacia. In caso di identificazione di nuovi rischi che richiedono interventi, deve essere attuato un processo di trattamento dei rischi conforme ai criteri definiti.
Infine, la sezione 8.4, "AI system impact assessment", tratta le valutazioni d'impatto dei sistemi di IA, che devono essere eseguite a intervalli pianificati o in risposta a cambiamenti significativi. Queste valutazioni sono cruciali per comprendere e mitigare gli impatti potenziali dei sistemi di IA sull'organizzazione e sui suoi stakeholder.
In conclusione, la sezione "Operation" stabilisce procedure dettagliate per la gestione e il controllo operativo dei sistemi di IA, assicurando che tutte le fasi del loro utilizzo siano monitorate e gestite in modo da minimizzare i rischi e massimizzare l'efficacia operativa. Questo è essenziale per la sicurezza, l'affidabilità e la conformità dei sistemi di IA all'interno delle organizzazioni.

Capitolo 9

Il capitolo 9 "Valutazione delle prestazioni" della norma è incentrato sull'importanza della misurazione, monitoraggio, analisi e valutazione per assicurare che il sistema di gestione dell'IA funzioni efficacemente e raggiunga gli obiettivi previsti.
9.1 Monitoraggio, misurazione, analisi e valutazione: l'organizzazione deve determinare cosa necessita di essere monitorato e misurato. Inoltre, deve stabilire i metodi appropriati per garantire la validità dei risultati. Le tempistiche per il monitoraggio e la misurazione sono definite, così come i momenti in cui i risultati devono essere analizzati e valutati. La documentazione dei risultati è essenziale come prova dell'efficacia delle misure prese.
9.2 Audit interno: gli audit interni sono condotti a intervalli pianificati per fornire informazioni sulla conformità del sistema di gestione dell'IA rispetto ai requisiti dell'organizzazione e della norma stessa.
È fondamentale stabilire un programma di audit che includa la frequenza, i metodi, le responsabilità, i

requisiti di pianificazione e la relazione dei risultati. Questo approccio assicura che il sistema di gestione dell'IA sia implementato e mantenuto efficacemente.
9.3 Revisione della direzione: la direzione deve esaminare il sistema di gestione dell'IA a intervalli pianificati per assicurare che sia adeguato, appropriato ed efficace rispetto agli scopi dell'organizzazione. Gli input per la revisione includono lo stato delle azioni da revisioni precedenti, i cambiamenti nei fattori esterni e interni che influenzano il sistema, e i risultati del monitoraggio e delle misurazioni. Le revisioni devono portare a decisioni relative al miglioramento continuo e a eventuali modifiche necessarie nel sistema di gestione dell'IA.

Capitolo 10

Il capitolo 10 della norma, denominato "Improvement", si focalizza sull'importanza del miglioramento continuo nel sistema di gestione dell'Intelligenza Artificiale. Questa sezione è essenziale per garantire che le organizzazioni siano in grado di adeguare e perfezionare continuamente il loro approccio alla gestione dell'IA in risposta a nuove sfide, scoperte e risultati operativi.
La sottosezione 10.1, "Continual improvement", evidenzia il requisito per le organizzazioni di migliorare continuamente l'adeguatezza, la pertinenza e l'efficacia del sistema di gestione dell'IA. Ciò implica una valutazione regolare delle prestazioni del sistema di gestione per identificare aree di potenziale miglioramento e l'attuazione di cambiamenti mirati a ottimizzare i processi e i risultati.
La sottosezione 10.2, "Nonconformity and corrective action", tratta della gestione delle non conformità quando si verificano. L'organizzazione è tenuta a reagire prontamente alle non conformità adottando azioni per controllarle e correggerle, nonché gestire le conseguenze. Inoltre, è necessaria un'analisi per determinare le cause delle non conformità e per evitare la loro ricorrenza. Le azioni correttive devono essere appropriate agli effetti delle non conformità incontrate.

Annexes

Gli allegati della norma offrono orientamenti implementativi e dettagli informativi su vari aspetti della gestione dei sistemi di Intelligenza Artificiale. Qui di seguito ne è riassunto il contenuto principale.
Annex A (normativo): fornisce obiettivi di controllo e controlli di riferimento che le organizzazioni possono utilizzare per soddisfare gli obiettivi organizzativi e affrontare i rischi legati alla progettazione e all'operatività dei sistemi di IA. Questo allegato elenca una serie di controlli e obiettivi di controllo che non sono obbligatori per tutte le organizzazioni, ma possono essere adattati e implementati a seconda delle necessità specifiche dell'organizzazione. Nella prossima sezione procederemo ad analizzare le modalità in cui INAIL potrebbe implementare questa norma tecnica proprio sulla base di questo annex.
Annex B (normativo): offre una guida all'implementazione dei controlli elencati nell'Annex A. Questa guida è intesa come un punto di partenza per lo sviluppo di un'implementazione specifica per l'organizzazione dei controlli per il trattamento dei rischi associati all'IA. La guida non è sempre sufficientemente adatta o esaustiva per tutte le situazioni e le organizzazioni possono estenderla o modificarla a seconda delle loro esigenze specifiche di controllo e trattamento dei rischi.
Annex C (informativo): fornisce esempi di obiettivi organizzativi relativi all'IA che possono essere utili per determinare gli obiettivi per l'uso dei sistemi di IA.
Questo allegato aiuta le organizzazioni a identificare fonti di rischio e obiettivi organizzativi che possono essere considerati nella gestione dei rischi.
Annex D (informativo): tratta dell'uso del sistema di gestione IA attraverso diversi domini o settori, suggerendo come le pratiche di gestione dell'IA possono essere applicate in vari contesti organizzativi e settoriali, fornendo una panoramica flessibile e adattabile all'uso dei sistemi di IA.

In conclusione, gli allegati forniscono strumenti essenziali per aiutare le organizzazioni a navigare le complessità della gestione dei sistemi di IA, offrendo sia linee guida normative che informative per l'implementazione e l'adattamento dei controlli di gestione dell'IA a seconda delle esigenze e del contesto specifico dell'organizzazione.

