EDPS

11 novembre 2025

EUROPEAN DATA PROTECTION SUPERVISOR

The EU's independent data protection authority

Linee guida per la gestione dei rischi dei sistemi di intelligenza artificiale

Indice

Sintesi **4**

1 Introduzione **5**

1.1 Obiettivo **5**

1.2 Ambito **5**

1.3 Destinatari **6**

2 Metodologia di gestione dei rischi **7**

3 Comprendere il ciclo di vita dell'IA **9**

3.1 Definizione di un sistema di IA **9**

3.2 Ciclo di vita di un sistema di IA **9**

3.3 Acquisizione di un sistema di IA **11**

4 Interpretabilità e spiegabilità come *condizioni sine qua non* .13

4.1.1 Rischio 1: sistema di IA non interpretabile o non spiegabile 14

5 Rischi associati ai principali principi di protezione dei dati. 15

5.1 Principio di correttezza **15**

5.1.1 Rischio 1: Distorsione causata dalla scarsa qualità dei dati utilizzati per l'addestramento dei dati personali 17

5.1.2 Rischio 2: Distorsione nei dati personali utilizzati per l'addestramento 19

5.1.3 Rischio 3: Overfitting nei dati personali utilizzati per l'addestramento 21

5.1.4 Rischio 4: Distorsione algoritmica 22

5.1.5 Rischio 5: Distorsione interpretativa 24

5.2 Principio di accuratezza **25**

5.2.1 Significato giuridico di accuratezza nell'EUDPR 25

5.2.2 Significato statistico dell'accuratezza nello sviluppo dell'IA 25

5.2.3 Rischio 1: Output di dati personali inesatti 25

5.2.4 Esempio specifico: risultati imprecisi dovuti alla deriva dei dati e al deterioramento della qualità dei dati personali inseriti 27

5.2.5 Rischio 2: Informazioni poco chiare da parte del fornitore del sistema di IA 28

5.3 Principio di minimizzazione dei dati **30**

5.3.1 Rischio 1: raccolta e conservazione indiscriminata dei dati personali 30

5.4 Principio di sicurezza **31**

5.4.1 Rischio 1: divulgazione dei dati personali utilizzati per l'addestramento del sistema di IA 32

5.4.2 Rischio 2: archiviazione dei dati personali e violazioni dei dati personali 34

5.4.3 Rischio 3: fuga di dati personali attraverso le interfacce di programmazione delle applicazioni 35

5.5 Diritti dell'interessato 36

5.5.1 Rischio 1: Identificazione incompleta dei dati personali trattati 36

5.5.2 Rischio 2: Rettifica o cancellazione incompleta 38

6 Conclusione 39

Allegato 1: Parametri di misurazione 41

Allegato 2: Panoramica delle preoccupazioni e dei rischi 47

Allegato 3: Lista di controllo per ogni fase del ciclo di vita dello sviluppo dell'IA 48

Sviluppo di un sistema di IA 48

Acquisizione di un sistema di IA 52

Sintesi

Lo sviluppo, l'approvvigionamento e l'implementazione di sistemi di intelligenza artificiale che comportano il trattamento di dati personali da parte delle istituzioni, degli organi e degli organismi dell'Unione europea (EUI) comportano rischi significativi per i diritti e le libertà fondamentali degli interessati, inclusi, a titolo esemplificativo ma non esaustivo, la privacy e la protezione dei dati. In quanto pietra miliare del regolamento 2018/1725 (EUDPR),[1] il principio di responsabilità sancito dall'articolo 4, paragrafo 2 (per i dati personali amministrativi) e dall'articolo 71, paragrafo 4 (per i dati personali operativi) impone alle EUI di identificare e mitigare tali rischi, nonché di dimostrare in che modo lo hanno fatto. Ciò è tanto più importante per i sistemi di IA che sono il prodotto di catene di approvvigionamento complesse che spesso coinvolgono più attori che trattano dati personali in diverse vesti.

La presente guida ha lo scopo di aiutare gli EUI che agiscono in qualità di titolari del trattamento dei dati a identificare e mitigare alcuni di questi rischi. Più specificamente, essa si concentra sul rischio di non conformità con alcuni principi di protezione dei dati enunciati nell'EUDPR per i quali le strategie di mitigazione che i titolari del trattamento devono attuare possono essere di natura tecnica, vale a dire correttezza, accuratezza, minimizzazione dei dati, sicurezza e diritti degli interessati. Pertanto, i controlli tecnici elencati nella presente guida non sono affatto esaustivi e non esentano gli EUI dal condurre una propria valutazione dei rischi derivanti dalle loro specifiche attività di trattamento. In tal modo, si astiene dal classificarne la probabilità e la gravità.

In primo luogo, il presente documento fornisce una panoramica della metodologia di gestione dei rischi secondo la norma ISO 31000:2018 (sezione 2). In secondo luogo, delinea il ciclo di vita tipico dello sviluppo dei sistemi di IA, nonché le diverse fasi coinvolte nel loro approvvigionamento (sezione 3). In terzo luogo, esplora i concetti di interpretabilità e spiegabilità come questioni trasversali che condizionano la conformità a tutte le disposizioni trattate nella presente guida (sezione 4). Infine, suddivide i quattro principi generali sopra elencati, ovvero equità, accuratezza, minimizzazione dei dati e sicurezza, in rischi specifici, ciascuno dei quali viene poi descritto e abbinato a misure tecniche che i responsabili del trattamento possono attuare per mitigare tali rischi (sezione 5).

Il GEPD pubblica le presenti linee guida nella sua veste di autorità di controllo della protezione dei dati e non nella sua veste di autorità di vigilanza del mercato ai sensi della legge sull'IA. Le presenti linee guida non pregiudicano la legge sull'intelligenza artificiale.

[1] Regolamento (UE) 2018/1725 del Parlamento europeo e del Consiglio, del 23 ottobre 2018, sulla protezione delle persone fisiche con riguardo al trattamento dei dati personali da parte delle istituzioni, degli organi e degli organismi dell'Unione, sulla libera circolazione di tali dati e che abroga il regolamento (CE) n. 45/2001 e la decisione n. 1247/2002/CE [2018] GU L 295/39 https://data.europa.eu/eli/reg/2018/1725/oj.

1 Introduzione

1.1 Obiettivo

Il presente documento ha lo scopo di guidare le istituzioni, gli organi, gli uffici e le agenzie dell'Unione europea (EUI) che agiscono in qualità di titolari del trattamento ai sensi dell'articolo 3, paragrafo 8, del regolamento (UE) 2018/1725 (EUDPR) **nell'identificare** e **mitigare** alcuni dei rischi per i diritti fondamentali degli interessati derivanti dal trattamento dei dati personali durante lo sviluppo, l'approvvigionamento e l'implementazione di sistemi di IA.[2] Esso è inteso a integrare la parte II del toolkit sulla responsabilità sul campo in materia di valutazioni d'impatto sulla protezione dei dati e consultazioni preventive. [3]

Esso integra inoltre gli Orientamenti del GEPD sull'uso dell'IA generativa da parte delle istituzioni dell'UE per garantire la conformità alla protezione dei dati nell'uso dei sistemi di IA generativa, pubblicati nel giugno 2024, che forniscono consigli pratici su come le istituzioni dell'UE possono garantire la conformità all'EUDPR nello sviluppo o nell'uso di sistemi di IA generativa.[4] Il presente documento è sia più ampio, in quanto comprende tutti i tipi di sistemi di IA, sia più ristretto, in quanto si concentra su strategie di mitigazione tecniche piuttosto che giuridiche (cfr. sezione 1.2).

Il presente documento fornisce un quadro analitico per identificare e trattare i rischi che possono insorgere nei sistemi di IA, strutturato in base ai principi di protezione dei dati potenzialmente interessati. Esso non costituisce e non deve essere considerato come una serie di linee guida di conformità. L'unico scopo del presente documento è quello di facilitare una valutazione sistematica dei rischi dal punto di vista della protezione dei dati. In altre parole, non sostituisce la necessaria valutazione di conformità di ciascun sistema di IA che deve essere effettuata dal responsabile del trattamento, il quale deve garantire che i rischi individuati (con il supporto del presente quadro) siano gestiti in modo adeguato per soddisfare tutti gli obblighi derivanti dall'EUDPR.

Il GEPD pubblica le presenti linee guida nella sua veste di autorità di controllo della protezione dei dati e non nella sua nuova funzione di autorità di vigilanza del mercato ai sensi della legge sull'IA.

1.2 Ambito di applicazione

Ai fini della presente Guida, e sulla base della terminologia utilizzata nella norma ISO 31000:2018,[5] il concetto di «rischio» è espresso in termini di «fonte di rischio», "evento", "conseguenza" e "controllo", dove la "fonte **di rischio**" si riferisce al trattamento dei dati personali nel contesto dell'approvvigionamento, dello sviluppo o dell'implementazione di un sistema di IA, l'"evento" si riferisce a una situazione in cui tale trattamento ostacolerebbe i diritti e le libertà fondamentali degli interessati, la "conseguenza" si riferisce al danno materiale o immateriale che ciò potrebbe causare.

---

[2] Regolamento (UE) 2018/1725 del Parlamento europeo e del Consiglio, del 23 ottobre 2018, relativo alla protezione delle persone fisiche in relazione al trattamento dei dati personali da parte delle istituzioni, degli organi e degli organismi dell'Unione e alla libera circolazione di tali dati, e che abroga il regolamento (CE) n. 45/2001 e la decisione n. 1247/2002/CE [2018] GU L 295/39 https://data.europa.eu/eli/reg/2018/1725/oj.

[3] GEPD, *Responsabilità sul campo Parte II: Valutazioni d'impatto sulla protezione dei dati e consultazione preventiva*, febbraio 2018, https://www.edps.europa.eu/sites/default/files/publication/18-02-06_accountability_on_the_ground_part_2_en.pdf

[4] Garante europeo della protezione dei dati, *IA generativa e EUDPR. Orientamenti per garantire la conformità alla protezione dei dati quando si utilizzano sistemi di IA generativa. (Versione 2)*, 28 ottobre 2025, https://www.edps.europa.eu/system/files/2025-10/25-10_28_revised_genai_orientations_en.pdf

[5] Organizzazione internazionale per la normazione, *ISO 31000:2018 Gestione del rischio — Linee guida*, Edizione 2, 2018, https://www.iso.org/standard/65694.html.

agli interessati,[6] mentre il termine "controllo" si riferisce alle strategie di mitigazione che i responsabili del trattamento possono mettere in atto per ridurre la probabilità che tale rischio si concretizzi e/o l'impatto che esso ha sugli interessati qualora si concretizzasse.[7]

Ai sensi dell'articolo 1, paragrafo 2, dell'EUDPR, l'obiettivo del regolamento è proteggere i diritti e le libertà delle persone fisiche, **inclusi, ma non limitati a,** la privacy e la protezione dei dati nel contesto del trattamento dei loro dati personali. Ai sensi degli articoli 4, paragrafo 2, 26, paragrafo 1, e 27, paragrafo 1, gli EUI hanno la responsabilità di individuare e mitigare i rischi per tali diritti e libertà derivanti dalle loro attività di trattamento e di dimostrare in che modo lo hanno fatto. Ciò è particolarmente importante quando si tratta dell'approvvigionamento, dello sviluppo e dell'implementazione di sistemi di IA, i cui effetti negativi non sono stati ancora valutati. È quindi fondamentale che i responsabili del trattamento identifichino e mitigino adeguatamente, per ciascuna delle loro attività di trattamento, i rischi che queste comportano per i diritti fondamentali di tutti gli interessati. Il rispetto delle disposizioni esplicitamente stabilite nell'EUDPR è un **indicatore** per raggiungere tale obiettivo. La presente guida si attiene quindi a una concettualizzazione del rischio in cui l'"evento" è una situazione di **non conformità** con una disposizione esplicitamente stabilita nel testo.

Più specificamente, la presente Guida si concentra sul rischio di non conformità con **alcuni principi selezionati in materia di protezione dei dati** per i quali i "controlli" che i titolari del trattamento devono attuare possono essere di natura tecnica, vale a dire correttezza, accuratezza, minimizzazione dei dati, sicurezza e determinati diritti degli interessati. Il GEPD sottolinea che l'elenco dei rischi e delle contromisure descritti nella presente guida **non** è **esaustivo,** ma riflette semplicemente alcune delle questioni più urgenti che i responsabili del trattamento devono affrontare quando acquistano, sviluppano e implementano sistemi di IA.

1.3 Destinatari

Il pubblico a cui è destinato il presente documento è il personale delle istituzioni dell'Unione europea coinvolto nell'acquisto, nello sviluppo e nell'implementazione di sistemi di IA, compresi sviluppatori di software, data scientist, ingegneri informatici, responsabili di progetti informatici, responsabili della protezione dei dati e coordinatori della protezione dei dati.

[6] Il considerando 46 dell'EUDPR specifica che "[i]l rischio per i diritti e le libertà delle persone fisiche può derivare dal trattamento dei dati personali che potrebbe causare danni fisici, materiali o immateriali, in particolare: quando il trattamento può dare luogo a discriminazione, furto di identità o frode, perdite finanziarie, danni alla reputazione, perdita di riservatezza dei dati personali protetti dal segreto professionale, inversione non autorizzata della pseudonimizzazione o qualsiasi altro svantaggio economico o sociale significativo; quando gli interessati potrebbero essere privati dei loro diritti e libertà o impediti di esercitare il controllo sui loro dati personali; quando vengono trattati dati personali che rivelano l'origine razziale o etnica, le opinioni politiche, la religione o le convinzioni filosofiche, l'appartenenza sindacale, nonché il trattamento di dati genetici, dati relativi alla salute o dati relativi alla vita sessuale o alle condanne penali e ai reati o alle misure di sicurezza correlate; quando vengono valutati aspetti personali, in particolare analizzando o prevedendo aspetti relativi alle prestazioni lavorative, alla situazione economica, alla salute, alle preferenze o agli interessi personali, all'affidabilità o al comportamento, all'ubicazione o agli spostamenti, al fine di creare o utilizzare profili personali; quando vengono trattati dati personali di persone fisiche vulnerabili, in particolare di bambini; o quando il trattamento riguarda una grande quantità di dati personali e interessa un numero elevato di interessati".

[7] La norma ISO 31000:2018 (sezione 3.8) utilizza il termine "controllo" definito come "misura che mantiene e/o modifica il rischio". L'EUDPR utilizza il termine "misura". Il resto del presente documento utilizzerà il termine misura.

2 Metodologia di gestione del rischio

Secondo la norma ISO 31000:2018, **la gestione del rischio** è un processo attraverso il quale un'organizzazione può controllare i rischi (cfr. figura 1). Il nucleo di tale attività è la parte **relativa alla valutazione del rischio**, durante la quale un'organizzazione identifica, analizza e valuta in modo progressivo i rischi.[8]

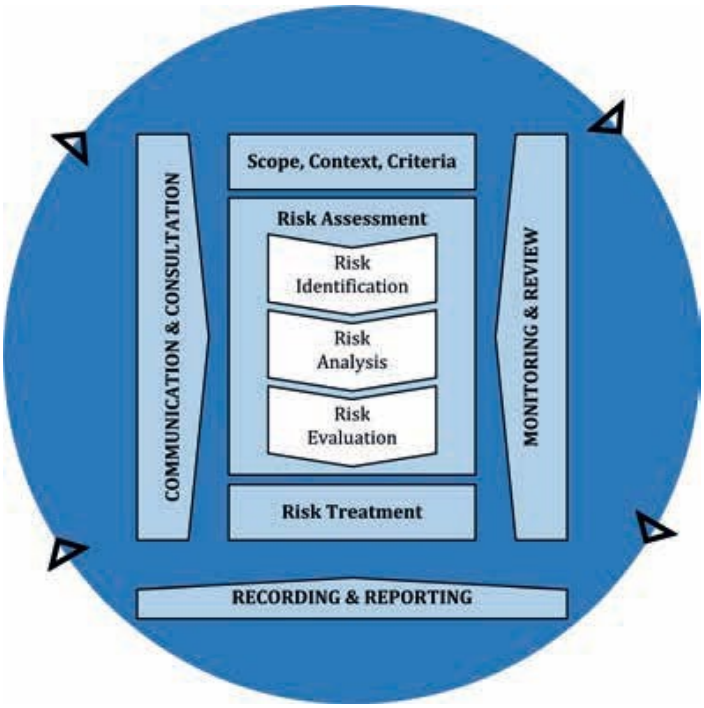


Figura 1: valutazione dei rischi[9]

L'identificazione dei rischi comporta il riconoscimento sistematico dei rischi che potrebbero potenzialmente influire sugli obiettivi dell'organizzazione. Questa fase si concentra sull'identificazione delle fonti di rischio, delle aree di impatto e degli eventi o delle situazioni che potrebbero causare incertezza. L'obiettivo è quello di creare un registro dei rischi completo che sarà ulteriormente analizzato nelle fasi successive. Come già accennato nella sezione 1.2, la presente guida parte dal presupposto che l'obiettivo perseguito dalle EUI sia quello di garantire che il trattamento dei dati personali di cui sono responsabili non leda i diritti fondamentali degli interessati.

L'analisi dei rischi è la fase successiva, durante la quale l'organizzazione esamina i rischi identificati per comprenderne la natura, le fonti, la probabilità e le potenziali conseguenze. Questa fase mira a determinare la probabilità che ciascun rischio si concretizzi, nonché il suo impatto sugli interessati qualora dovesse verificarsi. Per **l'analisi** qualitativa **dei rischi**, i livelli di probabilità e impatto possono essere classificati su una scala che va da "Molto basso", "Basso", "Medio",

[8] Isabel Barberá, *AI Possibili Rischi e Mitigazioni - Entità Entità riconoscimento,* settembre *2023,* https://www.edpb.europa.eu/system/files/2024-07/ai-risks_d1named-entity-recognition_edpb-spe-programme_en.pdf Isabel Barberá, *AI Possibili Rischi e Mitigazioni -Ottica Riconoscimento riconoscimento,* settembre 2023, https://www.edpb.europa.eu/system/files/2024-06/ai-risks_d2optical-character-recognition_edpb-spe- programme_en_2.pdf

[9] Da ISO 31000:2018.

"Elevato" e "Molto elevato".[10] Una volta definiti ciascuno di questi elementi nella scala,[11] il rischio può essere valutato come il prodotto della sua probabilità e della gravità del suo impatto (Rischio = Probabilità x Impatto). Questo è tipicamente rappresentato tramite una matrice di rischio (vedi Figura 2).

Probabilità	**Molto alta**	Media	Alta	Molto alta	Molto alta
	Alto	Basso	Alto	Molto alto	Molto alto
	Basso	Basso	Medio	Alto	Molto alto
	Improbabile	Basso	Basso	Medio	Molto alto
		Molto limitato	**Limitato**	**Significativo**	**Molto significativo**
		Gravità			

Figura 2: matrice qualitativa per il rischio[12]

La valutazione del rischio è la fase finale della **valutazione del rischio**, in cui i risultati **dell'analisi del rischio** vengono confrontati con i criteri di rischio dell'organizzazione (come la propensione al rischio e la tolleranza) per determinare se ciascun rischio è accettabile o richiede un intervento. Il risultato di questa valutazione aiuta l'organizzazione a decidere se evitare, mitigare, trasferire o accettare i rischi, a seconda della loro gravità e degli obiettivi organizzativi.

Dopo la **valutazione dei rischi** segue il **trattamento dei rischi**, il cui scopo è quello di selezionare e attuare misure volte a mitigare tali rischi in modo efficace. Si tratta di un processo iterativo che prevede diverse fasi chiave. Innanzitutto, vengono formulate e selezionate le opzioni **di trattamento dei rischi**. Successivamente, viene elaborato e attuato un piano per affrontare i rischi individuati. Dopo l'attuazione, viene valutata l'efficacia delle misure per determinare se hanno mitigato il rischio in misura sufficiente. Se il rischio residuo è ritenuto accettabile, non sono necessarie ulteriori azioni. Tuttavia, se il rischio è ancora inaccettabile, vengono adottate misure aggiuntive per ridurlo ulteriormente.

