

# **MITIGATING DEMOGRAPHIC BIAS IN SOFT FACIAL ATTRIBUTE RECOGNITION THROUGH SYNTHETIC DATA GENERATION**



Dott.ssa Chiara PERO  
[cpero@unisa.it](mailto:cpero@unisa.it)



# INDICE

**Introduzione** - Bias demografici: definizione e impatti nei sistemi biometrici.

---

**Attributi Facciali Soft** - Definizione e uso.

---

**Analisi sperimentale in corso** - Identificare i bias demografici presenti nei database noti in letteratura, caratterizzati dall'etichettatura di *soft attributi*.

---

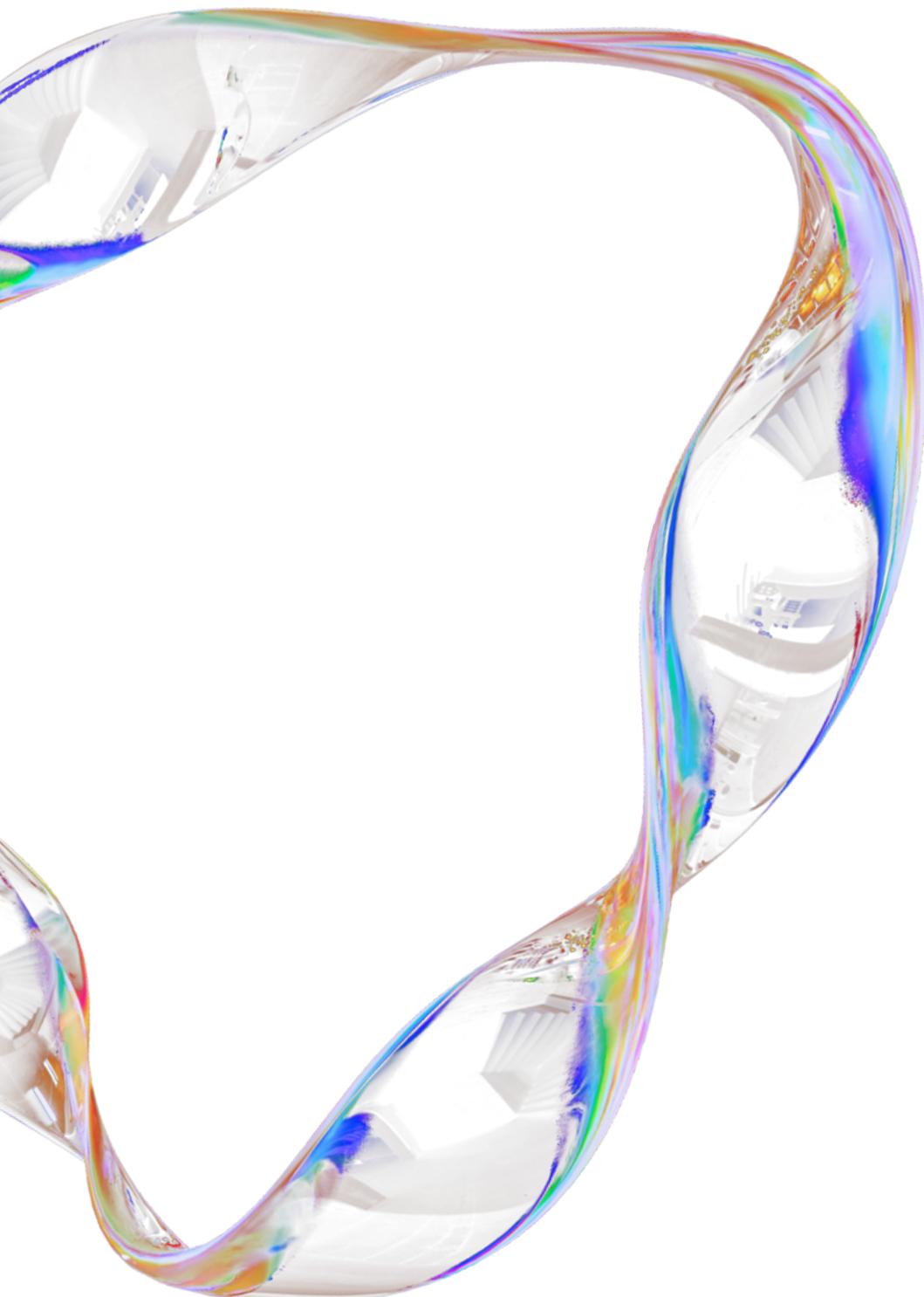
**Generazione di dati sintetici** - Uso di Stable Diffusion Models per creare volti umani realistici.