5.2.2 Descrizione del sistema oggetto del PoC con INAIL: il Progetto Antifrode

Nel seguito presenteremo una panoramica delle attività del Progetto Antifrode, evidenziando la creazione e l'implementazione di un sistema informatico-amministrativo finalizzato alla rilevazione e al trattamento di comportamenti anomali anche come presupposto di possibili frodi. Questo sistema è progettato per implementare tempestivamente misure preventive e correttive per affrontare non conformità, sia interne che esterne, compresa la collusione. La struttura del documento è articolata in tre sezioni principali che trattano il contesto, la gestione dei dati e la descrizione dettagliata della soluzione adottata, incluso uno studio di un caso relativo al rischio di infortuni per illustrare l'applicazione delle tecniche di apprendimento automatico. In linea con gli obiettivi strategici e istituzionali, l'INAIL ha adottato una serie di interventi proattivi per la prevenzione e il contrasto delle frodi. Il progetto ha dato vita a un modello operativo e organizzativo, integrando analisi avanzate di fonti informative e tecnologie innovative. Il lavoro svolto richiede un approccio trasversale che coinvolga tutti gli aspetti dell'organizzazione: dal personale ai processi, dalle informazioni alla tecnologia. È stata istituita una Cabina di Regia per stabilire linee guida e strategie operative, supportata da un Nucleo Operativo e da fornitori esterni che si occupano dell'analisi dei dati e della loro presentazione. Questo team multidisciplinare si dedica alla ricerca e alla prevenzione di attività fraudolente, implementando soluzioni personalizzate che utilizzano tecniche di IA e machine learning per identificare sistematicamente le anomalie. L'obiettivo principale del Progetto Antifrode è quello di sviluppare un potente strumento analitico in grado di elaborare grandi volumi di dati, utilizzando indicatori di rischio specifici e algoritmi di machine learning personalizzati.
Questo strumento non solo identifica anomalie e comportamenti atipici ma è anche capace di attribuire un livello di rischio ad ogni entità analizzata. Le finalità specifiche del sistema includono:

- Valorizzazione del patrimonio informativo dell'istituto, anche attraverso l'integrazione con fonti dati esterne.
- Impiego di algoritmi di Intelligenza Artificiale e di machine learning per automatizzare il rilevamento e l'analisi delle anomalie.
- Riduzione della discrezionalità umana, migliorando l'automazione e la tracciabilità delle operazioni e garantendo il rispetto degli standard.
- Classificazione dei comportamenti di ogni entità interagente con l'istituto (aziende, enti, cittadini, ecc.), rendendo disponibili informazioni rilevanti per i processi interni e consentendo controlli integrati.
- Orientamento della pianificazione degli audit.
- Generazione di un effetto deterrente contro comportamenti anomali.
- Analisi dei dati sui comportamenti agiti nelle prassi operative.

Il sistema Antifrode è una soluzione personalizzata che mira a identificare e gestire eventi anomali attraverso un processo strutturato che utilizza diverse fasi chiave e una vasta gamma di algoritmi di Intelligenza Artificiale. Questo sistema è progettato per generare allarmi automatici e gestire le

segnalazioni attraverso una piattaforma centralizzata, facilitando l'analisi e ulteriori verifiche da parte degli analisti. Una volta completate le valutazioni, i risultati vengono utilizzati per affinare continuamente gli algoritmi del sistema, e le anomalie confermate possono essere oggetto di audit.

Le fasi del Processo Antifrode sono:

- Definizione dell'ambito di analisi: la Cabina di Regia determina gli ambiti di analisi e gli indicatori di rischio da incorporare nel sistema per la rilevazione delle anomalie.
- Data discovery: identificazione delle fonti dati rilevanti, analisi del loro contenuto, delle variabili, del tipo di informazioni contenute e delle interazioni tra diverse fonti.
- Addestramento e valutazione degli algoritmi: gli algoritmi utilizzati variano da quelli deterministici a quelli di IA, inclusi algoritmi non supervisionati come il clustering K-Means, algoritmi supervisionati che apprendono da set di dati etichettati, e algoritmi predittivi che stimano la probabilità di eventi futuri.
- Estrazione dei dati e attivazione delle fasi di verifica con ritorni ciclici sui sistemi di riferimento.

Gli algoritmi di IA sono:

- Motore a regole: utilizza criteri specifici da analisi statistiche per identificare e estrarre casi sospetti, in maniera automatica, mediante codice python in grado di attribuire anche un valore di rischio all'entità in esame.
- Motore analitico: impiega algoritmi non supervisionati per individuare alcune caratteristiche intrinseche dei dati e dei nuovi possibili pattern di frodi, senza l'utilizzo di regole a priori. Un esempio è l'algoritmo di clustering, che organizza i dati in gruppi omogenei, e in una fase di post-modeling si individuano le difformità e si attribuisce un Risk Score.
- Motore supervisionato e predittivo: utilizza dati etichettati per mappare relazioni tra input e output, aiutando a fare previsioni accurate. Include fasi di preprocessing, generazione di dati aggiuntivi tramite tecniche come Bootstrapping o simili, e si suddivide in modelli supervisionati che consentono di individuare delle anomalie sulla base di dati etichettati e di definire un risk score a ciascuna entità analizzata, e modelli predittivi, che non solo classificano ma anche predicono nuovi eventi utilizzando dati reinterpretati, ad esempio attraverso l'uso di reti neurali per analisi temporali.

Il sistema non solo identifica le anomalie ma contribuisce anche attivamente alla prevenzione di atti fraudolenti, proteggendo le risorse finanziarie dell'istituto e mantenendo l'integrità delle operazioni. L'approccio proattivo e l'uso di tecnologie avanzate permettono di adattarsi dinamicamente a nuove strategie di frode, garantendo che le prestazioni e i servizi siano erogati in modo equo e conforme alle leggi. Il Sistema Antifrode è progettato per identificare e gestire rischi di frode specifici, classificati in varie aree di attività come Aziende e Lavoratori. Ogni area ha scenari di rischio ben definiti che mirano a mitigare le vulnerabilità settoriali. Nell'Area Aziende, gli scenari includono gestione di compensazioni indebite, problematiche in rapporti assicurativi, rimborsi, e variazioni nei rapporti assicurativi. Questi scenari sono essenziali per prevenire abusi finanziari o amministrativi che possono danneggiare l'istituto. Per esempio, il sistema controlla le compensazioni attraverso il modello F24 per assicurarsi che rispettino le normative vigenti, identificando eventuali benefici a entità inadeguate.
Nell'Area Lavoratori, i rischi sono focalizzati sulla corretta gestione delle pratiche di infortunio e delle opposizioni amministrative o sanitarie, cruciali per garantire l'erogazione equa delle prestazioni e la protezione contro abusi o errori di gestione. L'analisi delle pratiche di infortunio, per esempio, verifica la corrispondenza tra i giorni di inabilità concessi e la gravità dell'infortunio, mentre l'opposizione amministrativa/sanitaria assicura che l'iter di valutazione sia stato eseguito correttamente.

Questo sistema non solo tutela le risorse finanziarie dell'INAIL ma salvaguarda anche l'integrità e la fiducia nelle sue operazioni, assicurando che le prestazioni e i servizi siano erogati in modo equo e conforme alle leggi, garantendo così trasparenza ed efficienza nell'ambito delle operazioni istituzionali.