La presente guida si concentra su **due aspetti specifici** del processo di gestione **dei rischi**, ovvero **l'identificazione** e **il trattamento dei rischi**; gli aspetti relativi **all'analisi** e **alla valutazione dei rischi** dipendono troppo dal contesto specifico di trattamento ed è preferibile che la loro valutazione sia lasciata a ciascuna organizzazione in linea con i propri criteri di rischio. Ciò significa che le EUI dovrebbero effettuare un'analisi approfondita per ciascun sistema di IA che intendono utilizzare, al fine di valutare anche la probabilità e l'impatto dei rischi e decidere le misure di mitigazione da adottare per affrontarli, nonché

(10) A differenza delle valutazioni quantitative del rischio, che richiedono misurazioni spesso difficili da raccogliere.

11 Isabel Barberá, *AI Possibili Rischi e Mitigazioni - Entità Entità riconoscimento,* settembre *2023,* https://www.edpb.europa.eu/system/files/2024-07/ai-risks_d1named-entity-recognition_edpb-spe-programme_en.pdf

12 ibid

sui rischi residui.[13] Questa analisi potrebbe anche portare alla conclusione che l'EUI non è in grado di mitigare con mezzi ragionevoli i rischi posti dal sistema di IA previsto e che occorre quindi trovare una soluzione diversa alle esigenze dell'organizzazione. In tal caso, l'EUI dovrebbe consultare preventivamente il GEPD ai sensi dell'articolo 40, paragrafo 1, dell'EUDPR.

3 Comprendere il ciclo di vita dell'IA

3.1 Definizione di sistema di IA

Ai fini della presente guida, per sistema di IA si intende, ai sensi dell'articolo 3, paragrafo 1, del regolamento 2024/1689 (AI Act) come «un sistema basato su macchine progettato per funzionare con diversi livelli di autonomia, che può mostrare capacità di adattamento dopo l'implementazione e che, per obiettivi espliciti o impliciti, deduce, dagli input che riceve, come generare output quali previsioni, contenuti, raccomandazioni o decisioni che possono influenzare ambienti fisici o virtuali»([14]).L'AI Act, tuttavia, non contiene una definizione di «modello di IA».[15] I termini "sistema di IA" e "modello di IA" sono spesso utilizzati come se fossero sinonimi, ma non lo sono.

I modelli di IA sono rappresentazioni matematiche che catturano, in una serie di parametri, i modelli alla base dei dati personali utilizzati per il loro addestramento [(16).]Sebbene i modelli di IA siano componenti essenziali dei sistemi di IA, essi non costituiscono di per sé sistemi di IA, in quanto richiedono sempre altri componenti software per poter funzionare e interagire con gli utenti e l'ambiente virtuale o fisico. Infatti, un sistema di IA può essere composto da più di un modello di IA. Ad esempio, un sistema di IA per la traduzione vocale potrebbe essere composto da un primo modello che trascrive i dati vocali in testo, un secondo modello che traduce il testo da una lingua all'altra e un terzo modello che produce come output i dati vocali dal testo tradotto.

3.2 Ciclo di vita di un sistema di IA

I rischi possono manifestarsi in diverse fasi del ciclo di vita dello sviluppo di un sistema di IA. È quindi necessario comprendere le specificità del ciclo di vita dello sviluppo di un sistema di IA rispetto a un ciclo di vita tradizionale (per sistemi non basati sull'IA).[17] Rischi diversi possono manifestarsi nelle diverse fasi del ciclo di vita dello sviluppo (cfr. sezioni 4 e 5). Il ciclo di vita dello sviluppo dell'IA comprende in genere le fasi descritte in dettaglio nella figura 3.([18])

---

[14] Regolamento (UE) 2024/1689 del Parlamento europeo e del Consiglio, del 13 giugno 2024, che stabilisce norme armonizzate in materia di intelligenza artificiale e modifica i regolamenti (CE) n. 300/2008, (UE) n. 167/2013, (UE) n. 168/2013, (UE) 2018/858, (UE) 2018/1139 e (UE) 2019/2144 e delle direttive 2014/90/UE, (UE) 2016/797 e (UE) 2020/1828 (Legge sull'intelligenza artificiale) CE [2024] GU L2024/1689 http://data.europa.eu/eli/reg/2024/1689/oj.

[15] Sebbene definisca un modello di IA generico come «un modello di IA, compreso un modello di IA addestrato con una grande quantità di dati utilizzando l'autosupervisione su larga scala, che mostra una significativa generalità ed è in grado di svolgere con competenza un'ampia gamma di compiti distinti indipendentemente dal modo in cui il modello è immesso sul mercato e che può essere integrato in una varietà di sistemi o applicazioni a valle» (articolo 3, paragrafo 63, della legge sull'IA)

[16] *La [norma] ISO/IEC 22989:2022 definisce un modello di IA come una «rappresentazione fisica, matematica o comunque logica di un sistema, un'entità, un fenomeno, un processo o dei dati».*

[17] Cfr. ISO/IEC 15288, ISO/IEC 12207 e ISO 24748:2024.

[18] I dettagli sono forniti nella norma ISO 22989:2022

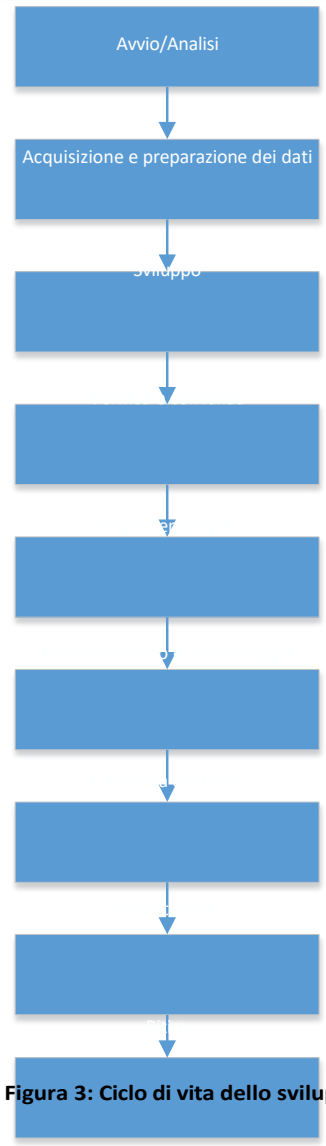


Figura 3: Ciclo di vita dello sviluppo dell'IA

1 Ideazione/Analisi: questa fase iniziale prevede la definizione chiara del problema che il sistema di IA è destinato a risolvere e la selezione dell'architettura del modello di IA.

2 Acquisizione e preparazione dei dati: i dati personali necessari per l'addestramento dipendono dagli obiettivi del sistema di IA. Ad esempio, un sistema di IA destinato all'elaborazione di immagini richiederà immagini come dati personali di addestramento. Queste immagini possono provenire da varie fonti (Internet, banche dati private, ecc.). I dati personali di addestramento da inserire in un sistema di IA specifico devono essere formattati, verificati rispetto ai requisiti di qualità e legali applicabili e normalizzati prima di poter essere utilizzati.

3 Sviluppo: i sistemi di IA possono essere programmati e addestrati per svolgere specifiche funzioni limitate predefinite. Questa fase comprende la selezione di algoritmi appropriati, l'addestramento del sistema di IA sui dati preparati, il collaudo (per verificare se il sistema di IA funziona ed è privo di bug) e la messa a punto dei suoi iperparametri (ad esempio il tasso di apprendimento) per migliorarne le prestazioni. La creazione del sistema di IA può essere effettuata combinando "librerie" (che possono essere acquistate), modelli pre-addestrati acquisiti e sviluppo interno.

4 Verifica e convalida: dopo la fase di sviluppo, il sistema di IA viene rigorosamente verificato ("Stiamo realizzando il prodotto correttamente?") e convalidato ("Stiamo realizzando il prodotto giusto?") per garantire che soddisfi i requisiti funzionali e non funzionali stabiliti nella fase iniziale. Ciò comporta la verifica dell'accuratezza statistica, della robustezza e della generalizzabilità del sistema di IA utilizzando set di dati di test e convalida. I problemi relativi al modello di IA durante questa fase vengono risolti riqualificando il sistema di IA.

5 Implementazione: il sistema di IA può essere implementato nel suo ambiente finale (che potrebbe essere costituito da dispositivi degli utenti finali, server, automobili, ecc.

6 Funzionamento e monitoraggio: una volta implementato, il sistema di IA viene utilizzato dagli utenti e necessita di un monitoraggio continuo per garantire che funzioni come previsto. Ciò include il monitoraggio delle prestazioni, l'aggiornamento del sistema di IA per soddisfare i nuovi requisiti e il suo perfezionamento sulla base del feedback.

7 Convalida continua: quando un sistema di IA utilizza l'apprendimento continuo,[19] la fase operativa e di monitoraggio viene estesa a una fase aggiuntiva di convalida continua. In questa fase, l'addestramento viene condotto in modo continuo mentre il sistema è in produzione. Le prestazioni del sistema vengono valutate regolarmente utilizzando dati di test per garantirne il corretto funzionamento. Inoltre, potrebbe essere necessario aggiornare periodicamente i dati di test per riflettere meglio i dati di produzione attuali, garantendo una valutazione più accurata delle capacità del sistema di IA.

8 Rivalutazione: dopo la fase operativa e di monitoraggio e l'eventuale convalida continua, potrebbe essere necessario rivalutare il sistema di IA sulla base dei suoi risultati prestazionali. I risultati operativi del sistema dovrebbero essere analizzati in modo approfondito e confrontati con i rischi identificati associati al sistema di IA per verificare se tali rischi siano stati adeguatamente mitigati. È possibile che, durante questa fase, emergano rischi che non erano stati identificati in precedenza. Questi dovrebbero essere trattati nel ciclo successivo del processo di gestione dei rischi presentato nella sezione 2.

9 Ritiro: un sistema di IA dovrebbe essere dismesso in modo responsabile ed efficiente quando non è più necessario o viene sostituito da una soluzione più avanzata.

3.3 Acquisizione di un sistema di IA

Inoltre, in molti casi, la creazione di un sistema di IA richiede competenze esterne o l'acquisizione di un prodotto commerciale che copra parte delle funzionalità, dei dati, della sicurezza, ecc. In questi casi è anche importante riconoscere, già nella fase di approvvigionamento, la necessità di effettuare una valutazione dei rischi prima di impegnare effettivamente qualsiasi budget per una soluzione che comporterebbe rischi indesiderati per l'organizzazione.

In questi casi, una delle diverse fasi viene assegnata al fornitore esterno di IA (che si occuperà della fornitura del sistema di IA in tutto o in parte). Il personale delle istituzioni dell'Unione europea coinvolto nell'appalto dei sistemi di IA dovrà coordinare gli sforzi con il personale delle istituzioni dell'Unione europea coinvolto nella diffusione di tali sistemi di IA (ad esempio ingegneri informatici, responsabili di progetti informatici, responsabili della protezione dei dati o coordinatori della protezione dei dati) al fine di redigere la parte tecnica del bando di gara, selezionare il prodotto giusto ed eseguire l'implementazione del sistema di IA.

È possibile adottare diversi approcci per integrare un sistema di IA acquistato in un'infrastruttura esistente. Ai sensi del regolamento 2024/2509[20] , il processo seguito dalle EUI è il seguente:

1. Pubblicazione e trasparenza (articolo 163): garantire i principi di sana gestione finanziaria, trasparenza e parità di trattamento.
2. Bando di gara (articoli 167-169): lanciare un bando di gara aperto che descriva tutte le specifiche necessarie. Il capitolato d'appalto deve contenere i requisiti relativi alla capacità del potenziale offerente di acquistare il sistema di IA previsto e tutti i

[19] Capacità di adattarsi e migliorare le proprie prestazioni nel tempo apprendendo da nuovi dati senza bisogno di essere riqualificato da zero.
[20] Regolamento (UE, Euratom) 2024/2509 del Parlamento europeo e del Consiglio, del 23 settembre 2024, recante norme finanziarie applicabili al bilancio generale dell'Unione (rifusione). https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A32024R2509

Garanzie di qualità tecniche e procedurali pertinenti. Tra le altre cose, le informazioni richieste sono descritte nella sezione 5.3.

3. Criteri di selezione e di aggiudicazione (articolo 170): valutare le offerte sulla base di criteri predefiniti (ad esempio prezzo, qualità, sostenibilità).[21]
4. Esecuzione (articolo 175): monitorare l'attuazione e garantire la conformità.

In questo caso, la "fase di esecuzione" comprenderà fasi simili ad alcune di quelle presentate nella sezione 3.2, ovvero:

5.1 Verifica e convalida

5.2 Implementazione

5.3 Funzionamento e monitoraggio

5.4 Convalida continua

5.5 Rivalutazione

5.6 Ritiro

Analogamente a quanto avviene nello sviluppo dei sistemi di IA, nelle diverse fasi del ciclo di vita degli appalti possono emergere rischi diversi (cfr. sezioni 4 e 5). Per ciascun rischio sono indicate le diverse fasi in cui esso può concretizzarsi (riquadri blu). Se del caso, occorre elaborare misure di mitigazione adeguate per ciascuna delle fasi indicate.

[21] In base al principio di responsabilità, spetta al responsabile del trattamento effettuare controlli in relazione alle preoccupazioni elencate nelle sezioni 4 e 5.

4 Interpretabilità e spiegabilità come *condizioni imprescindibili*

L'interpretabilità e **la spiegabilità** sono aspetti trasversali nell'acquisizione, nello sviluppo e nell'implementazione dei sistemi di IA. In quanto tali, costituiscono prerequisiti per gli EUI che agiscono in qualità di titolari del trattamento dei dati personali nel contesto dei sistemi di IA, al fine di garantire il rispetto degli obblighi previsti dall'EUDPR. Tuttavia, l'interpretabilità e la spiegabilità non devono essere confuse con la trasparenza. I primi due concetti si riferiscono alla misura in cui il titolare del trattamento comprende il funzionamento del proprio sistema di IA. Il terzo, invece, si riferisce all'obbligo del titolare del trattamento di fornire informazioni significative agli interessati. Sebbene l'interpretabilità e la spiegabilità siano fondamentali per il titolare del trattamento nel fornire tali informazioni, esse non sono di per sé sufficienti a soddisfare la soglia di trasparenza. Il presente documento affronta i primi due concetti, come definiti di seguito.

L'interpretabilità si riferisce al grado di comprensibilità umana di un determinato modello o decisione "black box". Equivale alla capacità di comprendere come un modello di IA prende le sue decisioni. Un modello interpretabile funziona in modo trasparente, rivelando le connessioni tra i suoi input e output. Quando un algoritmo è interpretabile, un essere umano può spiegarne il funzionamento in modo chiaro e comprensibile. Ciò rende l'interpretabilità fondamentale per garantire che gli utenti possano comprendere e fidarsi dei modelli di IA.

Ad esempio, un modello di IA che utilizza la regressione lineare[22] per stimare il prezzo degli immobili che appare come "Prezzo = 100.000 + (50 × superficie_in_metri_quadrati) + (10.000 × stanze) + (30.000 × punteggio_codice_postale)" è altamente interpretabile in quanto possiamo comprendere chiaramente il calcolo effettuato.

La spiegabilità nell'IA si concentra sul fornire spiegazioni chiare e coerenti per previsioni o decisioni specifiche del modello. Si riferisce alla capacità di chiarire in che modo un modello di IA prende decisioni in modo accessibile agli utenti finali. Un modello spiegabile offre spiegazioni chiare e intuitive per i suoi risultati, aiutando gli utenti a comprendere le ragioni alla base di un particolare risultato. In sostanza, la spiegabilità sottolinea perché un algoritmo ha raggiunto una decisione specifica e come tale decisione può essere giustificata. La spiegabilità può includere tecniche di analisi post-hoc che riassumono o visualizzano come le caratteristiche influenzano le previsioni, anche se il modello stesso non è intrinsecamente interpretabile.

Ad esempio, una rete neurale convoluzionale (CNN)[23] utilizzata per diagnosticare la polmonite dalle radiografie toraciche non è intrinsecamente interpretabile a causa della complessità del funzionamento interno del modello. Tuttavia, è possibile utilizzare strumenti di spiegabilità come LIME (Local Interpretable Model-Agnostic Explanations)[24] per generare una mappa termica che mostra quali

---

[22] In parole povere, la regressione lineare è un modello che stima la relazione tra le variabili di input e di output.

[23] Tipo di modello di deep learning che utilizza filtri scorrevoli per rilevare modelli e caratteristiche nei dati.

[24] LIME funziona perturbando i dati di input, osservando come cambiano le previsioni e apprendendo un modello interpretabile localmente attorno alla previsione. Ciò contribuisce a costruire un modello di IA semplice e comprensibile dall'uomo. Vedi Marco Tulio Ribeiro, Sameer Singh e Carlos Guestrin. *"Why Should I Trust You?": Explaining the Predictions of Any Classifier.*, 13 agosto 2016, https://doi.org/10.1145/2939672.2939778

parti della radiografia su cui il modello si è concentrato per prendere la sua decisione. Questa spiegazione aiuta i radiologi a comprendere il ragionamento del modello, anche se il modello stesso è una scatola nera.

La **differenza** tra interpretabilità e spiegabilità è che la prima riguarda la comprensione del funzionamento interno dei modelli di IA, mentre la seconda si concentra sulla spiegazione delle decisioni prese da tali modelli. I modelli di IA complessi, come le reti neurali profonde, possono essere difficili da interpretare a causa delle loro strutture intricate e delle interazioni tra i diversi componenti. In questi casi, la spiegabilità può essere più pratica, poiché dà la priorità alla spiegazione delle decisioni piuttosto che alla comprensione del modello. Infine, l'interpretabilità è tipicamente rivolta agli esperti e ai ricercatori di IA, mentre la spiegabilità si concentra sulla comunicazione efficace delle decisioni del modello agli utenti finali. Pertanto, la spiegabilità richiede una presentazione delle informazioni più semplice e intuitiva. Ciò è necessario per garantire, tra le altre considerazioni, che:

- L'organizzazione possa fidarsi del sistema di IA affinché funzioni come previsto;
- Gli errori e i pregiudizi dei modelli possano essere facilmente identificati;
- L'organizzazione è in grado di rilevare quando un sistema di IA viene utilizzato in modo improprio;
- I criteri decisionali sono in linea con gli obiettivi dell'organizzazione;
- Il sistema di IA è verificabile.

4.1.1 Rischio 1: Sistema di IA non interpretabile o non spiegabile

4.1.1.1 Descrizione

I sistemi di IA non interpretabili o non spiegabili comportano rischi significativi perché funzionano come "scatole nere" in cui il funzionamento interno e i processi decisionali rimangono opachi per gli utenti umani, rendendo difficile comprendere come e perché sono stati generati determinati risultati o decisioni.