---

**Proposta progetto FVAB 2024/2025** - Utilità dei dataset sintetici per superare i bias demografici.

---

# BIAS DEMOGRAFICO



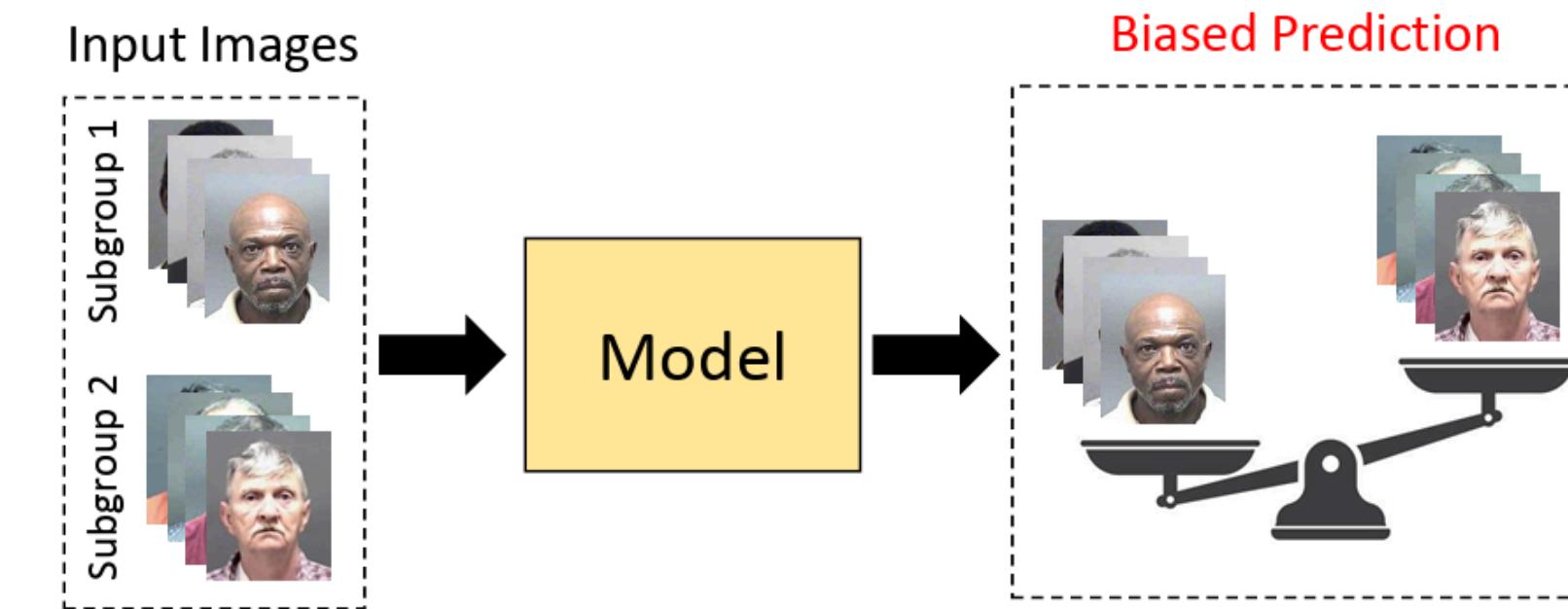
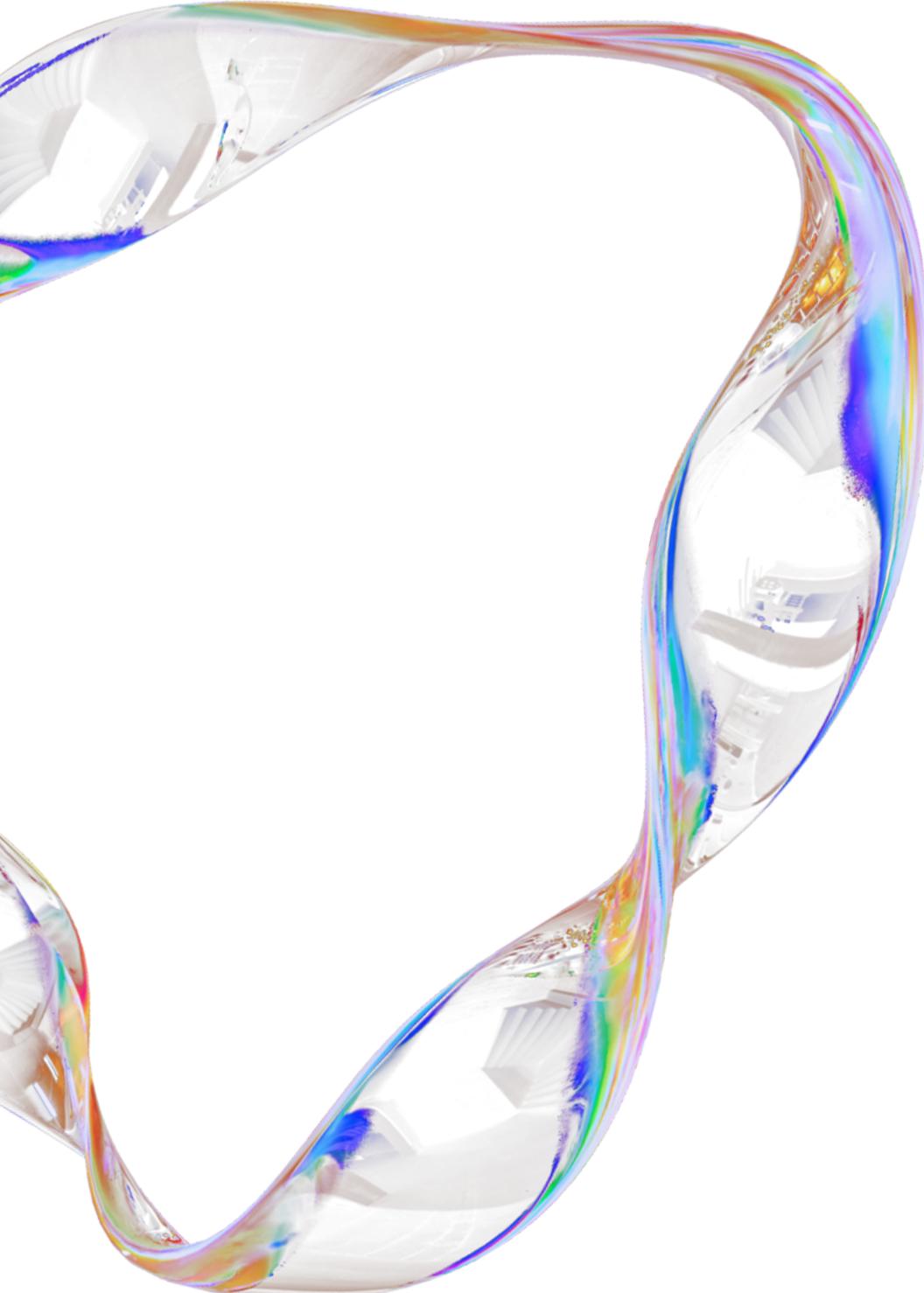
## Bias Demografico

Il **bias demografico** si verifica quando un sistema di intelligenza artificiale produce risultati differenziati o discriminatori per specifici gruppi demografici (ad esempio, basati su sesso, razza o età). Questo fenomeno emerge tipicamente da correlazioni nei dati di addestramento, che portano il modello a replicare e amplificare pregiudizi presenti nei dataset.

## Problema

Sebbene la misurazione del bias nei modelli sia stata oggetto di numerosi studi, l'analisi del bias nei **dataset di origine** è rimasta in gran parte trascurata.

