La governance dell'AI: novità legislative e applicazioni settoriali

UNIVERSITÀ
DI TORINO

UNIVERSITÀ DI TORINO

Regolamento (UE) 1689/2024

Obiettivi del regolamento

- Garantire un elevato livello di tutela dei diritti fondamentali
- Assicurare la sicurezza dei sistemi di IA
- Favorire l'innovazione e lo sviluppo del mercato interno
- Evitare frammentazioni normative tra Stati membri

UNIVERSITÀ DI TORINO

Ambito di applicazione

Si applica a:

- Fornitori di sistemi di IA
- Utilizzatori (deployers)
- Importatori e distributori

Rileva anche per operatori extra-UE se i sistemi producono effetti nell'UE

UNIVERSITÀ DI TORINO

Approccio basato sul rischio

- Il regolamento adotta una classificazione dei sistemi di IA in base al rischio:

1. **Rischio inaccettabile** (vietati)
2. **Alto rischio**
3. **Rischio limitato**
4. **Rischio minimo**

UNIVERSITÀ DI TORINO

Entrata in vigore e applicazione

- **Entrata in vigore:** 1° agosto 2024
- **Applicazione graduale (phased approach):**
  - 2 febbraio 2025 → divieti e disposizioni generali
  - 2 agosto 2025 → governance e modelli di IA general purpose
  - 2 agosto 2026 → maggior parte degli obblighi (es. sistemi ad alto rischio)
  - 2 agosto 2027 → piena applicazione completa

UNIVERSITÀ DI TORINO

Pratiche vietate

Sono vietati i sistemi che comportano rischi inaccettabili, tra cui:

- Manipolazione subliminale
- Sfruttamento di vulnerabilità
- Social scoring pubblico
- Identificazione biometrica remota in tempo reale (con eccezioni limitate)

UNIVERSITÀ DI TORINO

Sistemi ad alto rischio

Requisiti:

- Sistema di gestione del rischio
- Qualità dei dati
- Documentazione tecnica
- Trasparenza e tracciabilità
- Supervisione umana
- Robustezza, accuratezza e sicurezza

UNIVERSITÀ DI TORINO

Dati per l'intelligenza artificiale

Comunicazione della Commissione, «L'intelligenza artificiale per l'Europa», 2018:

- Sono necessari ingenti volumi di dati per sviluppare l'IA. L'apprendimento automatico, un tipo di IA, opera mediante l'individuazione di modelli a partire dai dati disponibili e la successiva applicazione di questa conoscenza ai dati nuovi35. Quanto più è grande il set di dati, tanto più accurata sarà l'individuazione delle relazioni anche impercettibili tra i dati. Quando si tratta di utilizzare l'IA, gli ambienti ad alto contenuto di dati offrono anche le maggiori opportunità, perché i dati sono il mezzo attraverso il quale l'algoritmo apprende e interagisce con il suo ambiente. Per esempio, se tutte le macchine e i processi in uno stabilimento producono continuamente dati, è probabile che con l'aiuto dell'IA si possano realizzare ulteriore automazione e ottimizzazione. In un contesto analogico, per esempio in un lavoro basato su documenti cartacei senza dati digitalizzati sulle operazioni in corso, tale automazione non è possibile. Alla luce di quanto sopra, l'accesso ai dati è un elemento fondamentale per un'IA competitiva, che l'UE dovrebbe agevolare.

UNIVERSITÀ DI TORINO

Dati per l'intelligenza artificiale

European Parliament, Resolution of 3 May 2022 on artificial intelligence in a digital age

- Recalls the essential link between the availability of high-quality data and the development of AI applications
- Stresses the key importance of opening data silos and fostering access to data for AI researchers and companies as outlined in Parliament's resolution on the European data strategy
- The areas for action set out in this White Paper are complementary to the plan presented in parallel under the European data strategy. Improving access to and the management of data is fundamental. Without data, the development of AI and other digital applications is not possible. The enormous volume of new data yet to be generated constitutes an opportunity for Europe to position itself at the forefront of the data and AI transformation. Promoting responsible data management practices and compliance of data with the FAIR principles will contribute to build trust and ensure re-usability of data. Equally important is investment in key computing technologies and infrastructures.

UNIVERSITÀ DI TORINO

Orientamenti

- EDPB, Report of the work undertaken by the ChatGPT Taskforce, 24 Maggio 2024
- EDPB, Parere 28/2024 su taluni aspetti relativi alla protezione dei dati ai fini del trattamento dei dati personali nel contesto dei modelli di IA

UNIVERSITÀ DI TORINO

Trattamento dati per l'addestramento di sistemi AI

- Il legittimo interesse può costituire una base giuridica appropriata
- Per categorie particolari di dati?