Si noti che questo rischio si applica alle seguenti fasi del ciclo di vita del sistema di IA:

- Selezione (per l'acquisto di un sistema di IA)
- Verifica e convalida (sia per lo sviluppo che per l'acquisto di un sistema di IA)
- Funzionamento e monitoraggio (sia per lo sviluppo che per l'approvvigionamento di un sistema di IA)
- Convalida continua (sia per lo sviluppo che per l'acquisto di un sistema di IA)
- Rivalutazione (sia per lo sviluppo che per l'approvvigionamento di un sistema di IA)

4.1.1.2 Possibili misure

Le possibili misure per questo rischio sono:

1. Documentazione: è necessario redigere una documentazione adeguata, che includa:
    a. Il tipo di architettura di IA utilizzata (albero decisionale, rete neurale, ecc.) con le sue specificità (dettagli sul tipo o sui tipi di algoritmi di IA utilizzati) e una spiegazione del motivo per cui sono stati scelti questo tipo di modello e algoritmo.
    b. Dettagli sulla provenienza dei dati personali utilizzati per l'addestramento e sul motivo per cui sono adatti all'attività in questione.

c. Informazioni su come agisce il sistema di IA e sulla sua accuratezza nei diversi gruppi identificabili nei dati.
d. Una descrizione dei potenziali pregiudizi, con una spiegazione delle differenze e delle misure adottate per migliorare la qualità complessiva e ridurre le possibilità di pregiudizio.
e. Una descrizione dei limiti del sistema, che chiarisca quali aspettative dovrebbero essere riposte in ciò che il sistema può e non può fare.

Questa documentazione è il punto di partenza per spiegare cosa fa l'IA e come lo fa; il responsabile del trattamento può leggere la documentazione per ottenere informazioni sul funzionamento del sistema di IA e sarà in grado di verificare se ciò che fa il sistema di IA è corretto per quanto riguarda il trattamento dei propri dati personali. La documentazione deve essere pertinente, utile e comprensibile per l'utente.

2. Considerazione di tecniche di spiegabilità quali LIME o SHAP (Shapley Additive Explanations)[25] .
3. Analisi statistica:[26] Analizzare statisticamente i risultati dell'IA e spiegare la logica alla base dei risultati o della loro mancanza.

5 Rischi associati ai principali principi di protezione dei dati

5.1 Principio di equità

Sebbene l'EUDPR non definisca esplicitamente il concetto di correttezza, esso costituisce un principio generale della normativa in materia di protezione dei dati sancito dall'articolo 8, paragrafo 2, della Carta dei diritti fondamentali dell'Unione europea (la Carta).[27]L'obbligo dei responsabili del trattamento di rispettare il principio di correttezza del trattamento è stabilito all'articolo 4, paragrafo 1, lettera a), dell'EUDPR e, nei casi riguardanti il trattamento di dati personali operativi, all'articolo 71, paragrafo 1, lettera a). Il principio di correttezza, sebbene intrinsecamente legato a quello di liceità e trasparenza, ha un significato indipendente e la sua osservanza può essere valutata su base autonoma, indipendentemente dal rispetto degli altri principi di protezione dei dati.(28)

L'EDPB ha chiarito che l'equità è un principio fondamentale che richiede che i dati personali non siano trattati in modo ingiustificatamente pregiudizievole, illegale

[25] Metodo basato sulla teoria dei giochi cooperativi che assegna valori a ciascuna caratteristica in un modello di IA. Quindi, calcola il contributo di ciascuna caratteristica alla previsione per un caso specifico, considerando tutte le possibili combinazioni di caratteristiche. Questa tecnica fornisce una misura unificata dell'importanza delle caratteristiche e aiuta a spiegare la decisione del modello di IA. Cfr. Lundberg, S. M., & Lee, S. I., A *unified approach to interpreting model predictions. Advances in neural information processing systems,* 25 novembre 2017, https://arxiv.org/pdf/1705.07874

[26] ICO, "Compito 4: Tradurre la logica dei risultati del proprio sistema in ragioni utilizzabili e facilmente comprensibili", sito web dell'ICO, 6 agosto 2025, https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/part-2-explaining-ai-in-practice/task-4-translate/

[27] L'articolo 8, paragrafo 2, della Carta stabilisce che «i dati devono essere trattati in modo lecito per finalità determinate e sulla base del consenso dell'interessato o di un altro fondamento giuridico previsto dalla legge».

[28] Comitato europeo per la protezione dei dati, *Linee guida 03/2022 sui modelli di progettazione ingannevoli nelle interfacce delle piattaforme di social media: come riconoscerli ed evitarli*, 14 febbraio 2023, paragrafo 9, https://www.edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-032022-deceptive-design-patterns-social-media_en. Cfr. anche Comitato europeo per la protezione dei dati, *decisione vincolante 2/2023 sulla controversia presentata dall'autorità irlandese di controllo in merito a TikTok Technology Limited (articolo 65 del GDPR),*
2 agosto 2023, paragrafi 100-107, https://www.edpb.europa.eu/our-work-tools/our-documents/binding-decision-board-art-65/binding-decision-22023-dispute-submitted_en.

discriminatorio, inaspettato o fuorviante per l'interessato.[29] Affinché un trattamento sia equo, gli interessati devono comprendere chiaramente in che modo saranno utilizzati i dati personali raccolti e quali saranno le conseguenze di tale trattamento. L'equità impone al responsabile del trattamento di garantire la trasparenza affinché il trattamento non superi le ragionevoli aspettative degli interessati.

Tuttavia, l'equità impone obblighi che vanno oltre i requisiti di trasparenza. Al fine di ottemperare all'obbligo di trattamento equo, occorre valutare in che modo il trattamento inciderà sugli interessi e sui diritti fondamentali delle persone interessate, sia come gruppo che individualmente, e i dati personali non dovrebbero essere utilizzati in modi che potrebbero avere effetti negativi ingiustificati su di esse.[30] Essa impone inoltre ai responsabili del trattamento di attuare garanzie procedurali relative alla raccolta e al trattamento dei dati, nonché all'esercizio dei diritti e degli interessi nell'ambito del quadro di protezione dei dati. In tal senso, l'equità è anche correlata al principio di buona amministrazione, che impone alle istituzioni dell'Unione europea di trattare gli affari delle persone «in modo imparziale, equo e in tempi ragionevoli» (articolo 41 della Carta).

In questo modo, il principio del trattamento equo è alla base dell'intero quadro normativo in materia di protezione dei dati e mira ad affrontare le asimmetrie di potere tra i responsabili del trattamento e gli interessati, al fine di annullare gli effetti negativi di tali asimmetrie e garantire l'effettivo esercizio dei diritti degli interessati. Ciò è particolarmente importante quando i dati personali sono trattati in sistemi di IA, il cui funzionamento e impatto potrebbero essere difficili da comprendere anche per gli stessi responsabili del trattamento. Affidarsi a sistemi di IA complessi per prendere decisioni che riguardano le persone rende anche più difficile per le istituzioni dell'Unione europea giustificare e motivare tali decisioni.

Un rischio importante che potrebbe portare al mancato rispetto del principio di equità in questo contesto è l'esistenza di pregiudizi. Come già sottolineato negli Orientamenti del GEPD sull'intelligenza artificiale generativa e la protezione dei dati personali, «le soluzioni di intelligenza artificiale tendono ad amplificare i pregiudizi umani esistenti e possibilmente a incorporarne di nuovi»([31]).Gli EUI che si basano su sistemi di IA rischiano quindi anche di replicare i pregiudizi contenuti nei set di dati utilizzati per addestrarli, il che a sua volta porterebbe a risultati discriminatori. Ciò è particolarmente problematico quando tali sistemi sono utilizzati per prendere decisioni che riguardano le persone.

Ai fini della presente guida, il principio di correttezza è inteso come l'obbligo per i responsabili del trattamento di identificare, misurare e mitigare tali distorsioni.[32] Per garantire che il trattamento sia corretto, i responsabili del trattamento che acquistano, sviluppano o implementano sistemi di IA che comportano il trattamento di dati personali, in particolare quelli utilizzati per prendere o assistere nella presa di decisioni in merito a

---

[29] Comitato europeo per la protezione dei dati, *Linee guida sulla protezione dei dati fin dalla progettazione e per impostazione predefinita*, 20 ottobre 2020, paragrafo 69, https://www.edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-42019-article-25-data-protection-design-and_en.

[30] Comitato europeo per la protezione dei dati, *Linee guida 2/2019 sul trattamento dei dati personali ai sensi dell'articolo 6, paragrafo 1, lettera b), del GDPR nel contesto di la disposizione di servizi online a dati soggetti*, Versione 2.0, 8 ottobre 2019, paragrafo 12, https://www.edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-22019-processing-personal-data-under-article-61b_en.

[31] Garante europeo della protezione dei dati, *Generative AI and the EUDPR. Prime linee guida del GEPD per garantire la conformità alla protezione dei dati nell'utilizzo dei sistemi di IA generativa (versione 2),* 28 ottobre 2025, sezione 13 (pag. 31), https://www.edps.europa.eu/system/files/2025-10/25-10_28_revised_genai_orientations_en.pdf.

[32] *L'AI Act include requisiti specifici nell'articolo 10 per affrontare questioni simili in termini di pregiudizi e set di dati rappresentativi.*

Gli individui dovrebbero identificare e misurare tali pregiudizi e attuare misure tecniche e organizzative per prevenire o correggere qualsiasi forma di risultato discriminatorio.

Non esiste una definizione unica e universalmente accettata di pregiudizio nell'IA. Tuttavia, esso si riferisce comunemente alla produzione di risultati ingiusti, pregiudizievoli o sistematicamente errati che favoriscono o discriminano determinati gruppi di persone o tipi di input.[33]

I pregiudizi nei sistemi di IA possono produrre risultati che integrano punti di vista discriminatori (ad esempio, concentrandosi in modo ingiusto su una popolazione razziale o etnica in un contesto di polizia) o preferenze ingiuste (ad esempio, approvando in modo sproporzionato le richieste di prestito provenienti da codici postali con redditi più elevati).

Alcune delle cause alla base dei pregiudizi nel contesto dell'IA possono essere:[34]

- Pregiudizio algoritmico: la progettazione stessa del sistema di IA può produrre risultati distorti. La decisione di utilizzare determinati modelli e algoritmi di IA e di includere determinate informazioni nello sviluppo del sistema di IA può portare a risultati ingiusti.
- I dati personali di addestramento: i sistemi di IA apprendono dai dati personali di addestramento. Per addestrare efficacemente un sistema di IA, i dati personali di addestramento sono solitamente disponibili in grandi quantità [(35)].Se i dati personali di addestramento sono distorti, il sistema di IA apprenderà tale distorsione e produrrà risultati altrettanto distorti. Ad esempio, storicamente gli uomini hanno occupato determinati posti di lavoro. Un sistema di IA addestrato con dati storici potrebbe conservare tale pregiudizio storico e apprendere che gli uomini sono i candidati più adatti per quel tipo di lavoro. Allo stesso modo, i sistemi di riconoscimento facciale addestrati sui volti di individui appartenenti a una determinata fascia demografica avranno difficoltà a raggiungere un'elevata accuratezza statistica quando si troveranno di fronte a volti di individui non rappresentati o sottorappresentati nel set di dati personali di addestramento.
- Altri pregiudizi umani: gli sviluppatori o le persone responsabili della formazione o dell'utilizzo di un sistema di IA possono introdurre i propri pregiudizi consci o inconsci nella progettazione o nell'implementazione dei sistemi di IA. Ad esempio, se una parte del processo di formazione richiede che un essere umano esamini alcuni risultati, l'individuo può scegliere di rifiutare un e o accettare alcuni risultati sulla base dei propri pregiudizi espliciti o inconsci.

5.1.1 Rischio 1: Distorsione causata dalla scarsa qualità dei dati utilizzati per l'addestramento dei dati personali

5.1.1.1 Descrizione

I sistemi di IA richiedono dati di qualità da utilizzare durante la fase di addestramento, poiché funzionano secondo il principio "garbage in, garbage out" (se entrano dati errati, escono risultati errati).[36] Set di dati personali di addestramento imprecisi o incompleti possono portare a risultati errati da parte dei sistemi di IA. Ad esempio, l'addestramento di un

[33] IBM, "What is AI bias?", sito web IBM, 6 agosto 2025, https://www.ibm.com/think/topics/ai-bias

[34] Esistono altre fonti di pregiudizio. Cfr. ad esempio Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. e Galstyan, A. (2021). A *survey on bias and fairness in machine learning. ACM computing surveys (CSUR),* 2022, https://arxiv.org/abs/1908.09635

[35] Ad esempio, la classificazione delle immagini può richiedere milioni di immagini; i modelli linguistici di grandi dimensioni vengono in genere addestrati su miliardi o trilioni di frammenti di testo chiamati token.

[36] Questo principio si riferisce all'idea che, in qualsiasi sistema, la qualità dell'output è determinata dalla qualità dell'input.

Un programma di riconoscimento delle immagini su un set di dati con informazioni errate porterebbe il programma a replicare tali errori e a fornire etichette errate. [37]

Questo rischio si applica alle seguenti fasi del ciclo di vita del sistema di IA:

- Acquisizione e preparazione dei dati (per lo sviluppo di un sistema di IA)
- Verifica e convalida (sia per lo sviluppo che per l'approvvigionamento di un sistema di IA)
- Rivalutazione (sia per lo sviluppo che per l'approvvigionamento di un sistema di IA)

5.1.1.2 Possibili misure

Per affrontare i rischi legati alla qualità dei dati, è essenziale garantire che i set di dati personali utilizzati nella fase di acquisizione e preparazione dei dati siano diversificati e rappresentativi della popolazione in cui verrà utilizzato il sistema di IA. Ciò richiede uno sforzo per raccogliere dati da un'ampia gamma di fonti e includere i gruppi sottorappresentati. Audit e aggiornamenti regolari dei set di dati personali utilizzati per l'addestramento possono contribuire a mantenerne la pertinenza e l'inclusività.

1. Definire una politica di garanzia della qualità per il set di dati personali di addestramento che:
    a. Descriva i tipi di dati da raccogliere e i metodi di acquisizione dei dati.
    b. Descriva le fasi seguite durante la preparazione dei dati personali di addestramento (pulizia,[38] etichettatura,[39] normalizzazione e ridimensionamento,[40] suddivisione[41]).
    c. Fornisce una definizione dei criteri e delle misure di qualità.
    d. Definisce la soglia di qualità.
2. Definisce e implementa una procedura per valutare il set di dati personali di formazione che, in conformità con la politica, campiona il set di dati, misura e valuta rispetto alla soglia di qualità concordata.
3. Effettua regolari controlli di qualità dei dati personali di formazione del sistema di IA per verificarne la qualità.
4. Utilizzo di tecniche statistiche per individuare i valori anomali, che dovrebbero essere verificati per determinare se sono validi (e quindi devono essere lasciati nei dati personali di addestramento) o se sono errati (e quindi devono essere rimossi). Ad esempio, se si tratta di dati personali di addestramento contenenti date di nascita, le date che indicano un individuo di età superiore ai 100 anni dovrebbero essere esaminate attentamente. Una volta identificati, questi errori possono essere corretti tramite intervento manuale (se necessario), tecniche statistiche (ad esempio stimando i valori potenziali calcolando la media o la mediana degli altri valori), tecniche di regressione (ad esempio lineari, polinomiali, ecc.) o tecniche statistiche più evolute come K-Nearest Neighbours, in cui vengono utilizzati dati personali di addestramento simili per correggere i valori devianti), o anche rimuovendoli dai dati personali di addestramento se ritenuto necessario.

[37] Processo di assegnazione di tag o annotazioni significative ai dati grezzi (come immagini, testo o audio) per indicare l'output o la categoria corretta, utilizzato per addestrare modelli di apprendimento automatico supervisionati al fine di effettuare previsioni accurate.
[38] Rimuovere o correggere errori, duplicati o informazioni irrilevanti nel set di dati, come valori mancanti, valori anomali o incongruenze.
[39] Verificare che le etichette siano accurate per garantire che l'IA apprenda le associazioni corrette.
[40] Standardizzare il set di dati ridimensionando le caratteristiche per garantire l'uniformità ed evitare che determinati attributi influenzino in modo sproporzionato il modello.
[41] Dividere il set di dati in set di addestramento, convalida e test per garantire che il modello generalizzi bene i dati non visti e non si adatti eccessivamente.

5. I dati personali utilizzati per l'addestramento possono essere verificati per confermare la validità delle informazioni e ridurre al minimo il rischio di distorsioni. Ciò comporta l'ottenimento dei dati da fonti affidabili e anche il controllo dei dati personali di formazione rispetto a fonti simili affidabili, se disponibili, effettuando una revisione umana da parte di alcuni individui con competenze specifiche nel settore, utilizzando tecniche di corrispondenza approssimativa (in cui si verifica la vicinanza delle voci dei dati personali di formazione), utilizzando tecniche statistiche per rilevare cluster di dati personali di formazione (per identificare potenziali distorsioni) e documentando la provenienza dei dati personali di formazione, giustificandone la correttezza,([42]) la validità e la tracciabilità.[43]
6. È possibile verificare la standardizzazione e la coerenza per garantire che tutte le voci dei dati personali utilizzati per l'addestramento siano formattate e rappresentate in modo identico (ad esempio, le date di nascita utilizzando lo stesso formato GG/MM/AAAA).

5.1.2 Rischio 2: Distorsioni nei dati personali di addestramento

5.1.2.1 Descrizione

Il risultato di un modello di apprendimento automatico può essere distorto anche quando viene addestrato con dati personali accurati e completi. Le fonti comuni di distorsione relative ai dati personali di addestramento sono:[44]

- Gli errori di campionamento durante la raccolta dei dati possono portare a distorsioni della popolazione. Questi errori si verificano quando il set di dati personali di addestramento non è rappresentativo della popolazione più ampia. Ad esempio, se un sistema di IA sanitario viene sviluppato utilizzando dati provenienti prevalentemente da ospedali urbani, potrebbe non funzionare altrettanto bene nelle zone rurali, dove le caratteristiche demografiche e le condizioni di salute dei pazienti potrebbero essere diverse. La mancanza di dati diversificati e rappresentativi significa che il sistema di IA non è in grado di generalizzare le sue previsioni su popolazioni e contesti diversi.
- Il pregiudizio storico si riferisce a pregiudizi preesistenti e questioni socio-tecniche presenti nella società che possono infiltrarsi nei dati, anche quando si utilizzano tecniche per prevenire i pregiudizi. Ad esempio, la storia dimostra che il numero di donne amministratrici delegate (CEO) era ed è tuttora inferiore rispetto a quello degli uomini. Un'intelligenza artificiale che cercasse immagini di amministratori delegati potrebbe essere influenzata da pregiudizi e mostrare principalmente immagini di amministratori delegati uomini.

Questo rischio si applica alle seguenti fasi del ciclo di vita del sistema di IA:

- Acquisizione e preparazione dei dati (per lo sviluppo di un sistema di IA)
- Verifica e convalida (sia per lo sviluppo che per l'approvvigionamento di un sistema di IA)
- Rivalutazione (sia per lo sviluppo che per l'approvvigionamento di un sistema di IA)

[42] I risultati del sistema di IA sono in linea con i risultati attesi o reali.

[43] Se il sistema di IA è appropriato e funziona come previsto nel contesto dato.

[44] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A, *A survey on bias and fairness in machine learning. ACM computing surveys (CSUR)*, 2022, https://arxiv.org/abs/1908.09635
I tipi di pregiudizi sono descritti in Mikołajczyk-Bareła, A., & Grochowski, *Un'indagine sui pregiudizi nella ricerca sull'apprendimento automatico*, 2023, https://arxiv.org/abs/2308.11254

5.1.2.2 Possibili misure

Quando l'organizzazione ha accesso ai dati personali di formazione, è possibile ricorrere a tecniche per individuare e correggere i pregiudizi all'interno dei dati personali di formazione. Questi metodi possono includere adeguamenti statistici per bilanciare la rappresentanza dei diversi gruppi e interventi algoritmici per mitigare l'impatto di eventuali pregiudizi individuati.