# BIAS DEMOGRAFICO



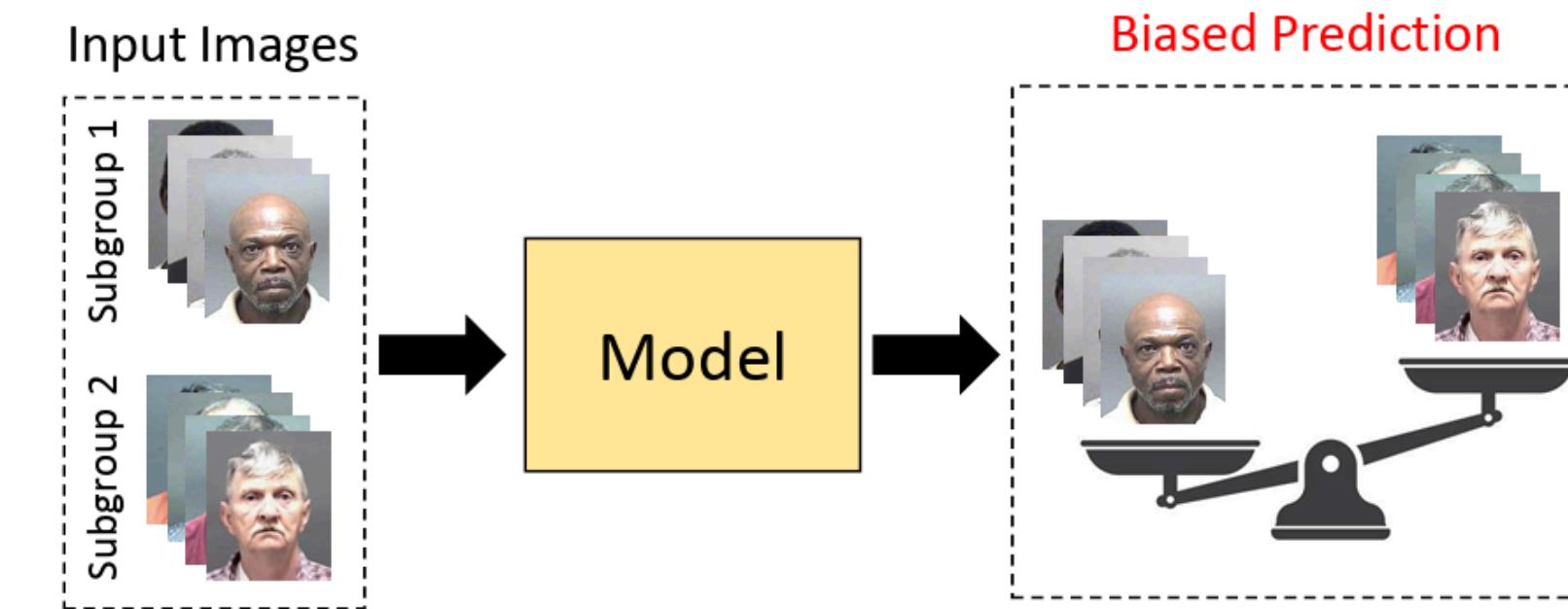
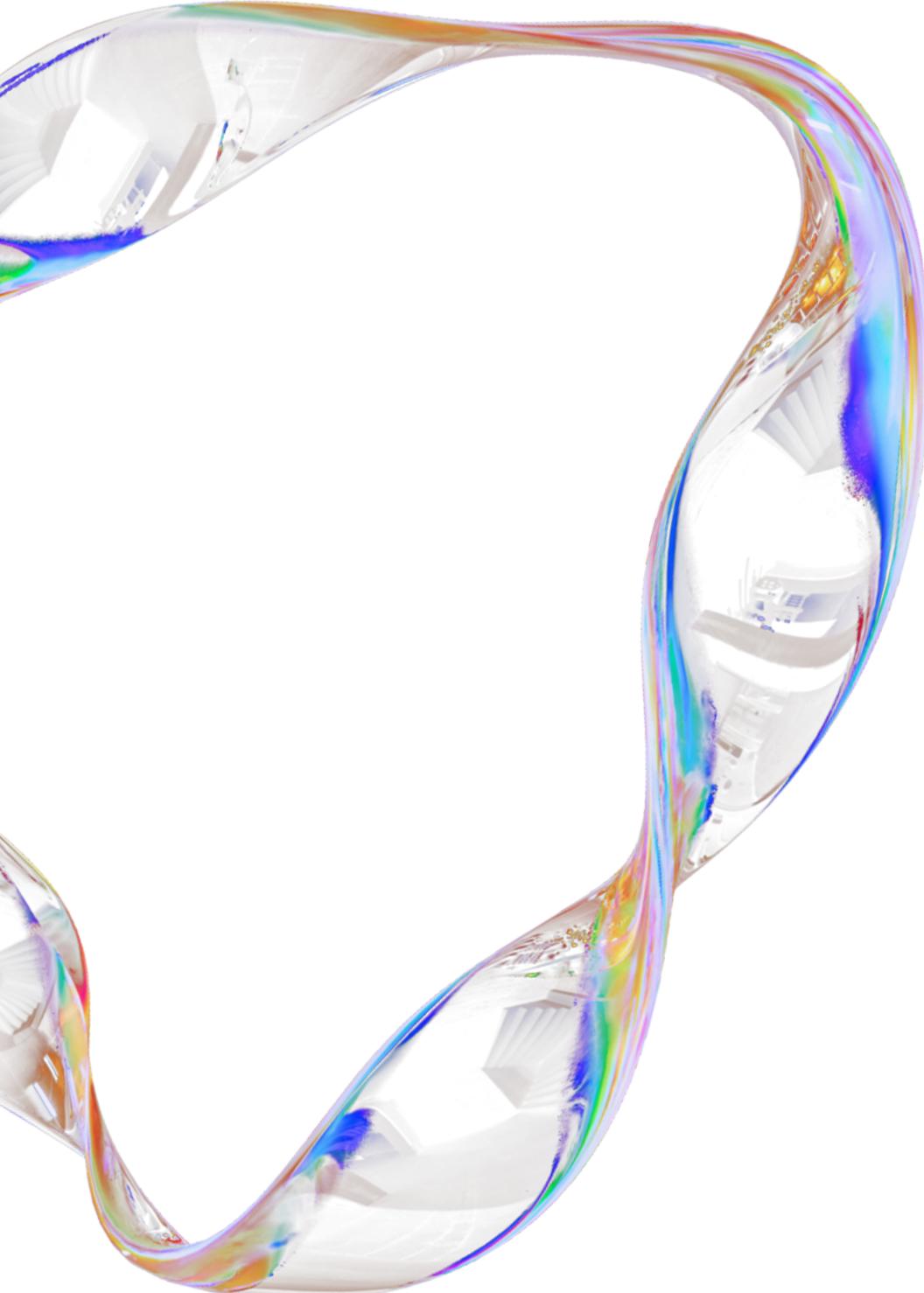
## 🔍 Cosa mostra questa immagine?

A sinistra: abbiamo due sottogruppi demografici di immagini di input (**Subgroup 1** e **Subgroup 2**), rappresentanti persone di etnia e aspetto differenti.

Al centro: un modello di intelligenza artificiale riceve le immagini ed effettua delle predizioni.

A destra: il modello produce una **predizione sbilanciata**, ovvero più favorevole a un sottogruppo rispetto all'altro.