UNIVERSITÀ DI TORINO

Trattamento dati per l'addestramento di sistemi AI

Art. 8, l. 132/2025

1. I trattamenti di dati, anche personali, eseguiti da soggetti pubblici e privati senza scopo di lucro, dagli Istituti di ricovero e cura a carattere scientifico (IRCCS), […] nonché da soggetti privati operanti nel settore sanitario nell'ambito di progetti di ricerca a cui partecipano soggetti pubblici e privati senza scopo di lucro o IRCCS, **per la ricerca e la sperimentazione scientifica nella realizzazione di sistemi di intelligenza artificiale per finalità di prevenzione, diagnosi e cura di malattie, sviluppo di farmaci, terapie e tecnologie riabilitative, realizzazione di apparati medicali, incluse protesi e interfacce fra il corpo e strumenti di sostegno alle condizioni del paziente, salute pubblica, incolumità della persona, salute e sicurezza sanitaria nonché studio della fisiologia, della biomeccanica e della biologia umana anche in ambito non sanitario**, in quanto necessari ai fini della realizzazione e dell'utilizzazione di banche di dati e modelli di base, sono dichiarati di rilevante interesse pubblico […].

2. Ai medesimi fini, fermo restando l'obbligo di informativa in favore dell'interessato, che può essere assolto anche mediante un'informativa generale messa a disposizione nel sito web del titolare del trattamento e senza ulteriore consenso dell'interessato ove inizialmente previsto dalla legge, **è sempre autorizzato l'uso secondario di dati personali privi degli elementi identificativi diretti**, anche appartenenti alle categorie [particolari], da parte dei soggetti di cui al comma 1, salvi i casi nei quali la conoscenza dell'identità degli interessati sia inevitabile o necessaria al fine della tutela della loro salute.

3. Negli ambiti di cui al comma 1 o per le finalità di [programmazione, gestione, controllo e valutazione dell'assistenza sanitaria], è sempre consentito, previa informativa all'interessato […], il **trattamento per finalità di anonimizzazione, pseudonimizzazione o sintetizzazione dei dati personali, anche appartenenti alle categorie particolari** […]. È consentito altresì il predetto trattamento finalizzato allo studio e alla ricerca sui gesti atletici, sui movimenti e sulle prestazioni nell'attività sportiva in tutte le sue forme, nel rispetto dei principi generali di cui alla presente legge e dei diritti di sfruttamento economico dei dati relativi alle attività agonistiche che spettano a chi le organizza.

UNIVERSITÀ DI TORINO

Trattamento dati per l'addestramento di sistemi AI

Legittimo interesse

1) Identificare il legittimo interesse

- Interesse è lecito
- Interesse è articolato in modo chiaro e preciso
- Interesse è reale e presente (ossia non speculativo)

«Tale interesse può riguardare, ad esempio, nell'ambito dello sviluppo di un modello di IA – la realizzazione di un servizio di agente conversazionale per l'assistenza agli utenti o, in fase di diffusione – il miglioramento del rilevamento delle minacce in un sistema informatico**»**

UNIVERSITÀ DI TORINO

Trattamento dati per l'addestramento di sistemi AI

Legittimo interesse

2) Test di necessità: analisi della necessità del trattamento ai fini dell'interesse legittimo (o degli interessi legittimi) perseguito/i

- Valutare se l'attività di trattamento consente il perseguimento dell'interesse legittimo e se non esiste un modo meno invasivo per perseguire tale interesse.
- Prestare particolare attenzione alla quantità di dati personali trattati e se questi siano proporzionati

UNIVERSITÀ DI TORINO

Trattamento dati per l'addestramento di sistemi AI

Legittimo interesse

3) Test di bilanciamento: verifica che sull'interesse legittimo (o interessi legittimi) non prevalgano gli interessi o i diritti e le libertà fondamentali degli interessati