1. I dati personali rappresentativi utilizzati per l'addestramento sono essenziali affinché un sistema di IA produca risultati affidabili e imparziali. Se i dati personali utilizzati per l'addestramento differiscono in modo significativo dai dati reali, il sistema rischia di fornire conclusioni errate. Ciò comporta:
    a. Corrispondenza della distribuzione: esaminare se le distribuzioni statistiche delle caratteristiche chiave sia nel set di dati di addestramento che in quello di input previsto sono simili. Strumenti come istogrammi, statistiche riassuntive e clustering possono identificare differenze e lacune.
    b. Diversità e copertura: analizzare se la gamma e la diversità dei dati di input sono adeguatamente rappresentate dai dati personali di addestramento. I dati personali di addestramento rappresentativi dovrebbero coprire tutti gli scenari, le classi e le variazioni significative presenti nei dati operativi del mondo reale.
    c. Convalida e convalida incrociata: utilizzare set di dati di convalida che imitano i dati operativi previsti. L'applicazione della convalida incrociata o la rotazione attraverso le fasi di addestramento/convalida aiutano a misurare se le prestazioni del modello sono coerenti e robuste per i dati non visti.
    d. Revisione da parte di esperti e verifiche degli scenari: gli esperti in materia dovrebbero definire le variabili chiave e verificare se i dati personali di addestramento racchiudono tutti gli aspetti operativi rilevanti. È inoltre fondamentale verificare l'eventuale squilibrio delle classi o i bias di campionamento.
2. Caratteristiche prive di pregiudizi: [45] Selezionare caratteristiche che siano meno soggette a introdurre pregiudizi. Evitare caratteristiche che codificano direttamente attributi sensibili come razza, genere o status socioeconomico, a meno che la loro inclusione non sia giustificata e gestita con attenzione. Cercare di individuare caratteristiche che fungano da proxy di attributi sensibili (ad esempio, il codice postale come proxy del reddito familiare).
3. Ingegneria delle caratteristiche:[46] Le caratteristiche selezionate per essere incluse nel modello di IA possono influire in modo significativo sul comportamento del modello stesso. Se alcune caratteristiche vengono scelte sulla base di ipotesi distorte, le previsioni del modello di IA che ne derivano rifletteranno tali distorsioni. Questo processo richiede un'attenta valutazione per garantire che le caratteristiche utilizzate siano pertinenti e non introducano inavvertitamente distorsioni. Inoltre, le caratteristiche possono essere trasformate in modo da ridurre le distorsioni. Ad esempio, la riponderazione o il riscalamento delle caratteristiche può contribuire a garantire che nessuna singola caratteristica influenzi in modo sproporzionato i risultati del modello di IA.([47])
4. Verifiche dei pregiudizi:[48]Condurre verifiche regolari dei dati personali di addestramento del sistema di IA per verificare la presenza di pregiudizi.

---

[45] Le caratteristiche sono gli attributi dei punti dati nel set di dati (ad esempio età, codice postale o altezza).

[46] L'ingegneria delle caratteristiche è un processo di selezione, trasformazione e creazione di variabili rilevanti dai dati grezzi per migliorare le prestazioni e l'interpretabilità dei modelli di apprendimento automatico.

[47] Riscale: processo di riadattamento dell'intervallo o della distribuzione dei valori dei dati, spesso prima di inserire i dati in un modello di apprendimento automatico.
Riponderazione: riassegnazione di valori numerici, o "pesi", a vari input, caratteristiche o connessioni in un modello di IA, in particolare nell'apprendimento automatico e nelle reti neurali.

[48] IBM, "Introducing AI Fairness 360", sito web IBM Research, 6 agosto 2025, https://research.ibm.com/blog/ai-fairness-360 Aequitas, "The Bias and Fairness Audit Toolkit for Machine Learning", sito web Aequitas, 6 agosto 2025. https://dssg.github.io/aequitas/

5. Tecniche di mitigazione del bias per i dati: le tecniche di mitigazione del bias per i dati, come il re- la ponderazione,[49] possono ridurre i bias identificati nei dati utilizzati dal sistema di IA.

5.1.3 Rischio 3: Overfitting dei dati personali di addestramento

5.1.3.1 Descrizione

Per un sistema di IA, l'overfitting si riferisce alla tendenza ad apprendere i dettagli e il rumore nei dati personali di addestramento a tal punto da riprodurre i dati su cui è stato addestrato e influire negativamente sulle prestazioni del sistema di IA su dati nuovi e non visti. In altre parole, l'overfitting si verifica quando il modello apprende i dettagli specifici e il rumore nei dati personali di addestramento in modo così approfondito da memorizzare effettivamente i dati. Ciò si verifica perché il sistema di IA diventa eccessivamente complesso, catturando modelli specifici del set di dati personali di addestramento ma che non si generalizzano bene ad altri dati, oppure quando il set di dati personali di addestramento non ha la dimensione minima adeguata.

Questo rischio si applica alle seguenti fasi del ciclo di vita del sistema di IA:

- Acquisizione e preparazione dei dati (per lo sviluppo di un sistema di IA)
- Verifica e convalida (sia per lo sviluppo che per l'approvvigionamento di un sistema di IA)
- Funzionamento e monitoraggio (sia per lo sviluppo che per l'approvvigionamento di un sistema di IA)
- Convalida continua (sia per lo sviluppo che per l'approvvigionamento di un sistema di IA)
- Rivalutazione (sia per lo sviluppo che per l'approvvigionamento di un sistema di IA)

5.1.3.2 Possibili misure

Le possibili misure per questo rischio sono:[50]

1. Interruzione anticipata: l'interruzione anticipata è una tecnica che consiste nell'interrompere il processo di addestramento non appena le prestazioni del modello di IA sul set di convalida iniziano a deteriorarsi, indicando un potenziale sovradattamento ai dati personali di addestramento.
2. Semplificazione: semplificare il modello di IA è anche un approccio pratico per mitigare l'overfitting. Ciò può comportare la selezione di un numero minore di caratteristiche più rilevanti o il potatura del modello di IA rimuovendo i parametri o i neuroni meno importanti.[51] Riducendo la complessità del modello di IA, diventa meno probabile che si verifichi un overfitting rispetto al rumore nei dati personali di addestramento, migliorando così la sua capacità di generalizzazione.

---

[49] Assegnare pesi diversi ai campioni provenienti da gruppi sottorappresentati per garantire che abbiano un'influenza equa sul modello. Vedi Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos e Yiannis Kompatsiaris. 2018. *Riponderazione sensibile adattiva per mitigare il bias nella classificazione attenta all'equità. In Atti della Conferenza mondiale sul World Wide Web 2018 (WWW '18). Comitato direttivo delle conferenze internazionali sul World Wide Web, Repubblica e Cantone di Ginevra, CHE,* 23 aprile 2018, https://doi.org/10.1145/3178876.3186133

[50] Ying, Xue. (2019), *Una panoramica dell'overfitting e delle sue soluzioni. Journal of Physics: Conference Series*, 2019, https://iopscience.iop.org/article/10.1088/1742-6596/1168/2/022022/pdf

[51] Proprietà o caratteristica individuale misurabile dei dati utilizzata da un modello per effettuare previsioni o classificazioni. Domino.ai, "Che cos'è una caratteristica nel machine learning e nella scienza dei dati?", sito web Domino.ai, 6 agosto 2025, https://domino.ai/data-science-dictionary/feature

3. Tecniche di regolarizzazione: le tecniche di regolarizzazione sono metodi utilizzati nell'apprendimento automatico per prevenire il sovradattamento aggiungendo una penalità alla complessità del modello, garantendo una buona generalizzazione ai nuovi dati. Le due tecniche più comuni sono la regolarizzazione L1 e L2. La regolarizzazione L1 (Lasso) aggiunge una penalità basata sui valori assoluti dei pesi del modello, incoraggiando la sparsità portando alcuni pesi a zero, il che effettua efficacemente la selezione delle caratteristiche. La regolarizzazione L2 (Ridge) aggiunge una penalità basata sul quadrato dei pesi, riducendoli verso lo zero senza eliminarli, promuovendo la fluidità e prevenendo l'overfitting senza scartare le caratteristiche. Elastic Net combina sia la regolarizzazione L1 che L2 per bilanciare la selezione delle caratteristiche e la stabilità dei pesi, il che è particolarmente utile quando si ha a che fare con caratteristiche altamente correlate.
4. Dropout: il dropout è un'altra tecnica di regolarizzazione comunemente utilizzata nelle reti neurali, in cui i neuroni selezionati casualmente vengono ignorati durante l'addestramento. Ciò impedisce al modello di IA di diventare eccessivamente dipendente da neuroni specifici e incoraggia la rete ad apprendere caratteristiche più robuste.

5.1.4 Rischio 4: Distorsione algoritmica

5.1.4.1 Descrizione

Il pregiudizio algoritmico è definito come il pregiudizio che emerge dalla progettazione stessa del sistema di IA, indipendentemente dai dati personali utilizzati per l'input e l'addestramento.[52]

Anche il modo in cui è progettato un algoritmo può portare a decisioni distorte. Ad esempio, la scelta delle funzioni matematiche utilizzate per ottimizzare le prestazioni dell'algoritmo, i metodi utilizzati per prevenire il sovradattamento e la decisione di applicare modelli statistici all'intero set di dati o a sottogruppi specifici possono tutti introdurre distorsioni. Il modello COMPAS,[53] utilizzato per prevedere i tassi di recidiva nel sistema giudiziario statunitense, è risultato avere una distorsione nei confronti degli imputati afroamericani. Una delle cause era che il modello ipotizzava erroneamente una relazione lineare tra alcune caratteristiche e la previsione.([54]) Inoltre, anche l'uso di metodi statistici soggetti a distorsioni può influenzare il risultato dell'algoritmo. Queste scelte di progettazione possono influenzare le decisioni dell'algoritmo, portando a risultati distorti che possono favorire o svantaggiare ingiustamente determinati gruppi.

Questo rischio si applica alle seguenti fasi del ciclo di vita del sistema di IA:

- Ideazione/analisi (per lo sviluppo di un sistema di IA)
- Verifica e convalida (sia per lo sviluppo che per l'approvvigionamento di un sistema di IA)
- Convalida continua (sia per lo sviluppo che per l'approvvigionamento di un sistema di IA)
- Rivalutazione (sia per lo sviluppo che per l'approvvigionamento di un sistema di IA)

[52] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. e Galstyan, A. (2021). *Indagine sulla parzialità e l'equità nell'apprendimento automatico. Indagini informatiche ACM (CSUR),* 25 gennaio 2022, https://arxiv.org/pdf/1908.09635

[53] Profilo di gestione dei detenuti per sanzioni alternative (COMPAS)

[54] Cynthia Rudin, Caroline Wang e Beau Coker, *L'era della segretezza e dell'ingiustizia nella previsione della recidiva,* 31 maggio 2020, https://hdsr.mitpress.mit.edu/pub/7z10o269/release/7

5.1.4.2 Possibili misure

È importante includere algoritmi attenti all'equità, funzioni obiettivo equilibrate, un'attenta progettazione delle caratteristiche, audit regolari, trasparenza, metriche di equità [55] e pratiche di sviluppo inclusive per creare sistemi di IA che siano equi e funzionino in modo corretto su popolazioni diverse [(56)].

1. Algoritmi sensibili all'equità:[57] Scegliere algoritmi progettati tenendo conto dei vincoli di equità. Alcuni algoritmi, come Two Naive Bayes,[58] sono sviluppati specificamente per affrontare la questione dell'equità e possono aiutare a mitigare i pregiudizi durante la fase di avvio/analisi.
2. Selezione della funzione obiettivo:[59] Gli algoritmi sono in genere ottimizzati per raggiungere obiettivi specifici definiti da funzioni obiettivo, come massimizzare l'accuratezza o minimizzare l'errore. Tuttavia, queste funzioni possono introdurre pregiudizi se non tengono conto dell'equità tra i diversi gruppi. Ad esempio, un algoritmo ottimizzato esclusivamente per l'accuratezza complessiva potrebbe trascurare la disparità di prestazioni tra i gruppi maggioritari e minoritari, portando a risultati distorti.
3. Verifiche dei pregiudizi: [60] Effettuare verifiche regolari del sistema di IA per verificare la presenza di pregiudizi. Ciò comporta la valutazione delle prestazioni dell'algoritmo su diversi gruppi demografici e l'identificazione e la valutazione di eventuali disparità.
4. Test con dati diversificati: testare l'algoritmo su set di dati diversificati che riflettono la varietà di scenari reali che incontrerà. Ciò contribuisce a garantire che il modello di IA funzioni in modo equo su popolazioni diverse.
5. Interpretabilità del modello di IA: [61] Utilizzare modelli o tecniche di IA interpretabili che rendano più comprensibili i modelli di IA complessi. Ciò consente di identificare e affrontare le fonti di bias all'interno del modello di IA.
6. Verificare se il problema in questione possa essere risolto in modo efficace ed efficiente utilizzando algoritmi diversi da quelli di machine learning o deep learning, oppure integrandoli con altri approcci, tra cui l'IA neurosimbolica.[62]

---

[55] Le metriche sono misure quantitative utilizzate per valutare le prestazioni e l'efficacia di un sistema di IA in vari compiti. Modelli di IA diversi (ad esempio classificazione, regressione, clustering) richiedono metriche diverse. Spesso, nessuna singola metrica fornisce un quadro completo delle prestazioni, pertanto si raccomanda di calcolare più metriche per valutare in modo completo i diversi aspetti delle prestazioni del sistema di IA.

[56] Costruire team diversificati e inclusivi. Vedi Dr. Moeed Yusuf, *Algorithmic Justice: Bias in Code, Bias in Society*, 2024, https://journalpsa.com/index.php/JPSA/article/view/14/16 e Hernández, E. G, *Towards an ethical and inclusive implementation of artificial intelligence in organizations: a multidimensional framework,* 2 maggio 2024, https://arxiv.org/abs/2405.01697

[57] Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P. e Roth, D, *Uno studio comparativo sugli interventi volti a migliorare l'equità nell'apprendimento automatico. In Atti della conferenza su equità, responsabilità e trasparenza (*pp. 329-338), gennaio 2019, https://arxiv.org/pdf/1802.04422

[58] Toon Calders e Sicco Verwer, *Tre approcci Naive Bayes per una classificazione priva di discriminazioni. Rivista Data Mining; numero speciale con articoli selezionati da ECML/PKDD*, 2010, https://www.cs.ru.nl/~sicco/papers/dmkd10.pdf

[59] Formulazione matematica utilizzata per misurare la differenza tra i risultati previsti e quelli effettivi, che guida l'ottimizzazione del modello

[60] IBM, "Introducing AI Fairness 360", sito web IBM Research, 6 agosto 2025, https://research.ibm.com/blog/ai-fairness-360 Aequitas, "The Bias and Fairness Audit Toolkit for Machine Learning", sito web Aequitas, 6 agosto 2025. https://dssg.github.io/aequitas/

[61] Carvalho DV, Pereira EM, Cardoso JS., *Machine Learning Interpretability: A Survey on Methods and Metrics,* 26 luglio 2019, https://doi.org/10.3390/electronics8080832

[62] Wan, Z., Liu, C. K., Yang, H., Li, C., You, H., Fu, Y., ... & Raychowdhury, A. *Verso sistemi di intelligenza artificiale cognitiva: indagine e prospettive sull'intelligenza artificiale neuro-simbolica,* 2 gennaio 2024, https://arxiv.org/abs/2401.01040

5.1.5 Rischio 5: Distorsione interpretativa

5.1.5.1 Descrizione

Il pregiudizio interpretativo si verifica quando gli analisti traggono conclusioni errate o distorte dai dati personali di addestramento e dai risultati del modello di IA, spesso influenzati da preconcetti o da una comprensione incompleta. Inoltre, l'interpretazione selettiva delle metriche di prestazione può oscurare le questioni relative all'equità, portando potenzialmente all'implementazione di sistemi di IA distorti. Ciò può portare a considerazioni errate che possono avere conseguenze durante la correzione, la messa a punto o il riaddestramento del modello.

Ad esempio, un'organizzazione sanitaria sviluppa uno strumento diagnostico basato sull'intelligenza artificiale per prevedere la probabilità che i pazienti abbiano una malattia specifica sulla base della loro storia clinica, dei sintomi e dei risultati dei test. Lo strumento fornisce un punteggio di probabilità compreso tra 0 e 1, dove 1 rappresenta un'alta probabilità di avere la malattia. Il bias di interpretazione si avrebbe se gli operatori sanitari che utilizzano lo strumento interpretassero erroneamente il risultato come una diagnosi definitiva piuttosto che come un punteggio di probabilità. Potrebbero supporre che un punteggio di 0,8 significhi che il paziente ha sicuramente la malattia, mentre un punteggio di 0,2 significhi che il paziente è sano.

Questo rischio si applica alle seguenti fasi del ciclo di vita del sistema di IA:

- Verifica e convalida (sia per lo sviluppo che per l'approvvigionamento di un sistema di IA)
- Operatività e monitoraggio (sia per lo sviluppo che per l'approvvigionamento di un sistema di IA)
- Rivalutazione (sia per lo sviluppo che per l'approvvigionamento di un sistema di IA)

5.1.5.2 Possibili misure

Le possibili misure per questo rischio sono:

1. Coinvolgimento di un team diversificato: il coinvolgimento di un team diversificato composto da data scientist, esperti del settore e parti interessate può fornire molteplici prospettive sui dati personali di addestramento e sui risultati del modello di IA.
2. Documentazione e comunicazione chiare: mantenere una documentazione chiara e completa delle fonti dei dati, della selezione delle caratteristiche, delle fasi di pre-elaborazione e delle decisioni di modellizzazione contribuisce a garantire la trasparenza del processo di analisi.
3. Tecniche di spiegabilità dei modelli di IA: l'integrazione di tecniche di spiegabilità dei modelli di IA come SHAP, LIME e l'analisi dell'importanza delle caratteristiche può fornire informazioni su come i modelli di IA prendono le decisioni. [63]
4. Formazione e sensibilizzazione: fornire formazione e sensibilizzazione su pregiudizi, equità e interpretabilità ai membri del team coinvolti nel processo di sviluppo dell'IA può migliorare la capacità del team di identificare e affrontare i pregiudizi di interpretabilità.

[63] GEPD, *TechDispatch n. 2/2023 sull'intelligenza artificiale spiegabile*, 16 novembre 2023, https://www.edps.europa.eu/data-protection/our-work/publications/techdispatch/2023-11-16-techdispatch-22023-explainable-artificial-intelligence_en

5. Verifiche dei pregiudizi:[64] Effettuare verifiche periodiche dell'interpretazione dei risultati da parte degli analisti per controllare eventuali pregiudizi ( ).

5.2 Principio di accuratezza

5.2.1 Significato giuridico di accuratezza nell'EUDPR

Ai sensi dell'articolo 4, paragrafo 1, lettera d), dell'EUDPR, i dati personali devono essere esatti e, se necessario, aggiornati. Devono essere adottate tutte le misure ragionevoli per garantire che i dati personali inesatti, tenuto conto delle finalità per le quali sono trattati, siano cancellati o rettificati senza indugio. Ai sensi dell'EUDPR, l'esattezza richiede che i responsabili del trattamento garantiscano che *i dati personali stessi* non siano errati o fuorvianti in merito a qualsiasi questione di fatto.

5.2.2 Significato statistico dell'accuratezza nello sviluppo dell'IA

Contrariamente al significato del principio di accuratezza nella protezione dei dati, nel contesto dell'IA l'accuratezza è un parametro di prestazione che misura la frequenza con cui un sistema di IA indovina la risposta corretta divisa per il numero totale di previsioni.

L'accuratezza nell'IA non si riferisce all'accuratezza dei dati personali inseriti o ai dati personali previsti in sé, ma alle prestazioni del *sistema di IA*.

Nella presente Guida, d'ora in poi utilizzeremo il termine "accuratezza" per riferirci al corrispondente principio di protezione dei dati e "accuratezza statistica" per riferirci all'accuratezza di un sistema di IA.

