



*Bias - Presidi tecnico-operativi a supporto della fiducia
dei sistemi di AI*

AI Legal 2026



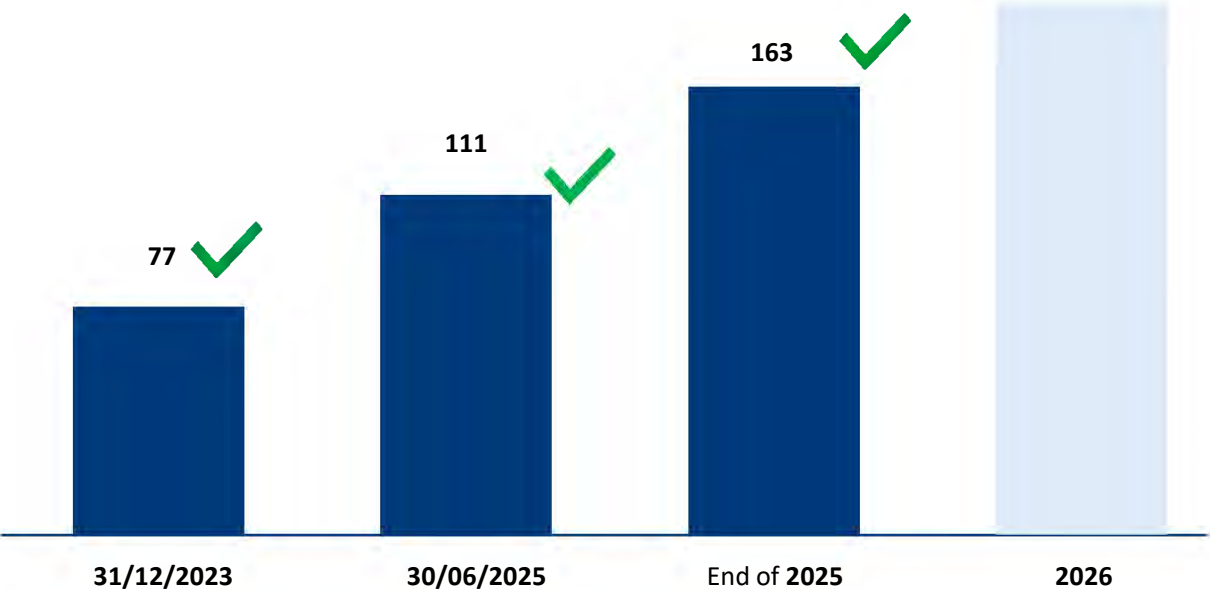
Roma, 8 aprile 2026

Alessandro Castelnovo

Nel 2024–2025, Intesa Sanpaolo ha raggiunto risultati significativi, con oltre 100 casi d'uso di intelligenza artificiale già implementati con successo in produzione

NUMERO DI AI/ GENAI USE CASE SVILUPPATI

Aggiornamento al 06/03/2026



38

Generative AI
use cases

+10

Iniziative
Agentic AI

€100M

Valore generato

IMPATTO DELL'IA NELLA BANCA

+250

Esperti di AI/ GenAI
(Data&AI Office)

+15

Aree di business bancario e di
governance coinvolte

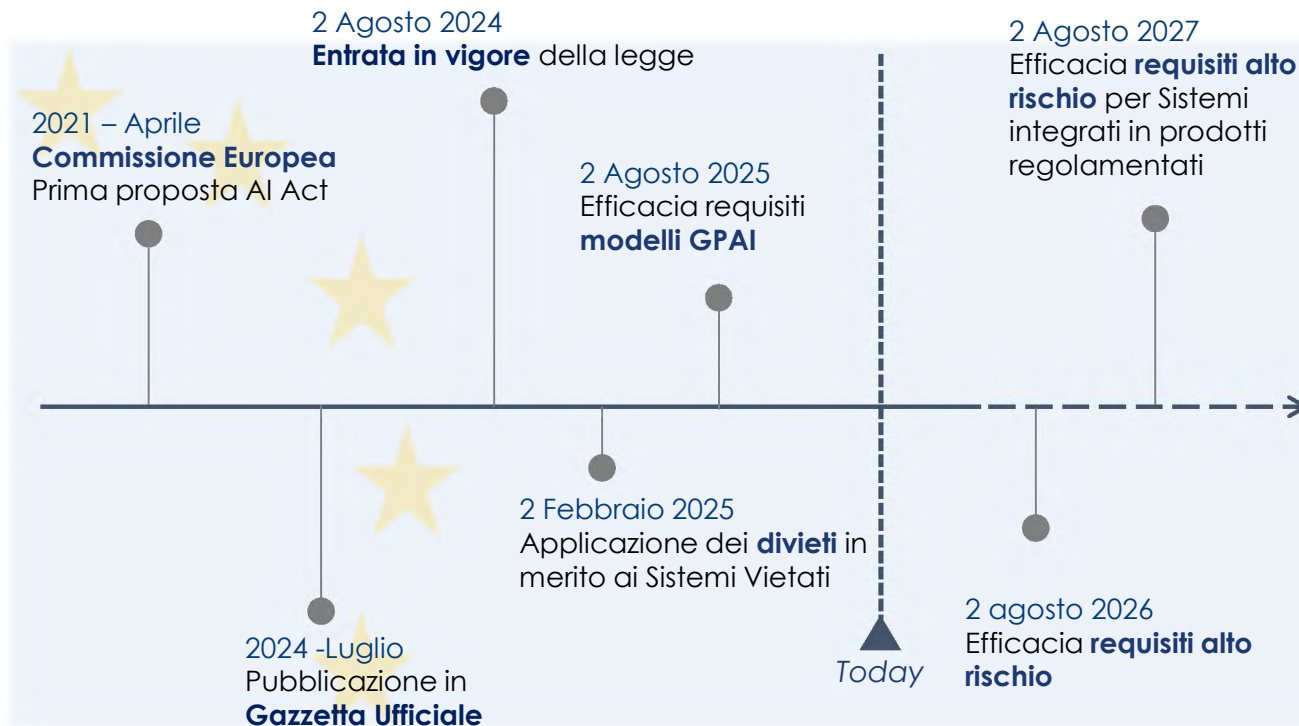
26

Use case classificati ad alto rischio

L'AI Act è stato introdotto per regolamentare l'uso dei Sistemi di I.A., garantendo un approccio antropocentrico che tuteli i diritti fondamentali

L'AI Act è ufficialmente in vigore dal **2 agosto 2024** e definisce obblighi specifici per Fornitori e Deployer

TIMELINE DI ADOZIONE AI ACT



Prossime scadenze:

- Ad Agosto 2026 saranno efficaci i requisiti per i Sistemi di I.A. ad alto rischio.

PRINCIPI CHIAVE

I Risk based



L'AI Act, per ogni Sistema di I.A. **prevede requisiti ed obblighi graduati in funzione del sottostante rischio per la salute, la sicurezza e la protezione dei diritti fondamentali della persona**, anche tenuto conto dei costi e benefici associati

II Principle based



L'AI Act prevede che le **misure di presidio siano declinate caso per caso nell'operatività concreta**

III Conformity by design



L'AI Act impone una **valutazione preliminare dei rischi e delle relative misure di mitigazione** secondo un approccio di conformity by design nonché l'esecuzione di un **self-assessment** dei Sistemi di I.A. volto ad applicare i requisiti in accordo con le previsioni normative e di auto regolamentazione

Per una piena applicazione dei presidi di Responsible AI, è stato creato un centro di eccellenza dedicato che svolge diverse attività



SUSTAINABILITY AI

Sviluppo di iniziative mirate a integrare l'AI in **ambiti di sostenibilità**, con il lancio di soluzioni **innovative** a supporto degli **obiettivi ESG**