Questionario per INAIL sui sistemi di gestione della qualità

In questa sezione vengono presentate le risposte al questionario sviluppato da CINI-Accredia per comprendere in che misura i requisiti della ISO/IEC 42001:2023 sono già rispettati, o sono già presi in considerazione, da INAIL nei suoi processi di implementazione di sistemi di IA. Le seguenti domande sono utili, nel contesto del progetto Accredia-CINI-INAIL, per comprendere le modalità con cui INAIL potrebbe, in futuro, applicare la norma tecnica ISO/IEC 42001 "Quality Management" all'interno della propria organizzazione. Va inteso che, comunque, INAIL dovrà operare in conformità allo AI Act e che l'adozione della norma ISO/IEC 42001 permetterà di strutturare il processo inteso a perseguire tale "compliance". Pertanto, l'interesse è quello di comprendere in che modo i requisiti di questa norma potranno essere applicati nei processi di governance già presenti all'interno di INAIL. Le domande, di carattere generale, sono riferite all'adozione dei sistemi di Intelligenza Artificiale di cui si è discusso nelle sessioni in presenza, in particolare in due focus group che hanno visto la partecipazione di rappresentanti di INAIL.

Domanda 1) Chi avrebbe la responsabilità, all'interno di INAIL, della definizione di processi di risk management, comprese le valutazioni di impatto, per l'individuazione e la mitigazione dei rischi relativi ai sistemi di Intelligenza Artificiale? Pur esistendo una funzione di Risk Management, per lo specifico campo dell'IA va considerato il coinvolgimento di diversi attori interni ad INAIL. Questo aspetto è richiesto anche dall'applicazione della norma ISO 31000 o dal "framework" COSO Report, al fine di garantire una maggiore consapevolezza e comprensione del contesto di rischio.

La gestione del rischio in INAIL è articolata secondo le seguenti responsabilità:

- Servizio Ispettorato e Sicurezza: cura la valutazione dei rischi e dei controlli su processi e prodotti, malversazioni o frodi potenziali ai danni dell'Istituto con eventuale proposta di azioni correttive nonché programmazione e coordinamento dell'attività di controllo operativo (Risk Management); è responsabile del coordinamento e monitoraggio delle strutture centrali e territoriali nelle attività connesse all'attuazione delle norme in materia di sicurezza e riservatezza delle informazioni e dati personali.
- Direzione Centrale Organizzazione Digitale: È responsabile del sistema informatico e dell'organizzazione dell'ente e, in particolare, gestisce i rischi dei progetti IT rispetto alla conformità al GDPR (in raccordo con il Responsabile della Protezione dei Dati e il SIS); alla cybersicurezza (sicurezza by design, piano della sicurezza); ai rischi standard dei progetti IT (tecnici, organizzativi, normativi, riferiti agli stakeholder interni ed esterni al progetto, di project management, ecc.).

Inoltre, l'Istituto ha conseguito le seguenti certificazioni:

- UNI CEI ISO/IEC 27001 (con estensione alla ISO/IEC 27017) - Sistema di gestione per la Sicurezza delle informazioni - per il campo di applicazione Servizi di conduzione tecnico-operativa del data center e relativo monitoraggio e Servizi ICT a supporto della denuncia telematica di infortuni, della registrazione della Domanda On-line per i bandi ISI e del suo invio durante l'evento click-day erogato in modalità cloud;
- UNI EN ISO 9001 - Sistema di gestione per la Qualità - per il campo di applicazione Servizi di conduzione tecnico-operativa del data center e relativo monitoraggio; Governo della fornitura ICT e dei relativi livelli di servizio; Gestione dei Centri protesi e riabilitazione sul territorio nazionale;

- ISO/IEC 20000-1 - Sistema di Gestione del Servizio - per il campo di applicazione Servizi di Housing per le Pubbliche Amministrazioni erogati dal Data Center INAIL.

Come invece richiesto dall'art. 9 dell'AI Act, che si riferisce ai sistemi ad Alto Rischio, è fondamentale istituire un modello di gestione dei rischi integrato continuo nel tempo e iterativo. Al momento, INAIL non dispone di un sistema appositamente dedicato all'IA, e questo rappresenta il primo punto da affrontare per anticipare i rischi che potrebbero emergere in futuro da nuove applicazioni ad alto rischio o rischi non monitorati. Per il sistema antifrode oggetto del POC i rischi di progetto sono stati gestiti secondo i processi standard in essere in INAIL, e declinati in base alla diversa natura del rischio: tecnologica, contrattuale, normativa, utenza, tempi, pianificazione, costi, interdipendenze con altri progetti, change management. È in corso la sperimentazione delle risultanze delle analisi. Tutte le operazioni sono gestite all'interno di una piattaforma proprietaria destinata a gestire tutte le tipologie di audit. Inoltre sono stati considerati gli standard ISO 31000 per la gestione dei rischi, ISO 37000 per anticorruzione e trasparenza.

Domanda 2) Quali sarebbero le principali conoscenze, da prendere in considerazione nella formazione dei dipendenti di INAIL, con riferimento all' utilizzo di sistemi di IA? Nel rispondere, si prenda particolarmente in considerazione la formazione degli utenti dei sistemi, e coloro che ne verrebbero direttamente impattati (se rilevante).

L'attività di formazione dovrà coinvolgere tutti coloro che si interfacceranno con il sistema, appartenenti sia ad un profilo tecnico che di business. In fase di definizione dei requisiti di tale attività, sarà necessario approfondire gli elementi fondanti e il fabbisogno formativo diversificato a seconda del personale coinvolto, nonché l'impatto e le ricadute organizzative sulle attività e sui ruoli da formare al fine di pervenire ad un efficace e mirato apprendimento. In linea generale, si potrebbero considerare le seguenti macro-aree:

- Formazione base sui sistemi di IA (per tutto il personale dell'Istituto): il personale dell'Istituto dovrebbe acquisire conoscenza di base sui sistemi di Intelligenza Artificiale. Questa formazione dovrebbe includere una comprensione dei principi fondamentali e generali sull'utilizzo di tali sistemi in INAIL.
- Comprensione dell'impatto operativo, etico e legale del sistema (per tutto il personale dell'Istituto): oltre a comprendere l'impatto operativo dei sistemi di IA sul business e il valore aggiunto che apportano, i dipendenti dovrebbero essere consapevoli delle questioni etiche e legali correlate all'uso di tali tecnologie.
- Capacità di comprendere come il sistema prende le decisioni e, se necessario, ignorare l'output del sistema (formazione specifica agli utenti finali dei sistemi di IA): è importante che gli utenti finali dei sistemi di IA in essere siano in grado di comprenderne il funzionamento interno e le loro implicazioni sul piano dei diritti fondamentali delle persone. Ciò è importante per sensibilizzarli a interpretare l'output del sistema e agire in modo critico, individuando anche i casi in cui l'output sia da eludere perché non rappresenta fedelmente la realtà fattuale o il risultato desiderato.
- Capacità di monitorare la performance del sistema (gruppo di progetto): i dipendenti che fanno parte del gruppo di progetto dovrebbero essere in grado di monitorare costantemente la performance dei sistemi di IA utilizzati. Questo include la capacità di raccogliere e analizzare dati sulle prestazioni del sistema attraverso specifici KPI, valutare l'accuratezza e l'affidabilità dell'output e identificare eventuali anomalie. Il lavoro di monitoraggio, che è alla base di ogni processo di gestione de rischi, è particolarmente importante, per agire tempestivamente di fronte a problemi o malfunzionamenti del sistema.

