Dimensione Etica
Temi Generali e Proposte Educative

Giovanni Della Lunga
giovanni.dellalunga@unibo.it

A lezione di Intelligenza Artificiale

Siena - Giugno 2025

Indice

Etica dell'IA: Questioni e Principi

Etica dell'IA: Questioni e Principi

- **Responsabilità**: chi risponde delle decisioni prese da sistemi automatizzati?
- **Discriminazione e pregiudizi**: quali rischi emergono dall'uso di tecnologie IA?
- **Privacy e sicurezza**: impatti dello sviluppo dell'IA sui dati personali.
- **Trasparenza**: quale ruolo gioca nella fiducia verso l'IA?
- **Fiducia e accettazione**: fino a che punto possiamo affidarci a questa tecnologia?

Secondo **Floridi (2022)**, l'etica dell'IA nasce negli anni '50-'60, ma l'interesse si è intensificato solo recentemente grazie ai progressi tecnologici (*Yang et al.*, 2018).

Müller (2020) colloca l'etica dell'IA nell'ambito dell'etica applicata, evidenziandone la natura dinamica e instabile.

La letteratura recente è ampia (Boddington, 2023; Stahl, Schroeder, Rodrigues).

IA, Privacy e Sorveglianza

- **Digitalizzazione e IA** hanno potenziato enormemente la raccolta e sorveglianza dei dati personali.
- **Dati scambiati tra attori diversi**, spesso a pagamento e senza controllo da parte dell'utente.
- **Processo opaco**: gli utenti raramente sono informati in modo adeguato.
- **Manipolazione sottile**: la gratuità apparente si basa sul modello "servizi in cambio di dati".
- **Perdita di autonomia**: molti utenti non riescono più a sottrarsi al controllo dei grandi attori tecnologici.

Focus: La maggioranza degli utenti ha perso la capacità di opporsi consapevolmente alla raccolta e alla monetizzazione dei propri dati personali.