# BIAS DEMOGRAFICO



## ⚖️ Cosa succede?

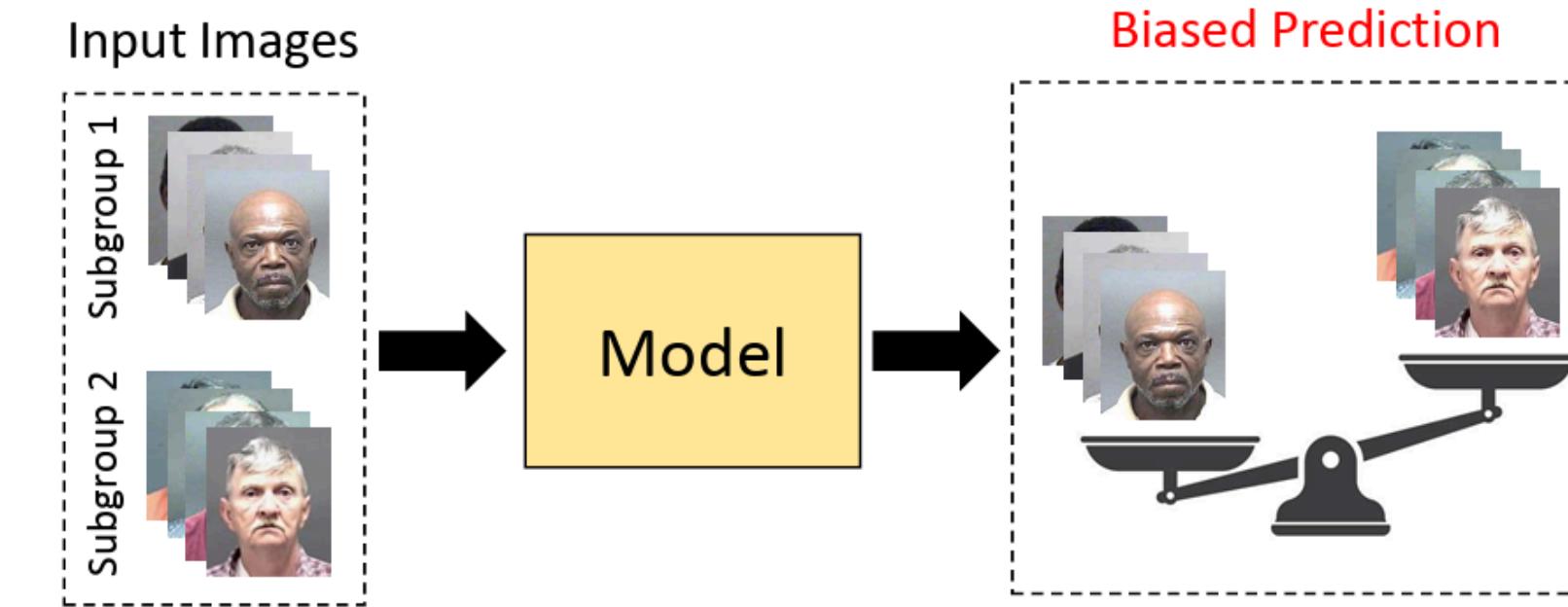
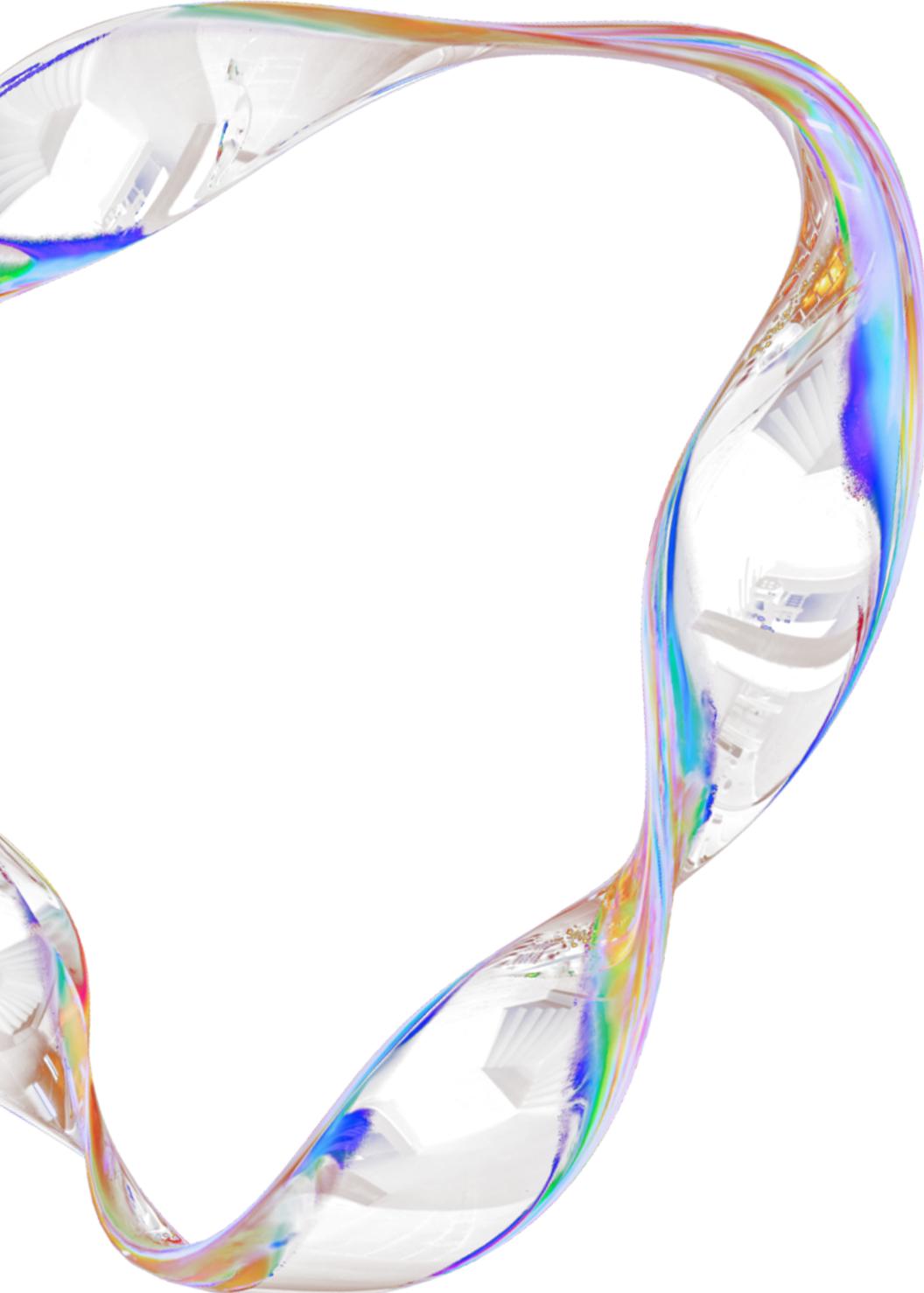
Il sistema sembra “**favorire**” il secondo sottogruppo (rappresentato visivamente da una bilancia inclinata).

Questo è un esempio di **bias algoritmico**, cioè un comportamento imparziale appreso dal modello a causa di dati non bilanciati o distorsioni nei pattern appresi.

### Messaggio chiave:

Anche se il modello tratta tutti i dati allo stesso modo in teoria, in pratica può imparare comportamenti ingiusti, penalizzando alcuni gruppi più di altri.

# BIAS DEMOGRAFICO



## PROBLEMA

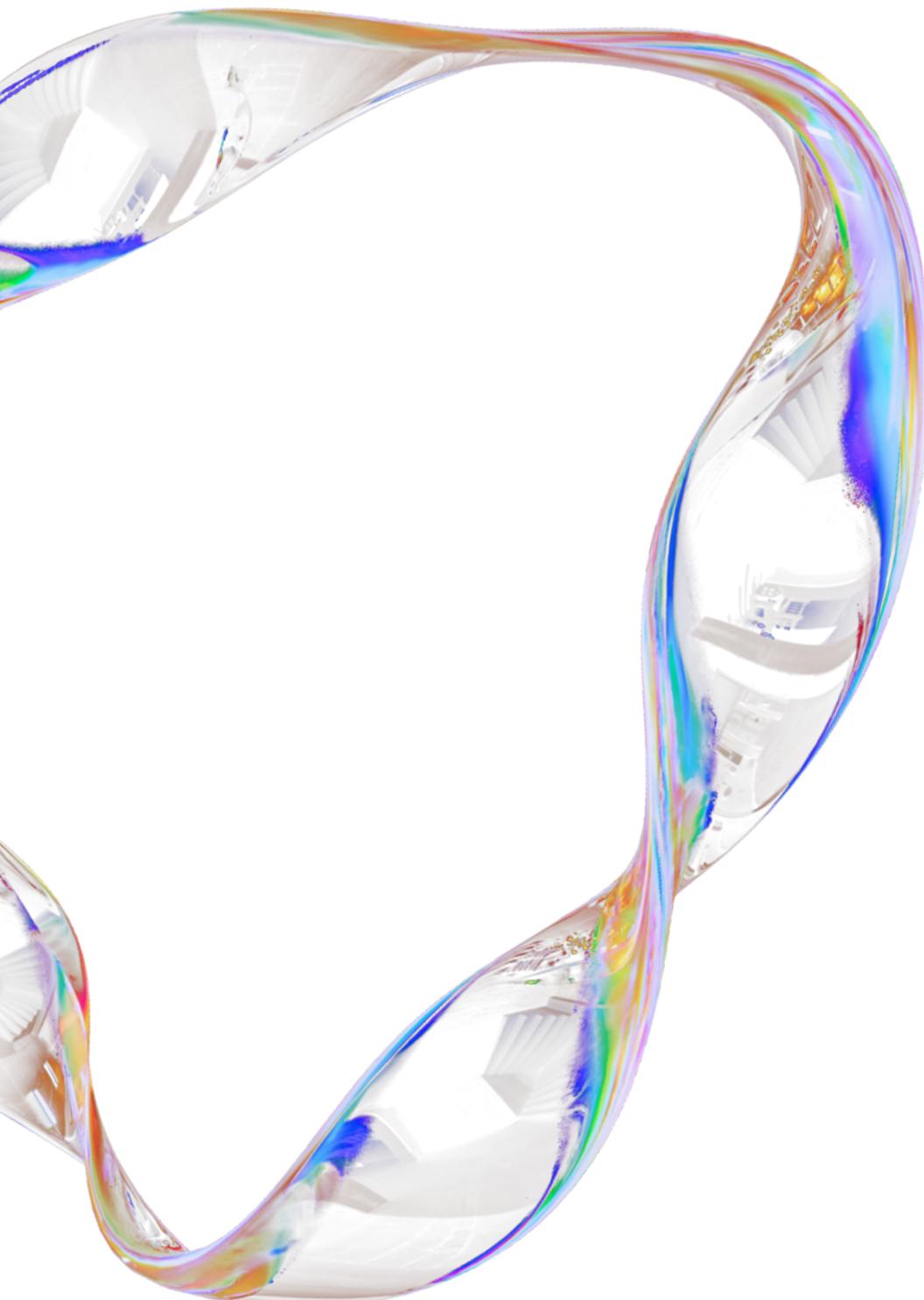
Studi recenti mostrano che i modelli di analisi facciale non sono equi:

- Maggiore accuratezza su individui **bianchi** o di pelle chiara
- Errori più frequenti su donne con pelle **scura**
- Performance variabili tra **gruppi etnici** e di genere

📌 Conseguenze:

- Discriminazione algoritmica
- Ridotta affidabilità in contesti critici
- Erosione della fiducia nell'IA

# SOFT FACIAL ATTRIBUTES



Gli **attributi facciali soft** sono caratteristiche percettive e non identificative estratte da immagini del volto.

Esempi comuni:

👉 gender, smiling, young, attractive, wearing makeup, has beard...

Non sono sempre binari, e sono spesso soggetti a interpretazione culturale e sociale.

🔍 Perché sono rilevanti?

Ampiamente utilizzati in:

- Sistemi di **personalizzazione** (recommender systems, advertising)
- Computer vision forense e **videosorveglianza**
- Analisi **comportamentale** e interazione uomo-macchina

# DATASET FACCIALI ETICHETTATI: SOFT-BIOMETRICS



## Labelled Faces in the Wild (LFW)

Dataset composto da 13.200 immagini di 5.700 identità in ambienti non controllati, con variabilità in posa, illuminazione, espressione e demografia. Include **74 attributi binari**, ma con annotazioni di accuratezza limitata (72% rispetto a quelle umane).



## CelebFaces Attributes Dataset (CelebA)

CelebA comprende oltre 202.000 immagini di 10.000 soggetti diversi. Il dataset include una vasta gamma di variazioni di posa e sfondi complessi, con annotazioni per **40 attributi binari** relativi a demografia, capelli, geometria del volto e accessori.



## MAAD-Face

Partendo da VGGFace2, il dataset include oltre 3,3 milioni di immagini di oltre 9.100 soggetti, con una grande varietà di pose, età ed etnie. Fornisce annotazioni per **47 attributi binari**, totalizzando oltre 123,9 milioni di annotazioni.

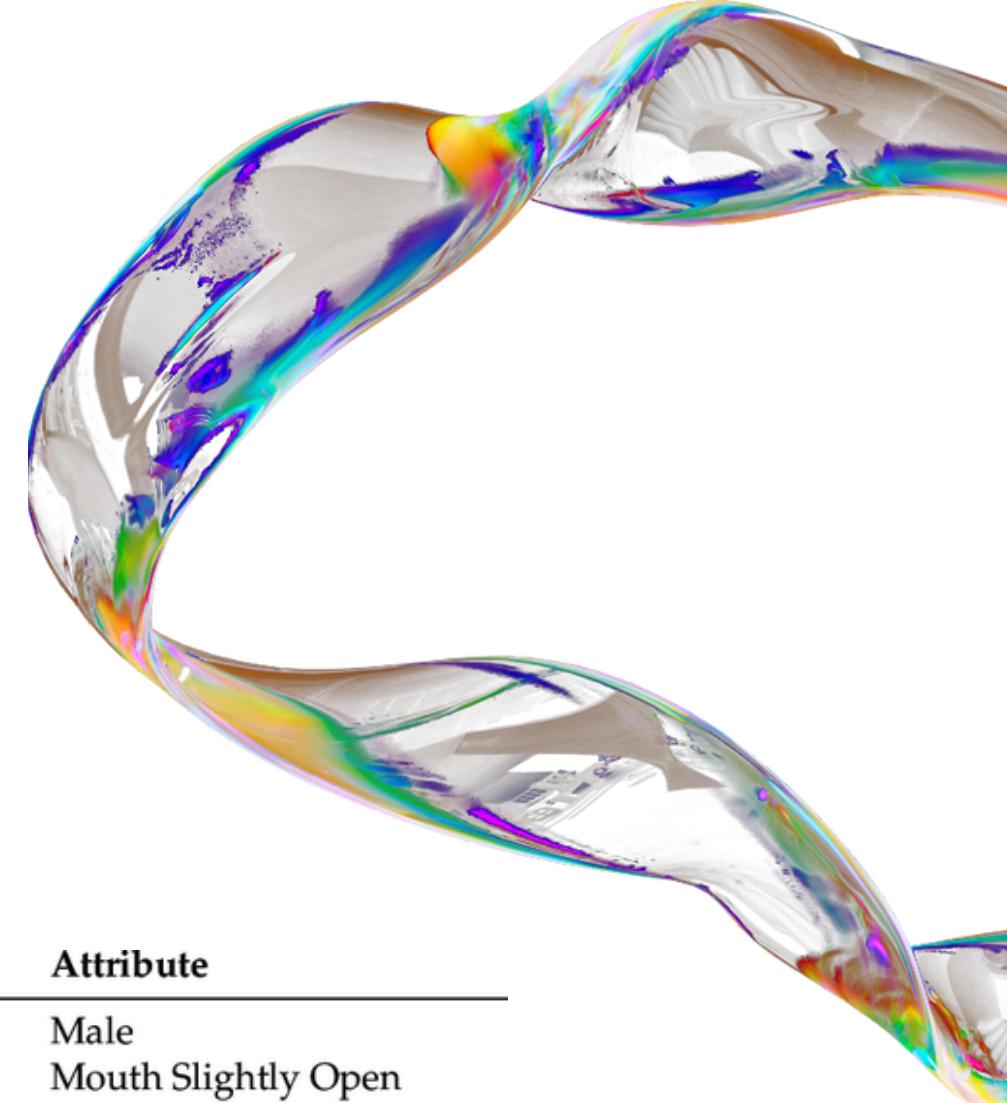
# CELEBFACES ATTRIBUTES DATASET (CELEBA)



## CelebFaces Attributes Dataset (CelebA)

CelebA comprende oltre 202.000 immagini di 10.000 soggetti diversi. Il dataset include una vasta gamma di variazioni di posa e sfondi complessi, con annotazioni per **40 attributi binari** relativi a demografia, capelli, geometria del volto e accessori.

Idx.	Attribute	Idx.	Attribute
1	5 O'Clock Shadow	21	Male
2	Arched Eyebrows	22	Mouth Slightly Open
3	Attractive	23	Mustache
4	Bags Under Eyes	24	Narrow Eyes
5	Bald	25	No Beard
6	Bangs	26	Oval Face
7	Big Lips	27	Pale Skin
8	Big Nose	28	Pointy Nose
9	Black Hair	29	Receding Hairline
10	Blond Hair	30	Rosy Cheeks
11	Blurry	31	Sideburns
12	Brown Hair	32	Smiling
13	Bushy Eyebrows	33	Straight Hair
14	Chubby	34	Wavy Hair
15	Double Chin	35	Wearing Earrings
16	Eyeglasses	36	Wearing Hat
17	Goatee	37	Wearing Lipstick
18	Gray Hair	38	Wearing Necklace
19	Heavy Makeup	39	Wearing Necktie
20	High Cheekbones	40	Young



# PREDIZIONE DI ATTRIBUTI FACCIALI

## STATO DELL'ARTE



### Slim-CNN

Una CNN leggera con moduli Slim per ridurre calcoli e memoria, ideale per dispositivi mobili.