Inoltre, come prassi già consolidata in INAIL, anche in relazione a queste tematiche, è opportuno sviluppare all'interno dell'Istituto competenze trasversali quali strumenti per la costruzione di una cultura identitaria a supporto dell'evoluzione tecnologica.

Domanda 3) Quale percorso di formazione sarà definito, in considerazione della creazione di consapevolezza e allo sviluppo delle necessarie competenze, nelle diverse fasi di tali progetti: definizione obiettivi, sviluppo tecnico, educazione e addestramento, manutenzione.

La formazione dei dipendenti, con i necessari approfondimenti sopra citati, dovrà essere strutturata in modo differenziato, considerando l'impatto dei sistemi sulle loro responsabilità e sulla natura specifica del loro ruolo. Ciò significa che la formazione deve soddisfare i fabbisogni formativi delle diverse figure implicate, considerando inoltre la differenza tra figure tecniche e non tecniche. Le competenze in materia di formazione del personale dell'Istituto sono in capo alla Direzione centrale risorse umane, che, in particolare, previo approfondimento di contesto e di tematica, nonché di impatti e ricadute organizzative, effettua l'analisi dei fabbisogni formativi e cura l'elaborazione degli interventi di formazione. Annualmente, a cura di questa Direzione, e in raccordo con le diverse strutture dell'INAIL, viene elaborato il Piano triennale della formazione.
Nel Piano della formazione 2023-2025, è già presente un'iniziativa formativa denominata "Data governance" che ha l'obiettivo di supportare la trasformazione digitale e gli obblighi che ne derivano, favorire l'evoluzione organizzativa dell'Istituto in ottica data-driven sviluppando la capacità di utilizzare i dati per migliorare il processo decisionale, rafforzare analisi e strategie, identificare problemi e trovare soluzioni, compiere scelte innovative e sostenere tecnologie "disruptive". A tal proposito dovranno essere analizzati i potenziali elementi di interrelazione tra governance dei dati e governance dell'IA. Pertanto, relativamente alla sezione formazione, nell'individuare iniziative mirate da includere nel PIAO 2025/2027 è necessario individuare i destinatari, i contenuti diversificati secondo i ruoli, nonché le modalità più idonee al trasferimento delle conoscenze.

Domanda 4) Quali dipartimenti sarebbero responsabili della definizione della governance e delle policies interne a INAIL per la gestione dei sistemi di IA? A questo scopo pare giusto notare che potrebbe esserci l'interessamento di diversi dipartimenti, che necessiteranno di formazione.

L'assegnazione di responsabilità tra le diverse strutture dell'Istituto riflette la complessità e l'interdisciplinarietà delle questioni coinvolte nella gestione dei sistemi di IA e dovrà pertanto garantire un approccio integrato e coordinato alla definizione delle politiche dell'istituto. Il relativo modello di governance dell'IA è ancora oggetto di approfondimento e analisi.

Domanda 5) Quale dipartimento/ufficio sarà responsabile delle politiche di comunicazione con le parti interessate, anche con riferimento alle comunicazioni in caso di criticità?

La Direzione centrale pianificazione e comunicazione, è responsabile della comunicazione dell'Istituto, realizza campagne e iniziative di comunicazione dirette ad informare l'utenza interna ed esterna. Le campagne informative, valoriali e/o di prodotto, sono promosse, oltre che dalla Direzione stessa, anche dalle Strutture dell'Istituto competenti per materia e in alcuni casi in collaborazione con i Ministeri vigilanti e altri Organismi del welfare. Riguardo alla "comunicazione con gli stakeholder" nella norma ISO/IEC 42001:2023 (7.4 Comunicazione) è precisato che "L'organizzazione deve determinare le comunicazioni interne ed esterne pertinenti al sistema di gestione dell'IA, inclusi: cosa comunicare; quando comunicare; con chi comunicare; come comunicare".

Domanda 6) Stante il fatto che l'introduzione di strumenti di IA in una qualsiasi organizzazione comporta dei cambiamenti organizzativi e/o di processo e nuove/diverse responsabilità, chi avrà in carico il processo di cambiamento?

La gestione del processo di cambiamento in INAIL è affidata, per le rispettive competenze a:

- DC organizzazione digitale, per gli aspetti di organizzazione e IT;
- DC risorse umane, per gli aspetti di formazione e di gestione del personale;
- DC pianificazione e comunicazione, per gli aspetti di comunicazione interna/esterna;
- Servizio Ispettorato e Sicurezza per gli aspetti di Risk Management, Sicurezza e Controllo cooperativo;
- Strutture centrali e territoriali coinvolte nelle specifiche progettualità.

Si valuterà se mutuare e specificare questo modello anche per il Governo del processo di cambiamento determinato dall'introduzione di sistemi di IA.

Domanda 7) La gestione dei requisiti normativi avrà un certo grado di affinità con gli altri processi di gestione già oggetto di certificazione. Ritenete che potrà essere considerata un'integrazione sistemica?

È fondamentale una definizione organica delle policy e della compliance alle diverse fonti normative (normativa di settore, privacy, norme ISO, ecc.). Saranno pertanto analizzate e valutate: aree di complementarità, coerenza, integrazione e omogeneità di gestione. In questo quadro potrebbe essere valutata l'eventuale estensione delle politiche di certificazione che progressivamente, in chiave di generazione di valore pubblico, dovrebbero riguardare tutte le attività dell'Istituto.