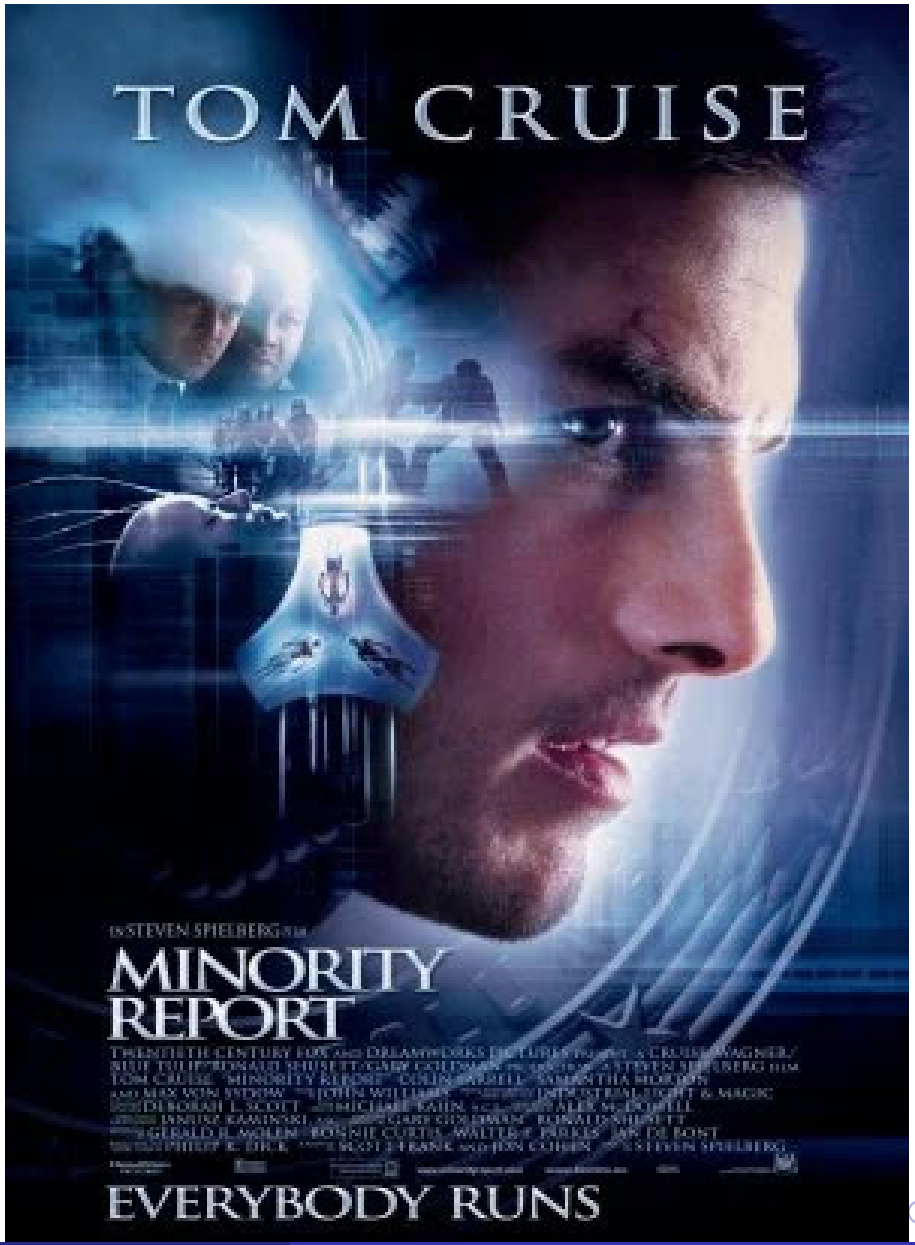
IA: Manipolazione, Opacità e Bias

- **Stimoli personalizzati** e **pattern oscuri** vengono usati per influenzare i comportamenti, specialmente nel marketing e nel gioco d'azzardo.
- **Opacità dei modelli**: molti sistemi di IA sono *black box*, nemmeno i programmatori riescono a spiegare come si formano le decisioni.
- **Mancanza di trasparenza** e partecipazione: utenti ed esperti spesso non comprendono le decisioni dell'IA.
- **Bias nei sistemi decisionali**: input distorti possono produrre risultati discriminatori.
- **Polizia predittiva**: esempio emblematico di rischio etico legato alla previsione algoritmica applicata alla sicurezza.

Focus: L'automazione delle decisioni comporta rischi di manipolazione, opacità e discriminazione sistematica.