5.2.3 Rischio 1: Output di dati personali inesatti

5.2.3.1 Descrizione

La mancata valutazione dell'accuratezza statistica può creare un rischio di non conformità al principio di accuratezza dei dati quando porta all'implementazione di modelli di IA che producono dati personali inesatti. Quando i risultati di un sistema di IA non sono accuratamente convalidati, gli errori possono passare inosservati. Data l'ampia varietà di modelli di IA, esiste anche un'ampia varietà di metriche che possono essere utilizzate per valutare l'accuratezza statistica dei modelli di IA. Alcuni esempi sono forniti nell'allegato 1.

Un modello di IA può generare informazioni errate o prive di senso (compresi dati personali) che non erano presenti né nei dati personali utilizzati per l'addestramento né negli input ricevuti. Ciò può verificarsi in modelli come i Large Language Models (LLM), che possono "inventare" fatti o fornire risposte sicure ma false. Queste "allucinazioni" derivano dalla natura probabilistica dei modelli di IA, che tentano di prevedere il risultato più probabile piuttosto che effettuare calcoli basati su regole deterministiche rilevate e convalidate in anticipo.

[64] IBM, "Introducing AI Fairness 360", sito web di ricerca IBM, 6 agosto 2025, https://research.ibm.com/blog/ai-fairness-360 Aequitas, "The Bias and Fairness Audit Toolkit for Machine Learning", sito web Aequitas, 6 agosto 2025. https://dssg.github.io/aequitas/

Inoltre, l'accuratezza statistica dei sistemi di IA dipende fortemente dalla qualità dei set di dati utilizzati per l'addestramento.[65] Se i dati personali utilizzati per l'addestramento sono inaccurati, incompleti o distorti, il sistema di IA potrebbe produrre risultati inaffidabili o errati. Poiché gli algoritmi di apprendimento automatico apprendono modelli, comportamenti e associazioni dai dati su cui vengono addestrati, eventuali errori o rappresentazioni errate in questi dati possono perpetuarsi nelle previsioni del sistema di IA. Si noti che anche i sistemi di IA addestrati con set di dati di buona qualità possono dare luogo ad allucinazioni.

Questo rischio si applica alle seguenti fasi del ciclo di vita del sistema di IA:

- Acquisizione e preparazione dei dati (per lo sviluppo di un sistema di IA)
- Verifica e convalida (sia per lo sviluppo che per l'approvvigionamento di un sistema di IA)
- Funzionamento e monitoraggio (sia per lo sviluppo che per l'acquisto di un sistema di IA)
- Rivalutazione (sia per lo sviluppo che per l'approvvigionamento di un sistema di IA)

5.2.3.2 Possibili misure

Le possibili misure per questo rischio sono:

1. Dati personali di alta qualità per l'addestramento: dati personali di alta qualità per l'addestramento sono fondamentali per sviluppare modelli di IA accurati e affidabili. Poiché i sistemi di IA apprendono dai dati su cui vengono addestrati, garantire che i dati siano ben preparati e puliti può migliorare significativamente l'accuratezza statistica del modello.
2. Casi limite: il sistema di IA dovrebbe essere verificato e convalidato con casi limite (valori anomali) ed esempi contraddittori per valutarne la resilienza e l'affidabilità in condizioni insolite o difficili. [66]
3. Dati diversificati e rappresentativi: è essenziale raccogliere dati da fonti diverse e garantire che rappresentino tutti i possibili scenari che il sistema di IA incontrerà nella produzione. Ad esempio, se si sta sviluppando un sistema di IA per il riconoscimento facciale in un aeroporto, il sistema di IA dovrebbe essere addestrato su immagini rappresentative (condizioni di illuminazione, espressioni facciali) e non solo su immagini frontali ben illuminate ad alta risoluzione.
4. Set di dati bilanciato: un set di dati bilanciato garantisce che ogni categoria o classe in un problema di classificazione sia rappresentata in modo equo. Ad esempio, in un modello di diagnosi medica, dovrebbe esserci un numero adeguato di casi positivi e negativi per evitare che il modello diventi parziale verso un risultato.
5. Ottimizzazione degli iperparametri (HPO):[67] L'HPO consiste nel trovare la migliore serie di iperparametri per migliorare le prestazioni di un modello su dati non visti. Gli iperparametri sono impostazioni di configurazione nei modelli di apprendimento automatico che vengono impostate prima dell'addestramento e controllano aspetti del processo di apprendimento, come la complessità del modello, il tasso di apprendimento e

[65] Zhou, Y., Tu, F., Sha, K., Ding, J. e Chen, H., *A Survey on Data Quality Dimensions and Tools for Machine Learning*, 28 giugno 2024, https://arxiv.org/abs/2406.19614v1
Budach, Lukas & Feuerpfeil, Moritz & Ihde, Nina & Nathansen, Andrea & Noack, Nele & Patzlaff, Hendrik & Harmouch, Hazar & Naumann, Felix, *The Effects of Data Quality on ML-Model Performance*, luglio 2022, https://www.researchgate.net/publication/362386427_The_Effects_of_Data_Quality_on_ML-Model_Performance

[66] Un caso limite è un problema o una situazione che si verifica solo in presenza di parametri operativi estremi (massimi o minimi).

[67] Morales-Hernández, A., Van Nieuwenhuyse, I. & Rojas Gonzalez, S. *A survey on multi-objective hyperparameter optimization algorithms for machine learning*, 24 dicembre 2022, https://doi.org/10.1007/s10462-022-10359-2

regolarizzazione (limitazione di pesi o parametri elevati), ma non vengono appresi dai dati.

6. Supervisione umana (collaborazione uomo-IA (HAIC) e Human in-the-loop (HITL)):[68] l'integrazione della revisione umana nel processo decisionale dell'IA garantisce che le previsioni del modello siano ricontrollate, riducendo le possibilità di errore. La revisione umana dei sistemi di IA può assumere varie forme, a seconda del contesto, della complessità dell'applicazione di IA e del livello di rischio associato alle sue decisioni.[69]
7. Verificare se il problema in questione possa essere risolto utilizzando in modo efficace ed efficiente algoritmi diversi da quelli di machine learning o deep learning, oppure integrandoli con altri approcci , tra cui l'IA neurosimbolica.[70]

5.2.4 Esempio specifico: risultati imprecisi dovuti alla deriva dei dati e al deterioramento della qualità dei dati personali inseriti

5.2.4.1 Descrizione

Per deriva dei dati si intendono i cambiamenti nel tempo delle proprietà statistiche dei dati di input [(71).] Essa può verificarsi a causa di vari fattori, quali cambiamenti nel comportamento degli utenti o modifiche nel contesto operativo del sistema di IA. La deriva dei dati può indurre il modello di IA a formulare previsioni o decisioni inaccurate. La qualità dei dati di input può deteriorarsi a causa di problemi quali aumento del rumore, valori mancanti o inesattezze. Ad esempio, un modello di IA per il credit scoring addestrato con dati relativi a crediti richiesti durante una situazione economica stabile e utilizzato nel contesto di una crisi economica con cambiamenti sostanziali nell'inflazione e nella disoccupazione è rappresentativo del data drift.

Questo rischio si applica alle seguenti fasi del ciclo di vita del sistema di IA:

- Operazioni e monitoraggio (sia per lo sviluppo che per l'approvvigionamento di un sistema di IA)
- Convalida continua (sia per lo sviluppo che per l'approvvigionamento di un sistema di IA)
- Rivalutazione (sia per lo sviluppo che per l'approvvigionamento di un sistema di IA)

5.2.4.2 Possibili misure

Le possibili misure per questo rischio sono:

---

[68] Fragiadakis, G., Diou, C., Kousiouris, G., & Nikolaidou, M., *Evaluating human-ai collaboration: A review and methodological framework*, 07 marzo 2025, https://arxiv.org/abs/2407.19098
Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, Liang He, *A survey of human-in-the-loop for machine learning, Future Generation Computer Systems*, ottobre 2022, https://doi.org/10.1016/j.future.2022.05.014

[69] La revisione pre-implementazione comporta l'esame approfondito del sistema di IA durante la sua fase di sviluppo, compresa la convalida della qualità dei dati personali di addestramento, la verifica della presenza di pregiudizi e la garanzia della conformità al quadro giuridico. La supervisione Human In The Loop (HITL) prevede l'intervento umano durante la fase operativa dell'IA, consentendo agli esseri umani di monitorare, modificare o approvare le decisioni in tempo reale. La revisione post-decisione si concentra sulla valutazione delle decisioni prese dall'IA dopo la loro esecuzione, identificando errori o aree di miglioramento.

[70] EDPS, *Techsonar 2025*, 15 novembre 2024, https://www.edps.europa.eu/data-protection/our-work/publications/reports/2024-11-15-techsonar-report-2025_en
Wan, Z., Liu, C. K., Yang, H., Li, C., You, H., Fu, Y., ... & Raychowdhury, A, *Towards cognitive ai systems: a survey and prospective on neuro-symbolic ai*, 2 gennaio 2024, https://arxiv.org/abs/2401.01040

[71] GeeksforGeeks. *Deriva dei dati nell'apprendimento automatico*, 23 luglio 2025, https://www.geeksforgeeks.org/machine-learning/data-drift-in-machine-learning/

1. Metodi di rilevamento della deriva dei dati:[72] Implementare metodi di rilevamento della deriva dei dati che monitorino i cambiamenti nella distribuzione dei dati nel tempo.
2. Monitoraggio della qualità dei dati: implementare sistemi di monitoraggio della qualità dei dati che tengano traccia di metriche quali completezza, accuratezza statistica e coerenza dei dati in entrata. Rivedere regolarmente queste metriche per garantire che i dati utilizzati per la convalida rimangano di alta qualità.
3. Riqualificazione regolare dei modelli: si tratta di un processo in cui i modelli di apprendimento automatico vengono aggiornati periodicamente con nuovi dati per garantire che le loro prestazioni rimangano elevate man mano che i modelli di dati evolvono nel tempo. La riqualificazione può essere effettuata secondo un programma fisso (ad esempio settimanale o mensile) o attivata da cali di prestazioni o da derive rilevate. Il processo di riqualificazione prevede in genere la raccolta di nuovi dati, la loro pre-elaborazione, l'aggiornamento del modello e la convalida della nuova versione per garantire prestazioni migliorate.
4. Feedback degli utenti: creare circuiti di feedback in cui gli utenti possano segnalare problemi o anomalie nelle inferenze. Utilizzare questo feedback per identificare potenziali problemi di qualità dei dati o derive e apportare le modifiche necessarie al modello di IA.

5.2.5 Rischio 2: Informazioni poco chiare da parte del fornitore del sistema di IA

5.2.5.1 Descrizione

Per gestire efficacemente i rischi relativi alla protezione dei dati quando si acquista un sistema di IA pre-addestrato, le organizzazioni dovrebbero concentrarsi sulla comprensione dei processi di sviluppo utilizzati dal fornitore di IA.

Ciò comporta porre domande relative al sistema di IA nel suo complesso e approfondire i dettagli relativi alla corretta gestione dei rischi rilevanti per le fasi di avvio/analisi, acquisizione e preparazione dei dati, sviluppo e verifica e convalida (vedere la sezione 3.2), al fine di comprendere se il prodotto finale soddisferà le esigenze dell'organizzazione.

Questo rischio si applica alle seguenti fasi del ciclo di vita del sistema di IA:

- Bando di gara (per l'acquisto di un sistema di IA)
- Selezione (per l'acquisto di un sistema di IA)

5.2.5.2 Possibili misure

L'organizzazione dovrebbe invitare il fornitore a produrre:

1. Documentazione generale che copra:
    a. Cosa fa il sistema di IA e come lo fa: Specifiche tecniche e documentazione sull'architettura che descrivono in dettaglio il funzionamento del sistema di IA, compreso il suo

[72] Gemaque RN, Costa AFJ, Giusti R, dos Santos EM. *Una panoramica dei metodi di rilevamento della deriva non supervisionati*, 21 luglio 2020, https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.1381
Andrés L. Suárez-Cetrulo, David Quintana, Alejandro Cervantes, *Indagine sull'apprendimento automatico per flussi di dati con deriva concettuale ricorrente, Expert Systems with Applications*, 1 marzo 2023, https://doi.org/10.1016/j.eswa.2022.118934

algoritmi sottostanti, metodi di elaborazione dei dati, caratteristiche e capacità di integrazione con altri sistemi esistenti.

b. Interfaccia utente e interfaccia di programmazione dell'applicazione (API)[73] dettagli: modalità di accesso e utilizzo del sistema di IA da parte dell'organizzazione dal punto di vista dell'utente e dello sviluppatore.

2. Documentazione su come il sistema di IA gestisce la trasparenza, l'interpretabilità e la spiegabilità: per evitare di utilizzare il sistema di IA come una "scatola nera", in cui i risultati vengono generati senza una chiara visibilità su come vengono ottenuti, l'organizzazione dovrebbe ottenere informazioni relative alla comprensione e alla spiegazione di come l'IA giunge alle sue conclusioni e quali fattori determinano le decisioni dell'IA, fornendo ragioni chiare e comprensibili alla base di risultati specifici.
3. Documentazione sulle misure di sicurezza informatica relative all'integrità del modello: come è stata garantita l'integrità del modello durante lo sviluppo e quali misure sono state adottate/dovrebbero essere adottate per garantirne l'integrità nel tempo.
4. Documentazione relativa alle pratiche di governance dei dati personali del fornitore, compresa la raccolta e il trattamento dei dati personali: che tipo di dati sono stati raccolti? Come sono stati ottenuti i dati personali? Come sono stati utilizzati i dati personali per addestrare il modello di IA e quali metodi sono stati impiegati per garantirne l'equità e l'accuratezza? Molti fornitori di IA potrebbero fornire informazioni vaghe o limitate al riguardo. Tuttavia, un certo livello di trasparenza è fondamentale, comprese le proprietà statistiche del set di dati personali di addestramento. I responsabili del trattamento dovrebbero valutare se i dati demografici del set di dati personali di addestramento sono vicini o lontani dai dati demografici dei dati personali che il sistema di IA acquisirà durante il suo funzionamento.
5. Procedure di convalida e test, e risultati: come è stato testato e convalidato il modello in vari scenari e quali sono stati i risultati? Quali dati sono stati utilizzati e come sono stati gestiti i casi limite? Inoltre, l'organizzazione dovrebbe richiedere al fornitore una serie di metriche che consentano di valutare il sistema di IA rispetto agli obiettivi dell'organizzazione.

   Sebbene le metriche dipendano dal contesto,[74] alcune metriche comuni sono:

   a. Tasso di falsi positivi (FPR) Parità: il numero di casi positivi errati può essere confrontato tra diversi gruppi. Questo parametro dovrebbe essere simile per questi gruppi.
   b. Parità del tasso di falsi negativi (FNR): il numero di veri positivi mancati può essere confrontato tra diversi gruppi. Questo parametro dovrebbe essere simile per questi gruppi.
   c. Equità della calibrazione: questa metrica confronta il risultato del modello con la realtà tra diversi gruppi. Il modello dovrebbe funzionare con un'accuratezza statistica simile tra i gruppi.

---

[73] Insieme di protocolli e strumenti che consente a diverse applicazioni software di comunicare e scambiare dati senza soluzione di continuità. Funge da intermediario, consentendo a vari componenti software di interagire senza bisogno di comprendere i dettagli di implementazione sottostanti. Le API semplificano il processo di sviluppo fornendo metodi predefiniti per accedere a funzionalità o dati specifici. Ad esempio, l'API OpenAI fornisce l'accesso a modelli avanzati per l'elaborazione del linguaggio naturale, la generazione di immagini e altre attività di intelligenza artificiale. Gli sviluppatori possono utilizzare questa API per integrare funzionalità come la generazione di testo, la sintesi e le capacità di conversazione nelle loro applicazioni.

[74] Isabel Barberá, *AI Possibili Rischi e misure di mitigazione - Entità Entità riconoscimento,* settembre *2023,* https://www.edpb.europa.eu/system/files/2024-07/ai-risks_d1named-entity-recognition_edpb-spe-programme_en.pdf Isabel Barberá, *AI Possibili Rischi e Mitigazioni - Ottica Riconoscimento riconoscimento,* settembre 2023, https://www.edpb.europa.eu/system/files/2024-06/ai-risks_d2optical-character-recognition_edpb-spe-programme_en_2.pdf

d. Pari opportunità (EOP): input individuali con caratteristiche simili dovrebbero avere le stesse possibilità di ottenere un risultato positivo nella previsione del modello.

Oltre a questi parametri comuni, esistono anche benchmark specifici per determinate attività (ad esempio, comprensione del linguaggio naturale o risoluzione di problemi matematici) che potrebbero consentire la valutazione del sistema di IA rispetto agli obiettivi dell'organizzazione. L'allegato I include un elenco di alcuni dei più noti.

5.3 Principio di minimizzazione dei dati

Per apprendere accuratamente i modelli, effettuare previsioni affidabili e generalizzare bene su dati nuovi e non visti, i sistemi di IA vengono spesso addestrati su grandi set di dati. Ciò è necessario per garantire che il sistema di IA riceva informazioni sufficienti per apprendere i modelli e creare output con proprietà statistiche vicine a quelle dei dati personali di addestramento che ha ricevuto. Se i dati personali di addestramento non sono sufficientemente rappresentativi dei dati di input che riceverà una volta implementato, il sistema di IA non sarà in grado di produrre output accurati per alcuni dei suoi dati di input.

Inoltre, i dati personali utilizzati per l'addestramento possono provenire da varie fonti, essere distribuiti in tutto il mondo e appartenere a diverse entità/organizzazioni/individui con requisiti di qualità dei dati diversi. Gli EUI devono assicurarsi di disporre di una base giuridica valida prima di utilizzare dati personali per addestrare i sistemi di IA [(75).]

Le EUI devono inoltre garantire il rispetto del principio di minimizzazione dei dati. L'articolo 4, paragrafo 1, lettera c), stabilisce che i dati personali devono essere «adeguati, pertinenti e limitati a quanto necessario rispetto alle finalità per le quali sono trattati (minimizzazione dei dati)». È quindi necessario trovare un equilibrio per fornire al sistema di IA dati personali sufficienti per funzionare in modo accurato, limitando al contempo la quantità di dati personali a quanto necessario per conseguire la finalità perseguita dal responsabile del trattamento.

5.3.1 Rischio 1: raccolta e conservazione indiscriminata di dati personali

5.3.1.1 Descrizione

Dato che i modelli di apprendimento automatico dipendono dai dati personali utilizzati per il loro addestramento, si tende a raccogliere e trattare il maggior numero possibile di dati personali di addestramento.

[75] Garante europeo della protezione dei dati, *IA generativa e EUDPR. Orientamenti per garantire la conformità alla protezione dei dati quando si utilizzano sistemi di IA generativa. (Versione 2)*, 28 ottobre 2025, https://www.edps.europa.eu/system/files/2025-10/25-10_28_revised_genai_orientations_en.pdf.
«Il GEPD ha già messo in guardia dall'uso di tecniche di web scraping per raccogliere dati personali, attraverso le quali gli individui possono perdere il controllo delle loro informazioni personali quando queste vengono raccolte a loro insaputa, contro le loro aspettative e per scopi diversi da quelli della raccolta originaria. Il GEPD ha inoltre sottolineato che il trattamento dei dati personali disponibili al pubblico rimane soggetto alla legislazione dell'UE in materia di protezione dei dati. A tale riguardo, l'uso di tecniche di web scraping per raccogliere dati dai siti web e il loro utilizzo a fini di formazione potrebbero non essere conformi ai principi pertinenti in materia di protezione dei dati, compresi quelli di minimizzazione dei dati e di accuratezza, nella misura in cui non vi è alcuna valutazione dell'affidabilità delle fonti».