Environmental



MITIGAZIONE RISCHI

Definizione di strumenti che facilitino lo sviluppo di sistemi di I.A. conformi all'**AI Act** per mitigare i rischi che **incidono sui diritti fondamentali** e, più in generale, quelli connessi all'utilizzo della **GenAI**

Social



GOVERNANCE E NORMATIVA INTERNA

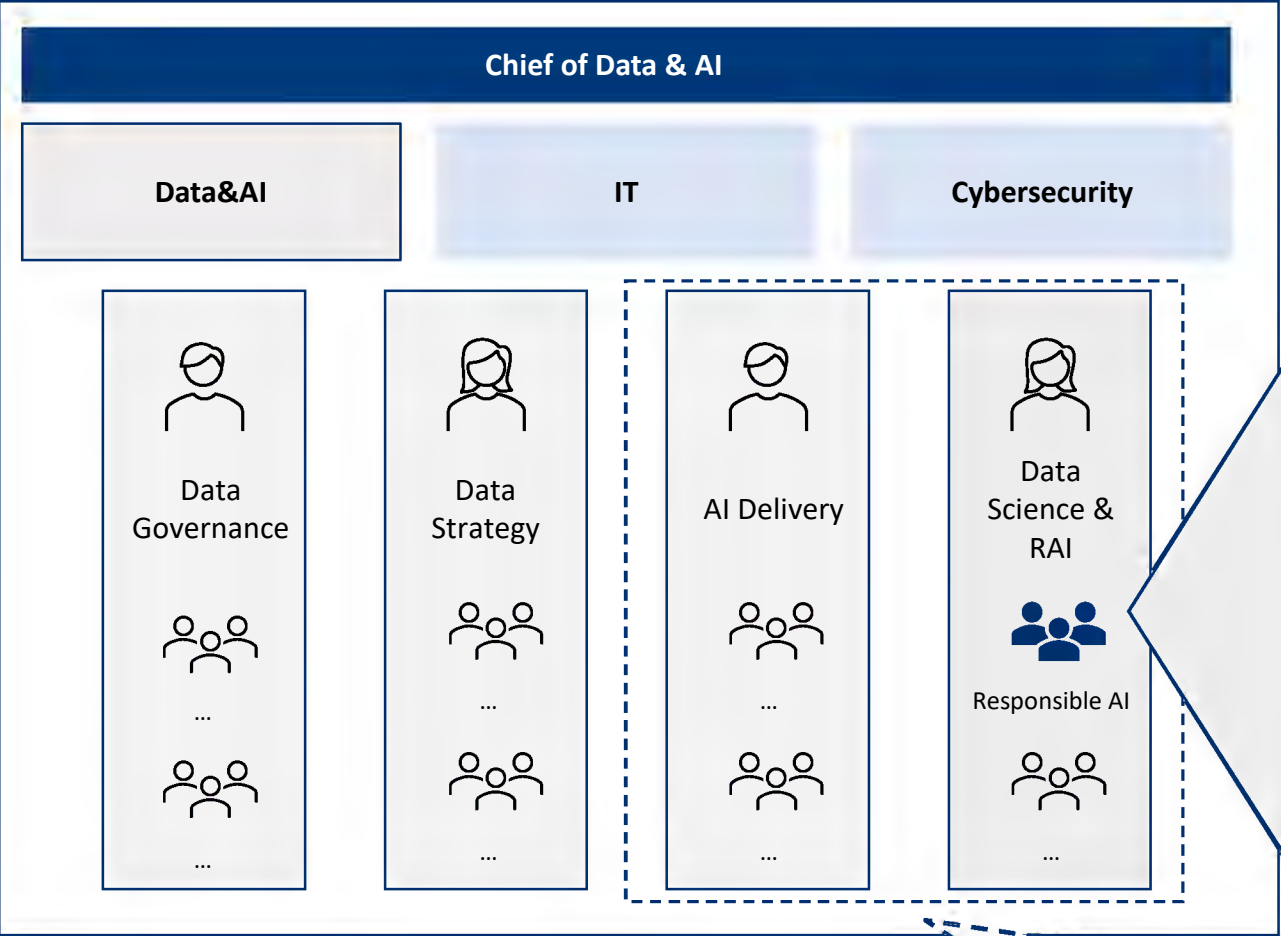
Definizione di **regole e processi per lo sviluppo** e l'adozione dei sistemi di intelligenza artificiale, accompagnate da meccanismi di **monitoraggio continuo**

Governance



Intesa Sanpaolo ha creato un centro di eccellenza dedicato all'intelligenza artificiale per gestire lo sviluppo delle soluzioni AI all'interno della banca

VISTA ORGANIZZATIVA



Più di 200 professionisti Lavoro con Intelligenza Artificiale

OBIETTIVI DEL COE DI RESPONSIBLE AI

- Definizione di **regole e standard tecnologici**
- Governance della Responsible AI**
- Sviluppo e condivisione del **Know-how in ambito AI**

19



**RESPONSIBLE AI
EXPERT**

9

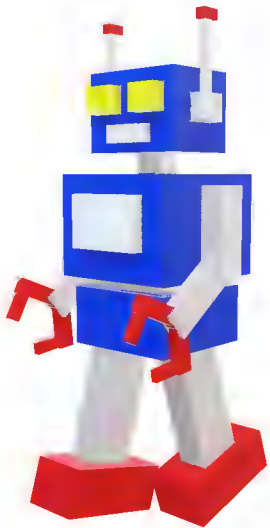


**DATA & AI ETHICS
SPECIALIST**

5



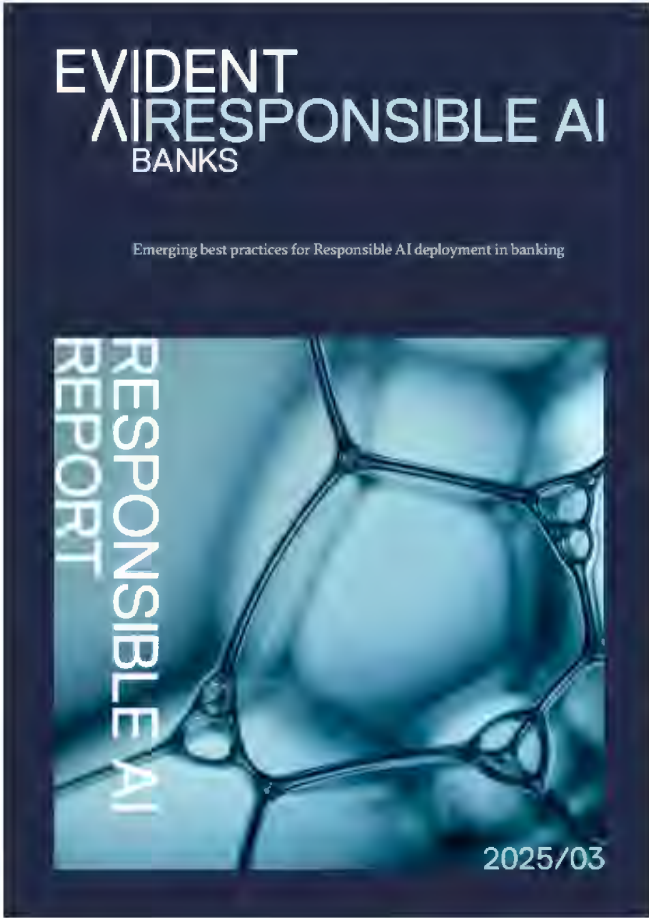
**AI GOVERNANCE &
PROCESSES**



EVIDENT AI certifica che Intesa Sanpaolo è tra le principali banche a livello mondiale per contributi sulla Responsible AI (analisi relativa al periodo 2022-2024)

INDICATORI CHIAVE	RANK	AZIENDE
3° <i>Intesa Sanpaolo ha raggiunto la terza posizione (a pari merito, marzo 2025) al mondo per pubblicazioni su temi di IA responsabile</i>	1	 JPMorganChase
	2	 Capital one
47% <i>Le pubblicazioni del Gruppo su Responsible AI rappresentano quasi la metà del totale dei contributi</i>	3	 Royal Bank of Canada
	3	 Intesa Sanpaolo
	5	 Wells Fargo
	6	 TD Bank
	7	 Standard Chartered
	8	 CommBank
	9	 UniCredit
	10	 Groupe BPCE