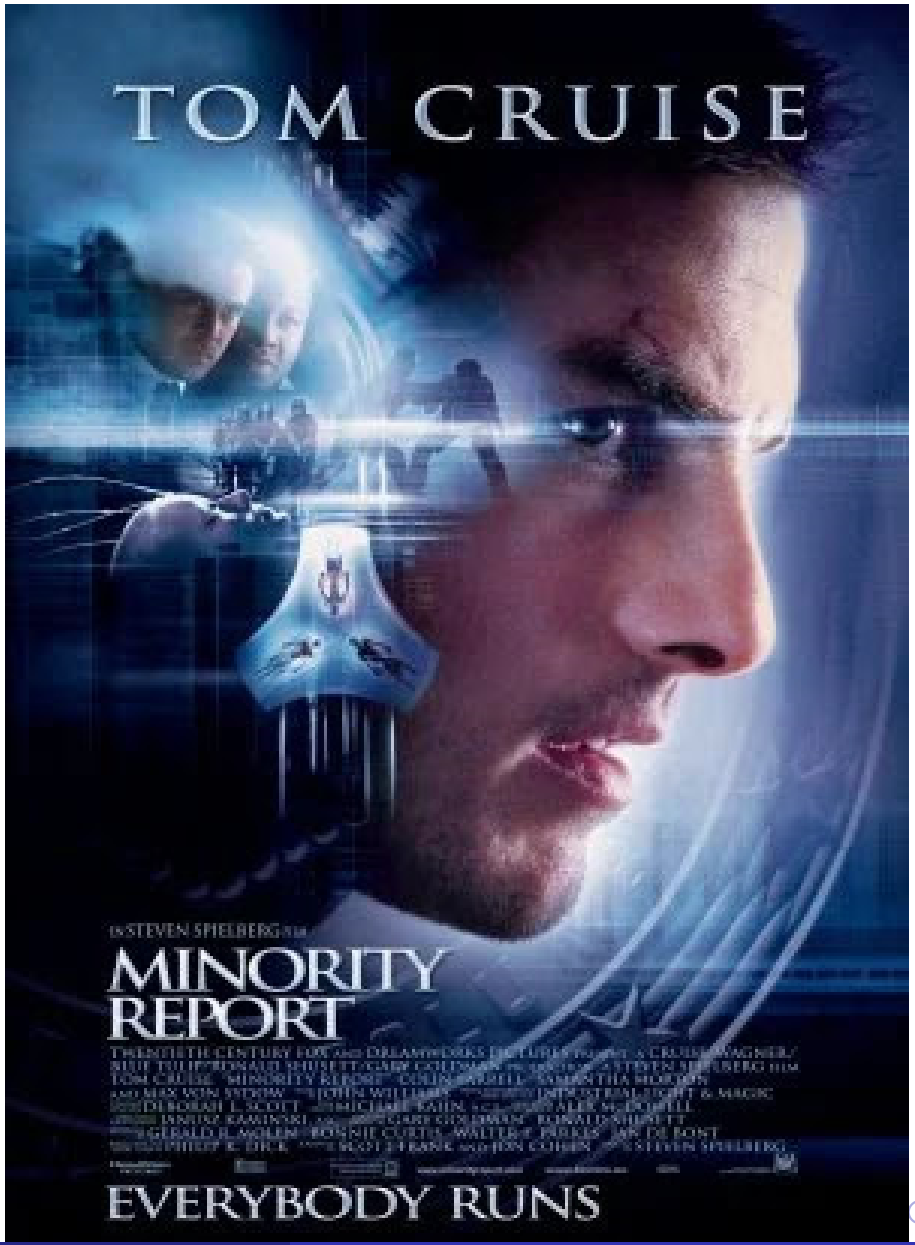
Bias nei dati e discriminazione

- I sistemi si basano su dati storici di arresti e denunce, quindi su dati spesso influenzati da pratiche di polizia già razziste o ingiuste.
- Le aree più pattugliate, tipicamente comunità di colore, generano più dati che alimentano previsioni ulteriormente sbilanciate: un classico feedback loop che rafforza il ciclo di sorveglianza .
- Conseguenza: interi quartieri, già stigmatizzati, rischiano di essere sorvegliati e criminalizzati ingiustamente .



Bias nei dati e discriminazione

- Quando la polizia interviene basandosi sulle previsioni, generano nuovi arresti in quell'area, rafforzando ulteriormente gli algoritmi. Questa spirale è ben documentata, sia teoricamente sia empiricamente .
- Molti algoritmi sono proprietari (black-box): la comunità o le autorità non possono verificarne il funzionamento o contrastare decisioni errate .
- Senza audit indipendenti, un programma può continuare a operare malgrado effetti negativi tangibili, come nel caso di PredPol, attivo per anni



Contesto

Nome del programma: IMPACT
Luogo: Distretto scolastico di Washington D.C.
Obiettivi dichiarati:

- Valutare la performance degli insegnanti in modo oggettivo e meritocratico.
- Premiare i docenti più efficaci con bonus economici.
- Licenziare chi non soddisfa gli standard di qualità.

Strumenti utilizzati:

- Osservazioni in aula da parte di supervisori.
- Analisi dei punteggi dei test standardizzati (modello VAM).
- Valutazione del contributo alla comunità scolastica.

Risultato: Decine di insegnanti licenziati ogni anno in base a punteggi VAM spesso instabili e poco trasparenti.

Contesto

Obiettivo del programma

- Migliorare la qualità dell'insegnamento nelle scuole pubbliche, premiando i docenti più efficaci.
- Introdurre criteri quantitativi per rendere meritocratico il sistema educativo.
- Razionalizzare l'impiego delle risorse e rimuovere gli insegnanti considerati poco performanti.

