On Human and Algorithmic Biases

NURIA OLIVER, PhD

Cofounder and Director of ELLIS Alicante – The Institute of Humanity-centric AI
Cofounder and Vicepresident of ELLIS, The European Laboratory for Learning and Intelligent Systems
Chief Data Scientist - Data-Pop Alliance
Independent Board Member – AESIA
Fellow at the Spanish Royal Academy of Engineering

Work funded by

GENERALITAT VALENCIANA
Conselleria de Innovación, Universidades, Ciencia y Sociedad Digital

EUROPEAN UNION

ELLIAS

intel

Sabadell Fundación

NIPPON GASES

ellis

ALICANTE unit

The Institute of Human(ity)-Centric Artificial Intelligence

https://ellisalicante.org

With funding from

GENERALITAT VALENCIANA
Conselleria de Innovación, Universidades, Ciencia y Sociedad Digital

EUROPEAN UNION

ELIAS

crue Universidades Españolas

Santander Fundación BBVA

intel

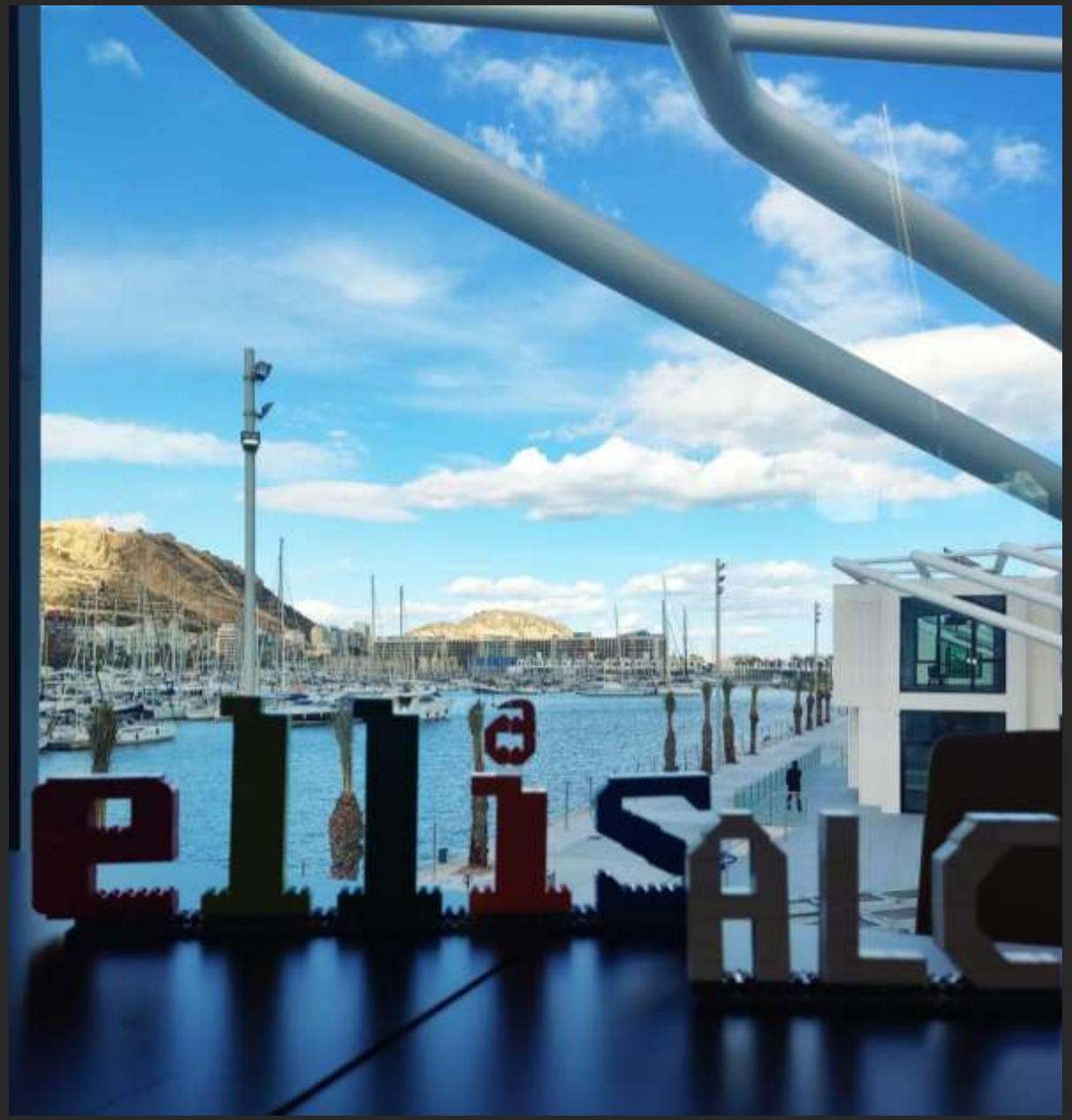
Sabadell Fundación

NIPPON GASES

FUNDACIÓ BALEARIA

41 ELLIS Units in 17 European Countries & Israel


AI FOR GOOD



ELLIS Unit in Alicante is unique: A science start-up

- **Only** ELLIS unit with a clear focus: **AI to drive positive societal impact**

Research on the development of novel machine learning (ML) techniques to automatically recognize, model, and predict individual and aggregate human behavior from data.
Aggregate behavior modeling with a special focus on how AI can help achieve the 17 Sustainable Development Goals and contribute to Social Good.

Development of novel Intelligent, interactive systems

Research on the development of new machine learning approaches to build intelligent user interfaces that interact with humans. Areas of focus include context-sensitive mobile services, novel mobile applications to help people, persuasive computing, wearables, and personal assistants (*chatbots*).

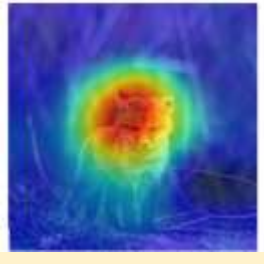
Research on how to address the **ethical challenges** of current AI systems, their risks, and the potential negative consequences stemming from the use of machine learning tools. Areas of work include research on algorithmic **discrimination**, lack of **transparency** and **veracity,** computational violations of **privacy**, subliminal **manipulation** of human behavior, **fragility** of AI systems, and the **social impact** of AI systems that are widely used on social platforms, services, and mobile *apps* used by billions of people.