Fonte: Evident AI Index, Marzo 2025



Agenda

- 01**  **Fairness nei modelli di Machine Learning**
- 02**  **Metriche di Fairness e incompatibilità**
- 03**  **Razionali alla base della scelta della metrica di Fairness**
- 04**  **Modelli di mitigazione dei bias**
- 05**  **Guardrail**



Southern State Parkway bridges – Robert Moses



100 humans - Netflix

Applicazioni di AI di successo



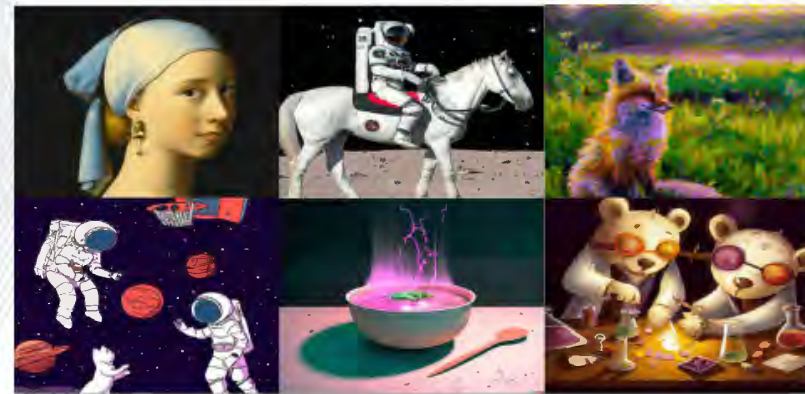
Navigazione



Relazioni



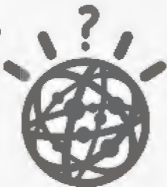
Automazione



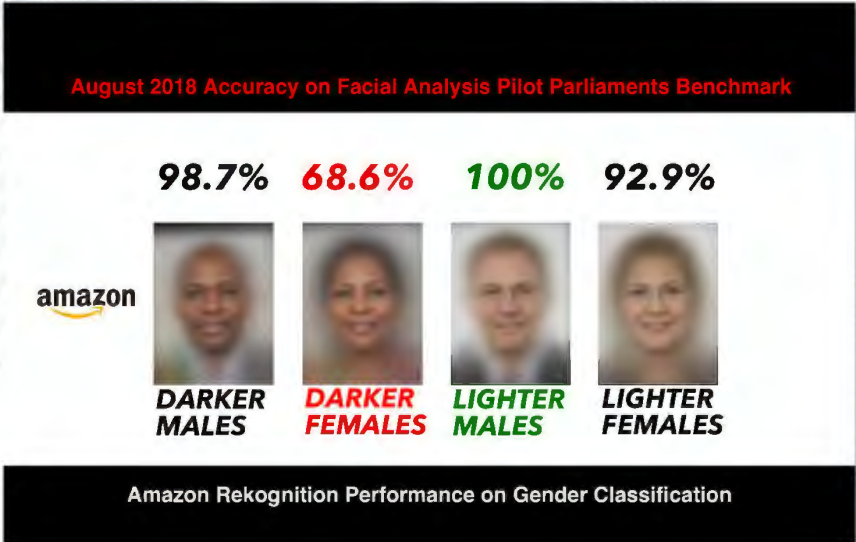
Creazione di contenuti

Soluzioni di AI fallite

IBM Watson Flops For
Cancer Treatment:
Why Did AI Fail?



IBM's **Watson** for Oncology è stato dismesso



Il **sistema di riconoscimento facciale** di **Amazon** funziona meglio per uomini bianchi

The dark side of Google Ads

AdFisher: tool to automate the creation of behavioral and demographic profiles.

Used to demonstrate that setting gender = female results in less ads for high-paying jobs.



Lo strumento **AdFisher** di **Google** ha mostrato significativamente meno annunci di lavori ben retribuiti alle donne rispetto agli uomini

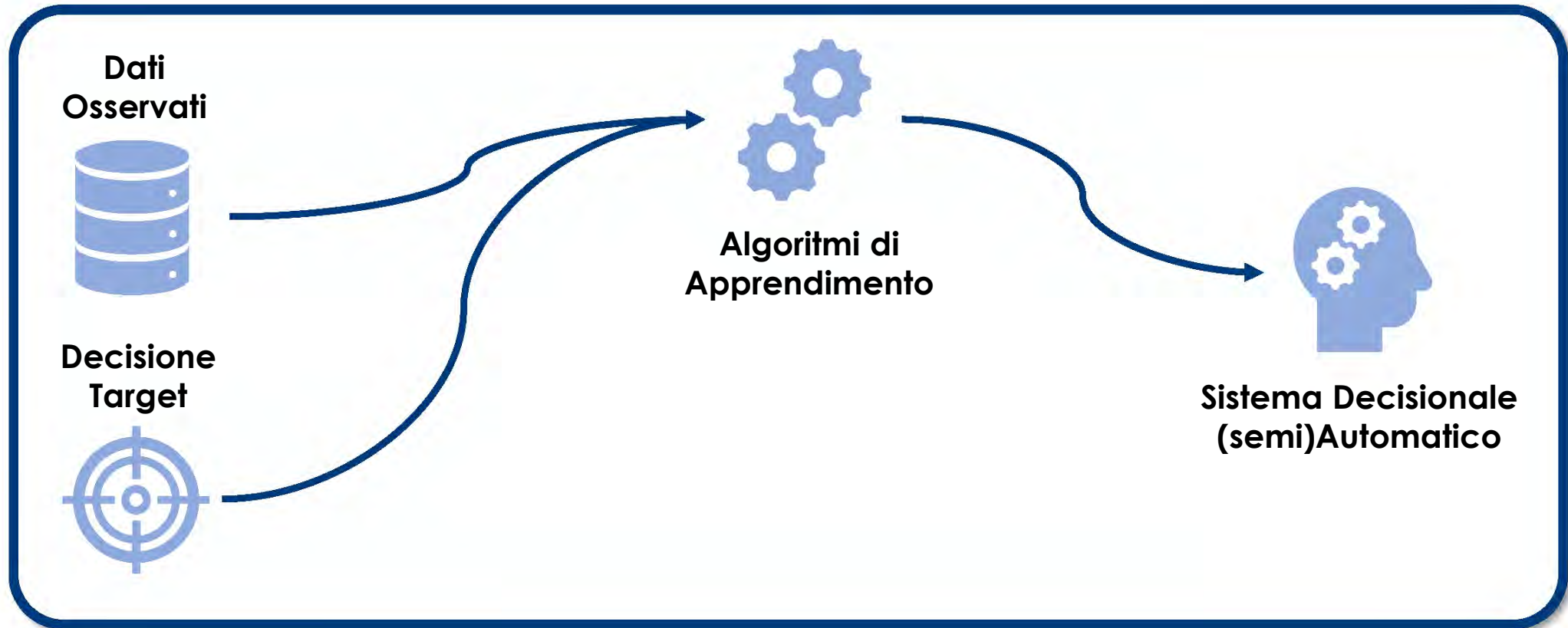


Il chatbot **Tay** di **Microsoft** è stato messo offline dopo aver pubblicato tweet razzisti

Come imparano gli algoritmi di AI «supervisionati»

Come «insegriamo» alle macchine

- «Il Machine Learning dà al computer l'abilità di imparare senza essere esplicitamente programmato» (Arthur Samuel, 1959)
- Nelle applicazioni di **Machine Learning supervisionato** le **logiche decisionali non** vengono **esplicitate** ma **dipendono dai dati** e **dall'algoritmo** di apprendimento