Simulazione di verifica di conformità alla norma ISO/IEC 42001:2023

Tabella 8. Tavola di controllo INAIL secondo ISO/IEC 42001 (sulla base di Annex A - Informativo)

Obiettivo	Controllo (ISO/IEC 42001)	Implementazione specifica per INAIL
A.2 Politiche relative all'IA		
Fornire direzione e supporto gestionale per i sistemi di IA secondo i requisiti aziendali.	**A.2.2 Documentazione della politica IA**	Formalizzare una politica IA che rifletta gli obiettivi del Progetto Antifrode e integrare normative sulla sicurezza dei dati e privacy. Questa azione, nell'implementazione della ISO/IEC 42001:2023 spetterebbe a tutti gli attori coinvolti, ciascuno per le proprie responsabilità.
	A.2.3 Allineamento con altre politiche organizzative	Verificare e garantire che la politica IA sia coerente con altre politiche esistenti, come quelle sulla sicurezza e sulla conformità normativa. Questa azione, nell'implementazione della ISO/IEC 42001:2023 spetterebbe a DC organizzazione digitale, per le policy relative alla digitalizzazione, e al Servizio Ispettorato e Sicurezza, per le policy relative ai processi di ispezione.
	A.2.4 Revisione della politica IA	Stabilire intervalli regolari per la revisione della politica IA per assicurare la sua adeguazione e efficacia nel tempo. Questa azione, nell'implementazione della ISO/IEC 42001:2023, spetterebbe a tutti gli attori coinvolti.
A.3 Organizzazione interna		
Stabilire responsabilità interne per mantenere un approccio responsabile nell'implementazione, gestione e operatività dei sistemi di IA.	**A.3.2 Definizione dei ruoli e delle responsabilità per l'IA**	Definire chiaramente i ruoli e le responsabilità all'interno del Progetto Antifrode, assicurando la copertura di tutte le fasi operative e decisionali. Questa azione, nell'implementazione della ISO/IEC 42001:2023, spetterebbe a DC organizzazione digitale e Servizio Ispettorato e Sicurezza per ripartire le responsabilità nei rispettivi ruoli.

Obiettivo	Controllo (ISO/IEC 42001)	Implementazione specifica per INAIL
	A.3.3 Segnalazione di preoccupazioni	Implementare un processo formale per la segnalazione di preoccupazioni riguardanti l'IA, incoraggiando una cultura di trasparenza e responsabilità. Questa azione, nell'implementazione della ISO/IEC 42001:2023, spetterebbe al Servizio Ispettorato e Sicurezza, responsabile del Risk Management e della Sicurezza.
A.4 Risorse per sistemi IA		
Assicurare che l'organizzazione consideri tutte le risorse necessarie per comprendere e gestire i rischi e gli impatti dei sistemi di IA.	**A.4.2 Documentazione delle risorse**	Documentare in modo completo tutte le risorse, inclusi hardware, software, dati e personale, impiegate nel sistema Antifrode. Questa azione, nell'implementazione della ISO/IEC 42001:2023, spetterebbe al Servizio Ispettorato e Sicurezza, che comunicherebbe i requisiti di risorse a DC organizzazione digitale.
	A.4.3 Risorse dati	Mantenere una documentazione dettagliata delle fonti di dati, inclusi i processi di acquisizione, utilizzo e gestione in conformità al GDPR. Questa azione, nell'implementazione della ISO/IEC 42001:2023, spetterebbe a DC organizzazione digitale, in collaborazione con il Servizio Ispettorato e Sicurezza.
	A.4.4 Risorse strumentali	Elencare e documentare gli strumenti analitici e di monitoraggio utilizzati nel sistema Antifrode, garantendo l'aggiornamento e la manutenzione. Questa azione, nell'implementazione della ISO/IEC 42001:2023, spetterebbe a DC organizzazione digitale.
	A.4.5 Risorse di sistema e di calcolo	Specificare le infrastrutture IT impiegate, comprese le capacità di calcolo necessarie per il processamento e l'analisi dei dati. Questa azione, nell'implementazione della ISO/IEC 42001:2023, spetterebbe a DC organizzazione digitale.
	A.4.6 Risorse umane	Identificare e qualificare il personale coinvolto, evidenziando competenze e necessità formative specifiche per il lavoro con l'IA. Questa azione, nell'implementazione della ISO/IEC 42001:2023, spetterebbe a DC risorse umane, in collaborazione con Servizio Ispettorato e Sicurezza.
A.5 Valutazione degli impatti dei sistemi di IA		
Valutare gli impatti del sistema di IA su parti interessate durante tutto il ciclo di vita del sistema.	**A.5.2 Processo di valutazione degli impatti del sistema di IA**	Stabilire metodologie per valutare gli impatti etici, legali e operativi del sistema Antifrode, con particolare attenzione alle implicazioni sui diritti individuali. Questa azione, nell'implementazione della ISO/IEC 42001:2023, spetterebbe a tutti gli attori coinvolti, in particolare al Servizio Ispettorato e Sicurezza, responsabile del risk management.
	A.5.3 Documentazione delle valutazioni d'impatto	Mantenere una documentazione completa delle valutazioni d'impatto, inclusi i metodi utilizzati e i risultati ottenuti.
	A.5.4 Valutazione degli impatti su individui e gruppi	Analizzare specificamente come le decisioni automatizzate possano influenzare individui e gruppi, specialmente in termini di equità e giustizia. Questa azione, nell'implementazione della ISO/IEC 42001:2023, spetterebbe alle procedure di risk management, e quindi al Servizio Ispettorato e Sicurezza.
	A.5.5 Valutazione degli impatti sociali	Valutare e documentare gli impatti potenziali del sistema Antifrode sulla società, compresi gli effetti a lungo termine sulla fiducia e sulla sicurezza pubblica. Questa azione, nell'implementazione della ISO/IEC 42001:2023, spetterebbe al risk management, e quindi al Servizio Ispettorato e Sicurezza.
A.6 Ciclo di vita del sistema di IA		
Definire i criteri e i requisiti per ciascuna fase del ciclo di vita del sistema di IA.	**A.6.1.2 Obiettivi per lo sviluppo responsabile del sistema di IA**	Identificare e documentare gli obiettivi per guidare lo sviluppo del sistema Antifrode, garantendo che le misure adottate rispettino principi di equità e trasparenza. Questa azione, nell'implementazione della ISO/IEC 42001:2023, spetterebbe a DC organizzazione digitale.