Overview of Projects

Human cognitive biases and AI: Impact of AI on cognitive biases; Cognitive biases in MLLMs

Fairness, biases, stereotyping: Algorithmic fairness
AI Regulation

Education: Maike, an LLM-based educational chatbot to promote critical thinking
Mental Health: Risks and limitations of LLMs in mental health



Explainable AI: Human-centric approaches to Explainable AI

(M)LLMs: Representation biases in training corpora; Safety risks of LLMs; Ethical implications in human-LLM interaction; Prompt moderation;

Social media: Technical study of AI-based beauty filters; Content moderation algorithms, and Text-to-Image generation

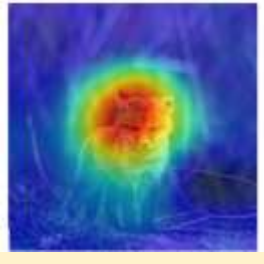
Privacy (Federated Learning or FL): Client selection in FL; Model integration in heterogeneous FL; Data heterogeneity in FL (FedDiverse)

Overview of Projects

Human cognitive biases and AI: Impact of AI on cognitive biases; Cognitive biases in MLLMs

Fairness, biases, stereotyping: Algorithmic fairness
AI Regulation

Education: Maike, an LLM-based educational chatbot to promote critical thinking
Mental Health: Risks and limitations of LLMs in mental health



Explainable AI: Human-centric approaches to Explainable AI

(M)LLMs: Representation biases in training corpora; Safety risks of LLMs; Ethical implications in human-LLM interaction; Prompt moderation;

Social media: Technical study of AI-based beauty filters; Content moderation algorithms, and Text-to-Image generation

Privacy (Federated Learning or FL): Client selection in FL; Model integration in heterogeneous FL; Data heterogeneity in FL (FedDiverse)

Outline

- Biases in social media
  - Beauty filters
- Cognitive biases
  - In humans in the context of AI
  - In ML algorithms

Outline

- Biases in social media
  - Beauty filters
- Cognitive biases
  - In humans in the context of AI
  - In ML algorithms

Biases in the Beautyverse

Piera Riccio

Graduating ELLIS PhD Student

We live in a world of Social Media

- There are **4.76 billion** social media users which are **59.4%** of the world's population.
- **Half** of our time on our phones in 2022 was spent on social media
- An average user spends **2 hours and 31 minutes** daily on social media.
- 26% of social media platform users are aged **18-29 years**
- Teens showed an increase in their **daily screen** time from 7 hours and 22 minutes to **8 hours and 39 minutes** in 2022
- China and India have the largest number of social media users with 1.02 billion and 722 million, respectively, followed by the US with 302 million users

Self-representation in the Digital World: The Beautyverse

- **Self-representation** is of paramount importance in social media platforms
- Thanks to AI techniques, we can **beautify ourselves** in real-time using AR-based filters applied to our selfies

Self-representation in the Digital World: The Beautyverse

- **85%** of girls have downloaded a filter or used an app to change how the look in photos by age **13** (2020, Dove self-esteem project)
- **90%** of young women reported using filters or editing their photos (Univ London Gender and Sexualistis Research Center, 2021)
- These filters contribute to the definition and adoption of **new facial aesthetics**, with significant social and cultural impact, such as an exponential increase in **teen plastic surgeries** and **mental health issues**

Beauty Filters: Challenges

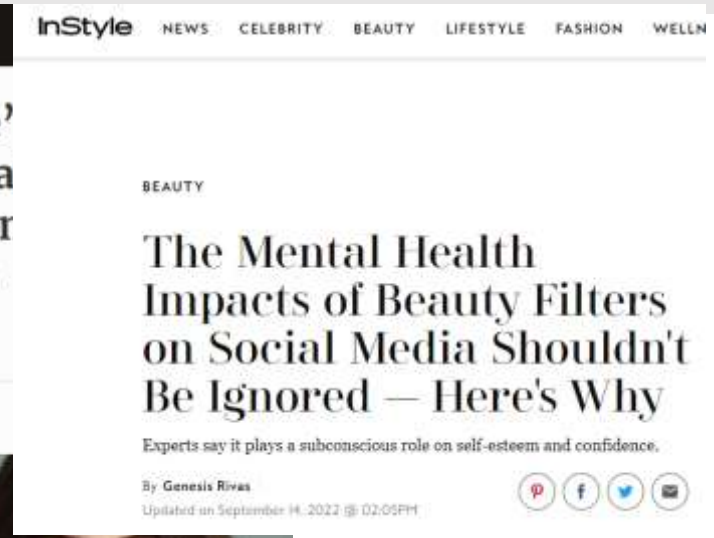
Societal:

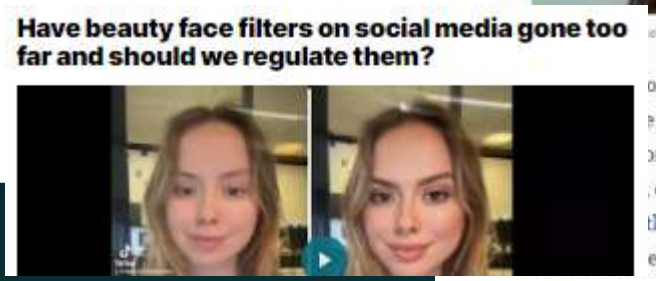
Impact of self-perception leading to dysmorphia, cognitive dissonance, plastic surgery, anxiety…

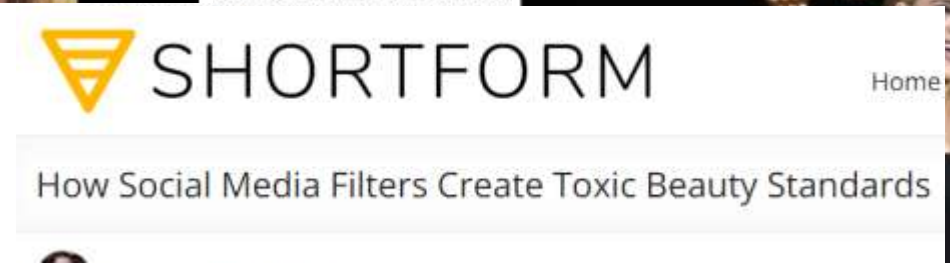
AI-enabled definition of canons of beauty

Technical:

Lack of research datasets to study the characteristics of these filters



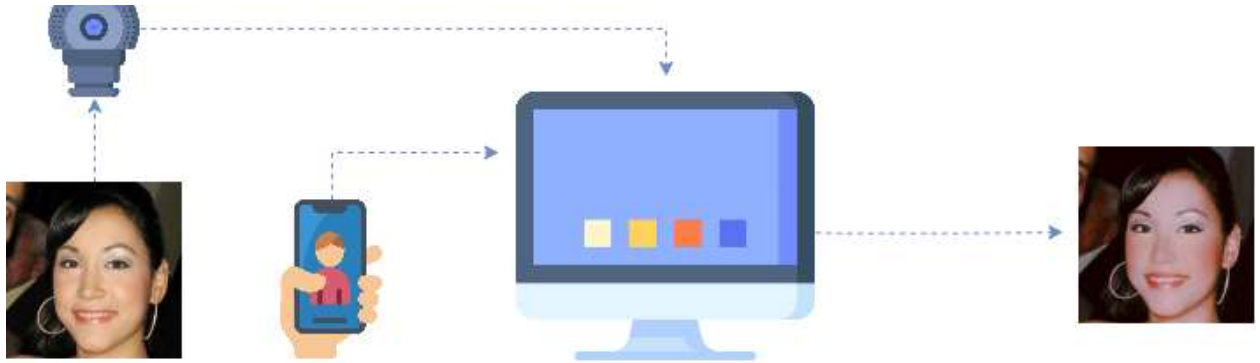


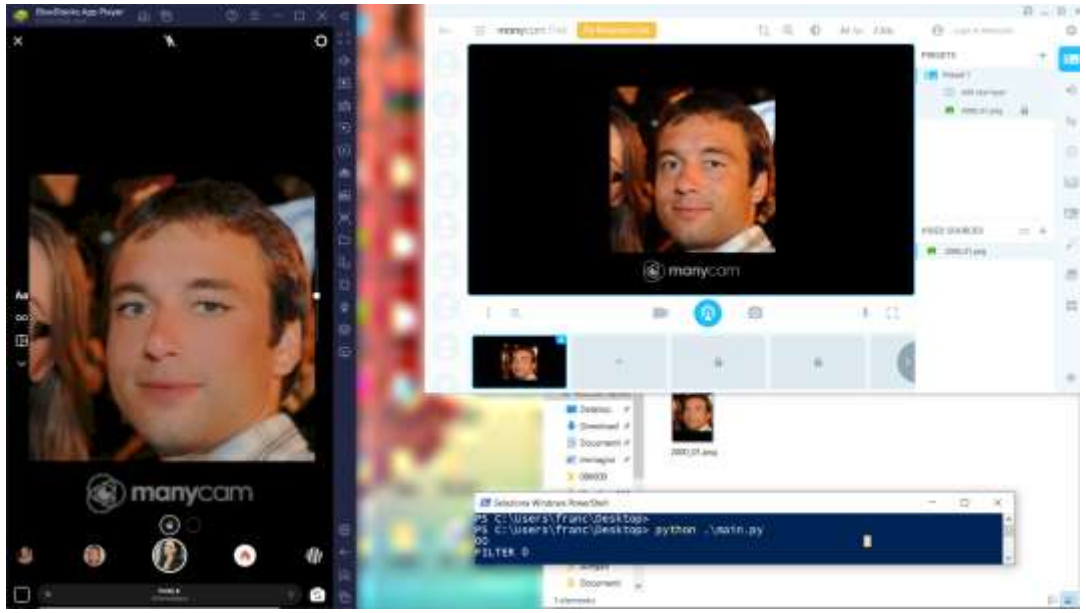


Open Filter

Beauty filters are very hard to study due to a lack of publicly available data

Open Filter: A custom framework to apply social media AR-based filters on existing public collections of faces





We created **two publicly available datasets** of **beautified faces:** FairBeauty and B-LFWA

Riccio, P., Psomas, B., Galati, F., Escolano, F., Hofmann, T., & Oliver, N. (2022). OpenFilter: a framework to democratize research access to social media AR filters. Advances in Neural Information Processing Systems, 35, 12491-12503.

Beauty Datasets

We selected **8 beauty filters** based on their popularity on social media and we have beautified two different existing datasets:

- FairFace [1], a diverse and fair collection of faces, generating **FairBeauty.**
- LFW [2], benchmark dataset for face verification, generating **B-LFW.**

Example of the eight different beauty filters applied to the left-most image [2]. From left to right and top to bottom: filter 0 "pretty" by herusugiarta; filter 1 "hari beauty" by hariani; filter 2 "Just Baby" by blondinochkavika; filter 3 "Shiny Foxy", filter 4 "Caramel Macchiato" and filter 5 "Cute baby face" by sasha_soul_art; filter 6 "Baby_cute_face_" by anya__ilicheva; filter 7 "big city life" by triutra.

https://ellisalicante.org/datasets/OpenFilter

[1] Kimmo Karkkainen and Jungseock Joo. 2021. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 1548–1558.

[2] Huang, G. B., Mattar, M., Berg, T., & Learned-Miller, E. (2008, October). Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In Workshop on faces in'Real-Life'Images: detection, alignment, and recognition.

Attractiveness and Beauty Filters: Research Questions

RQ1 – Do beauty filters reduce diversity?

RQ2 – Do beauty filters impact recognizability?

"OpenFilter: A Framework to democratize Research Access to Social Media AR filters"
Piera Riccio, Bill Psomas, Francesco Galati, Francisco Escolano, Thomas Hofmann, Nuria Oliver, NeurIPS 2022 – Datasets and Benchmark track

RQ3 – Are there racial biases encoded in the beauty filters?

"Mirror, Mirror on the Wall, Who Is the Whitest of All? Racial Biases in Social Media Beauty Filters
P Riccio, J Colin, S Ogolla, N Oliver, Social Media+ Society 10 (2), 20563051241239295"

RQ4 – What is the interplay between beauty filters and cognitive biases?

"What is beautiful is still good", *Gulati et al.* Royal Society Open Science, 2024

RQ5 – Do MLLMs and T2I models exhibit an attractiveness bias?

"Beauty and the Bias: Exploring the impact of attractiveness on MLLMs", *Gulati et al.* AIES, 2025