Rischi derivanti da AI: Persone (1/3)

Quando i **sistemi di AI** vengono usati per prendere **decisioni** possono avere un **effetto** sulla **vita delle persone**, rappresentando dunque un **rischio**



I Dati, specchio della società

Il **Machine Learning** potrebbe **amplificare** e **replicare** i **bias già presenti** nei dati, ma su più ampia scala



Squilibri nelle dimensioni del campione

Il **Machine Learning** potrebbe **trascurare** le **minoranze**, **generando** a tutti gli effetti **bias** originariamente **assenti** nei dati

Rischi derivanti da AI: Aziende (2/3)

Quando i **sistemi di AI non funzionano responsabilmente** possono portare **danni reputazionali e economici all'azienda**

HOME > TECH

Amazon built an AI tool to hire people but had to shut it down because it was discriminating against women

Isobel Asher Hamilton Oct 10, 2018, 11:47 AM

INSIDER

PRO PUBLICA

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

GOOGLE IS POISONING ITS REPUTATION WITH AI RESEARCHERS

The firing of top Google AI ethics researchers has created a significant backlash

By James Vincent | Apr 13, 2021, 9:30am EDT

THE VERGE

WILL KNIGHT

BUSINESS 11.19.2019 09:15 AM

WIRED

The Apple Card Didn't 'See' Gender—and That's the Problem

The way its algorithm determines credit lines makes the risk of bias more acute.

Rischi derivanti da AI (3/3)

Per **mitigare** i **rischi** associati all'**utilizzo** di **sistemi di intelligenza artificiale**, in particolare quelli classificati ad **alto rischio**, viene applicato un quadro normativo specifico come l'**AI Act**, che stabilisce **regole** e **requisiti** per garantirne un **utilizzo sicuro** e **responsabile**



Brussels, 21.4.2021
COM(2021) 206 final
2021/0106 (COD)

Proposal for a
REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL
LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE
(ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION
LEGISLATIVE ACTS

Article 13
Transparency and provision of information to users

Article 10 *Data and data governance*

- (f) examination in view of possible biases;
- (g) the identification of any possible data gaps or shortcomings, and how those gaps and shortcomings can be addressed.

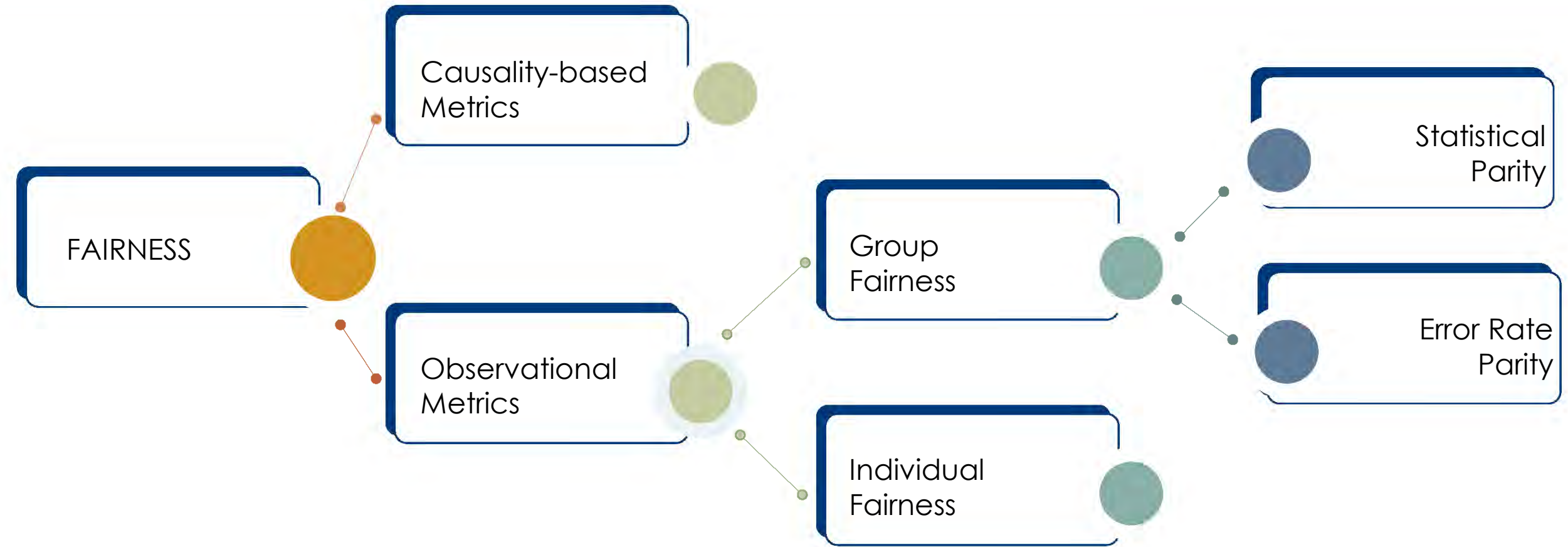
Training, validation and testing data sets shall be relevant, representative, free of errors and complete. They shall have the appropriate statistical properties, including, where applicable, as regards the persons or groups of persons on which the high-risk AI system is intended to be used. These characteristics of the data sets may be met at the level of individual data sets or a combination thereof.

Agenda

- 01  Fairness nei modelli di Machine Learning
- 02  Metriche di Fairness e incompatibilità
- 03  Razionali alla base della scelta della metrica di Fairness
- 04  Modelli di mitigazione dei bias
- 05  Guardrail

Lo Zoo delle Metriche di Fairness

Esistono **numerosa definizioni** di **fairness** che, nella maggior parte dei casi, non sono tra loro compatibili