Obiettivo	Controllo (ISO/IEC 42001)	Implementazione specifica per INAIL
	A.6.1.3 Processi per la progettazione e sviluppo fidato del sistema di IA	Definire e documentare i processi specifici per la progettazione e lo sviluppo del sistema Antifrode, includendo misure di sicurezza e revisioni periodiche. Questa azione, nell'implementazione della ISO/IEC 42001:2023, spetterebbe a DC organizzazione digitale.
	A.6.2.2 Requisiti e specifiche del sistema di IA	Specificare e documentare i requisiti per i nuovi sistemi di IA o per miglioramenti sostanziali ai sistemi esistenti, focalizzandosi sulle esigenze specifiche dell'Antifrode. Questa azione, nell'implementazione della ISO/IEC 42001:2023, spetterebbe a DC organizzazione digitale, che prenderebbe i requisiti non funzionali dal Servizio Ispettorato e Sicurezza.
	A.6.2.3 Documentazione di progettazione e sviluppo del sistema di IA	Mantenere documentazione dettagliata delle fasi di progettazione e sviluppo del sistema Antifrode, inclusi i cambiamenti e le modifiche apportate nel tempo. Questa azione, nell'implementazione della ISO/IEC 42001:2023, spetterebbe a DC organizzazione digitale.
	A.6.2.4 Verifica e validazione del sistema di IA	Definire e documentare le misure di verifica e validazione per il sistema Antifrode, assicurando che il sistema funzioni come previsto e sia privo di bias dannosi. Questa azione, nell'implementazione della ISO/IEC 42001:2023, spetterebbe a DC organizzazione digitale, in collaborazione con il Servizio Ispettorato e Sicurezza.
	A.6.2.5 Implementazione del sistema di IA	Documentare un piano di implementazione che includa requisiti di test e criteri di accettazione prima del rilascio effettivo nel contesto operativo. Questa azione, nell'implementazione della ISO/IEC 42001:2023, spetterebbe a tutti gli attori coinvolti.
	A.6.2.6 Operazione e monitoraggio del sistema di IA	Definire e documentare i componenti necessari per l'operatività continua del sistema Antifrode, includendo monitoraggio delle prestazioni e supporto tecnico. Questa azione, nell'implementazione della ISO/IEC 42001:2023, spetterebbe a DC organizzazione digitale e al Servizio Ispettorato e Sicurezza, ognuno per le proprie mansioni.
	A.6.2.7 Documentazione tecnica del sistema di IA	Determinare la documentazione tecnica necessaria per le diverse categorie di parti interessate e fornirla in forma appropriata. Questa azione, nell'implementazione della ISO/IEC 42001:2023, spetterebbe a DC organizzazione digitale.
	A.6.2.8 Registrazione dei log degli eventi del sistema di IA	Stabilire in quali fasi del ciclo di vita del sistema di IA sia necessario abilitare la registrazione dei log degli eventi, specialmente durante l'utilizzo del sistema. Questa azione, nell'implementazione della ISO/IEC 42001:2023, spetterebbe a DC organizzazione digitale.
A.7 Dati per i sistemi IA		
Assicurare la comprensione del ruolo e degli impatti dei dati nei sistemi IA durante le fasi di applicazione e sviluppo, fornitura o uso dei sistemi di IA.	**A.7.2 Dati per lo sviluppo e il miglioramento del sistema di IA**	Definire, documentare e implementare processi di gestione dei dati che supportano lo sviluppo del sistema Antifrode, assicurando l'integrità e la protezione dei dati. Questa azione, nell'implementazione della ISO/IEC 42001:2023, spetterebbe a DC risorse umane e DC organizzazione digitale.
	A.7.3 Acquisizione dei dati	Determinare e documentare i dettagli relativi all'acquisizione e alla selezione dei dati utilizzati nei sistemi IA, garantendo la loro rilevanza e qualità. Questa azione, nell'implementazione della ISO/IEC 42001:2023, spetterebbe a DC organizzazione digitale.
	A.7.4 Qualità dei dati per i sistemi di IA	Definire e documentare i requisiti di qualità dei dati, assicurando che i dati utilizzati nello sviluppo e nell'operatività del sistema Antifrode soddisfino tali standard. Questa azione, nell'implementazione della ISO/IEC 42001:2023, spetterebbe a tutti gli attori coinvolti.
	A.7.5 Provenienza dei dati	Definire e documentare un processo per registrare la provenienza dei dati utilizzati nei sistemi di IA di INAIL, inclusi i dettagli su come i dati vengano raccolti, usati e conservati. Questa azione, nell'implementazione della ISO/IEC 42001:2023, spetterebbe al servizio Ispettorato e Sicurezza, se proprietario delle banche dati, e a DC organizzazione digitale nelle sue funzioni di verifica della qualità dei dati.

Obiettivo	Controllo (ISO/IEC 42001)	Implementazione specifica per INAIL
	A.7.6 Preparazione dei dati	Definire e documentare le esigenze e le metodologie per la preparazione dei dati utilizzati nel sistema Antifrode, assicurando l'accuratezza e l'adeguatezza dei dati per l'analisi specifica. Questa azione, nell'implementazione della ISO/IEC 42001:2023, spetterebbe a DC organizzazione digitale.
A.8 Informazioni per le parti interessate		
Assicurare che le parti interessate abbiano le informazioni necessarie per comprendere e valutare i rischi e i loro impatti.	**A.8.2 Documentazione e informazioni di sistema per gli utenti**	Determinare e fornire le informazioni necessarie agli utenti del sistema, comprese guide e FAQ dettagliate sul funzionamento e le misure di sicurezza del sistema Antifrode. Questa azione, nell'implementazione della ISO/IEC 42001:2023, spetterebbe a DC organizzazione digitale per la creazione, a DC Pianificazione e comunicazione per la redazione, e DC risorse umane per la pubblicizzazione.
	A.8.3 Reporting esterno	Fornire meccanismi per le parti interessate per segnalare impatti avversi del sistema, includendo canali di comunicazione chiari e accessibili.
	A.8.4 Comunicazione degli incidenti	Stabilire un piano documentato per la comunicazione degli incidenti, assicurando che le informazioni raggiungano tutte le parti rilevanti in modo tempestivo e accurato. Questa azione, nell'implementazione della ISO/IEC 42001:2023, spetterebbe a tutti gli attori coinvolti, nelle loro mansioni.
	A.8.5 Informazioni per le parti interessate	Documentare e divulgare obbligazioni e responsabilità relative al sistema di IA alle parti interessate, mantenendo trasparenza e conformità normativa. Questa azione, nell'implementazione della ISO/IEC 42001:2023, spetterebbe a DC pianificazione e comunicazione.
A.9 Uso dei sistemi di IA		
Garantire che l'organizzazione utilizzi i sistemi di IA in modo responsabile e secondo le politiche organizzative.	**A.9.2 Processi per l'uso responsabile dei sistemi di IA**	Definire e documentare i processi per l'uso responsabile del sistema Antifrode, includendo protocolli per l'interazione umana laddove necessario. Questa azione, nell'implementazione della ISO/IEC 42001:2023, spetterebbe al Servizio Ispettorato e Sicurezza per la redazione dei processi di supervisione, a DC organizzazione digitale per la loro implementazione tecnica.
	A.9.3 Obiettivi per l'uso responsabile del sistema di IA	Identificare e documentare gli obiettivi per guidare l'uso responsabile del sistema di IA, assicurando che questi riflettano gli standard etici e legali. Questa azione, nell'implementazione della ISO/IEC 42001:2023, spetterebbe ai ruoli coinvolti nella pianificazione strategica e al Servizio Ispettorato e Sicurezza.
	A.9.4 Uso inteso del sistema di IA	Assicurarsi che il sistema di IA venga utilizzato solo per gli scopi previsti, con controlli adeguati per prevenire usi impropri o non autorizzati. Questa azione, nell'implementazione della ISO/IEC 42001:2023, spetterebbe al Servizio Ispettorato e Sicurezza e al sistema dei controlli.
A.10 Relazioni con terze parti e clienti		
Assicurare che l'organizzazione comprenda le sue responsabilità e rimanga responsabile quando le terze parti sono coinvolte in qualsiasi fase del ciclo di vita del sistema di IA.	**A.10.2 Assegnazione delle responsabilità**	Garantire che le responsabilità siano chiaramente assegnate tra INAIL e le terze parti coinvolte nel ciclo di vita del sistema di IA, stabilendo accordi chiari e trasparenti. Non è chiaro a chi spetterebbe tale responsabilità.
	A.10.3 Fornitori	Stabilire processi per assicurare che l'uso di servizi, prodotti o materiali forniti dai fornitori sia in linea con l'approccio responsabile di INAIL allo sviluppo e uso dei sistemi di IA. Non è chiaro a chi spetterebbe tale responsabilità.
	A.10.4 Clienti	Assicurarsi che l'approccio di INAIL allo sviluppo e uso dei sistemi di IA consideri le aspettative e le esigenze dei clienti, integrando feedback e osservazioni per miglioramenti continui. Questa azione, nell'implementazione della ISO/IEC 42001:2023, spetterebbe a DC Pianificazione e Comunicazione.