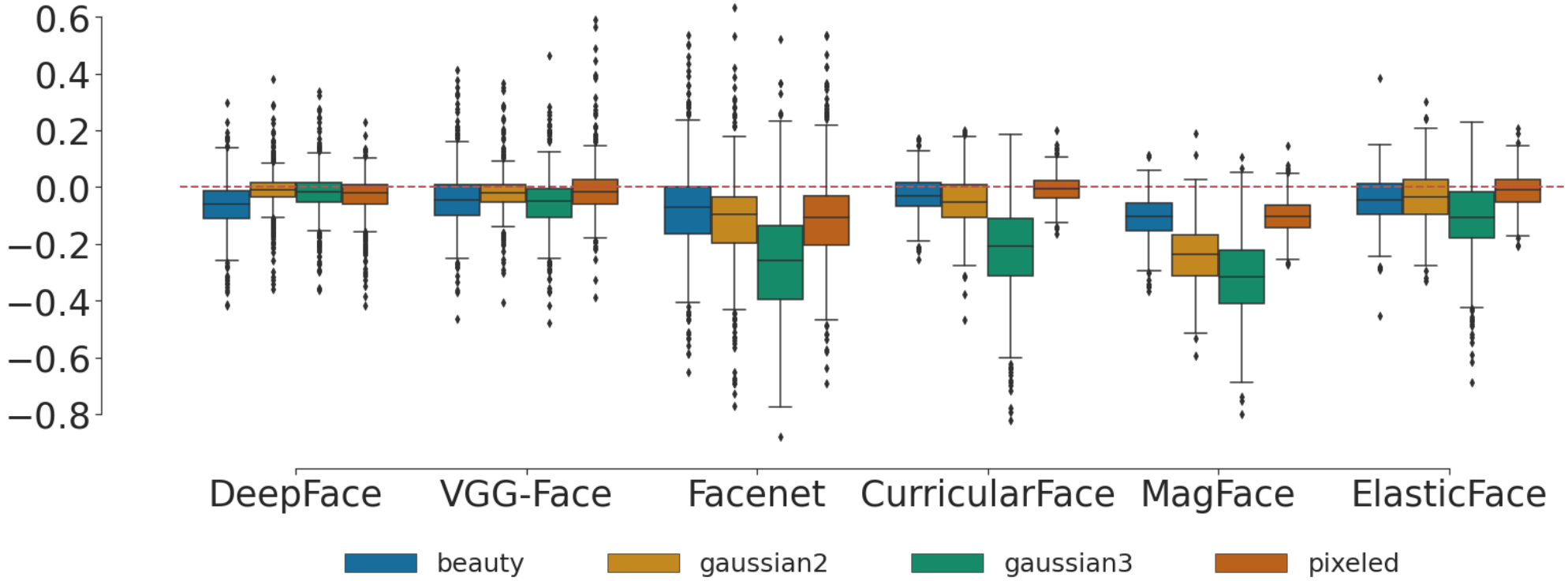
"The Aesthetics of Harm: Algorithmic Lookism in Generative AI and its Systemic Propagation", *Doh et al.* submitted 2025

RQ1: Do beauty filters reduce diversity?

We analyzed **five different versions** of the faces to address RQ1.

From left to right:
A **beautified** version using OpenFilter,
The **original** version,
A **blurred** version with a Gaussian filter at radius 2,
A **blurred** version with Gaussian filter at radius 3,
A **down-sampled** (pixelated) version to 64x64 pixels.

RQ1: Do beauty filters reduce diversity?



Boxplots of the differences in the **distance metric** obtained for filtered image pairs versus the metric obtained for the corresponding original pairs of images.
A value of **0** means that there is **no difference** between the original distance and the distance after applying a transformation to the pairs of faces
A **negative** value indicates that an **image pair** was **more similar** (lower distance metric) when **beautified** as compared to the **original pair**.

RQ1: Do beauty filters reduce diversity?

Paired t-test results comparing the similarity distributions of the original and the beautified faces. Each column corresponds to a different sample of 500 pairs of images, processed with a different model. **All the differences** are **statistically significant**

	DeepFace	VGG-Face	Facenet	CurricularFace	MagFace	ElasticFace
t-statistic	-15.09	-8.428	-10.32	-9.775	-30.63	-11.94
p-value	1.200e-42	3.776e-16	9.561e-23	9.110e-21	1.070e-116	4.400e-29

In all cases, the measurements obtained on the **beautified** pairs of faces have lower average distance than those of the original dataset.

In other words, according to these experiments, the **beautified faces** are **statistically more similar to each other than the original faces**

RQ2: Do beauty filters impact recognizability?

We tested **three state-of-the-art face verification** algorithms (CurricularFace, MagFace and ElasticFace) on the non-beautified and beautified versions of the faces

	CurricularFace	MagFace	ElasticFace
LFW	99.80	99.82	99.80
w/f0	98.93	99.47	99.17
w/f1	99.33	99.42	99.50
w/f2	98.90	99.37	99.35
w/f3	99.13	99.45	99.33
w/f4	99.13	99.45	99.43
w/f5	99.18	99.49	**99.67**
w/f6	98.08	98.42	98.38
w/f7	**96.06**	**96.23**	**96.18**
B-LFW	**99.38**	**99.63**	99.57

RQ3: Do beauty filters encode a racial bias?

- We select two datasets of faces: the first one has original faces (FairFace) and the second the beautified version of the same faces (FairBeauty).

- We test two algorithms, DeepFace[1] and FairFace [2], for race detection on the datasets.

[1] Serengil, Sefik Ilkin, and Alper Ozpinar. "Lightface: A hybrid deep face recognition framework." 2020 innovations in intelligent systems and applications conference (ASYU). IEEE, 2020.
[2] Kimmo Karkkainen and Jungseock Joo. 2021. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 1548–1558.

RQ3: Do beauty filters encode a racial bias?

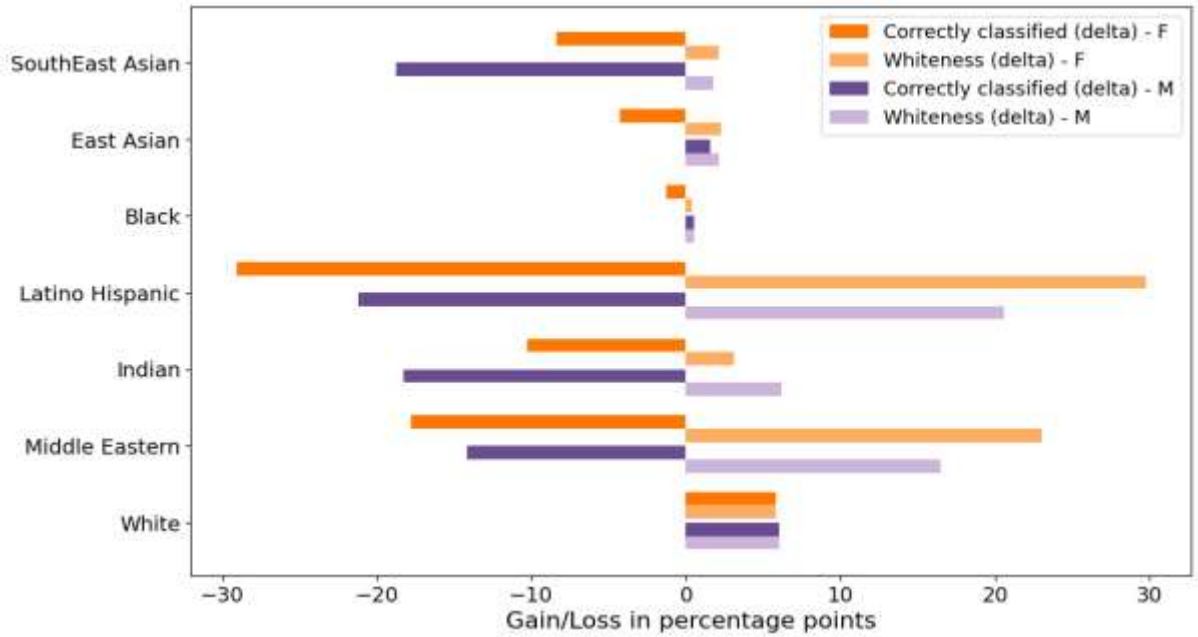
- In addition to the beautified version, we include two blurred versions of the original image

- We compare the performance of the algorithms on the datasets. Is there a shift in the performance on different races? Do these filters enhance characteristics of specific races?