Approccio adottato

- Utilizzo di test standardizzati annuali per misurare l'apprendimento degli studenti.
- Adozione di modelli statistici (Value-Added Models, VAM) per isolare il contributo dell'insegnante al miglioramento dei risultati.
- Applicazione diffusa in distretti come Washington D.C., New York City, Los Angeles.
- Decisioni di carriera (promozioni, licenziamenti) basate in larga parte sul punteggio algoritmico.

Contesto

Contesto storico

- Anni 2000: forte spinta alla responsabilità dei docenti, in particolare con la legge *No Child Left Behind*.
- Crescente fiducia nei dati e negli algoritmi come strumenti oggettivi di valutazione.

Il modello algoritmico

Come funziona il VAM?

- Si calcola la differenza tra il punteggio atteso e quello reale dello studente.
- La media dei delta definisce il "valore aggiunto" dell'insegnante.

Come si stima il punteggio atteso?

- Attraverso un modello di regressione che tiene conto di:
  - Punteggi precedenti dello studente;
  - Fattori demografici (età, livello socioeconomico, lingua madre);
  - Caratteristiche della scuola o della classe.
- L'obiettivo è stimare quale sarebbe stato il punteggio senza l'intervento dell'insegnante attuale.

Il modello algoritmico

Problemi principali:

- Alta variabilità da un anno all'altro.
- Nessun controllo sui dati e sull'equazione utilizzata.
- Uso di proxy approssimativi per misurare concetti complessi come "qualità dell'insegnamento".
- Basso numero di osservazioni (una sola classe per insegnante) genera risultati statisticamente fragili.
- Rumore e variabili non controllabili (es. condizioni socioeconomiche, eventi esterni) distorcono i punteggi.

Il problema dei proxy nei modelli VAM

Cosa sono i proxy?

- Variabili surrogate utilizzate per rappresentare concetti non direttamente misurabili (es. efficacia didattica).
- Nei VAM: si presume che il miglioramento nei punteggi dei test rifletta l'influenza dell'insegnante.

Problemi riscontrati

- I punteggi degli studenti dipendono da molteplici fattori esterni non legati all'insegnamento.
- I modelli non riescono a distinguere tra progresso reale e fluttuazioni casuali.
- Insegnanti valutati su piccoli gruppi (25-30 studenti): stime instabili.
- Proxy deboli incentivano comportamenti opportunistici (teaching to the test).

Conseguenze pratiche

- Licenziamenti arbitrari di insegnanti competenti.
- Focus eccessivo sull'insegnamento orientato ai test.
- Riduzione della fiducia e della motivazione tra i docenti.
- Esempio emblematico: Sarah Wysocki, insegnante apprezzata, licenziata a Washington D.C.

Conclusione

- L'uso scorretto di proxy in contesti ad alto impatto può generare effetti distorti e ingiusti.