Conclusioni

Esiste un ritardo fisiologico tra il ritmo dello sviluppo tecnologico e quello del diritto. La regolamentazione di fenomeni in rapida evoluzione non è compito facile. L'esigenza di disciplinare realtà e attività complesse come l'evoluzione dei sistemi di IA è quella di porre limiti agli effetti sui profili economici etici e sociali, oltre che sui diritti fondamentali dei cittadini, utilizzatori o meno di tecnologia. L'impostazione, in Europa, rispetto alla regolamentazione dell'IA, risponde a un preciso obiettivo di bilanciamento tra tutela dei diritti e sviluppo tecnologico. Rispetto ad altri dispositivi normativi e linee guida promosse in altre aree economiche, l'AI Act, pubblicato il 12 luglio 2024, prevede un impianto sanzionatorio che lo rende vincolante per i Paesi membri. Oltre a definire una classificazione dei sistemi di IA per classi di rischio, prevedendone la mitigazione, si occupa di determinarne la governance a livello europeo.

La decisione della Commissione europea del 24 gennaio 2024, volta a istituire l'Ufficio sull'Intelligenza Artificiale (AI Office) a partire dal 21 febbraio 2024, rappresenta un primo passo significativo verso la creazione di una struttura istituzionale dedicata alla regolamentazione e alla promozione dell'IA nell'Unione europea. Tra i compiti principali dell'AI Office, vi sono il contributo a un approccio coerente, strategico ed efficace dell'Unione europea verso le iniziative internazionali sull'IA, oltre alla promozione dell'adozione di sistemi affidabili e al monitoraggio dell'evoluzione del mercato settoriale. Inoltre, l'AI Office collabora con gli stakeholder e le Autorità degli Stati membri per conto della Commissione. È inoltre incaricato di sostenere la Commissione nell'elaborazione di decisioni, atti di attuazione e atti delegati, nonché di protocolli standardizzati e migliori pratiche relativi all'AI Act. Sarà coinvolto nella preparazione di richieste di standardizzazione, nella valutazione degli standard esistenti e nella definizione di specifiche comuni per l'attuazione dell'AI Act. Fornirà supporto tecnico e consulenza, implementando strumenti per la creazione e la gestione delle sandbox per l'IA, coordinandosi, se necessario, con le Autorità nazionali competenti responsabili della loro istituzione. Infine, condurrà valutazioni, revisioni e preparerà rapporti relativi all'attuazione dell'AI Act.

Oltre all'AI Office sono previste ulteriori figure istituzionali che integreranno la governance europea sull'Intelligenza Artificiale. In particolare:

- European Artificial Intelligence Board (EAIB), composto da rappresentanti nominati da ciascuno Stato membro. Il ruolo principale dell'EAIB sarà quello di garantire un'attuazione armonizzata dell'AI Act tra gli Stati membri, offrendo la propria consulenza sia alla Commissione che agli Stati stessi. L'EAIB dovrà anche organizzare sottogruppi permanenti per facilitare la cooperazione tra le Autorità di sorveglianza del mercato e gli organismi notificati.

- European Data Protection Supervisor (EDPS) parteciperà come osservatore.
- Forum Consultivo rappresenterà un ulteriore meccanismo di consultazione e coinvolgimento degli stakeholder nell'implementazione del regolamento sull'IA.
- Comitato Scientifico di esperti indipendenti, composto da specialisti selezionati dalla Commissione sulla base delle loro competenze scientifiche o tecniche aggiornate nel campo dell'Intelligenza Artificiale, fornirà consulenza e supporto all'AI Office, concentrandosi in particolare sui modelli e sistemi di IA ad uso generale (General Purpose AI).