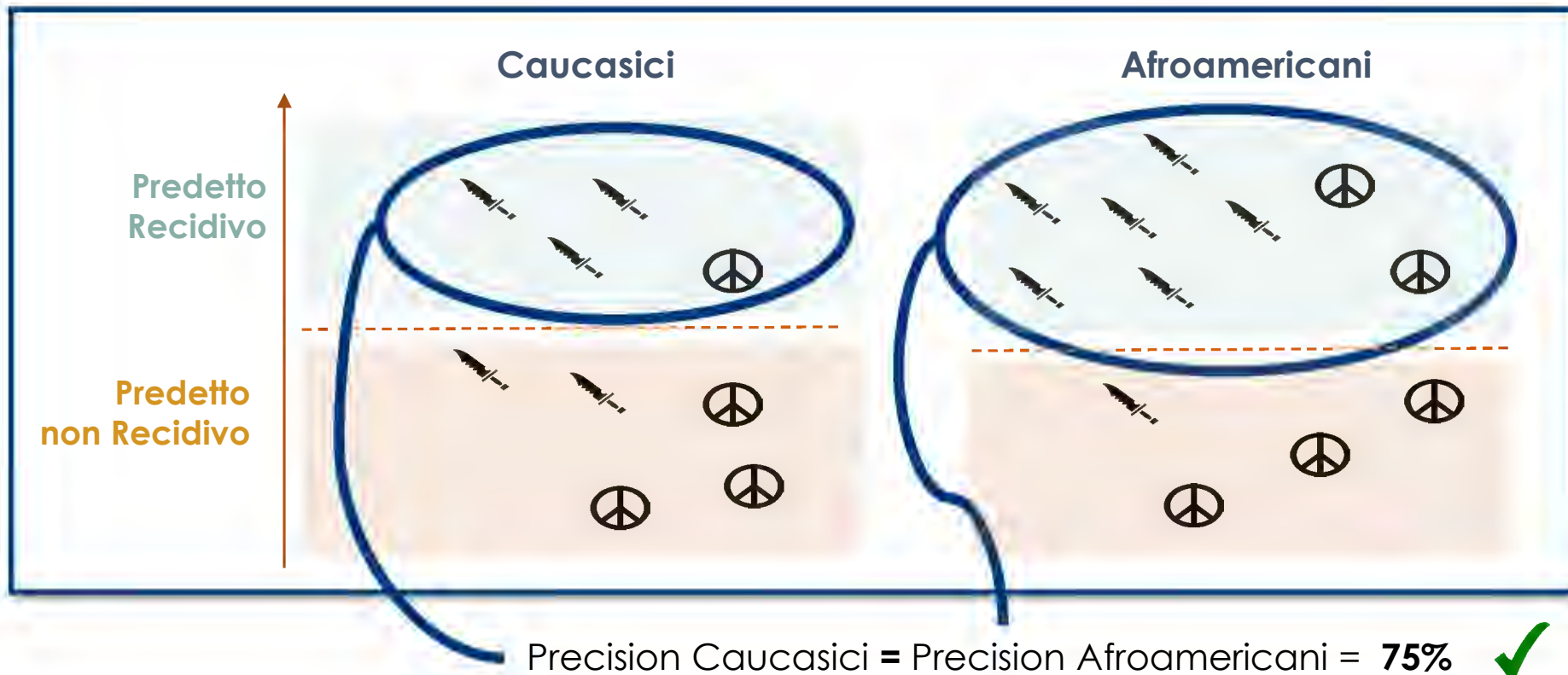


*A clarification of the nuances in the fairness metrics landscape, Castelnovo, Crupi, Greco, Regoli, Penco, Cosentini. **Scientific Reports 2022***

Incompatibilità tra metriche di Error Rate Parity (1/3)

Il dibattito COMPAS

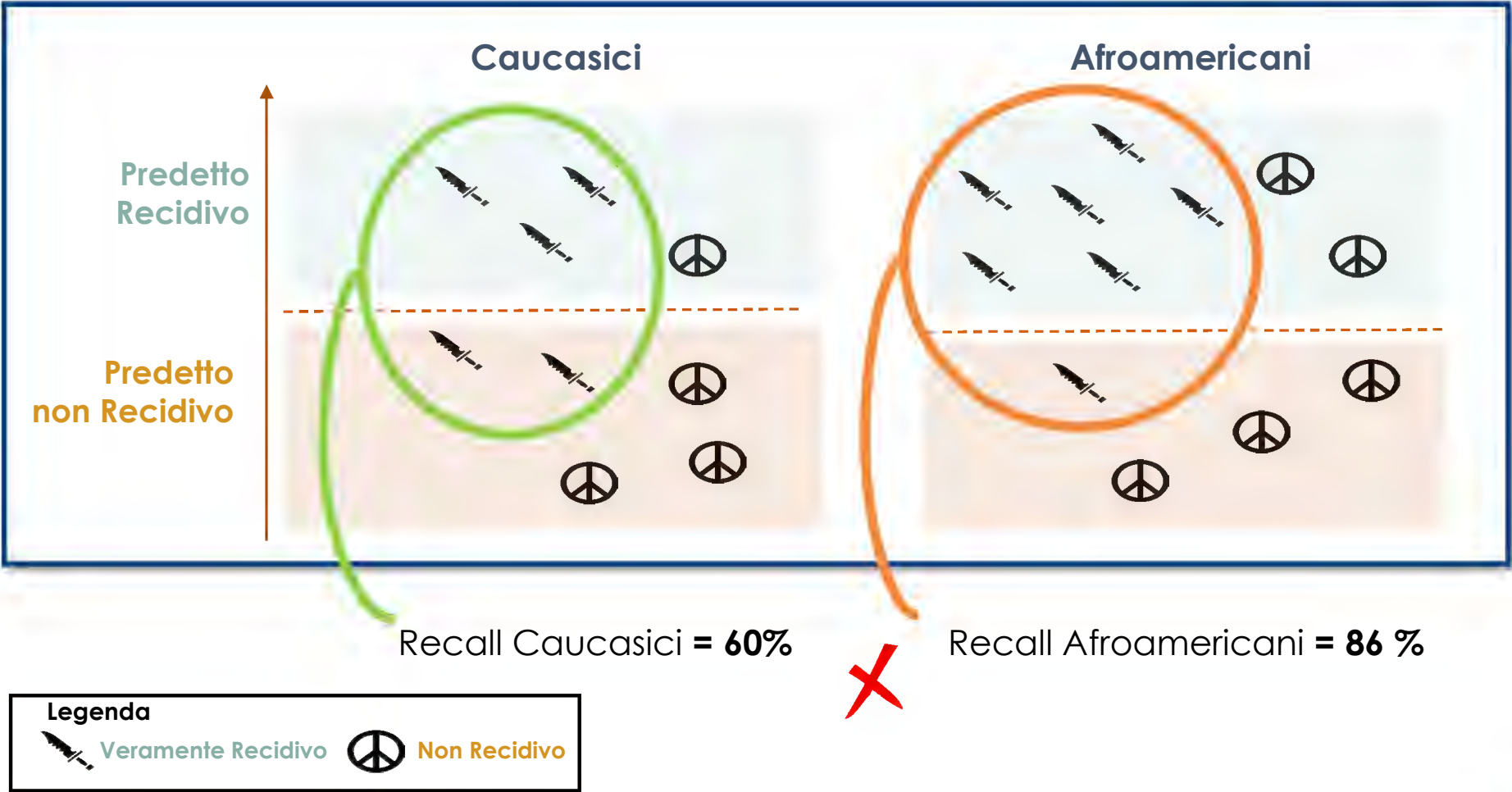
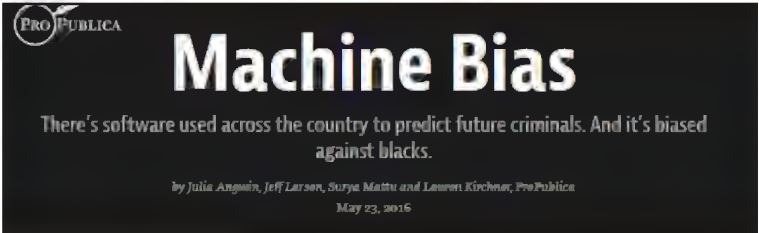
Negli **USA** era stato sviluppato un software per **prevedere la recidività dei criminali**. Tale software, tenendo conto della **Fairness**, era stato **sviluppato** affinché soddisfacesse **Precision Parity**



Incompatibilità tra metriche di Error Rate Parity (2/3)

Il dibattito COMPAS

COMPAS fu **criticato mediaticamente** come **discriminatorio** perché **non soddisfaceva Equal Opportunity**



Incompatibilità tra metriche di Error Rate Parity (3/3)

Il dibattito COMPAS

Sviluppare un **sistema** che predica la **recidiva rispettando** simultaneamente **Prediction Parity** ed **Equal Opportunity non è possibile**, come dimostrato dal *teorema di impossibilità*

Proposition 4. Assume that all events in the joint distribution of (A, R, Y) have positive probability, and assume $A \not\perp Y$. Then, separation and sufficiency cannot both hold.

Proof. A standard fact²⁷ about conditional independence shows

$$A \perp R \mid Y \text{ and } A \perp Y \mid R \implies A \perp (R, Y).$$

Moreover,

$$A \perp (R, Y) \implies A \perp R \text{ and } A \perp Y.$$

Taking the contrapositive completes the proof.