La raccolta di grandi volumi di dati, comprese tutte le informazioni possibili senza criteri chiari o rilevanza, può portare all'accumulo di informazioni che potrebbero non essere necessarie per gli obiettivi del sistema di IA e potrebbero essere contrarie al principio di minimizzazione dei dati. La minimizzazione dei dati garantisce che vengano memorizzate solo informazioni pertinenti e aggiornate, impedendo la conservazione di dati obsoleti o inesatti che potrebbero distorcere i processi decisionali o danneggiare le persone.

Questo rischio si applica alle seguenti fasi del ciclo di vita del sistema di IA:

- Acquisizione e preparazione dei dati (per lo sviluppo di un sistema di IA)
- Verifica e convalida (sia per lo sviluppo che per l'acquisto di un sistema di IA)
- Rivalutazione (sia per lo sviluppo che per l'approvvigionamento di un sistema di IA)

5.3.1.2 Possibili misure

Le misure possibili per questo rischio sono:

1. Utilizzare le informazioni esistenti sull'argomento per una valutazione preliminare del tipo di dati personali di formazione che possono essere utili per trarre le conclusioni necessarie. Convalidare la pertinenza dei tipi di dati personali di formazione pianificati prima della formazione completa e dell'operatività.
2. Campionamento dei dati:[76] Campionare un sottoinsieme rappresentativo dei dati personali di addestramento invece di utilizzare l'intero set di dati. Questo approccio, noto come campionamento dei dati, consiste nel selezionare una porzione più piccola e ben bilanciata dei dati che rifletta accuratamente la diversità e le caratteristiche chiave dell'intero set di dati. Progettando attentamente il campione in modo da includere tutte le categorie rilevanti ed evitare sovrarappresentazioni o distorsioni, le organizzazioni possono addestrare efficacemente i modelli di IA, riducendo al minimo la quantità di dati che elaborano.
3. Anonimizzazione/pseudonimizzazione: il sistema di IA dovrebbe essere sviluppato con dati anonimizzati, ove possibile. Se sono necessari dati personali, si dovrebbe prendere in considerazione l'uso di dati pseudonimizzati.

5.4 Principio di sicurezza

L'obbligo di garantire la sicurezza dei dati personali è sancito dall'articolo 4, paragrafo 1, lettera f), dell'EUDPR: «I dati personali sono [...] trattati in modo da garantire un'adeguata sicurezza dei dati personali, compresa la protezione da trattamenti non autorizzati o illeciti e dalla perdita, dalla distruzione o dal danneggiamento accidentali, mediante misure tecniche o organizzative adeguate (integrità e riservatezza)».

I sistemi informatici che integrano componenti di IA devono tenere conto delle minacce alla sicurezza relative ai sistemi informatici in generale (come attacchi di phishing, attacchi malware), ma anche delle minacce alla sicurezza specifiche relative a tali componenti di IA.

[76] Daitaku, «ML Models on a Data Diet: How Training Set Size Impacts Performance», sito web Daitaku, 19 luglio 2023, 6 agosto 2025, https://blog.dataiku.com/ml-models-on-a-data-diet
Andrea Montanari, 4 marzo 2024, «Improving AI via optimal selection of training samples», sito web Granica.ai, https://granica.ai/blog/improving-ai-via-optimal-selection-of-training-samples

Come accennato in precedenza, il presente documento affronta le preoccupazioni, i rischi e le misure in materia di protezione dei dati derivanti specificamente dallo sviluppo e dall'uso dei sistemi di IA. Pertanto, tratterà solo i rischi per la sicurezza specifici dell'IA in materia di protezione dei dati (e non i rischi generali per la sicurezza dei sistemi informatici).[77]

Ad esempio, data la quantità di dati necessari per addestrare il sistema di IA, questi dati personali di addestramento sono preziosi e, se la loro riservatezza viene compromessa, potrebbero essere utilizzati per prendere di mira le persone i cui dati sono stati divulgati. La riservatezza potrebbe essere compromessa sfruttando specifiche vulnerabilità dell'IA, come gli attacchi di inversione del modello.[78]

Un altro esempio potrebbe essere quello di manipolare i dati personali utilizzati per l'addestramento (data poisoning) o il modello di IA stesso (model poisoning) al fine di introdurre errori nel sistema di IA, come ad esempio introdurre distorsioni o far sì che l'IA produca risultati privi di senso.[79] Inoltre, il modello di IA stesso potrebbe essere rubato e poi utilizzato per scopi dannosi.

Pertanto, in termini di riservatezza e integrità (per garantire che il sistema di IA funzioni come previsto e per proteggere i dati personali degli individui), è necessario proteggere i dati personali di addestramento, i dati di input, i dati di output e il modello di IA stesso.

5.4.1 Rischio 1: divulgazione dei dati personali di addestramento da parte del sistema di IA

5.4.1.1 Descrizione

Quando i modelli di IA vengono addestrati su set di dati contenenti informazioni personali, esiste la possibilità che i risultati del modello rivelino involontariamente dettagli sulle persone incluse nel set di addestramento. Questo fenomeno può verificarsi attraverso vari attacchi alla privacy,[80] quali l'inversione del modello, l'inferenza di appartenenza o il rigurgito dei dati personali di addestramento. Negli attacchi di inversione del modello, un avversario può ricostruire informazioni sensibili analizzando i risultati del modello, rivelando di fatto i dati personali associati alle persone presenti nel set di dati personali di addestramento. Gli attacchi di inferenza di appartenenza sfruttano i punteggi di affidabilità generati dai modelli di IA.([81]) Se un modello mostra una maggiore affidabilità nelle previsioni relative a persone specifiche che facevano parte dei dati personali di addestramento, gli aggressori possono dedurre che tali persone erano incluse nel set di dati. Estratti dai set di dati personali di addestramento o dati in essi inclusi potrebbero anche essere rigurgitati alla lettera quando un modello di IA inavvertitamente

[77] Yupeng Hu, Wenxin Kuang, Zheng Qin, Kenli Li, Jiliang Zhang, Yansong Gao, Wenjia Li e Keqin Li, *Sicurezza dell'intelligenza artificiale: minacce e contromisure*, 23 novembre 2021, https://doi.org/10.1145/3487890

[78] Tipo di minaccia alla sicurezza dell'IA in cui un aggressore sfrutta i risultati di un modello di apprendimento automatico per dedurre informazioni sensibili sui dati personali utilizzati per l'addestramento, effettuando di fatto il reverse engineering del modello per rivelare gli attributi riservati dei dati su cui è stato addestrato.

[79] Avvelenamento dei dati: attacco informatico in cui un avversario manipola intenzionalmente un set di dati personali di addestramento utilizzato da un modello di intelligenza artificiale o di apprendimento automatico, con l'obiettivo di comprometterne le prestazioni o alterarne il comportamento iniettando informazioni false, modificando i dati esistenti o cancellando punti dati critici.
Avvelenamento del modello: manipolazione intenzionale dei parametri o dell'architettura di un modello di IA da parte di avversari per ottenere risultati dannosi specifici durante la fase di inferenza del modello di IA.

[80] Fang, H., Qiu, Y., Yu, H., Yu, W., Kong, J., Chong, B., ... & Xia, S. T., *Privacy leakage on DNNs: A survey of model inversion attacks and defenses*, 11 settembre 2024, https://arxiv.org/abs/2402.04013

[81] Hu, H., Salcic, Z., Sun, L., Dobbie, G., Yu, P. S., & Zhang, X., *Attacchi di inferenza di appartenenza sull'apprendimento automatico: un'indagine*, 3 febbraio 2022, https://arxiv.org/abs/2103.07853

riproduce nei suoi risultati esempi o dettagli identificabili relativi a persone incluse nei dati personali utilizzati per l'addestramento.

Questo rischio si applica alle seguenti fasi del ciclo di vita del sistema di IA:

- Funzionamento e monitoraggio (sia per lo sviluppo che per l'approvvigionamento di un sistema di IA)
- Convalida continua (sia per lo sviluppo che per l'approvvigionamento di un sistema di IA)
- Rivalutazione (sia per lo sviluppo che per l'approvvigionamento di un sistema di IA)

5.4.1.2 Possibili misure

Le misure possibili per questo rischio sono:

1. Formazione sulla minimizzazione dei dati personali: devono essere raccolti e utilizzati solo i dati personali necessari. Ciò riduce al minimo il rischio che le identità possano essere ricostruite.
2. Tecniche di perturbazione dei dati: è possibile utilizzare diverse tecniche per modificare i dati personali utilizzati per l'addestramento al fine di rendere più difficile la reidentificazione, mantenendo al contempo i dati sufficientemente accurati per gli scopi del sistema di IA:
    a. Generalizzazione: alcuni input possono essere generalizzati con intervalli più ampi per ridurre le possibilità di reidentificazione. Ad esempio, è possibile utilizzare i codici postali invece degli indirizzi, le province invece dei nomi delle città.
    b. Aggregazione: i punti dati possono essere raggruppati. Ad esempio, è possibile utilizzare fasce d'età ampie invece di età specifiche, fasce di reddito invece di redditi esatti.
    c. Privacy differenziale/aggiunta di rumore: è possibile introdurre una casualità controllata nei dati personali di addestramento (mantenendo le proprietà statistiche dei dati personali di addestramento).
3. Generazione di dati sintetici:[82] I sistemi di IA potrebbero essere addestrati, almeno in parte, utilizzando dati personali di addestramento generati artificialmente. Questi dati sintetici riflettono le proprietà statistiche dei dati reali senza essere attribuibili a un individuo specifico.[83] Se ritenuta adeguata, questa misura dovrebbe essere attuata con la dovuta cautela, poiché potrebbe comportare ulteriori difficoltà.[84] Pertanto, se si prevede di adottare questa misura, essa dovrebbe essere attuata in combinazione con le misure aggiuntive sopra presentate, data la possibilità di ulteriori attacchi (ad esempio attacchi di inferenza dell'appartenenza).[85]
4. Attuare misure per impedire la replica esatta dei dati personali di addestramento durante la produzione di un output, ad esempio utilizzando la decodifica MEMFREE.[86]

---

[82] Lu, Y., Shen, M., Wang, H., Wang, X., van Rechem, C., & Wei, W., *Machine learning for synthetic data generation: a review*, 04 aprile 2025, https://arxiv.org/abs/2302.04062v9

[83] La creazione di dati sintetici comporta un compromesso tra privacy e utilità. Un quadro pratico per valutare questo compromesso è disponibile su reslbesl, tandriamil Nampoina Andriamilanto, emidec Emiliano De Cristofaro, bristena-op "Privacy evaluation framework for synthetic data publishing", 23 giugno 2021, https://github.com/spring-epfl/synthetic_data_release

[84] Hao, S., Han, W., Jiang, T., Li, Y., Wu, H., Zhong, C., ... & Tang, H., *Synthetic data in AI: Challenges, applications, and ethical implications*, 3 gennaio 2024, https://arxiv.org/abs/2401.01629v1

[85] Van Breugel, B., Sun, H., Qian, Z., & van der Schaar, M., *Attacchi di inferenza dell'appartenenza contro dati sintetici attraverso il rilevamento dell'overfitting*., 24 febbraio 2023, https://arxiv.org/abs/2302.12580

[86] . Daphne Ippolito, Florian Tramèr, Milad Nasr, Chiyuan Zhang, Matthew Jagielski, Katherine Lee, Christopher A. Choquette-Choo, Nicholas Carlini, *Prevenire la generazione di memorizzazione letterale nei modelli linguistici dà un falso senso di privacy*, 11 settembre 2023, *https://arxiv.org/abs/2210.1/546*

5.4.2 Rischio 2: archiviazione dei dati personali e violazioni dei dati personali

5.4.2.1 Descrizione

Le enormi quantità di dati necessarie per i sistemi di IA aumentano i rischi per la sicurezza. Se i dati personali utilizzati per l'addestramento vengono in qualche modo compromessi (in termini di riservatezza e/o integrità), il sistema di IA può essere gravemente danneggiato e causare una violazione dei dati ai sensi dell'articolo 3, paragrafo 16, dell'EUDPR [87].L'effetto sull'operazione di trattamento complessiva dipenderà dal modo in cui i dati sono stati compromessi. Ad esempio, se l'integrità dei dati è compromessa (ad esempio, avvelenamento dei dati o attacchi di evasione), il sistema di IA potrebbe funzionare male o fornire risultati errati, mentre, se è compromessa la riservatezza, i dati personali compromessi avranno ripercussioni sugli individui (a seconda dei dati personali compromessi, gli effetti possono essere di natura finanziaria, sanitaria, ecc.

Questo rischio si applica alle seguenti fasi del ciclo di vita del sistema di IA:

- Acquisizione e preparazione dei dati (per lo sviluppo di un sistema di IA)
- Verifica e convalida (sia per lo sviluppo che per l'approvvigionamento di un sistema di IA)
- Funzionamento e monitoraggio (sia per lo sviluppo che per l'approvvigionamento di un sistema di IA)
- Convalida continua (sia per lo sviluppo che per l'approvvigionamento di un sistema di IA)
- Rivalutazione (sia per lo sviluppo che per l'approvvigionamento di un sistema di IA)

5.4.2.2 Possibili misure

Le misure possibili per questo rischio sono:

1. Utilizzo dell'anonimizzazione e/o della pseudonimizzazione ove possibile: ciò garantirà che, in caso di violazione della riservatezza dei dati, l'impatto sulle persone sia ridotto al minimo. Ciò deve essere bilanciato con il fatto che un sistema di IA necessita di dati di qualità sufficiente per essere efficace.
2. Crittografia: crittografare i dati quando non sono attivamente utilizzati dal sistema di IA per evitare la fuga di informazioni e proteggere l'integrità dei dati.
3. Dati personali sintetici per la formazione: [88] L'uso di dati personali sintetici per la formazione (in contrapposizione ai dati reali) garantirà che nessun dato reale possa essere compromesso in termini di riservatezza. I dati personali sintetici per l'addestramento dovrebbero essere costruiti in modo tale da essere rappresentativi dei dati reali (ad esempio, stesse caratteristiche statistiche) affinché la fase di sviluppo sia il più possibile in linea con l'uso reale del prodotto finale (garantendo la qualità del risultato). Questa misura dovrebbe essere attuata con la dovuta cautela in quanto potrebbe comportare ulteriori difficoltà. ([89]) Pertanto, se si prevede di adottare questa misura, essa dovrebbe essere attuata in combinazione con la

[87] Ovvero «una violazione della sicurezza che comporti la distruzione, la perdita, la modifica, la divulgazione non autorizzata o l'accesso accidentale o illecito a dati personali trasmessi, conservati o trattati in altro modo».
[88] Lu, Y., Shen, M., Wang, H., Wang, X., van Rechem, C., & Wei, W., *Machine learning for synthetic data generation: a review*, 04 aprile 2025, https://arxiv.org/abs/2302.04062v9
[89] Hao, S., Han, W., Jiang, T., Li, Y., Wu, H., Zhong, C., ... & Tang, H., *Dati sintetici nell'IA: sfide, applicazioni e implicazioni etiche*, 3 gennaio 2024, https://arxiv.org/abs/2401.01629v1

misure aggiuntive presentate sopra, data la possibilità di ulteriori attacchi (ad esempio attacchi di inferenza di appartenenza).[90]

4. Pratiche di sviluppo sicure**:** seguire pratiche di codifica sicure durante lo sviluppo di modelli di IA per impedire agli aggressori di sfruttare le vulnerabilità nel codice o nell'infrastruttura dell'IA.
5. Autenticazione a più fattori (MFA): implementare l'autenticazione a più fattori per l'accesso a sistemi sensibili di intelligenza artificiale per impedire a utenti non autorizzati di manipolare o rubare modelli.

5.4.3 Rischio 3: fuga di dati personali attraverso le interfacce di programmazione delle applicazioni

5.4.3.1 Descrizione

Molti sistemi di IA sono realizzati utilizzando modelli di IA forniti da terzi accessibili tramite chiamate API. Le API possono essere vulnerabili allo sfruttamento se non adeguatamente protette. L'accesso non autorizzato a queste API può portare a violazioni dei dati. Ad esempio, un'API potrebbe inavvertitamente esporre più dati del previsto se i controlli di accesso non sono definiti correttamente o se gli endpoint di debug che forniscono informazioni dettagliate sul sistema vengono lasciati abilitati in un ambiente di produzione.

Questo rischio si applica alla seguente fase del ciclo di vita del sistema di IA:

- Funzionamento e monitoraggio (sia per lo sviluppo che per l'approvvigionamento di un sistema di IA)

5.4.3.2 Possibili misure

Le misure possibili per questo rischio sono:

1. Accesso alle API: implementare meccanismi di autenticazione forte, come l'autenticazione a più fattori (MFA), per garantire che solo gli utenti e i sistemi autorizzati possano accedere alle API.
2. Controllo degli accessi basato sui ruoli (RBAC): applicare il RBAC per limitare chi può accedere, modificare o interagire con il sistema di IA in base al proprio ruolo nell'organizzazione.
3. Limitazione: [91] Si tratta di una tecnica utilizzata per controllare il numero di richieste che un client può inviare a un'API entro un determinato periodo di tempo, al fine di prevenirne l'uso improprio e mitigare il rischio di attacchi automatizzati, come i tentativi di forza bruta.
4. Crittografia delle comunicazioni: utilizzare HTTPS (TLS) per crittografare i dati trasmessi tra client e API, garantendo che i dati siano protetti da intercettazioni e spionaggio durante il transito.
5. Registrazione e monitoraggio: implementare la registrazione e il monitoraggio delle chiamate API. Utilizzare sistemi SIEM (Security Information and Event Management) per analizzare e rispondere alle potenziali minacce quasi in tempo reale.

---

[90] Van Breugel, B., Sun, H., Qian, Z., & van der Schaar, M., *Attacchi di inferenza dell'appartenenza contro dati sintetici attraverso il rilevamento dell'overfitting*, 24 febbraio 2023, https://arxiv.org/abs/2302.12580

[91] Tecnica utilizzata per controllare il numero di richieste che un client può effettuare a un'API entro un determinato periodo di tempo, gestendo in modo efficace il traffico e prevenendo il sovraccarico del server. Quando un client supera la frequenza di richiesta consentita, il throttling blocca o rallenta temporaneamente le sue richieste, garantendo un'allocazione equa delle risorse e mantenendo le prestazioni complessive del sistema.

6. Controlli di sicurezza: eseguire regolarmente controlli di sicurezza e test di penetrazione delle API per identificare e risolvere le vulnerabilità. Questi controlli dovrebbero includere revisioni del codice, verifiche della configurazione e scansioni delle vulnerabilità, e dovrebbero utilizzare strumenti automatizzati per la scansione continua delle vulnerabilità comuni.
7. Sviluppo sicuro: seguire le migliori pratiche di sicurezza nella progettazione delle API, come la convalida e la sanificazione degli input.
8. Applicazione di patch: mantenere aggiornati il software API e l'infrastruttura sottostante con le ultime patch di sicurezza e gli ultimi aggiornamenti.

5.5 Diritti dell'interessato

L'EUDPR conferisce agli interessati diversi diritti individuali, ovvero il diritto di accesso (articolo 17), di rettifica (articolo 16), di cancellazione (articolo 19), di limitazione (articolo 20), alla portabilità dei dati (articolo 22) e di opposizione (articolo 23). La natura complessa dei sistemi di IA potrebbe rendere più difficile per gli EUI agire in base a tali diritti, soprattutto quando gli interessati esercitano tali diritti in relazione ai dati personali contenuti nei set di dati personali di addestramento, che il modello potrebbe poi "memorizzare" e riproporre nella fase di inferenza. Anziché affrontare tutti i rischi legati al mancato rispetto delle disposizioni che disciplinano i diritti degli interessati, la presente sezione si concentra su questioni tecniche trasversali che condizionano la possibilità stessa per i responsabili del trattamento di esercitare tali diritti.