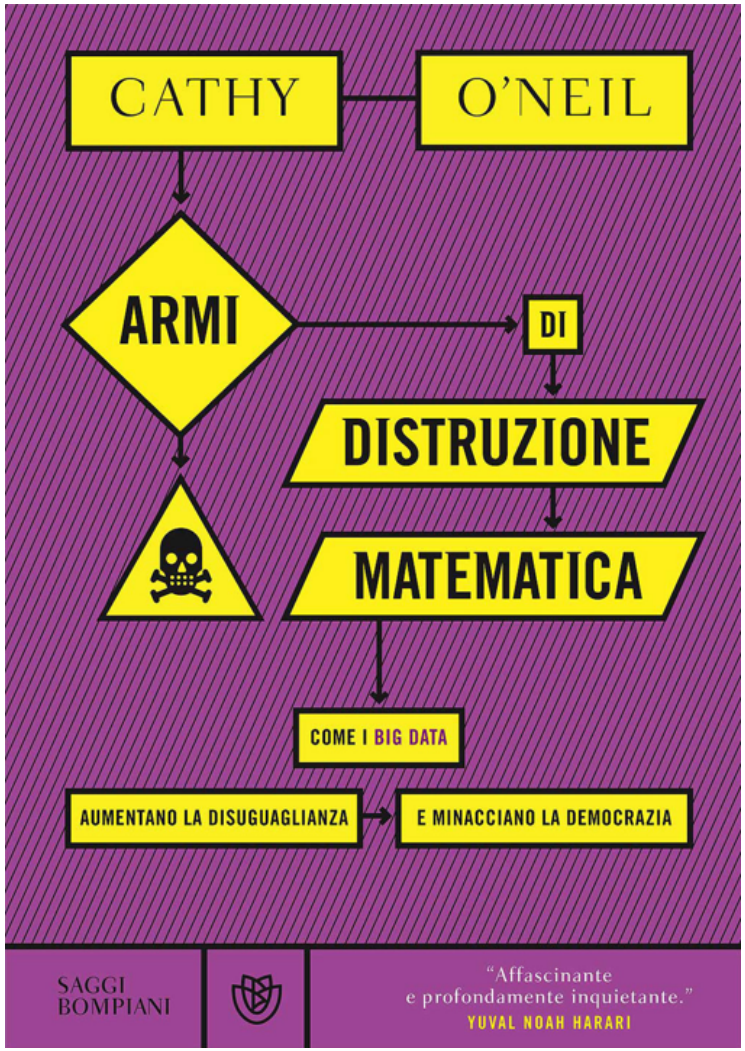
Critiche etiche

Perché è un Algoritmo di Distruzione di Massa (WMD)?

- **Opacità**: non spiegabile né contestabile.
- **Scala**: applicato su larga scala.
- **Danno**: ha effetti reali e negativi su persone innocenti.
- **Feedback negativo**: incentiva strategie opportunistiche.

Cosa ci insegna questo episodio

- L'intelligenza artificiale e gli algoritmi non sono neutri.
- Le metriche devono essere validate e comprensibili.
- L'educazione è un contesto complesso, non completamente riducibile a numeri.
- Serve trasparenza, partecipazione e responsabilità etica.



IA: Lavoro, Sostenibilità e Sistemi Autonomi

- **Interazione uomo-macchina**: l'IA può essere usata impropriamente per influenzare comportamenti umani.
- **Mercato del lavoro**: l'automazione riduce la necessità di manodopera, generando **sfide inedite** e polarizzazione tra mestieri altamente specializzati e lavori facilmente sostituibili.
- **Sostenibilità ambientale**: i sistemi di IA consumano molte risorse energetiche e materiali, ponendo interrogativi etici sull'impatto ambientale.
- **Sistemi autonomi**: pongono dilemmi inediti in termini di responsabilità, controllo e decisione.

Focus: L'IA non impatta solo dati e decisioni, ma anche lavoro, ambiente e dinamiche sociali.

Veicoli autonomi ed etica delle macchine

- **Veicoli autonomi**: potenziale riduzione di incidenti e inquinamento, ma emergono dilemmi su responsabilità e decisioni normative.
- **Equilibrio etico**: tra interesse individuale e bene comune, spostando responsabilità da utenti a produttori e sistemi.
- **Etica delle macchine**: riguarda le macchine come soggetti morali, non solo strumenti.
- Alcuni autori propongono che l'IA debba garantire che il proprio comportamento verso gli umani sia **eticamente accettabile**.
- Altri propongono IA in grado di **ponderare valori e interessi** in modo trasparente (Dignum, 2018).

Nota: Floridi (2022) sottolinea che la moltiplicazione dei principi etici rischia di generare confusione ed effetti collaterali.

Principi etici dell'IA secondo Floridi

Cinque principi fondamentali:

- **Beneficenza**: l'IA deve essere sviluppata per il bene comune.
- **Non maleficenza**: evitare usi impropri e danni (es. violazione privacy, sicurezza).
- **Autonomia**: bilanciare il potere decisionale umano e quello delegato alle macchine.
- **Giustizia**: promuovere equità, combattere le discriminazioni.
- **Esplicabilità**: garantire trasparenza e accountability.