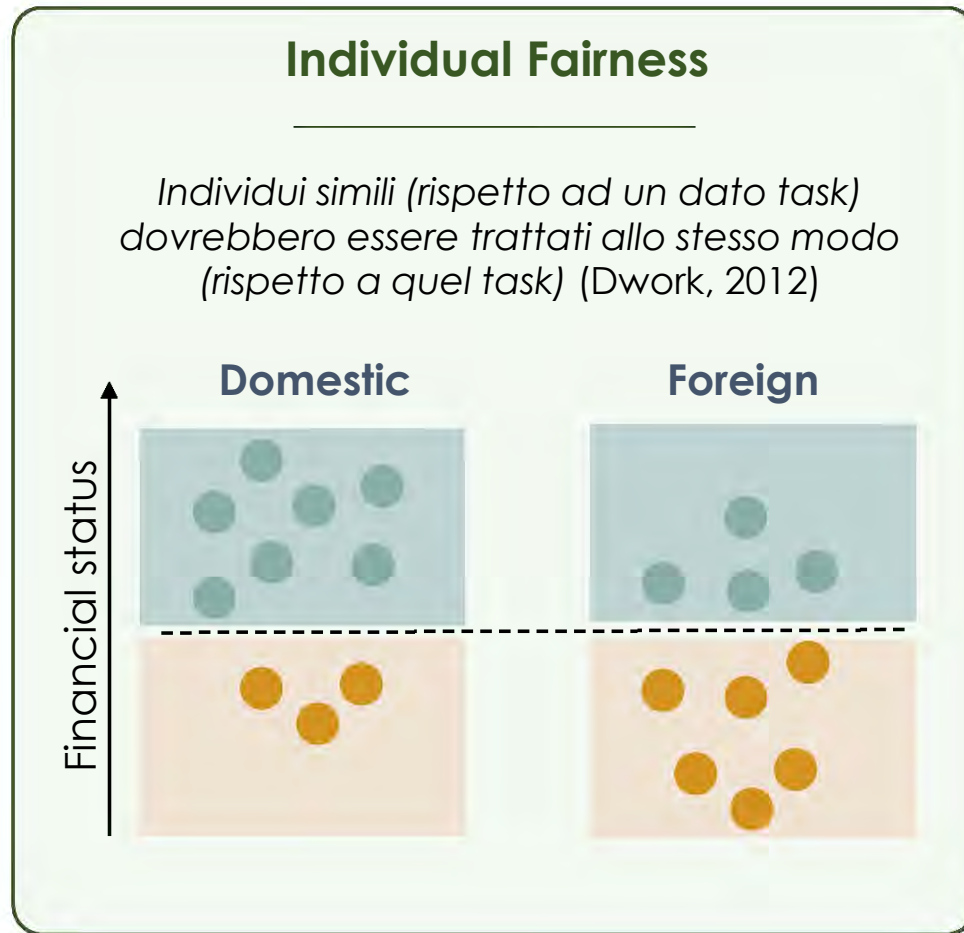
□

²⁷ See Theorem 17.2 in L. Wasserman, *All of Statistics: A Concise Course in Statistical Inference* (Springer, 2010)

Fairness and Machine Learning, Solon Barocas, Moritz Hardt, Arvind Narayanan (2019). <https://fairmlbook.org>

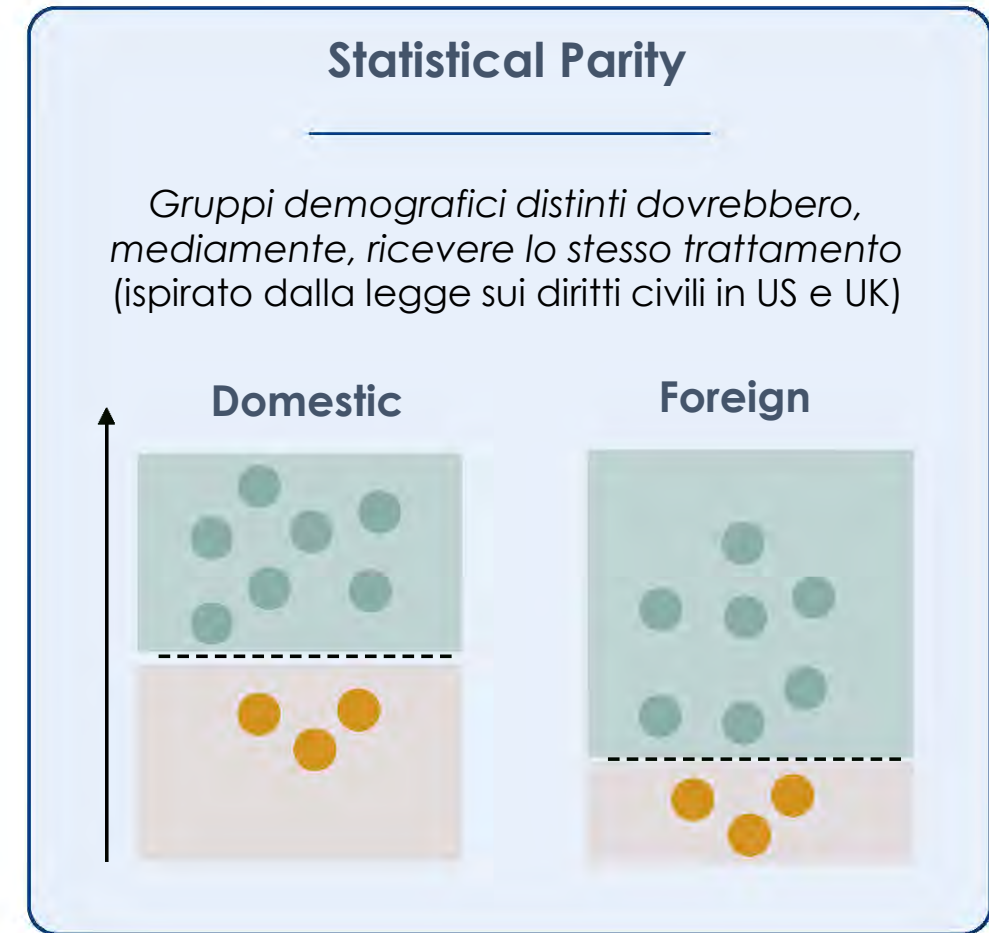
Incompatibilità tra Individual e Group Fairness (1/2)