In primo luogo, l'attuazione di tali diritti richiede l'identificazione dei dati personali trattati. Ad esempio, per consentire agli interessati di accedere ai propri dati personali, questi ultimi devono prima essere individuati nel sistema. Analogamente, per cancellare alcuni dati personali, è necessario prima identificarli. Solo quando il titolare del trattamento ha individuato i dati personali nel sistema può fornire l'accesso, rettificarli, cancellarli, limitarne il trattamento e trasferirli. In secondo luogo, e guardando in particolare ai diritti di rettifica e cancellazione, il titolare del trattamento deve attuare misure per rettificare o cancellare effettivamente i dati personali che ha identificato. Ciò è particolarmente difficile nel contesto dei sistemi di IA, in cui i set di dati personali di addestramento sono stati assorbiti nei parametri del modello.

5.5.1 Rischio 1: Identificazione incompleta dei dati personali trattati

5.5.1.1 Descrizione

Quando i dati personali sono trattati da sistemi di IA, gli interessati hanno il diritto di accedere ai propri dati personali, compresi quelli contenuti nei set di dati personali utilizzati per l'addestramento e potenzialmente conservati nel modello risultante. Nel rispondere alla richiesta di un interessato di esercitare i propri diritti, il titolare del trattamento deve sia identificare se e dove i dati personali sono utilizzati durante la fase di addestramento, sia valutare se il modello, nel rispondere a determinati prompt, possa divulgare tali dati nel formulare inferenze.

Per i set di dati strutturati,[92] identificare dove vengono trattati i dati personali dell'interessato nel set di dati personali di addestramento può essere relativamente facile se i dati personali di addestramento sono stati conservati. Per i set di dati non strutturati,[93] la situazione sarà più complessa poiché non esiste uno schema o un formato fisso che possa essere sfruttato.

Quando si considerano i dati personali contenuti nei modelli di IA stessi, la complessità di molti modelli di IA (in particolare i modelli di deep learning, in cui i dati sono rappresentati e memorizzati in modo molto complesso) rende difficile l'accesso a specifici punti di dati. Inoltre, data la loro stessa natura, in cui i risultati possono cambiare a parità di richiesta (i modelli di IA sono, nella loro essenza, macchine statistiche che forniscono un risultato selezionato da una serie di possibili risposte), non è possibile garantire l'estrazione delle informazioni complete.

Ad esempio, se Jane Doe desidera che i suoi dati vengano cancellati da un modello costruito utilizzando reti neurali profonde, sarà impossibile identificare quali siano i parametri del modello che rappresentano i dati relativi alla signora Doe. Anche se fosse possibile individuare tali parametri, essi potrebbero rappresentare anche dati relativi ad altre persone che condividono alcune caratteristiche con la signora Doe; potrebbe quindi non essere possibile modificare i parametri per cancellare i dati della signora Doe.

Questo rischio si applica alle seguenti fasi del ciclo di vita del sistema di IA:

- Sviluppo (per lo sviluppo di un sistema di IA)
- Funzionamento e monitoraggio (sia per lo sviluppo che per l'approvvigionamento di un sistema di IA)

5.5.1.2 Possibili misure

Le misure possibili per questo rischio sono:[94]

1. Conservazione dei metadati per facilitare l'identificazione dei dati personali: i set di dati personali utilizzati per l'addestramento dovrebbero includere metadati in modo che i record o i file pertinenti contenenti dati personali possano essere identificati più facilmente nel caso in cui un interessato invochi il proprio diritto di accedere ai propri dati personali. I metadati dovrebbero includere informazioni dettagliate sulle fonti dei dati e sui metodi utilizzati per la loro raccolta. Inoltre, dovrebbero documentare tutte le fasi di pre-elaborazione, come la pulizia dei dati, la pseudonimizzazione o l'aumento, per garantire una traccia chiara di come i dati sono stati preparati per l'addestramento.
2. Strumenti di recupero dei dati: gli strumenti di recupero dei dati dovrebbero essere creati dagli sviluppatori di IA o dai fornitori di set di dati personali di addestramento al fine di fornire agli interessati i propri dati personali quando esercitano il loro diritto di accesso. Tali strumenti dovrebbero consentire agli interessati di richiedere e ottenere una visione chiara e completa dei propri dati personali nell'addestramento

[92] Dati organizzati e formattati in modo predefinito, che li rendono facilmente ricercabili, archiviabili ed elaborabili dai computer. IBM, "Dati strutturati vs. dati non strutturati: qual è la differenza?", sito web IBM, 7 febbraio 2025, 6 agosto 2025, https://www.ibm.com/think/topics/structured-vs-unstructured-data?utm_source=chatgpt.com

[93] Dati che non hanno un modello predefinito o non sono organizzati in modo strutturato (come righe e colonne). IBM, "Dati strutturati vs. dati non strutturati: qual è la differenza?", sito web IBM, 7 febbraio 2025, 6 agosto 2025, https://www.ibm.com/think/topics/structured-vs-unstructured-data?utm_source=chatgpt.com

[94] Dr. Kris Shrishak, "AI: algoritmi complessi ed efficace supervisione della protezione dei dati", sito web dell'EDPB, marzo 2024, https://www.edpb.europa.eu/our-work-tools/our-documents/support-pool-experts-projects/ai-complex-algorithms-and-effective-data_en

set di dati personali. Tali strumenti devono essere progettati per identificare e recuperare in modo sicuro i singoli record di dati senza esporre le informazioni degli altri utenti. I dati personali forniti da questi strumenti dovrebbero essere in un formato leggibile da macchina e di facile utilizzo.

3. Strumenti come MemHunter (strumento automatizzato per rilevare la memorizzazione LLM) potrebbero essere implementati .[95]

5.5.2 Rischio 2: Rettifica o cancellazione incomplete

5.5.2.1 Descrizione

Quando i dati personali sono trattati da sistemi di IA, gli interessati hanno il diritto di richiedere la rettifica dei dati personali errati trattati dal sistema di IA e/o la cancellazione dei propri dati personali dal sistema di IA. Qualora gli interessati ritengano che un sistema di IA contenga dati errati o incompleti che li riguardano, sia nel set di dati personali di addestramento che nell'output del sistema di IA, possono richiedere all'organizzazione di correggerli.

L'esercizio di questi due diritti presenta le stesse difficoltà dell'esercizio del diritto di accesso. I dati non possono essere cancellati o rettificati se non possono essere identificati prima nei set di dati o nel modello. Per i set di dati strutturati, la rettifica o la cancellazione dei dati personali degli interessati nel set di dati personali di addestramento può essere relativamente facile da realizzare se i dati personali di addestramento sono stati conservati (data la natura strutturata dei dati). Per i set di dati non strutturati, la situazione sarà più complessa poiché non esiste uno schema o un formato fisso che possa essere sfruttato.

Correggere l'output del sistema di IA presenta delle sfide, soprattutto se i dati sono già stati incorporati in modelli complessi come le reti neurali profonde o gli LLM. Il diritto alla cancellazione pone complicazioni simili per i sistemi di IA. Garantire che queste richieste di rettifica/cancellazione siano pienamente implementate nei modelli di IA può essere complesso a causa della natura stessa di questi modelli. I modelli di IA spesso apprendono da vasti set di dati e, una volta addestrati, possono conservare modelli o informazioni che possono essere difficili da isolare, rettificare e/o cancellare.

Questo rischio si applica alle seguenti fasi del ciclo di vita del sistema di IA:

- Sviluppo (per lo sviluppo di un sistema di IA)
- Funzionamento e monitoraggio (sia per lo sviluppo che per l'approvvigionamento di un sistema di IA)

5.5.2.2 Possibili misure

Le misure possibili per questo rischio sono:

1. Strumenti di recupero dei dati: analogamente alla sezione 5.6.1.2, gli sviluppatori o i fornitori dovrebbero creare strumenti di rettifica e cancellazione dei dati per poter rettificare o cancellare i dati personali nel set di dati personali di addestramento.

[95] Zhenpeng Wu, Jian Lou, Zibin Zheng, Chuan Chen, *MemHunter: Automated and Verifiable Memorization Detection at Dataset-scale in LLMs,* 10 dicembre 2024, https://arxiv.org/html/2412.07261v1

2. Disimparare la macchina:[96] Il disimparare la macchina è un processo che consente ai modelli di apprendimento automatico di dimenticare in modo selettivo specifici punti di dati precedentemente appresi, consentendo al modello di comportarsi come se non fosse mai stato addestrato su quei dati. Il disimparare automatico può essere classificato in due approcci principali: il disimparare esatto, che comporta il rieducare il modello da zero per rimuovere completamente l'influenza dei dati specificati, e il disimparare approssimativo, che cerca di minimizzare l'impatto dei dati attraverso aggiornamenti limitati dei parametri del modello. Sebbene il disimparare esatto fornisca forti garanzie di rimozione dei dati, spesso è costoso (dal punto di vista computazionale o monetario a causa delle modifiche richieste). Al contrario, l'unlearning approssimativo offre un'alternativa più efficiente, ma potrebbe non eliminare completamente l'influenza dei dati.

3. Quando il disimparare automatico non è fattibile, è possibile ricorrere al filtraggio dell'output. Il filtraggio dell'output comporta la scansione in tempo reale delle risposte del modello di IA per rilevare e bloccare le informazioni personali prima che raggiungano gli utenti. Il sistema potrebbe impiegare il riconoscimento di modelli o il rilevamento di entità denominate per identificare i dati personali e bloccarli prima che raggiungano l'utente.

6 Conclusione

La messa in funzione di sistemi di IA che trattano dati personali comporta rischi significativi per gli interessati, che gli EUI, in qualità di titolari del trattamento, hanno il dovere legale ed etico di identificare, valutare e mitigare. La posta in gioco è alta: i sistemi di IA possono amplificare i rischi per i diritti fondamentali su una scala e a una velocità senza precedenti se non vengono integrate fin dall'inizio adeguate misure di salvaguardia. Per questo motivo, il presente documento ha applicato una metodologia di gestione dei rischi in linea con la norma ISO 31000:2018, contestualizzata ai requisiti specifici dell'EUDPR, per contribuire a garantire che i rischi siano affrontati in modo sistematico e responsabile.

Il capitolo 2 ha introdotto la metodologia di gestione dei rischi come elemento fondamentale per affrontare le sfide relative alla protezione dei dati, fornendo una base strutturata per analizzare le minacce e attuare misure di salvaguardia proporzionate. Il capitolo 3 ha esaminato la definizione e il ciclo di vita dei sistemi di IA, sottolineando che l'approvvigionamento è una fase decisiva in cui è possibile e necessario anticipare i rischi prima che i sistemi siano messi in funzione. Il capitolo 4 ha esaminato cinque principi di protezione dei dati – trasparenza, correttezza, accuratezza, minimizzazione dei dati e sicurezza – e ha analizzato come possono manifestarsi i rischi per garantire il rispetto di tali principi, insieme ad alcuni rischi legati all'effettivo esercizio dei diritti degli interessati. Per ciascun principio, il documento ha individuato scenari di rischio specifici e ha suggerito contromisure tecniche non esaustive per illustrare come i rischi possono essere gestiti nella pratica.

L'analisi non intende presentare un elenco esaustivo di tutti i rischi e delle strategie di mitigazione. Fornisce invece un quadro pratico per aiutare le EUI a sviluppare un approccio sistematico alla gestione dei rischi che potrebbe dover essere integrato alla luce di altri rischi e requisiti di conformità. Si tratta quindi di un quadro che può essere adattato alla diversità dei contesti in cui vengono utilizzati i sistemi di IA. In ultima analisi, i responsabili del trattamento rimangono pienamente responsabili dell'adempimento dei propri obblighi di conformità

[96] GEPD, *Techsonar 2025*, 15 novembre 2024, https://www.edps.europa.eu/data-protection/our-work/publications/reports/2024-11-15-techsonar-report-2025_en

valutazioni, integrando tale quadro con le necessarie analisi giuridiche e garantendo che le loro decisioni siano coerenti con l'EUDPR.

In conclusione, affrontare i rischi sollevati dai sistemi di IA non è un compito secondario, ma un obbligo fondamentale per le EUI. La conformità all'EUDPR, la protezione dei diritti fondamentali e il mantenimento della fiducia pubblica dipendono tutti dalla capacità dei responsabili del trattamento di identificare, valutare e mitigare in modo proattivo i rischi durante l'intero ciclo di vita dell'IA. Ciò richiede più di una valutazione una tantum: richiede una cultura della responsabilità, un monitoraggio continuo e un miglioramento adattivo. Integrando queste pratiche nella loro governance dell'IA, le EUI possono non solo affrontare le complessità dello sviluppo dell'IA e utilizzarla in modo responsabile, ma anche dimostrare la loro leadership nel garantire che l'innovazione sia saldamente ancorata al rispetto dei diritti fondamentali e dei principi di protezione dei dati.

Allegato 1: Parametri di misurazione

Le metriche per valutare le prestazioni dell'IA variano in modo significativo tra i diversi tipi di sistemi di IA, riflettendo la natura diversificata delle applicazioni di IA e i loro obiettivi specifici.[97]

Ad esempio, nelle attività di elaborazione del linguaggio naturale, metriche come BLEU, ROUGE, GLEU e METEOR sono comunemente utilizzate per valutare la qualità del testo o delle traduzioni generate. Queste metriche si concentrano sul confronto tra l'output generato dall'IA e i testi di riferimento, misurando aspetti quali precisione, richiamo e somiglianza semantica.

Al contrario, i compiti di classificazione e identificazione nei modelli di IA si basano spesso su metriche quali accuratezza statistica, precisione, richiamo e punteggio F1. Queste metriche valutano la capacità di un modello di IA di classificare i dati in classi o individui predefiniti, il che è fondamentale per applicazioni quali il rilevamento dello spam o il riconoscimento delle immagini.

Per i problemi di regressione, in cui l'IA prevede valori continui, entrano in gioco metriche diverse, come l'errore assoluto medio (MAE), l'errore quadratico medio (MSE) e l'errore quadratico medio radice (RMSE).

I benchmark forniscono metodi standardizzati per valutare e confrontare diversi modelli di IA [(98)].Offrendo set di dati coerenti, compiti predefiniti e metriche di valutazione, i benchmark consentono ai ricercatori e agli sviluppatori di valutare oggettivamente le prestazioni, l'efficienza e l'accuratezza statistica delle soluzioni di IA. Inoltre, i benchmark svolgono un ruolo fondamentale nell'identificare i potenziali rischi e limiti dei modelli di IA prima della loro implementazione, individuando le aree critiche che devono essere affrontate durante l'intero ciclo di vita dell'utilizzo del sistema di IA.

La tabella seguente fornisce un elenco di alcuni benchmark disponibili al momento della pubblicazione della presente relazione che possono essere utilizzati per valutare i sistemi di IA. L'elenco non intende essere esaustivo, ma offre un buon punto di partenza.

Tipo di IA	Possibile benchmark	Breve descrizione
Elaborazione del linguaggio naturale (NLP) e Grande	Recall-Oriented Understudy per la valutazione sintetica (ROUGE)[100]	Insieme di metriche e pacchetto software progettato per valutare la qualità dei riassunti generati automaticamente e delle traduzioni automatiche nell'elaborazione del linguaggio naturale.
	Valutazione bilingue	Benchmark che valuta la qualità del testo generato automaticamente , in particolare in macchina

[97] OECD.ai, "Catalogo di strumenti e metriche per un'IA affidabile", 6 agosto 2025, sito web OECD.ai, https://oecd.ai/en/catalogue/metrics
[98] Github, "Documenti con codice", sito web Github, 6 agosto 2025, https://paperswithcode.com/sota
[100] Neri Van Otten, "Metrica ROUGE nell'NLP: guida completa e tutorial pratico in Python", 12 agosto 2024, 06 agosto 2025, https://spotintelligence.com/2024/08/12/rouge-metric-in-nlp/

Modelli Modelli (LLM)[99]	Sostituto (BLEU)[101]	compiti di traduzione. Calcola questa somiglianza misurando la precisione degli n-grammi (sequenze di n parole consecutive) che compaiono sia nel testo generato che nei testi di riferimento.
	Comprensione generale del linguaggio (GLUE)[102] e SuperGLUE[103]	Set di dati di riferimento progettati per valutare le prestazioni dei modelli NLP in vari compiti di comprensione del linguaggio. GLUE, introdotto per primo, consiste in nove diversi compiti NLP, tra cui la classificazione delle frasi, l'analisi del sentiment e l'implicazione testuale. SuperGLUE, sviluppato come successore più impegnativo di GLUE, si basa sul suo predecessore introducendo compiti più complessi che richiedono ragionamento avanzato, conoscenza del senso comune e comprensione contestuale. Mentre GLUE si concentra su sfide linguistiche più semplici, SuperGLUE incorpora compiti come la risposta alle domande, la risoluzione delle correferenze([104]) e la comprensione della lettura, spingendo i modelli a dimostrare capacità cognitive di livello superiore
	Valutazione olistica d ei modelli linguistici (HELM)[105]	Struttura di riferimento progettata per valutare le capacità, i limiti e i potenziali rischi dei modelli linguistici in un'ampia gamma di scenari e metriche. La struttura valuta i modelli su 16 scenari principali e 26 scenari mirati, misurando sette metriche chiave: accuratezza statistica, calibrazione, robustezza, equità, distorsione, tossicità ed efficienza.
	Comprensione linguistica multitasking su larga scala	MMLU consiste in circa 16.000 domande a scelta multipla che coprono 57 materie accademiche, tra cui matematica, filosofia, diritto e medicina. Il benchmark

[99] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang e Xing Xie. 2024. *Indagine sulla valutazione dei modelli linguistici di grandi dimensioni,* 29 marzo 2024, https://doi.org/10.1145/3641289

[101] https://spotintelligence.com/2024/08/13/bleu-score-in-nlp/

[102] Wang, A., *Glue: Una piattaforma multi-task di benchmark e analisi per la comprensione del linguaggio naturale*, 22 febbraio 2019, https://arxiv.org/abs/1804.07461

[103] Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., ... & Bowman, S., *Superglue: un benchmark più aderente per i sistemi di comprensione del linguaggio generici. Advances in neural information processing systems,* 13 febbraio 2020, https://arxiv.org/abs/1905.00537

[104] Identificazione e collegamento delle espressioni nel testo che si riferiscono alla stessa entità o evento

[105] Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., ... & Koreeda, Y.. *Valutazione olistica dei modelli linguistici*, 1° ottobre 2023, https://arxiv.org/abs/2211.09110

	(MMLU)[106] e MMLU-Pro[107]	mira a valutare le conoscenze generali e le capacità di risoluzione dei problemi dei modelli di IA, con livelli di difficoltà che vanno da elementare a professionale. MMLU-Pro presenta domande più impegnative, incentrate sul ragionamento, e aumenta il numero di opzioni di scelta da quattro a dieci.
Riconoscimento delle immagini[108]	ImageNet[109]	Database visivo progettato per essere utilizzato nella ricerca sul riconoscimento visivo degli oggetti. Contiene oltre 14 milioni di immagini etichettate che coprono migliaia di categorie di oggetti. Il set di dati è strutturato in 1.000 classi distinte, con circa 1,2 milioni di immagini utilizzate per l'addestramento, 50.000 per la convalida e 100.000 per il test.
	CIFAR-10 e CIFAR-100[110]	CIFAR-10 è composto da 60.000 immagini a colori 32x32 suddivise in 10 classi mutuamente esclusive. Il set di dati è suddiviso in 50.000 immagini di addestramento e 10.000 immagini di test, con ciascuna classe rappresentata in modo equo. CIFAR-100, pur mantenendo lo stesso numero totale di immagini e le stesse dimensioni delle immagini di CIFAR-10, amplia la sfida di classificazione dividendo il set di dati in 100 classi. Queste classi sono ulteriormente raggruppate in 20 superclassi, fornendo un ulteriore livello di categorizzazione. Ogni immagine in CIFAR-100 è associata sia a un'etichetta "fine" (classe specifica) che a un'etichetta "grossolana" (superclasse).