- Per valutare tale l'impatto, prendere in considerazione la natura dei dati trattati dai modelli, il contesto del trattamento e le possibili ulteriori conseguenze del trattamento
- Ruolo delle ragionevoli aspettative degli interessati: tra gli elementi da considerare possono rientrare sia informazioni fornite agli interessati che il contesto del trattamento. Per quanto riguarda il contesto, questo può includere: il fatto che i dati personali siano pubblicamente disponibili, la natura del rapporto tra l'interessato e il titolare del trattamento (e se esista un collegamento tra i due), la natura del servizio, il contesto in cui i dati personali sono stati raccolti, la fonte da cui i dati sono stati raccolti (ossia il sito web o il servizio in cui i dati personali sono stati raccolti e le impostazioni di privacy previste), i potenziali ulteriori utilizzi del modello e se gli interessati siano effettivamente consapevoli del fatto che i loro dati personali sono disponibili online
- Ruolo delle misure di attenuazione per limitare l'impatto del trattamento su tali interessati (diverse da quelle obbligatorie per legge)

UNIVERSITÀ DI TORINO

Trattamento dati per l'addestramento di sistemi AI

Misure di attenuazione nella fase di sviluppo

- Misure tecniche: selezionare le fonti dei dati, preparare e minimizzare i dati, misure per ridurre l'identificabilità
- Misure che facilitano l'esercizio di diritti individuali: osservare un periodo di tempo ragionevole tra la raccolta di un insieme di addestramento e il suo utilizzo, consentire agli interessati di opporsi, consentire agli interessati di presentare segnalazioni
- Misure di trasparenza: campagne mediatiche per informare gli interessati

UNIVERSITÀ DI TORINO

Trattamento dati per l'addestramento di sistemi AI

Dati di input

- Raccolti da fornitore o deployer
- Inseriti dall'utente: «Rather, if ChatGPT is made available to the public, it should be assumed that individuals will sooner or later input personal data. If those inputs then become part of the data model and, for example, are shared with anyone asking a specific question, OpenAI remains responsible for complying with the GDPR and should not argue that the input of certain personal data was prohibited in first place»

UNIVERSITÀ DI TORINO

Trattamento illecito di dati nella fase di sviluppo

Quali sono le conseguenze del trattamento illecito di dati personali nella fase di sviluppo di un modello di IA sul successivo trattamento o funzionamento del modello stesso?

1. I dati personali sono conservati nel modello di IA (quindi il modello non può essere considerato anonimo) e sono successivamente trattati dallo stesso titolare del trattamento (ad esempio nel contesto della diffusione del modello)

- Occorre valutare caso per caso, a seconda del contesto di riferimento, se le fasi di sviluppo e diffusione comportino finalità distinte (costituendo, così, attività di trattamento separate) e la misura in cui l'assenza di una base giuridica per l'attività di trattamento iniziale incida sulla liceità del trattamento successivo
- Quando il trattamento successivo è basato su un interesse legittimo, il fatto che il trattamento iniziale fosse illecito dovrebbe essere preso in considerazione nella valutazione dell'interesse legittimo (ad esempio, in relazione ai rischi per gli interessati o al fatto che gli interessati potrebbero non aspettarsi tale trattamento successivo)

UNIVERSITÀ DI TORINO

Trattamento illecito di dati nella fase di sviluppo

Quali sono le conseguenze del trattamento illecito di dati personali nella fase di sviluppo di un modello di IA sul successivo trattamento o funzionamento del modello stesso?

2. I dati personali sono conservati nel modello e sono trattati da un altro titolare del trattamento nella fase di diffusione del modello

- Valutare se il titolare del trattamento che utilizza il modello abbia condotto un'adeguata valutazione, nell'ambito dei suoi obblighi di responsabilità, per accertare che il modello di IA non sia stato sviluppato mediante il trattamento illecito di dati personali
- Chi usa il modello dovrebbe ad es. tenere conto della fonte dei dati personali, del fatto che il trattamento nella fase di sviluppo sia stato oggetto di una constatazione di violazione, soprattutto se sia stata accertata da un'autorità di controllo o giurisdizionale, e il suo accertamento dovrebbe essere più o meno dettagliato a seconda dei rischi derivanti dal trattamento nella fase di diffusione

UNIVERSITÀ DI TORINO

Trattamento illecito di dati nella fase di sviluppo

Quali sono le conseguenze del trattamento illecito di dati personali nella fase di sviluppo di un modello di IA sul successivo trattamento o funzionamento del modello stesso?