**91.24%** di accuratezza su CelebA.



### FaRL (Facer)

Framework per rappresentazioni universali del volto, basato su apprendimento contrastivo immagine-testo e mascheramento immagini.

**91.39%** di accuratezza su CelebA.



### FaceXFormer

Modello Transformer per **9** compiti di analisi facciale (es. parsing, landmark detection, attributi).

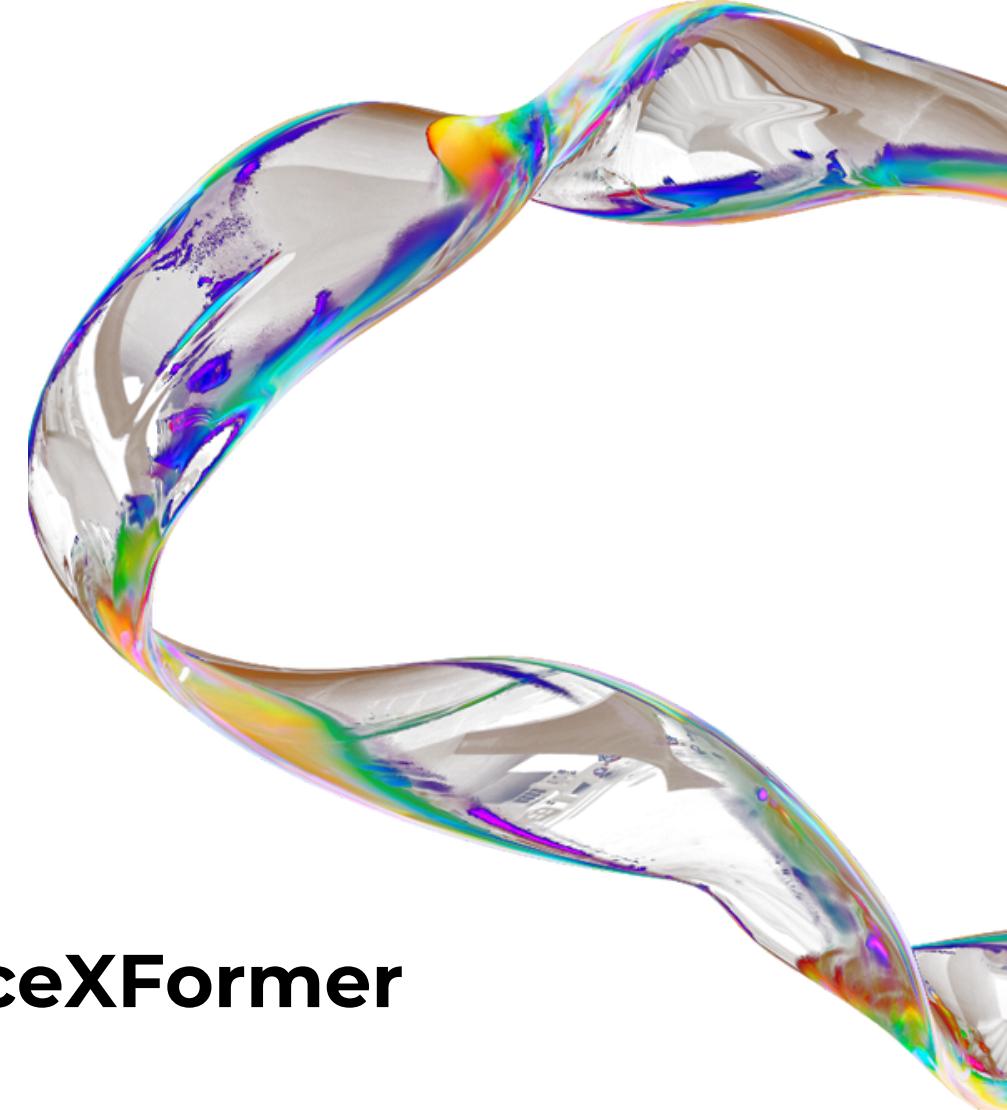
**91.83%** di accuratezza su CelebA.



Sharma, Ankit Kumar, and Hassan Foroosh. "Slim-cnn: A light-weight cnn for face attribute prediction." 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020). IEEE, 2020.

Zheng, Yinglin, et al. "General facial representation learning in a visual-linguistic manner." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

Narayan, Kartik, et al. "FaceXFormer: A Unified Transformer for Facial Analysis." arXiv preprint arXiv:2403.12960 (2024).





Utilizzare **dati sintetici**, ovvero immagini generate artificialmente, per:

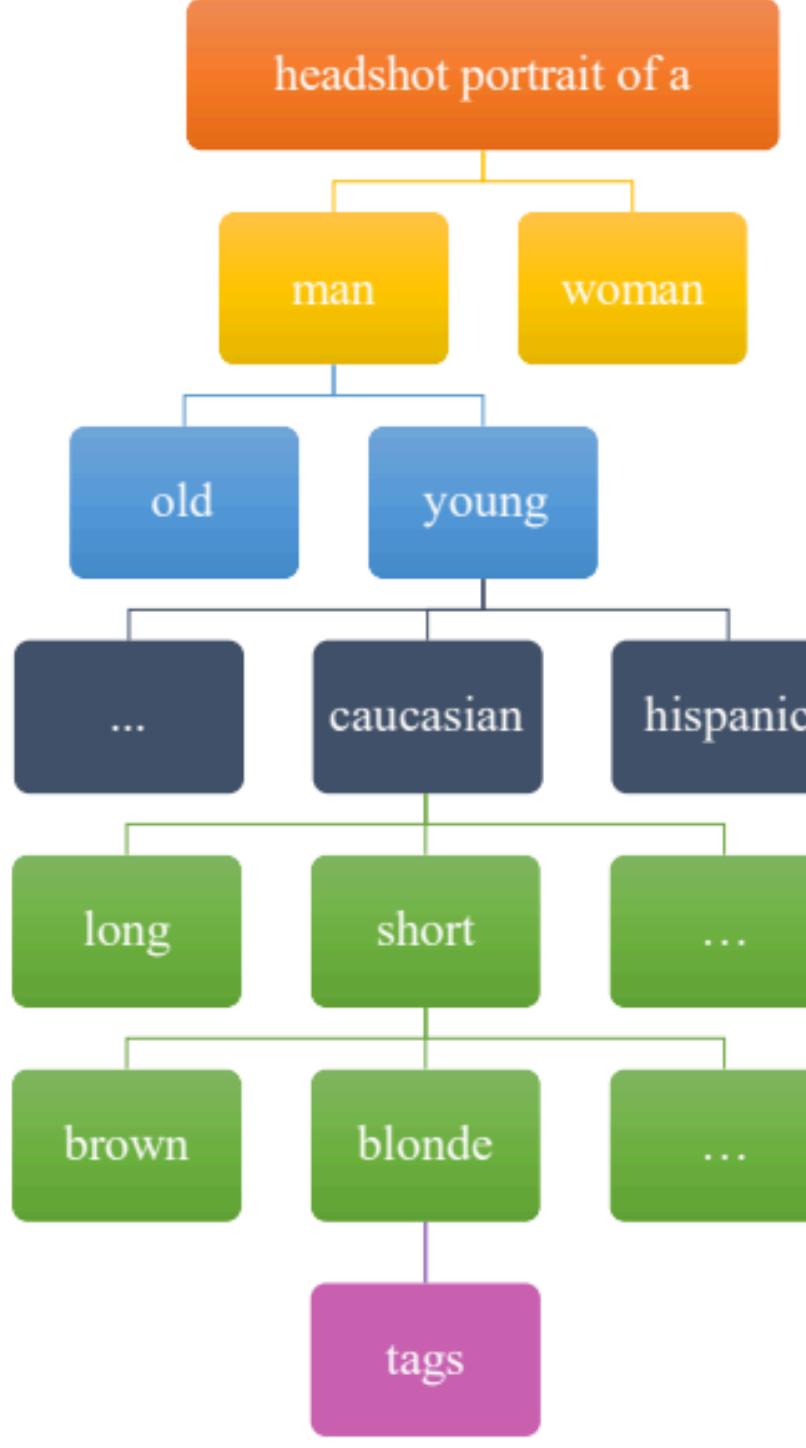
- **Correggere** lo sbilanciamento nei dataset reali
- **Allenare** meglio i modelli di riconoscimento degli attributi facciali
- Studiare il bias in modo più controllato e sistematico