[106] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J., *Measuring massive multitask language understanding*, 12 gennaio 2021, https://arxiv.org/abs/2009.03300

[107] Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., ... & Chen, W., *Mmlu-pro: A more robust and challenging multi-task language understanding benchmark,* 06 novembre 2024, https://arxiv.org/abs/2406.01574v4

[108] Li, L., Chen, G., Shi, H., Xiao, J. e Chen, L. (2024). *Indagine sui benchmark multimodali: nell'era dei grandi modelli di intelligenza artificiale*, 21 settembre 2024, https://arxiv.org/abs/2409.18142
Rangel, Gabriela, Cuevas-Tello, Juan C., Nunez-Varela, Jose, Puente, Cesar, Silva-Trujillo, Alejandra G., *Indagine sulle reti neurali convoluzionali e sui loro limiti prestazionali nei compiti di riconoscimento delle immagini*, 12 luglio 2024, https://doi.org/10.1155/2024/2797320

[109] Image Net, "Imagenet Database", sito web Image Net, 11 marzo 2021, 6 agosto 2025, https://www.image-net.org/

[110] Alex Krizhevsky, "Il dataset CIFAR-10", home page di Alex Krizhevsky, 6 agosto 2025, https://www.cs.toronto.edu/~kriz/cifar.html

	MNIST[111]	Il set di dati è composto da 70.000 immagini in scala di grigi di cifre scritte a mano, ciascuna delle dimensioni di 28x28 pixel. È suddiviso in due sottoinsiemi: un set di addestramento di 60.000 immagini e un set di test di 10.000 immagini.

Categoria	Benchmark	Descrizione	Modelli Testati
Elaborazione del linguaggio naturale (NLP)	**GLUE (Valutazione generale della comprensione del linguaggio)**	Una raccolta di compiti progettati per testare la capacità di comprensione generale del linguaggio dei modelli.	Modelli basati sul testo, modelli linguistici (ad es. BERT, GPT)
	SuperGLUE	Un'estensione di GLUE con compiti più impegnativi.	Modelli NLP avanzati (ad es. T5, RoBERTa)
	SQuAD (Stanford Question Answering Dataset)	Verifica la comprensione della lettura e la capacità di rispondere a domande basate su un passaggio.	Modelli QA (ad esempio, BERT, T5)
	CoNLL-03[112]	Dataset di riconoscimento delle entità denominate (NER) per la valutazione delle prestazioni di riconoscimento delle entità.	Modelli NER (ad es. LSTM, CRF)
	MNLI (Multi-Genre Natural Language Inference)	Valuta la capacità dei modelli di determinare se una premessa implica, contraddice o è neutrale rispetto a un'ipotesi.	Modelli di inferenza (ad es. BERT, RoBERTa, XLNet)
	TREC (Text REtrieval Conference)	Si concentra sui compiti di classificazione delle domande.	Classificatori di testo, riconoscimento dell'intento modelli
Visione artificiale (CV)	**ImageNet**	Benchmark di classificazione di immagini su larga scala con un numero elevato di categorie (1000).	CNN, trasformatori di visione (ad es. ResNet, EfficientNet)

[111] Hojjat Khodabakhsh, "MNIST Dataset", sito web Kaggle, 6 agosto 2025, https://www.kaggle.com/datasets/hojjatk/mnist-dataset

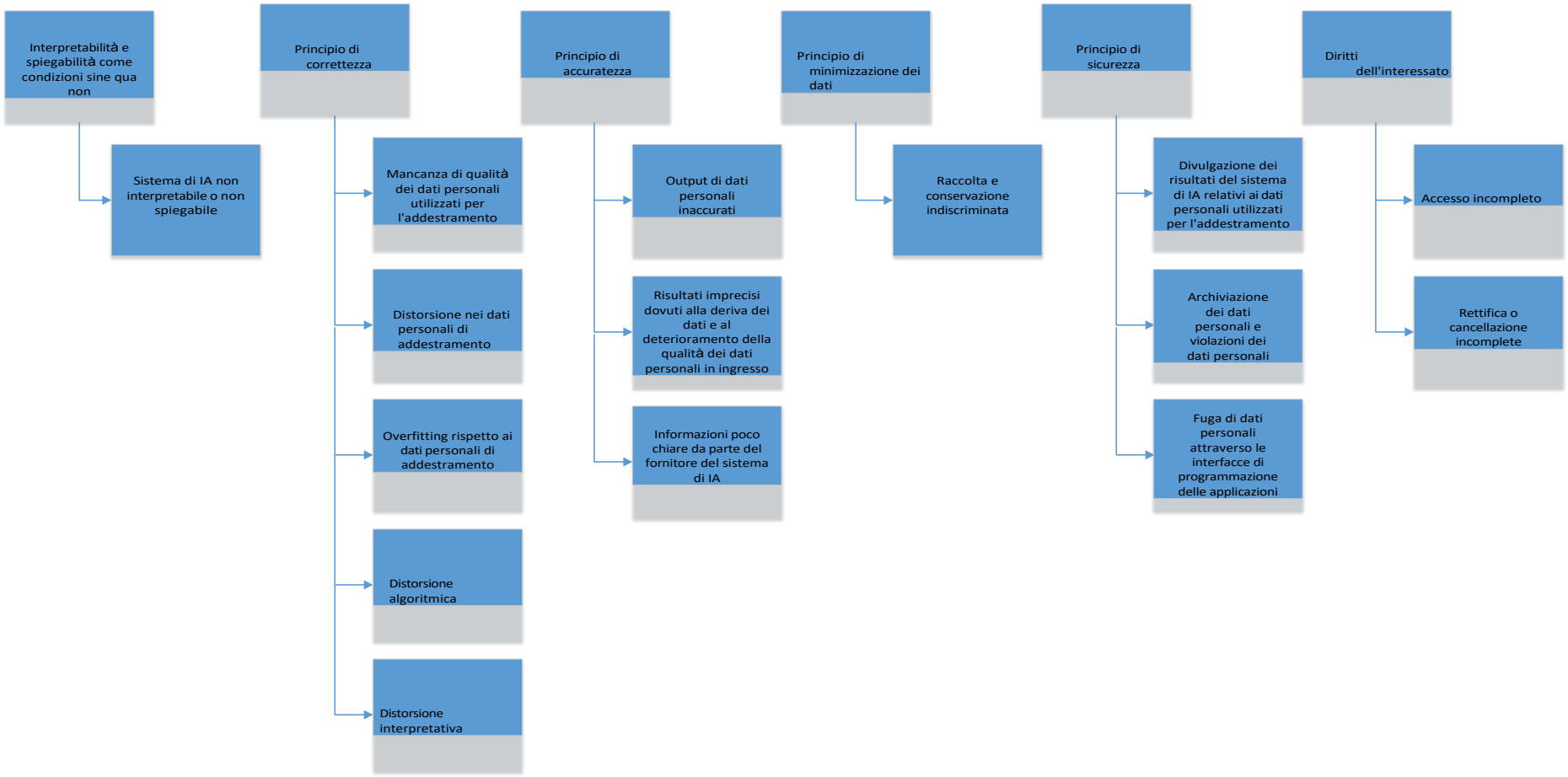
[112] Github, "Papers with code", sito web Github, 6 agosto 2025, https://paperswithcode.com/coll-2023

	COCO (Common Objects in Context)[113]	Un benchmark per attività di rilevamento, segmentazione e didascalia di oggetti.	Modelli di rilevamento e segmentazione degli oggetti (ad esempio, YOLO, Mask R-CNN, Faster R-CNN)
	PASCAL VOC	Si concentra su attività di classificazione delle immagini, rilevamento di oggetti e segmentazione.	Modelli di rilevamento e segmentazione degli oggetti
	ADE20K	Un benchmark di segmentazione semantica per annotazioni dense a livello di pixel.	Modelli di segmentazione (ad es. DeepLabV3+, U-Net)
	KITTI	Incentrato sulla guida autonoma, comprese attività quali corrispondenza stereo, flusso ottico e rilevamento di oggetti .	Rilevamento di oggetti, modelli di flusso ottico (ad es. CNN, RNN)
Discorso e audio	**LibriSpeech**	Benchmark di riconoscimento vocale incentrato sulla trascrizione di audiolibri in inglese.	Modelli di conversione da voce a testo (ad es. DeepSpeech, Wav2Vec)
	VoxCeleb	Riconoscimento e identificazione del parlante in clip audio.	Modelli di riconoscimento del parlante (ad es. ECAPA-TDNN, VGGVox)
	TIMIT	Un corpus per il riconoscimento acustico-fonetico del parlato continuo.	Modelli di riconoscimento vocale
	CHiME	Valuta il riconoscimento vocale in ambienti rumorosi.	Modelli di riconoscimento vocale robusti (ad es. RNN, LSTM)
Apprendimento rinforzato	**OpenAI Gym**	Una piattaforma per lo sviluppo e il confronto di algoritmi di apprendimento rinforzato in un'ampia gamma di ambienti.	Agenti RL (ad es. PPO, DDPG, A2C)

[113] Cocodataset, "Cocodataset", sito web Cocodataset, 6 agosto 2025, https://cocodataset.org

	MuJoCo	Motore fisico utilizzato per simulare ambienti per compiti di controllo continuo.	Agenti RL per il controllo continuo (ad es. DDPG, TRPO)
	Risposte visive alle domande (VQA)	Verifica la capacità di un modello di rispondere a domande in linguaggio naturale relative alle immagini.	Modelli di visione + NLP (ad es. LXMERT, ViLBERT)
AI multimodale	**MS COCO Captioning**	Valuta la generazione di descrizioni in linguaggio naturale delle immagini.	Modelli di sottotitolazione delle immagini (ad esempio, Show and Tell, Image Transformer)
	AI2 Reasoning Challenge (ARC)	Uno standard di riferimento per testare le capacità di ragionamento generale in domande a scelta multipla su un'ampia gamma di argomenti.	Ragionamento generale AI (ad es. GPT, T5)
Prestazioni dell'IA generale	**CLIP (Contrastive Language-Image Pretraining)**	Misura la capacità dei modelli di allineare input di testo e immagini per attività come la classificazione zero-shot.	Modelli di IA multimodali (ad es. CLIP, Flamingo)
	WinoBias	Misura i pregiudizi nei modelli linguistici analizzando le loro risposte ai pronomi di genere.	Modelli NLP (ad es. GPT, BERT)
Equità e pregiudizi	**FairFace**	Valuta i modelli di riconoscimento facciale in termini di equità tra diversi gruppi demografici (ad es. carnagione , età, genere).	Modelli di riconoscimento facciale (ad es. FaceNet, ArcFace)
	MedNLI	Un benchmark per la valutazione dell'inferenza del linguaggio naturale nel settore sanitario.	Modelli NLP per l'assistenza sanitaria (ad es. BioBERT, ClinicalBERT)
AI per l'assistenza sanitaria	**ChestX-ray14**	Utilizzato per valutare modelli di rilevamento di 14 malattie toraciche comuni dalle immagini radiografiche.	Modelli di classificazione delle immagini mediche (ad es. CNN)
	MIMIC-CXR	Un dataset su larga scala di radiografie toraciche per la valutazione dell'accuratezza statistica diagnostica.	Modelli di immagini mediche (ad esempio ResNet, DenseNet)

Allegato 2: Panoramica delle preoccupazioni e dei rischi



Allegato 3: Lista di controllo per ogni fase del ciclo di vita dello sviluppo dell'IA

Sviluppo di un sistema di IA

Fase del ciclo di vita dello sviluppo dell'IA	Principio	Rischio
1. Avvio/Analisi	5.1 Principio di equità	5.1.4 Distorsione algoritmica
2. Acquisizione e preparazione dei dati	5.1 Principio di correttezza	5.1.1 Mancanza di qualità dei dati nei dati personali di addestramento
		5.1.2 Distorsione nei dati personali utilizzati per l'addestramento
		5.1.3 Overfitting dei dati personali utilizzati per l'addestramento
	5.2 Principio di accuratezza	5.2.3 Output di dati personali inesatti
	5.3 Principio di minimizzazione dei dati	5.3.1 Raccolta e conservazione indiscriminata
	5.4 Principio di sicurezza	5.4.2 Conservazione dei dati personali e violazioni dei dati personali
3. Sviluppo	5.5 Diritti dell'interessato	5.5.1 Accesso incompleto
		5.5.2 Rettifica o cancellazione incompleta

4. Verifica e convalida	4 Interpretabilità e spiegabilità	4.1 Sistema di IA non interpretabile o non spiegabile
	5.1 Principio di equità	5.1.1 Mancanza di qualità dei dati nella formazione dei dati personali
		5.1.2 Distorsione nei dati personali di addestramento
		5.1.3 Overfitting nei dati personali utilizzati per l'addestramento
		5.1.4 Distorsione algoritmica
		5.1.5 Distorsione interpretativa
	5.2 Principio di accuratezza	5.2.3 Output di dati personali inesatti
	5.3 Principio di minimizzazione dei dati	5.3.1 Raccolta e conservazione indiscriminata
	5.4 Principio di sicurezza	5.4.2 Conservazione dei dati personali e violazioni dei dati personali
5. Implementazione	Nessuna	Nessuna
6. Funzionamento e monitoraggio	4 Interpretabilità e spiegabilità	4.1 Sistema di IA non interpretabile o non spiegabile
	5.1 Principio di equità	5.1.3 Overfitting rispetto ai dati personali di addestramento
		5.1.5 Distorsione interpretativa

	5.2 Principio di accuratezza	5.2.3 Output di dati personali inaccurati
		5.2.4 Risultati imprecisi dovuti alla deriva dei dati e al deterioramento della qualità dei dati personali inseriti
	5.4 Principio di sicurezza	5.4.1 Divulgazione dei risultati del sistema di IA relativi ai dati personali utilizzati per l'addestramento
		5.4.2 Conservazione dei dati personali e violazioni dei dati personali
		5.4.3 Fuga di dati personali attraverso le interfacce di programmazione delle applicazioni
	5.5 Diritti dell'interessato	5.5.1 Accesso incompleto
		5.5.2 Rettifica o cancellazione incompleta
7. Convalida continua	4 Interpretabilità e spiegabilità	4.1 Sistema di IA non interpretabile o non spiegabile
	5.1 Principio di equità	5.1.3 Overfitting rispetto ai dati personali di addestramento
		5.1.4 Distorsione algoritmica
	5.2 Principio di accuratezza	5.2.4 Risultati inaccurati dovuti alla deriva dei dati e al deterioramento della qualità dei dati personali in ingresso
	5.4 Principio di sicurezza	5.4.1 Divulgazione dei risultati del sistema di IA relativi ai dati personali utilizzati per l'addestramento

		5.4.2 Conservazione dei dati personali e violazioni dei dati personali
8. Rivalutazione	4 Interpretabilità e spiegabilità	4.1 Sistema di IA non interpretabile o non spiegabile
	5.1 Principio di equità	5.1.1 Mancanza di qualità dei dati nella formazione dei dati personali
		5.1.2 Distorsione nei dati personali utilizzati per l'addestramento
		5.1.3 Overfitting dei dati personali utilizzati per l'addestramento
		5.1.4 Distorsione algoritmica
		5.1.5 Distorsione interpretativa
	5.2 Principio di accuratezza	5.2.3 Output di dati personali inaccurati
		5.2.4 Output impreciso dovuto alla deriva dei dati e al deterioramento della qualità dei dati personali in ingresso
	5.3 Principio di minimizzazione dei dati	5.3.1 Raccolta e conservazione indiscriminata
	5.4 Principio di sicurezza	5.4.1 Divulgazione dei risultati del sistema di IA relativi ai dati personali utilizzati per l'addestramento
		5.4.2 Conservazione dei dati personali e violazioni dei dati personali

9. Pensionamento	Nessuna	Nessuno

Acquisizione di un sistema di IA

Fase dello sviluppo del ciclo di vita dell'IA	Preoccupazione	Rischio
1. Preparazione	Nessuna	Nessuna
2. Bando di gara	5.2 Principio di accuratezza	5.2.5 Informazioni poco chiare fornite dal fornitore del sistema di IA
3. Selezione	4 Interpretabilità e spiegabilità	4.1 Sistema di IA non interpretabile o non spiegabile
	5.2 Principio di accuratezza	Informazioni poco chiare fornite dal fornitore del sistema di IA
4. Premio e contratto	Nessuno	Nessuno
5. Esecuzione	Nessuna	Nessuna
6 Verifica e convalida	4 Interpretabilità e spiegabilità	4.1 Sistema di IA non interpretabile o non spiegabile
	5.1 Principio di equità	5.1.1 Mancanza di qualità dei dati nella formazione dei dati personali
		5.1.2 Distorsione nei dati personali utilizzati per l'addestramento
		5.1.3 Overfitting nei dati personali di addestramento
		5.1.4 Distorsione algoritmica

		5.1.5 Distorsione interpretativa
	5.2 Principio di accuratezza	5.2.3 Output di dati personali inesatti
	5.3 Principio di minimizzazione dei dati	5.3.1 Raccolta e conservazione indiscriminata
	5.4 Principio di sicurezza	5.4.2 Conservazione dei dati personali e violazioni dei dati personali
7 Implementazione	Nessuna	Nessuna
8 Funzionamento e monitoraggio	4 Interpretabilità e spiegabilità	4.1 Sistema di IA non interpretabile o non spiegabile
	5.1 Principio di equità	5.1.3 Overfitting rispetto ai dati personali di addestramento
		5.1.5 Distorsione interpretativa
	5.2 Principio di accuratezza	5.2.3 Output di dati personali inaccurati
		5.2.4 Output impreciso dovuto alla deriva dei dati e al deterioramento della qualità dei dati personali in ingresso
	5.4 Principio di sicurezza	5.4.1 Divulgazione dei dati personali di addestramento da parte del sistema di IA
		5.4.2 Conservazione dei dati personali e violazioni dei dati personali

		5.4.3 Fuga di dati personali attraverso le interfacce di programmazione delle applicazioni
	5.5 Diritti dell'interessato	5.5.1 Accesso incompleto
		5.5.2 Rettifica o cancellazione incompleta
9 Convalida continua	4 Interpretabilità e spiegabilità	4.1 Sistema di IA non interpretabile o non spiegabile
	5.1 Principio di equità	5.1.3 Overfitting rispetto ai dati personali di addestramento
		5.1.4 Distorsione algoritmica
		5.2.4 Risultati imprecisi dovuti alla deriva dei dati e al deterioramento della qualità dei dati personali inseriti
	5.4 Principio di sicurezza	5.4.1 Divulgazione dei risultati del sistema di IA relativi ai dati personali utilizzati per l'addestramento
		5.4.2 Conservazione dei dati personali e violazioni dei dati personali
10 Rivalutazione	4 Interpretabilità e spiegabilità	4.1 Sistema di IA non interpretabile o non spiegabile
	5.1 Principio di equità	5.1.1 Mancanza di qualità dei dati nella formazione dei dati personali
		5.1.2 Distorsione nei dati personali utilizzati per l'addestramento

		5.1.3 Overfitting dei dati personali utilizzati per l'addestramento
		5.1.4 Distorsione algoritmica
		5.1.5 Distorsione interpretativa
	5.2 Principio di accuratezza	5.2.3 Output di dati personali inaccurati
		5.2.4 Output impreciso dovuto alla deriva dei dati e al deterioramento della qualità dei dati personali in ingresso
	5.3 Principio di minimizzazione dei dati	5.3.1 Raccolta e conservazione indiscriminata
	5.4 Principio di sicurezza	5.4.1 Divulgazione dei dati personali di addestramento da parte del sistema di IA
		5.4.2 Conservazione dei dati personali e violazioni dei dati personali
11 Pensionamento	Nessuna	Nessuna