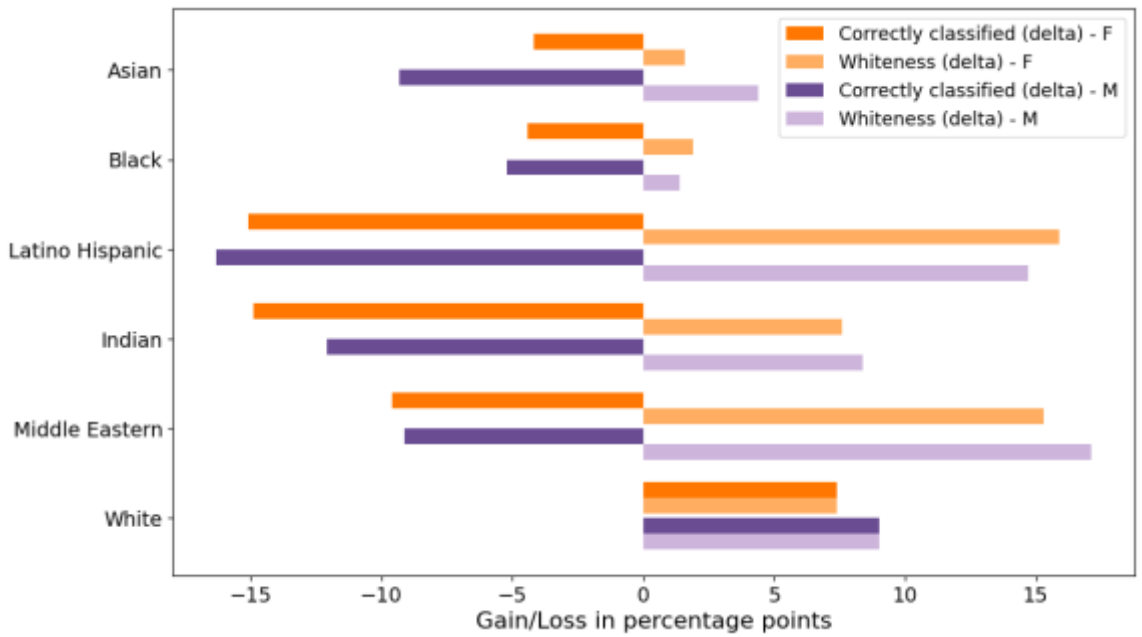
RQ3: Do beauty filters encode a racial bias?

YES: Beauty filters make people look "**whiter**"

DeepFace



FairFace

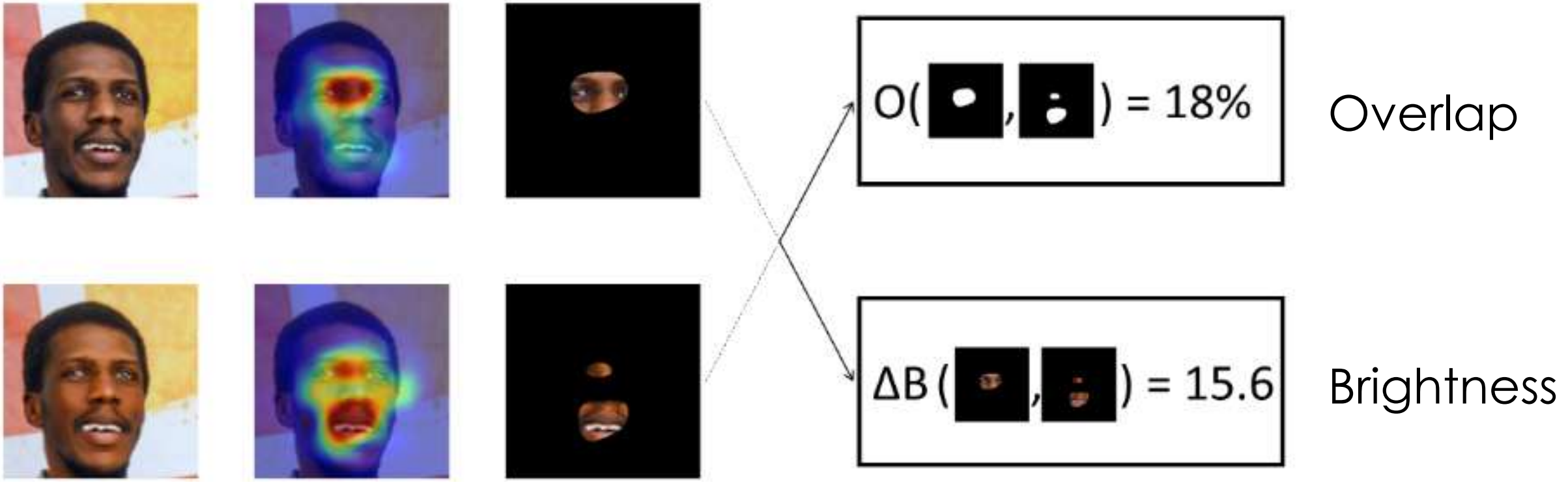


Riccio, P., & Oliver, N. (2022, October). Racial bias in the beautyverse: evaluation of augmented-reality beauty filters. In European Conference on Computer Vision (pp. 714-721). Cham: Springer Nature Switzerland.