Jobin, Ienca, Vayena (2019): analizzando 100 documenti, identificano 11 principi chiave, tra cui emergono con maggiore frequenza: **trasparenza**, **giustizia ed equità**, **non maleficenza**, **responsabilità** e **privacy**.

Intelligenza Artificiale in Educazione

IA e educazione: lettura pedagogica delle sfide etiche

- L'IA sta introducendo **forme di automazione** nei sistemi educativi (es. tutor intelligenti, valutazione automatica, percorsi personalizzati).
- Tuttavia, le **politiche educative** si concentrano sugli adulti, trascurando bambini e adolescenti.
- **Linee guida etiche** sull'uso dell'IA in classe risultano scarse o vaghe.
- Le questioni etiche legate all'uso dell'IA con i più giovani possono essere **ancora più critiche** rispetto ad altri contesti sociali.
- Organizzazioni come **UNICEF**, **UNESCO** e **UE** stanno iniziando a colmare il divario.

Nota: La governance etica dell'IA in ambito educativo è ancora poco strutturata ma di crescente rilevanza.

Il debate per esplorare i dilemmi etici dell'IA

- Il **debate** è una tecnica didattica utile per affrontare l'etica dell'IA senza cadere nella trasmissione dogmatica dei valori.
- Metodo: confronto fra due tesi opposte, con argomentazioni pro e contro, repliche e sintesi.
- Favorisce **pensiero critico**, capacità argomentativa e metacognizione.
- Scopo non è vincere, ma **comprendere i punti di vista**, analizzare criticamente e saper argomentare.
- Fasi: scelta del tema, raccolta dati, esposizione in classe, sintesi e valutazione finale.

Focus: Il debate aiuta studenti e docenti ad affrontare l'etica dell'IA in modo attivo, dialogico e critico.

European AI Act

Che cos'è l'AI Act

- L'AI Act (Regolamento UE 2024/1689) è il primo quadro normativo globale per regolamentare l'intelligenza artificiale, entrato in vigore il 1° agosto 2024. Questo regolamento europeo stabilisce regole armonizzate per lo sviluppo, l'immissione sul mercato e l'utilizzo di sistemi di intelligenza artificiale nell'Unione Europea.
- Il regolamento nasce dall'esigenza di bilanciare l'innovazione tecnologica con la protezione dei diritti fondamentali, della sicurezza e della salute dei cittadini europei. L'obiettivo principale è creare un ambiente digitale sicuro e affidabile, dove l'intelligenza artificiale possa svilupparsi nel rispetto dei valori europei.
- L'AI Act si propone di posizionare l'Europa come leader mondiale nella regolamentazione dell'IA, fornendo certezza giuridica alle imprese e garantendo al contempo protezione ai consumatori. Il regolamento copre tutti i settori e le applicazioni dell'intelligenza artificiale, dal riconoscimento facciale ai sistemi di raccomandazione, dai veicoli autonomi agli assistenti virtuali.

Approccio basato sul rischio

- L'AI Act adotta un approccio innovativo basato sulla classificazione dei sistemi di intelligenza artificiale in quattro categorie di rischio: inaccettabile, alto, limitato e minimo. Questa classificazione determina gli obblighi e le restrizioni applicabili a ciascun sistema.
- I sistemi a rischio inaccettabile includono pratiche che violano i diritti fondamentali, come il punteggio sociale generalizzato, la manipolazione comportamentale subliminale e alcuni tipi di sorveglianza biometrica in tempo reale. Questi sistemi sono completamente vietati nell'Unione Europea.
- I sistemi ad alto rischio, utilizzati in settori critici come sanità, trasporti, giustizia e servizi pubblici essenziali, devono rispettare requisiti rigorosi. Questi includono valutazioni del rischio, documentazione tecnica dettagliata, trasparenza algoritemica, supervisione umana e sistemi di gestione della qualità. L'obiettivo è garantire che questi sistemi siano sicuri, accurati e non discriminatori prima della loro immissione sul mercato.