### **FOCUS:**

Generazione controllata di immagini sintetiche → Utilizzo di modelli generativi (es. **StyleGAN3, Stable Diffusion**) per creare volti:

- Con tratti desiderati (etnia, età, genere).
- Con annotazioni coerenti.

# STABLE DIFFUSION

Taxonomy	Sample	Prompt
	      	headshot portrait of a headshot portrait of a man headshot portrait of a young man headshot portrait of a young caucasian man headshot portrait of a young caucasian man with short hair headshot portrait of a young caucasian man with short blonde hair headshot portrait of a young caucasian man with short blonde hair, real life, realistic background, 50mm, Facebook, Instagram, shot on iPhone, HD, HDR color, 4k, natural lighting, photography

- Prompt Analysis
- Stable Diffusion v1.x/v2.x/v3.x



# LA NOSTRA PROPOSTA

## Caratterizzazione dei Profili di Bias

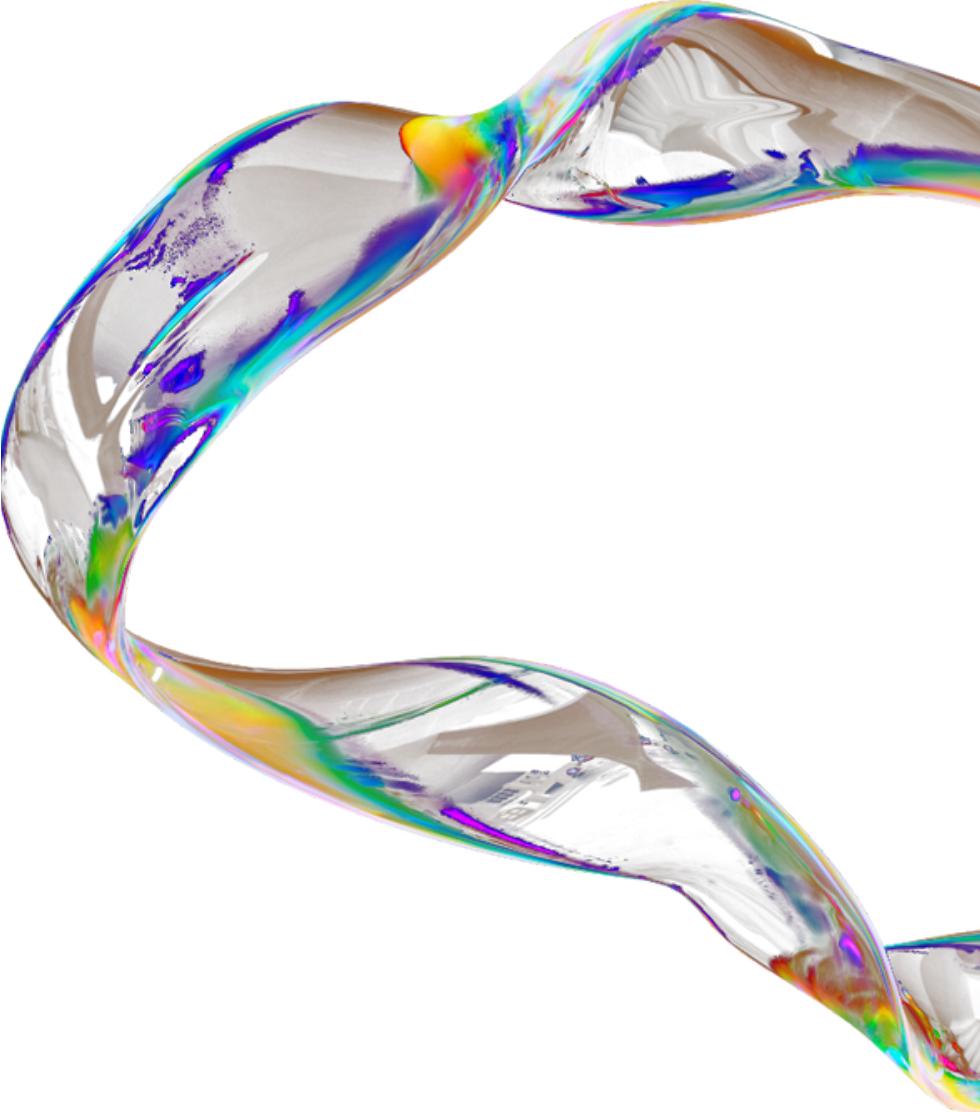
Questa fase del progetto si propone di analizzare in modo sistematico i profili di bias demografico presenti all'interno dei principali dataset utilizzati per la predizione di attributi facciali soft. L'**obiettivo** è identificare e descrivere le disparità di trattamento tra sottogruppi (ad esempio per etnia, genere, età) che emergono durante la fase di apprendimento o inferenza dei modelli.

## Dataset Sintetici come Strumento di Mitigazione

La seconda fase esplora l'utilizzo di **dati sintetici** generati artificialmente come approccio innovativo per la mitigazione dei bias rilevati. L'idea alla base è che tali dati, se opportunamente controllati nelle loro caratteristiche demografiche e semantiche, possano offrire una fonte alternativa e bilanciata per l'addestramento di modelli più equi.