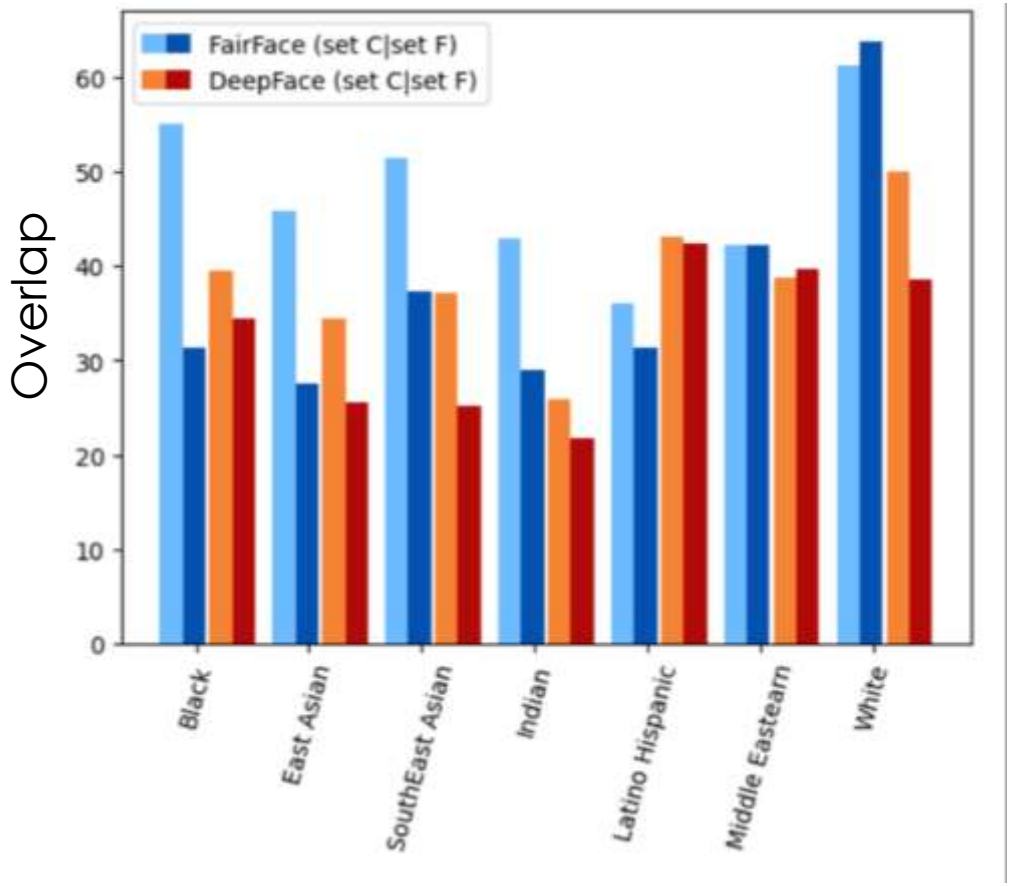
***Riccio, P.**, **Colin, J.**, Ogolla, S., & **Oliver, N.** (2024). Mirror, Mirror on the wall, who is the whitest of all? Racial biases in social media beauty filters. Social Media and Society .*

RQ3: How do beauty filters encode a racial bias?

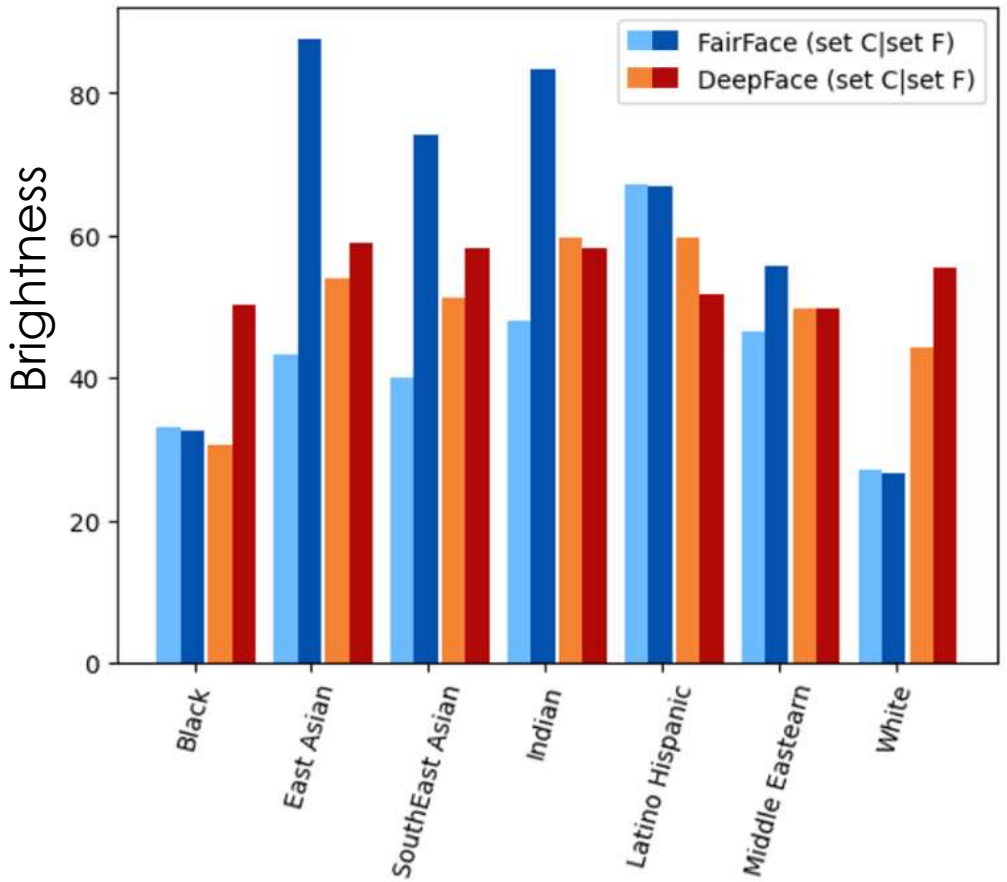
$C \subset X$ is the set of images for which x and xb are classified **correctly**
$F \subset X$ is the set of images for which
(1) x is classified correctly as non-white but xb is classified **incorrectly** as **white** or
(2) x is classified **incorrectly** as **non-white** and xb is classified correctly as white



RQ3: How do beauty filters encode a racial bias?



As a result of the **beautification** process, the race detection algorithms –and especially FairFace– focus on ***different*** facial parts than those used when analyzing the original image x



The parts of the images analyzed by the race classification algorithms to **wrongly** determine the race of the beautified faces (set F) tend to be ***brighter*** than the parts used in the correctly classified images (set C), especially in the case of the **FairFace** algorithm

RQ3: How do beauty filters encode a racial bias?

The Figure highlights two examples from the set F where the FairFace algorithm focuses on the **same** facial features both in x and xb and the focus area is **not brighter** yet xb is misclassified as white but x is correctly classified

The changes made by beauty filters encompass **complex modifications** to the facial features and skin texture or color, beyond a simple brightening of the face.

To Sum Up

RQ1 – Do beauty filters reduce diversity?