Un dilemma frequente è scegliere fra group e/o individual fairness



Risultato
positivo

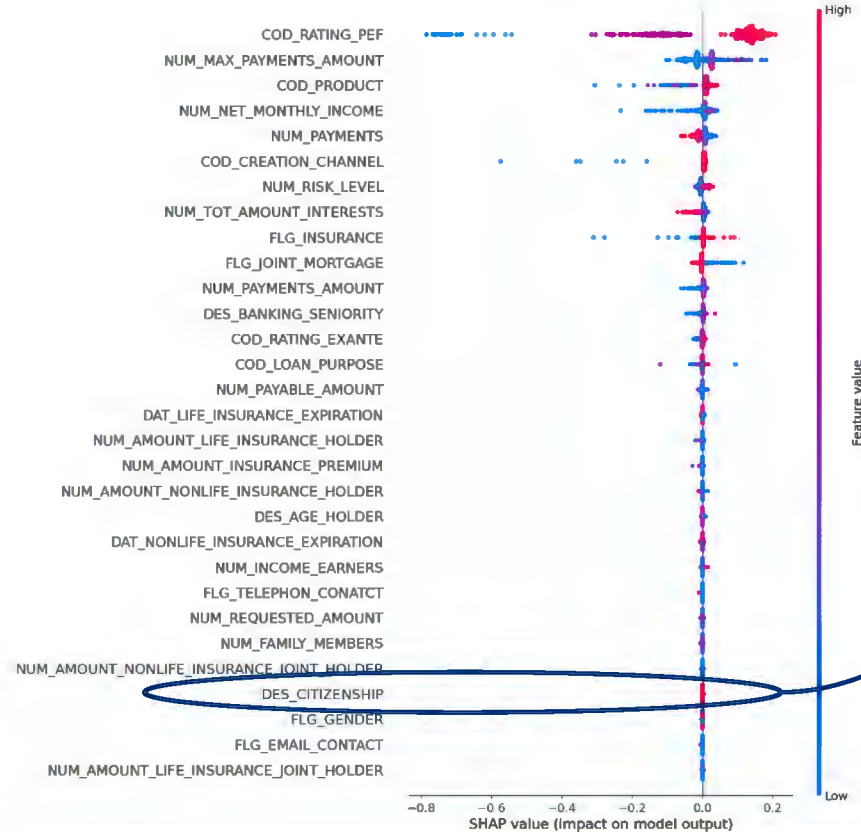
Risultato
negativo



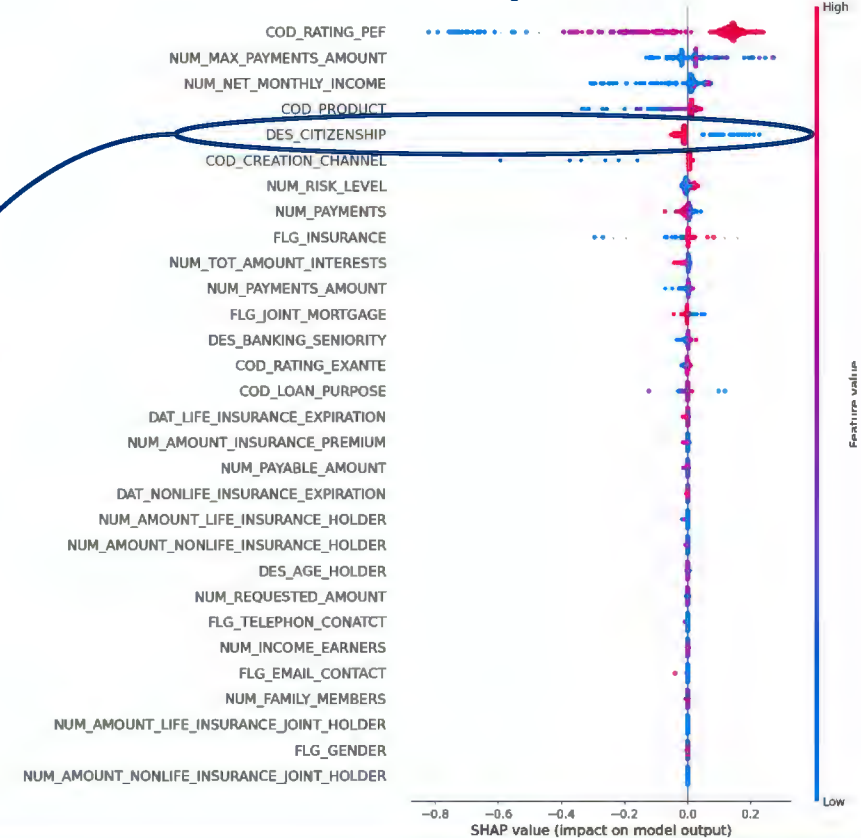
Incompatibilità tra Individual e Group Fairness (2/2)

Un dilemma frequente è scegliere fra group e/o individual fairness

Individual Fair Model



Statistical Parity Model



Quando la target è correlata con una sensibile, l'unico modo per creare un output che soddisfi la **Statistical Parity** è utilizzare la variabile sensibile nel modello, **perdendo la Individual Fairness**

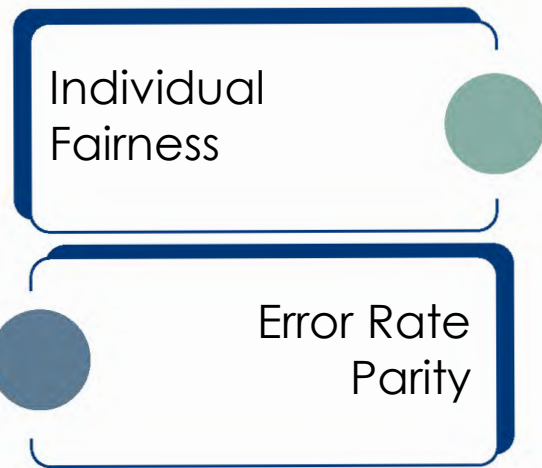
Agenda

- 01  Fairness nei modelli di Machine Learning
- 02  Metriche di Fairness e incompatibilità
- 03  Razionali alla base della scelta della metrica di Fairness
- 04  Modelli di mitigazione dei bias
- 05  Guardrail

Effetti delle metriche di Fairness (1/2)

La scelta della metrica di Fairness dipende innanzitutto dall'intenzione, o meno, di modificare lo status quo

Alcune metriche che **mantengono** lo status quo:



Queste metriche vanno tipicamente utilizzate quando si «assume» che i dati a disposizione e le loro relazioni siano una rappresentazione fedele della realtà che il sistema decisionale in cui sono impiegate deve cercare di non alterare

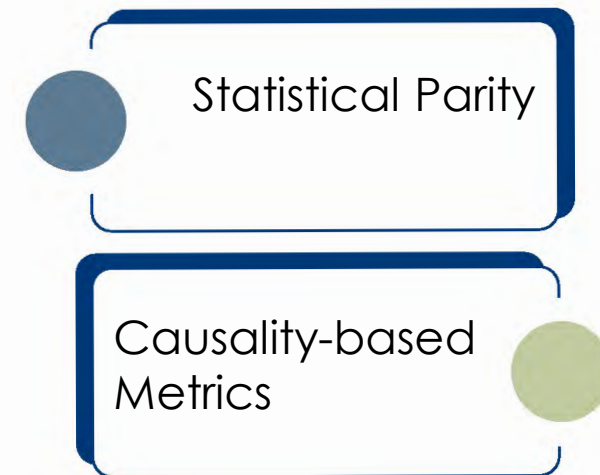


L'obiettivo principale di queste metriche è **non creare** nuovi bias e **non amplificare** quelli già presenti



Un modello perfetto è anche perfettamente fair

Alcune metriche che **modificano** lo status quo:



Queste metriche vanno tipicamente utilizzate quando si «assume» che i dati a disposizione, e in particolare la variabile target, contengono bias che si vogliono eliminare



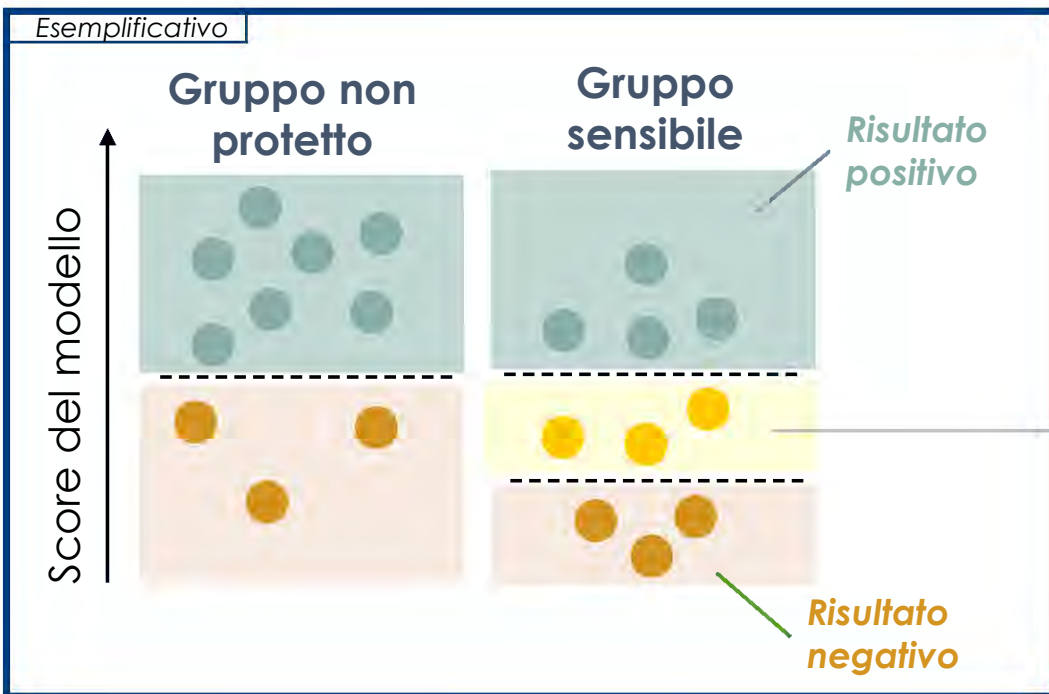
L'obiettivo principale di queste metriche è **rimuovere** bias già presenti nei dati