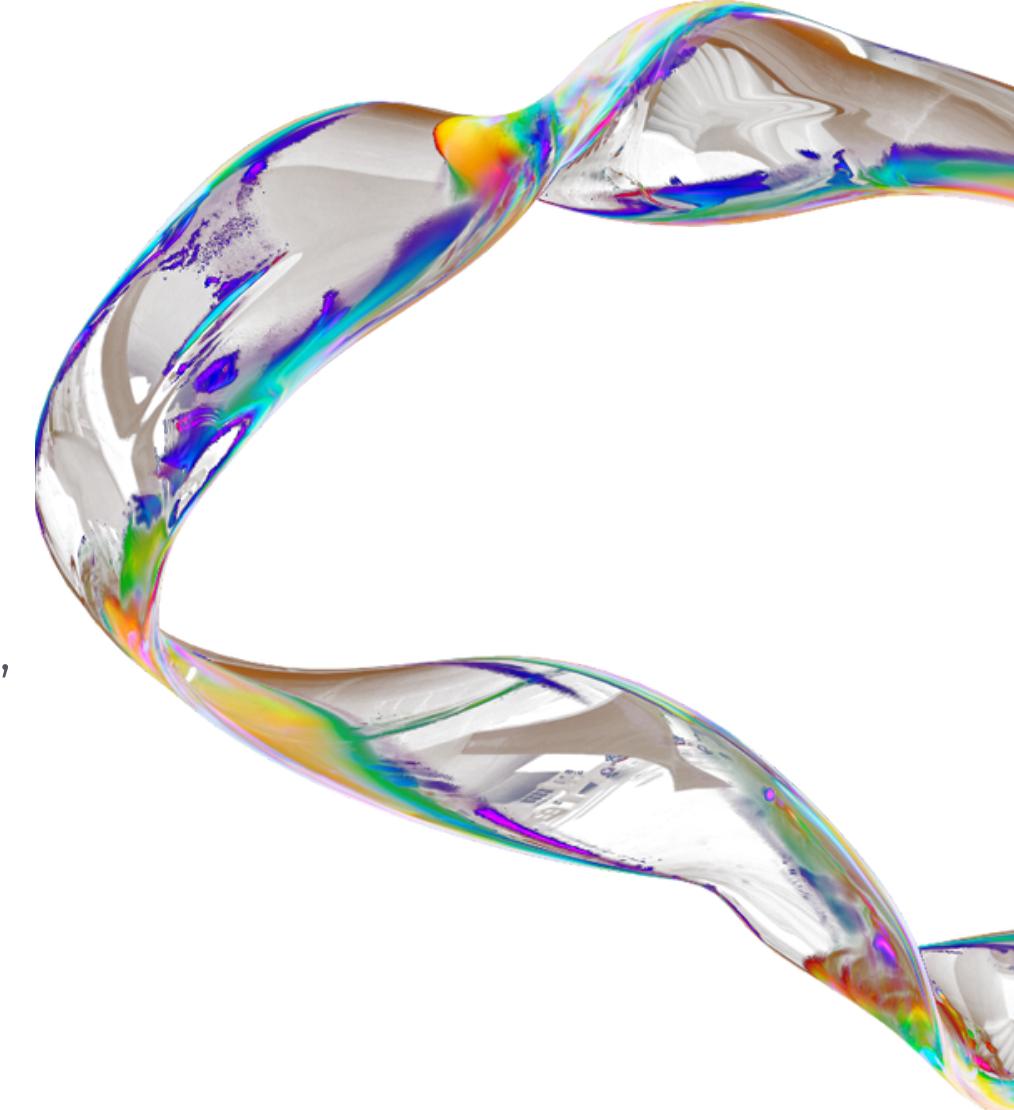
## Obiettivo generale

Ridurre le distorsioni demografiche nei modelli di predizione di attributi facciali soft, tramite l'integrazione di dataset sintetici generati in modo controllato.



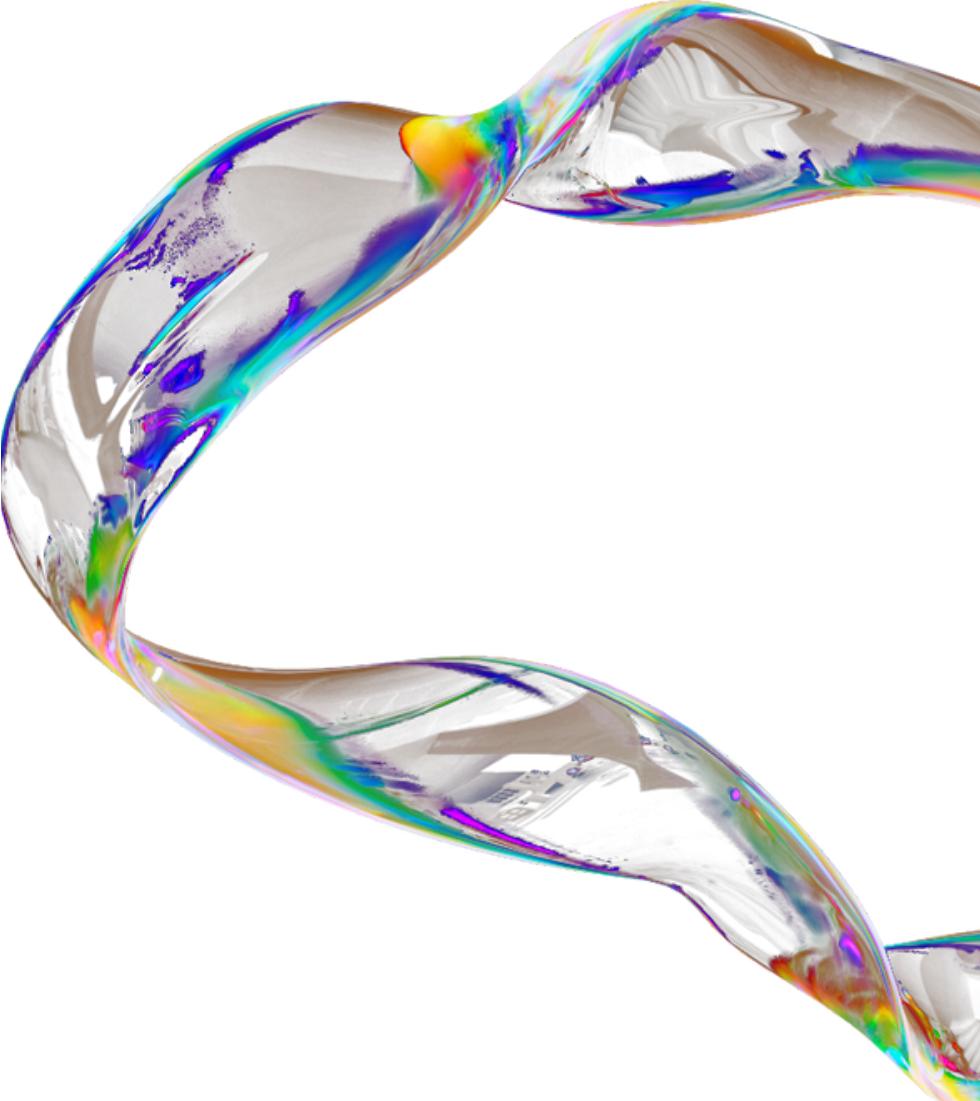
# STEP BY STEP

- Analisi preliminare del dataset CelebA, uno dei più utilizzati per la predizione di attributi facciali. L'obiettivo è identificare la distribuzione demografica dei soggetti (etnia, genere, età), squilibri e bias.
- **Generazione controllata di immagini sintetiche:** creazione di volti artificiali tramite modelli generativi avanzati, come StyleGAN3 e Stable Diffusion. Le immagini sintetiche saranno:
  - Demograficamente controllate (es. variazione di etnia, età, genere),
  - Anotate automaticamente con attributi facciali soft (es. smiling, wearing glasses, young),
  - Utili per coprire i gruppi minoritari sottorappresentati nel dataset reale.
- **Costruzione di dataset synthetic-aware:** integrazione delle immagini sintetiche nei dataset reali per creare insiemi di dati bilanciati e rappresentativi. Questa fase mira a:
  - Ampliare la copertura semantica e demografica,
  - Preparare i dati per un nuovo addestramento del modello.
- **Riallenamento del modello di attributi facciali (FAR):** esecuzione del retraining dei modelli di predizione degli attributi facciali utilizzando il nuovo dataset bilanciato.
- **Valutazione multi-gruppo:** valutazione comparativa delle prestazioni ottenute prima e dopo l'introduzione dei dati sintetici.



## STEP BY STEP: DETAILS

- **Generazione controllata di immagini sintetiche:** StyleGAN3, Stable Diffusion.
- **Costruzione di dataset synthetic-aware:** Prompt analysis.
- **Riallenamento del modello di attributi facciali (FAR):** SLIM-CNN (<https://github.com/gtamba/pytorch-slim-cnn>).



**GRAZIE**