✓ YES, they reduce facial diversity

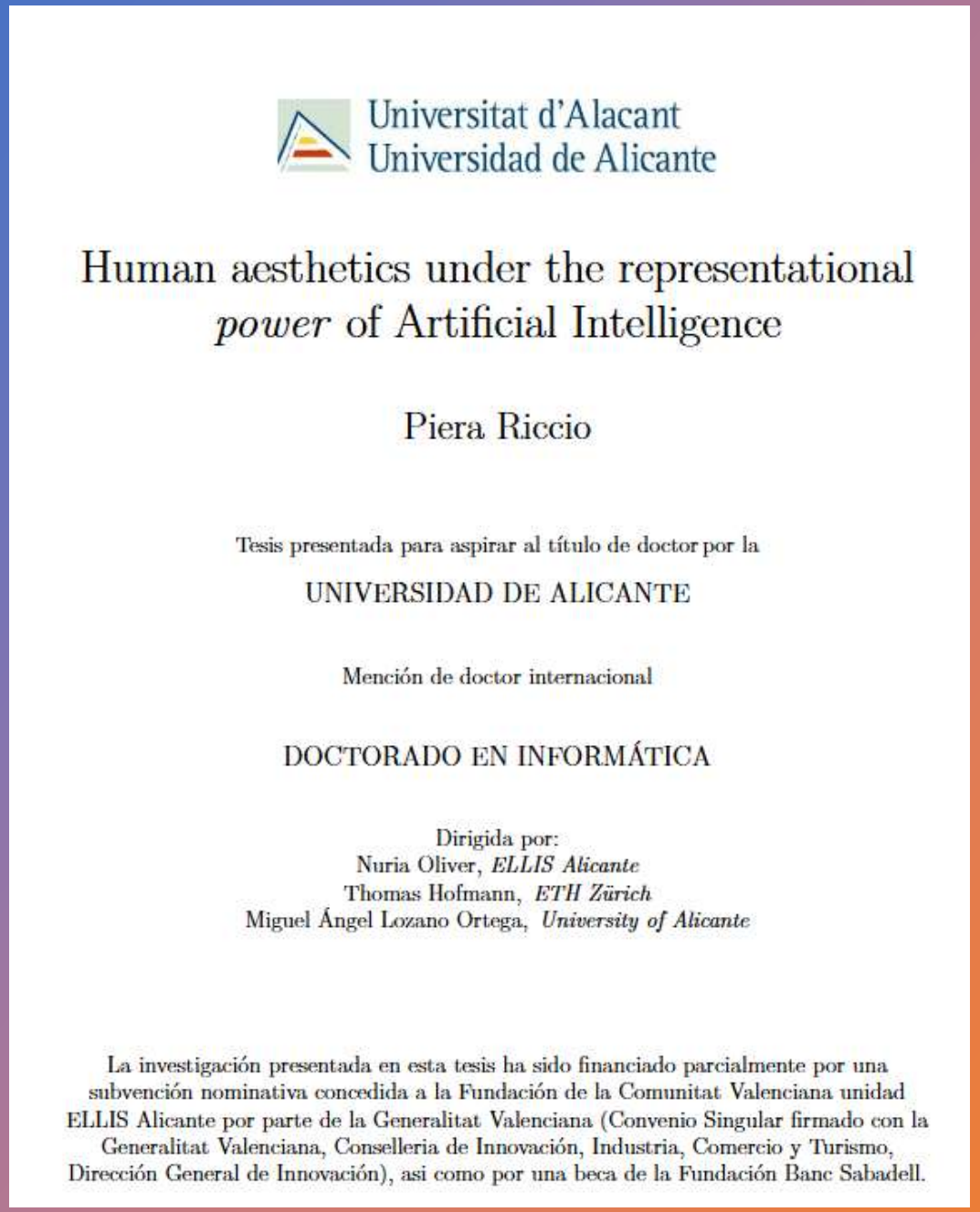
RQ2 – Do beauty filters impact recognizability?

✗ NO, they do not significantly impact recognizability

RQ3 – Is there a racial bias in the beauty filters?

✓ YES, there is a "white" racial bias

The reasons for such bias are complex



Universitat d'Alacant
Universidad de Alicante

Human aesthetics under the representational *power* of Artificial Intelligence

Piera Riccio

Tesis presentada para aspirar al título de doctor por la

UNIVERSIDAD DE ALICANTE

Mención de doctor internacional

DOCTORADO EN INFORMÁTICA

Dirigida por:
Nuria Oliver, *ELLIS Alicante*
Thomas Hofmann, *ETH Zürich*
Miguel Ángel Lozano Ortega, *University of Alicante*

La investigación presentada en esta tesis ha sido financiado parcialmente por una subvención nominativa concedida a la Fundación de la Comunitat Valenciana unidad ELLIS Alicante por parte de la Generalitat Valenciana (Convenio Singular firmado con la Generalitat Valenciana, Conselleria de Innovación, Industria, Comercio y Turismo, Dirección General de Innovación), asi como por una beca de la Fundación Banc Sabadell.

Thesis Defense on **September 22nd!**

International Committee Members:

Prof Nanne van Noord, Univ Amsterdam

Dr. Mia Cha, Max Planck Institute for Security and Privacy

Outline

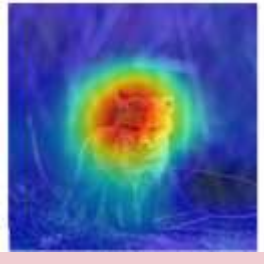
- Biases in social media
  - Beauty filters
- Cognitive biases
  - In humans in the context of AI
  - In ML algorithms

Overview of Projects

Human cognitive biases and AI: Impact of AI on cognitive biases; Cognitive biases in MLLMs

Fairness, biases, stereotyping: Algorithmic fairness
AI Regulation

Education: Maike, an LLM-based educational chatbot to promote critical thinking
Mental Health: Risks and limitations of LLMs in mental health



Explainable AI: Human-centric approaches to Explainable AI

(M)LLMs: Representation biases in training corpora; Safety risks of LLMs; Ethical implications in human-LLM interaction; Prompt moderation;

Social media: Technical study of AI-based beauty filters; Content moderation algorithms and art censorship

Privacy (Federated Learning or FL): Client selection in FL; Model integration in heterogeneous FL; Data heterogeneity in FL (FedDiverse)