Un modello perfetto non può essere fair

Effetti delle metriche di Fairness (2/2)

La scelta di cambiare lo status quo può dipendere sia da ragioni strategiche, sia di beneficio per la popolazione. In alcuni casi, decisioni positive possono creare danni



Si dovrebbero classificare **positivamente** quando:
la decisione porta **benefici** che migliorano l'intero gruppo di appartenenza e non ci sono ragioni per pensare che le classi siano trattate diversamente

Per esempio ammettere più donne in facoltà università tecniche, può ispirare sempre più donne a intraprendere studi tecnici. Inoltre, uomini e donne hanno pari predisposizione verso tali materie

Nelle situazioni reali bisogna decidere se classificare **positivamente** o **negativamente** i migliori individui del gruppo sensibile che permetterebbero di pareggiare le «quote» ma hanno caratteristiche individuali che tipicamente portano a un rifiuto

Si dovrebbero classificare **negativamente** quando:
la decisione può portare **danni** a chi la riceve o all'intera società

Per esempio concedere un prestito a chi non può ripagarlo può aggravargli la situazione finanziaria, o non controllare frodi laddove si verificano con più probabilità fa effettivamente aumentare le frodi contro i clienti

Agenda

- 01  Fairness nei modelli di Machine Learning
- 02  Metriche di Fairness e incompatibilità
- 03  Razionali alla base della scelta della metrica di Fairness
- 04  **Modelli di mitigazione dei bias**
- 05  Guardrail

Tecniche di mitigazione del Bias

Le tecniche di mitigazione si dividono nelle seguenti famiglie, che dipendono dalla fase nel processo in cui viene rimosso il bias

Pre-Processing:



In-Processing:



Post-Processing:



Agenda

- 01  Fairness nei modelli di Machine Learning
- 02  Metriche di Fairness e incompatibilità
- 03  Razionali alla base della scelta della metrica di Fairness
- 04  Modelli di mitigazione dei bias
- 05  Guardrail

L'evoluzione tecnologica porta nuove tipologie di rischi che richiedono nuovi modelli di gestione e tecnologie innovative per mitigarli

+ Entità del rischio -	 AI Agentica  GEN AI  AI	Lesione dei diritti fondamentali dell'uomo	Rischi reputazionali	Rischi di sicurezza	Rischi di Inefficienze operative
		 Coincide con il rischio di Conformità all'AI ACT es. prendendo una decisione discriminatoria in ambito concessione crediti	 NON Considerati nella classificazione del rischio richiesta dall'AI ACT es. un chatbot aziendale che da una risposta a un Cliente o con del contenuto inappropriato causando danni d'immagine	 es. Manipolazioni dei prompt per ottenere dati sensibili e informazioni non autorizzate	 es. gli agenti sono a rischio di eseguire operazioni ridondanti con decuplicazione dei costi

Questi rischi sono gestiti attraverso **strumenti, misure governance** e controlli di **Responsible AI**

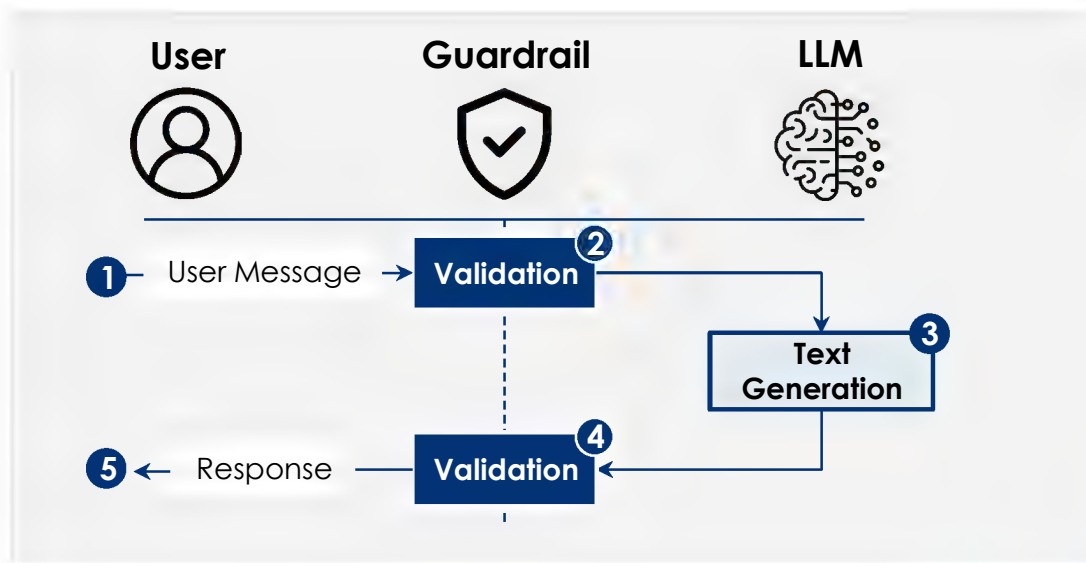


Take-away: la strategia del Gruppo è mitigare e gestire i rischi legati all'Intelligenza Artificiale, che diventano sempre più complessi con l'evoluzione tecnologica

L'adozione della Generative AI negli use case è in rapida crescita, rendendo sempre più centrale la mitigazione dei rischi legati all'uso dei LLM

Rischio	Presidi	Descrizione
Allucinazioni e disinformazione	Allucinazioni	Prevenire le informazioni fabbricate o errate prodotte da un modello di intelligenza artificiale quando manca di contesto o conoscenza sufficienti
Pregiudizi e risposte inappropriate	Tossicità	Blocco dell'uso di un linguaggio che incita all' odio, molesto o dannoso
Gestione e rischi di sicurezza	Prompt Injection	Rileva le tecniche di attacco in cui vengono inserite istruzioni dannose nell'input del modello per costringerlo a produrre output non intenzionali e ne previene gli effetti
Utilizzi a finalità generali	Gestione contenuti	Impedire al modello di rispondere a domande fuori contesto
Gestione dei dati sensibili e privacy	Rilevamento PII	Rileva informazioni di identificazione personale che possono essere utilizzate per riconoscere o tracciare un individuo e ne protegge la privacy

Come funziona un Guardrail



I Guardrail operano su due livelli:

Pre-guard: è la prima linea di difesa e si occupa della **validazione dell'input dell'utente** prima che venga elaborato dal modello LLM del caso d'uso

Post-guard: entra in azione **dopo** la **generazione della risposta** da parte dell'LLM e ne valida il contenuto, fungendo da **ultimo controllo** prima della consegna all'utente. Serve a intercettare eventuali violazioni di policy, linee guida o inesattezze

MODELLO DI CONTROLLO DEI GUARDRAIL

- 1 L'utente invia un messaggio al chatbot**
- 2 Pre-Guard: controllo del messaggio dell'utente**

Prima che il messaggio venga inviato all'LLM, il sistema di Pre-Guard lo analizza per verificare che sia conforme alle regole (sicurezza, policy, contenuti ammessi)

 - Se il messaggio è valido, viene inoltrato all'LLM
 - Se il messaggio non è conforme, viene bloccato e l'utente riceve una risposta standard o una richiesta di modifica
- 3 L'LLM genera la risposta**
- 4 Post-Guard: controllo della risposta dell'LLM**

Prima di essere mostrata all'utente, la risposta dell'LLM viene verificata dal Post-Guard per assicurare che sia sicura, corretta e conforme

 - Se la risposta è accettabile, viene mostrata all'utente
 - Se la risposta non è conforme, il sistema la blocca e fornisce un messaggio alternativo (ad esempio chiedendo di riformulare o segnalando il problema)
- 5 Il messaggio di risposta arriva all'utente**