Approccio basato sul rischio

Obblighi e requisiti principali

- Per i sistemi ad alto rischio, l'AI Act stabilisce obblighi specifici che devono essere rispettati durante tutto il ciclo di vita del prodotto. I fornitori devono implementare sistemi di gestione del rischio, garantire la qualità dei dati di addestramento e condurre test approfonditi prima del rilascio.
- La documentazione tecnica deve essere completa e aggiornata, includendo informazioni sul funzionamento del sistema, sui dati utilizzati e sulle misure di mitigazione dei rischi. È richiesta anche la registrazione automatica degli eventi per consentire la tracciabilità e l'audit delle decisioni del sistema.
- La supervisione umana è un elemento centrale: deve essere garantito un controllo umano significativo sui sistemi ad alto rischio, con la possibilità di intervenire, interrompere o annullare le decisioni automatizzate. Inoltre, i sistemi devono essere progettati per essere robusti, accurati e sicuri, con particolare attenzione alla prevenzione di bias e discriminazioni nei risultati prodotti.

Governance e controlli

- L'AI Act istituisce un sistema di governance multilivello per garantire l'applicazione effettiva del regolamento. A livello europeo, è stato creato l'European AI Office, responsabile del coordinamento e della supervisione dell'implementazione, particolarmente per i modelli di AI generale ad alto impatto.
- Ogni Stato membro deve designare autorità nazionali competenti per la vigilanza e il controllo sui sistemi di intelligenza artificiale. Queste autorità hanno il potere di condurre ispezioni, richiedere documentazione e imporre misure correttive quando necessario.
- Il regolamento prevede anche la creazione di organismi notificati indipendenti per la valutazione di conformità dei sistemi ad alto rischio prima della loro immissione sul mercato. Inoltre, è istituito un sistema di segnalazione degli incidenti gravi, che permette alle autorità di monitorare e rispondere rapidamente a problemi di sicurezza. La cooperazione tra Stati membri è facilitata attraverso meccanismi di assistenza reciproca e condivisione delle informazioni.

Sanzioni e tempistiche

- L'AI Act prevede un sistema sanzionatorio proporzionato ma severo per garantire il rispetto delle norme. Le sanzioni amministrative possono raggiungere i 35 milioni di euro o il 7% del fatturato annuo globale per le violazioni più gravi, come l'uso di sistemi vietati o la non conformità dei sistemi ad alto rischio.
- Per violazioni meno gravi, come la mancanza di trasparenza o problemi nella documentazione, le sanzioni possono arrivare a 15 milioni di euro o al 3% del fatturato annuo. Le sanzioni più basse, fino a 7,5 milioni di euro o all'1,5% del fatturato, si applicano per la fornitura di informazioni incomplete o inesatte alle autorità.
- L'implementazione del regolamento segue un calendario graduale: dal 2 febbraio 2025 sono in vigore i divieti per i sistemi a rischio inaccettabile, mentre i requisiti completi per i sistemi ad alto rischio entreranno in vigore ad agosto 2026. Questa tempistica permette alle aziende di adeguarsi progressivamente alle nuove norme, garantendo una transizione controllata verso il nuovo quadro normativo.

Impatto e prospettive future

- L'AI Act rappresenta un cambiamento paradigmatico nel panorama tecnologico globale, stabilendo standard che influenzeranno lo sviluppo dell'intelligenza artificiale ben oltre i confini europei. Il cosiddetto "Brussels Effect" potrebbe spingere aziende internazionali ad adottare gli standard europei come riferimento globale per la conformità.
- Per le imprese europee, il regolamento offre certezza giuridica e un vantaggio competitivo nel mercato globale dell'IA etica e responsabile. Tuttavia, comporta anche sfide significative in termini di costi di conformità e complessità amministrativa, particolarmente per le piccole e medie imprese.
- L'AI Act segna l'inizio di una nuova era nella regolamentazione tecnologica, dove l'innovazione deve procedere di pari passo con la protezione dei diritti e valori fondamentali. Il successo di questa normativa dipenderà dalla capacità di mantenere un equilibrio tra promozione dell'innovazione e gestione dei rischi, influenzando il futuro sviluppo dell'intelligenza artificiale a livello mondiale e definendo nuovi standard per la tecnologia responsabile.