3. Un titolare del trattamento tratta illecitamente i dati personali per sviluppare il modello di IA, quindi si assicura che sia anonimizzato, prima che lo stesso, o un altro titolare del trattamento, avvii un altro trattamento di dati personali in fase di diffusione

- Qualora possa essere dimostrato che il successivo funzionamento del modello di IA non comporta il trattamento di dati personali, il GDPR non troverebbe applicazione e quindi l'illiceità del trattamento iniziale non dovrebbe incidere sul successivo funzionamento del modello
- Se, successivamente, i titolari del trattamento effettuano il trattamento dei dati personali raccolti durante la fase di diffusione, dopo l'anonimizzazione del modello, si tratta di un trattamento autonomo

UNIVERSITÀ DI TORINO

Sistemi AI e diffusione di dati

Premessa

- alcuni modelli di IA sono **specificamente progettati per fornire dati personali relativi alle persone i cui dati personali sono stati utilizzati per addestrare il modello**, o in qualche modo, per rendere disponibili tali dati (ad es., un modello generativo affinato attraverso le registrazioni vocali di una persona per imitare la sua voce; o un qualsiasi modello progettato per rispondere attingendo ai dati personali usati per l'addestramento quando vengono richieste informazioni su una persona specifica): questi modelli NON sono anonimi
- modelli di IA, indipendentemente dal fatto che siano addestrati o meno con dati personali, sono solitamente progettati per **fare previsioni o trarre conclusioni**
- modelli di IA addestrati con dati personali sono spesso progettati per fare **inferenze su individui diversi** da quelli i cui dati personali sono stati utilizzati per addestrare il modello di IA

UNIVERSITÀ DI TORINO

Sistemi AI e diffusione di dati

Un modello di IA addestrato usando dati personali può essere considerato anonimo?

- Valutare la probabilità di estrazione diretta (anche probabilistica) di dati personali relativi a persone i cui dati personali sono stati utilizzati per sviluppare il modello
- Valutare la probabilità di ottenere, intenzionalmente o meno, tali dati personali dalle interrogazioni, che dovrebbero essere insignificanti, considerando «tutti i mezzi di cui può ragionevolmente avvalersi»

Dare atto di misure adottate e test effettuati in documentazione tecnica (principio di responsabilizzazione)

Qualità dei dati

Art. 10, regolamento 2024/1689 (AIA)

- Per sistemi AI ad alto rischio, i set di dati di addestramento, convalida e prova devono soddisfare criteri di qualità

UNIVERSITÀ DI TORINO

Qualità dei dati

Art. 10, AIA

Pratiche di governance e gestione dei dati riguardanti

a) le scelte progettuali pertinenti;

b) i processi di raccolta dei dati e l'origine dei dati, nonché la finalità originaria della raccolta nel caso di dati personali;

c) le operazioni di trattamento pertinenti ai fini della preparazione dei dati, quali annotazione, etichettatura, pulizia, aggiornamento, arricchimento e aggregazione;

d) la formulazione di ipotesi, in particolare per quanto riguarda le informazioni che si presume che i dati misurino e rappresentino;

e) una valutazione della disponibilità, della quantità e dell'adeguatezza dei set di dati necessari;

f) un esame atto a valutare le possibili distorsioni suscettibili di incidere sulla salute e sulla sicurezza delle persone, di avere un impatto negativo sui diritti fondamentali o di comportare discriminazioni vietate dal diritto dell'Unione, specie laddove gli output di dati influenzano gli input per operazioni future;

g) le misure adeguate per individuare, prevenire e attenuare le possibili distorsioni individuate conformemente alla lettera f;

h) l'individuazione di lacune o carenze pertinenti nei dati tali da pregiudicare il rispetto del presente regolamento e il modo in cui tali lacune e carenze possono essere colmate.

UNIVERSITÀ DI TORINO

Qualità dei dati

Art. 10, AIA

- I set di dati di addestramento, convalida e prova sono **pertinenti, sufficientemente rappresentativi** e, nella misura del **possibile, esenti da errori e completi** nell'ottica della finalità prevista. Essi possiedono le proprietà statistiche appropriate anche, ove applicabile, per quanto riguarda le persone o i gruppi di persone relativamente ai quali il sistema di IA ad alto rischio è destinato a essere usato

- I set di dati tengono conto, nella misura necessaria per la finalità prevista, delle **caratteristiche o degli elementi particolari dello specifico ambito** geografico, contestuale, comportamentale o funzionale all'interno del quale il sistema di IA ad alto rischio è destinato a essere usato

UNIVERSITÀ DI TORINO

Qualità dei dati

Trattamento di dati per assicurare la qualità dei dati

- Art. 10, par. 5, AIA: «Nella misura in cui ciò sia strettamente necessario al fine di garantire il rilevamento e la correzione delle distorsioni in relazione ai sistemi di IA ad alto rischio […] i fornitori di tali sistemi possono eccezionalmente trattare categorie particolari di dati personali, fatte salve le tutele adeguate per i diritti e le libertà fondamentali delle persone fisiche» (minimizzazione, finalità e sicurezza)