Secondo quanto stabilito dall'AI Act, ogni Stato membro sarà tenuto a istituire Autorità nazionali competenti per garantire il rispetto del regolamento. Dette Autorità avranno il compito di applicare le sanzioni previste in caso di violazioni dell'AI Act.
In Italia, le aree prioritarie e le politiche di intervento, già definite dal Programma Strategico per l'Intelligenza Artificiale 2022-2024, recentemente aggiornato dalla Strategia italiana per l'Intelligenza Artificiale 2024-2026, prevedono interventi per la creazione e il potenziamento di competenze e ricerca (fondamentale e applicata). Al contempo, la Strategia include politiche per promuovere corsi e carriere nelle materie STEM, per rafforzare le competenze digitali e promuovere progetti per incentivare il rientro in Italia di professionisti del settore. Allo stesso modo, per garantire che nessun lavoratore sia lasciato indietro, è contemplata la predisposizione di programmi di reskilling e upskilling strutturati sia nel settore pubblico che nel settore privato, al fine di aggiornare le competenze e riqualificare i lavoratori per l'utilizzo delle nuove tecnologie. Le politiche sono volte inoltre ad ampliare l'applicazione dell'IA nelle industrie e nella Pubblica Amministrazione promuovendo partenariati pubblico-privati. In tale ambito, le azioni proposte mirano al consolidamento di un ecosistema italiano di ricerca sul tema dell'IA che faciliti lo scambio di conoscenze tra Università, centri di ricerca e imprese. La Strategia nazionale, ricordando la rilevanza della valutazione della conformità e della marcatura CE nell'ambito dell'AI Act, evidenzia inoltre l'opportunità di sostenere le imprese attraverso una riduzione degli oneri relativi alla compliance normativa e alle certificazioni.
Gli interventi per la Pubblica Amministrazione sono volti certamente alla semplificazione e al rafforzamento del rafforzamento dell'ecosistema cosiddetto GovTech. Dal punto di vista della governance e del monitoraggio un gruppo di lavoro permanente sull'IA è stato istituito presso il Comitato Interministeriale per la Transizione Digitale, mentre, più recentemente il Comitato di Coordinamento (Comitato di Esperti) ha contribuito all'aggiornamento della Strategia nazionale sull'IA per il periodo 2024-2026 garantendo lo sviluppo, l'uso e la regolamentazione dell'IA in modo responsabile, etico e sicuro.
La strategia nazionale sull'IA è utile a inquadrare in un contesto più ampio il disegno di legge in materia di IA presentato dal Governo. Il DDL, attualmente in una versione avanzata ma non definitiva, si pone come obbiettivo la regolazione di una tecnologia che interseca i propri effetti in molteplici domini: dall'etica al diritto, al dibattito pubblico.
Già l'AI Act a livello europeo si qualifica quale strumento regolatorio che tenta di risolvere una molteplicità di problematiche complesse. Queste sono, da un lato, molto precise e dipendenti da ambito di applicazione, tecnologia e obiettivi legati ai sistemi di IA, dall'altro, richiedono un impianto regolatorio ampio che sia la base per un'attività di normazione tecnica per la definizione di regole conformi ai requisiti dell'AI Act.
Il ruolo della normazione tecnica, della valutazione di conformità alle norme e dell'accreditamento è fondamentale per lo sviluppo e la diffusione dei sistemi di IA.

Questi strumenti garantiscono che le tecnologie siano sicure e affidabili e promuovano un'innovazione etica e responsabile. All'interno dell'AI Act, infatti, le modalità per raggiungere la conformità sono essenzialmente tre: la valutazione di conformità effettuata da terze parti, la certificazione sulla base di norme armonizzate secondo il "New Legislative Framework" e la procedura basata sul controllo interno.
Lo studio fin qui esposto è un primo tentativo nella direzione di fornire indicazioni sui processi che dovranno essere stabiliti per garantire l'accreditamento, e la certificazione, dei sistemi di IA. Lo sviluppo dei PoC presentati nel capitolo 5 è una simulazione di quei processi di verifica che dovranno essere stabiliti e fornisce un primo approfondimento sull'effettivo utilizzo delle valutazioni di conformità accreditate per garantire profili di sicurezza ed efficacia dei sistemi di IA. In particolare, l'accreditamento sarà utile a dimostrare i requisiti di indipendenza, competenza e imparzialità richiesti ai soggetti coinvolti nelle attività di verifica nel caso dei sistemi di IA classificati ad Alto Rischio.
E allora il ruolo degli Enti nazionali di accreditamento in questa materia è quello di contribuire a rendere operative le indicazioni dell'AI Act, tutelando i diritti fondamentali e supportando le imprese nello sviluppo di sistemi di IA in un quadro regolatorio complesso.
In Italia, il già menzionato DDL attribuisce la qualifica di Autorità nazionale per l'IA a due soggetti:

- Agenzia per l'Italia digitale (AgID);
- Agenzia per la Cybersicurezza Nazionale (ACN).

Mentre all'ACN sono attribuite la responsabilità per la vigilanza, la promozione e lo sviluppo dei sistemi di IA relativamente ai profili di cybersicurezza, ad AgID sono attribuite le funzioni e i compiti in materia di notifica, valutazione, accreditamento e monitoraggio dei soggetti incaricati di verificare la conformità dei sistemi di IA.
Accredia, in qualità di Ente Unico di accreditamento designato dal Governo in applicazione del Regolamento europeo 765/2008 e vigilato dal Ministero delle Imprese e del Made in Italy, si rende disponibile a supportare l'Agenzia nello svolgere le funzioni attribuite.
Una collaborazione basata su un principio di sussidiarietà che, da sempre, ha ispirato l'azione dell'Ente. La cooperazione con le Pubbliche Amministrazioni non è infatti cosa nuova per Accredia, che svolge i propri compiti sulla base di Convenzioni e Protocolli di Intesa per le attività di accreditamento.
A titolo esemplificativo, si possono menzionare le convenzioni con ACN, GPDP e numerosi Ministeri (Salute, Interno, Agricoltura, Ambiente, Lavoro, Infrastrutture e Trasporti, Imprese e Made in Italy). In questa efficace relazione, mettendo a disposizione del pubblico le proprie competenze, Accredia alleggerisce il carico amministrativo delle Amministrazioni e, nel caso di specie, potrebbe rendere maggiormente efficienti le attività di accreditamento e monitoraggio delle attività di valutazione della conformità nell'applicazione dei sistemi di IA.

Osservatorio Accredia

Direttore editoriale
Gianluca Di Giulio

Coordinamento editoriale
Alessandro Nisi
Francesca Nizzero

Realizzazione grafica
ZERO ONE

Lo studio è stato realizzato dall'Osservatorio congiunto "Cybersecurity e Certificazione" costituito da Accredia e dal Laboratorio Nazionale di Artificial Intelligence and Intelligent Systems (AIIS) del Consorzio Interuniversitario Nazionale per l'Informatica (CINI).

Per Accredia: gruppo di lavoro coordinato dall'area Relazioni Istituzionali ed Esterne - Studi e Statistiche e composto da Riccardo Bianconi, Cettina Garufi, Lorenza Guglielmi, Sergio Guzzi, Rosalba Mugno, Alessandro Nisi, Maria Teresa Ruffo, Guglielmo Tozzi.

Per CINI: gruppo di lavoro diretto da Daniele Nardi e composto da Piercosma Bisconti, Stefano Marrone, Lidia Marrassi, Carlo Sansone, Domenico Bloisi.

ACCREDIA
L'Ente Italiano di Accreditamento

Via Guglielmo Saliceto, 7/9
00161 Roma

Tel. +39 06 844099.1
Fax. +39 06 8841199

info@accredia.it
www.accredia.it

Osservatorio
ACCREDIA

Via Guglielmo Saliceto, 7/9
00161 Roma

Tel. +39 06 844099.1
Fax. +39 06 8841199

info@accredia.it
www.accredia.it

ACCREDIA
L'ENTE ITALIANO DI ACCREDITAMENTO