- Proposta di regolamento Omnibus digitale sull'IA: estende la possibilità agli sviluppatori e deployer di qualunque modello IA (v. EDPB-EDPS joint opinion 1/2026 on the Digital Omnibus on AI proposal)

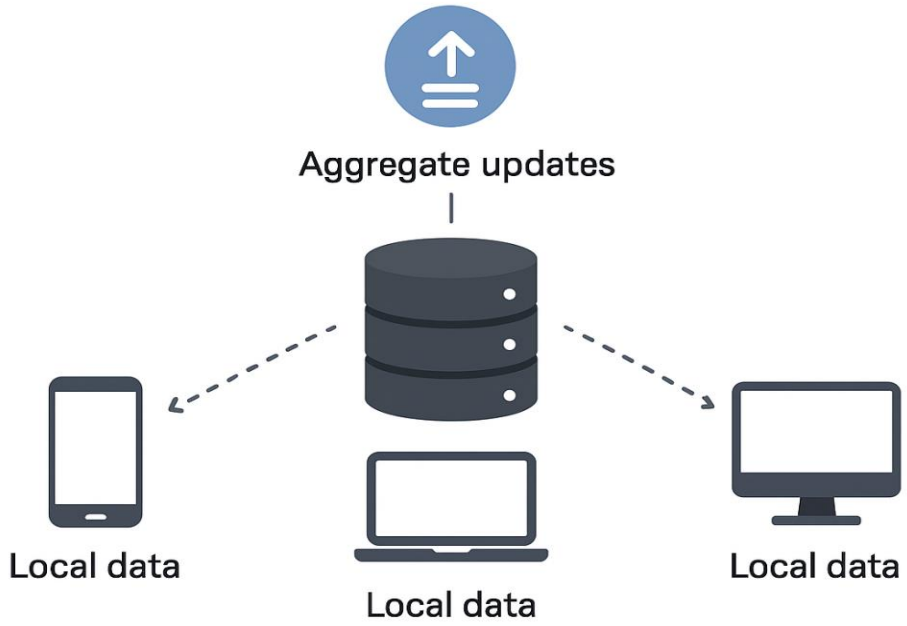
UNIVERSITÀ DI TORINO

Federated learning

- **È** un paradigma di **apprendimento automatico distribuito** in cui un modello di intelligenza artificiale viene addestrato **collaborativamente** d a più soggetti (ad esempio ospedali, banche, dispositivi mobili), **senza che i dati personali lascino la loro sede di origine**.

UNIVERSITÀ DI TORINO

FEDERATED LEARNING



UNIVERSITÀ DI TORINO

Federated learning e compliance

- Privacy by design
- Maggiore sicurezza
- Base giuridica?

UNIVERSITÀ DI TORINO

Il problema degli attacchi

- Attacchi sperimentali che ricostruiscono dati di addestramento a partire da gradienti/aggiornamenti (es. Deep Leakage from Gradients, Inverting Gradients, GradInversion).
- Questi attacchi sono efficaci in molti setting di laboratorio.

UNIVERSITÀ DI TORINO

Original 

Extracted 

Original ive read a few of the reviews and im kinda sad that a lot of the story seems [UNK] ...

Extracted ive read a few of the reviews and im kinda sad that a lot of the story seems [UNK] ...

When the Curious Abandon Honesty: Federated Learning Is Not Private◇

Franziska Boenisch*, Adam Dziedzic*†§, Roei Schuster*§, Ali Shahin Shamsabadi*‡§, Ilia Shumailov*§, and Nicolas Papernot*†
*Vector Institute †University of Toronto ‡The Alan Turing Institute

UNIVERSITÀ DI TORINO

Federated learning

- Non esiste (oggi) una garanzia teorica che impedisca la comparsa di nuovi metodi d'attacco: il rischio è basso ma **non nullo.**

UNIVERSITÀ DI TORINO

Possibili soluzioni?

- Valutazione d'impatto della protezione dei dati (DPIA)
- Convenzione tra i partecipanti al federated learning
- Consultazione preventiva al Garante della Privacy (art. 36 GDPR)

UNIVERSITÀ DI TORINO

Nuove prospettive normative

- EHDS
- Legge italiana sull'IA

UNIVERSITÀ DI TORINO

Nuove modifiche all'orizzonte?

- Digital Omnibus
- Digital Omnibus on AI

UNIVERSITÀ
DI TORINO