Beauty Filters and Cognitive Biases

Attractiveness Halo Effect

Aditya Gulati

Graduating ELLIS PhD Student

Human Cognitive Biases

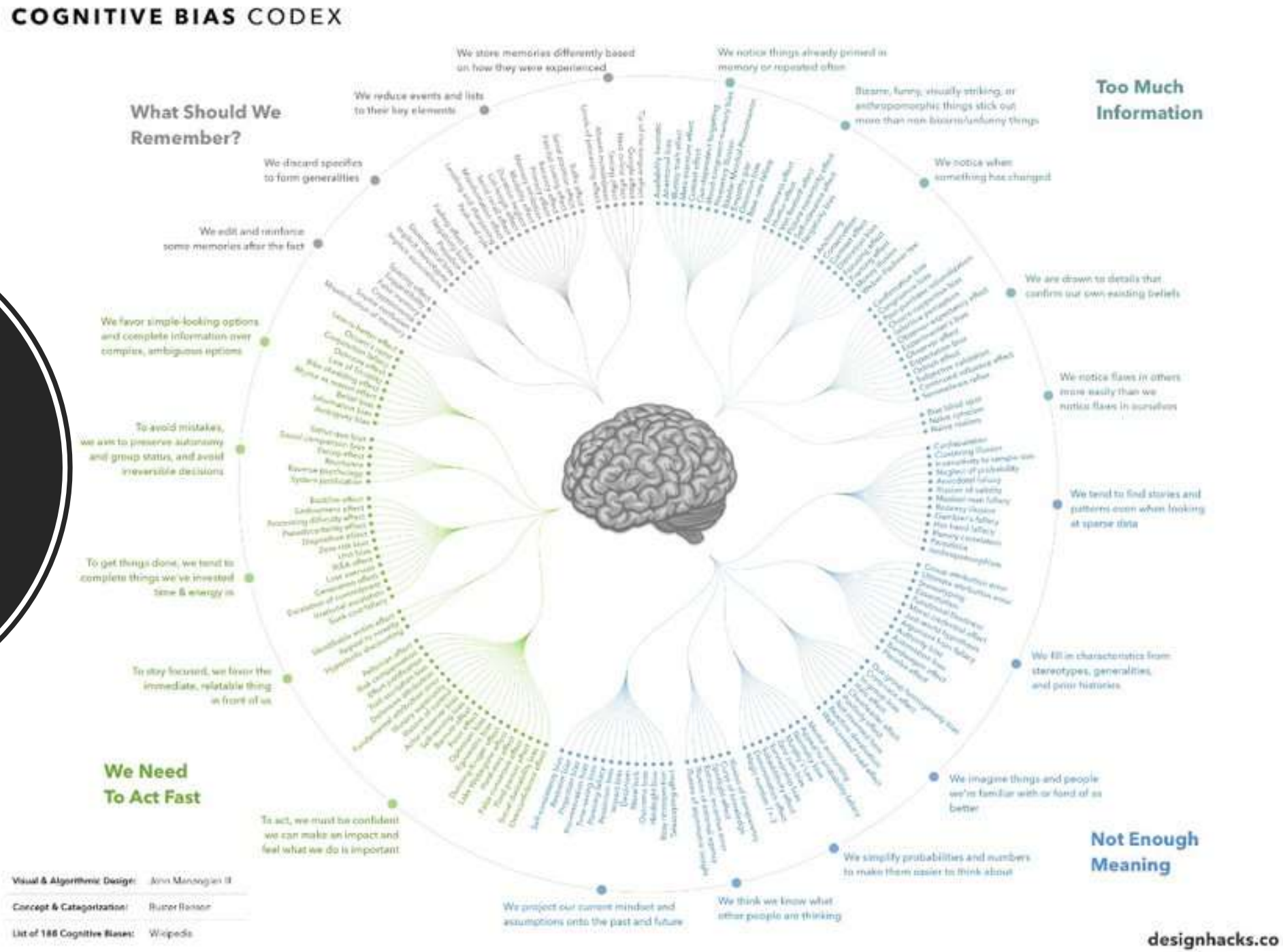
A **cognitive bias** is a **systematic** pattern of deviation from norm or rationality in judgment. Individuals create their own "subjective reality" from their perception of the input.

An individual's construction of reality, not the objective input, guides their behavior in the world.

Cognitive biases may sometimes lead to **perceptual distortion**, **inaccurate judgment**, **illogical interpretation**, or what is broadly called irrationality.

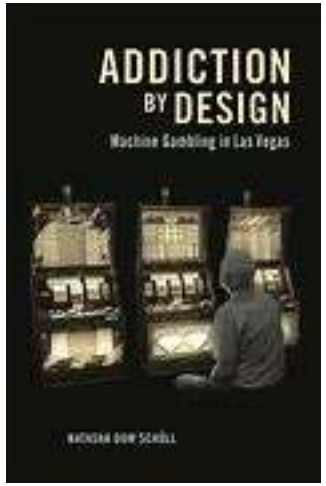
*Kahneman D, Tversky A (1972). "Subjective probability: A judgment of representativeness" (PDF). Cognitive Psychology. **3** (3): 430–454. doi:10.1016/0010-0285(72)90016-3.*

Human Cognitive Biases



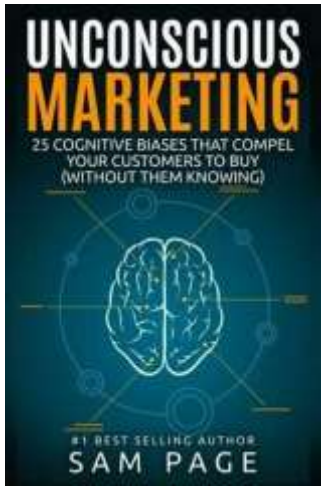
Impact of Cognitive Biases

Casinos



- Gambler's Fallacy
- Denomination Effect
- Travis Syndrome
- Dunning-Kruger Effect
- Illusion of Control
- Illusory Correlation

Marketing



- Loss Aversion
- Status Quo Bias – default
- Anchoring Bias
- Framing Bias
- Ingroup Bias
- Bandwagon Effect

Impact of Cognitive Biases

Elections

- Confirmation Bias
- Coverage Bias
- Concision Bias
- Authority Bias
- Dunning-Kruger Effect
- Availability Cascade
- Framing Effect
- InGroup Effect
- False consensus
- Declinism
- **Halo Effect**

Impact of Cognitive Biases

Judges


TEMPLE LAW
POLITICAL & CIVIL RIGHTS SOCIETY
ABSENTEE BALLOT REGISTER TO VOTE SAMPLE BALLOT SUBMISSIONS ARTICLES ABOUT US
WHO ARE MY REPS?
JUN 05 2020
LEAVE A COMMENT
BY YUNICA JIANG
UNCATEGORIZED
MISJUDGING IN JUDGING: THE ROLE OF COGNITIVE BIASES IN SHAPING JUDICIAL DECISIONS

- Racial Bias
- Anchoring Bias
- Framing Effect
- Gambler's fallacy
- Tiredness and hunger

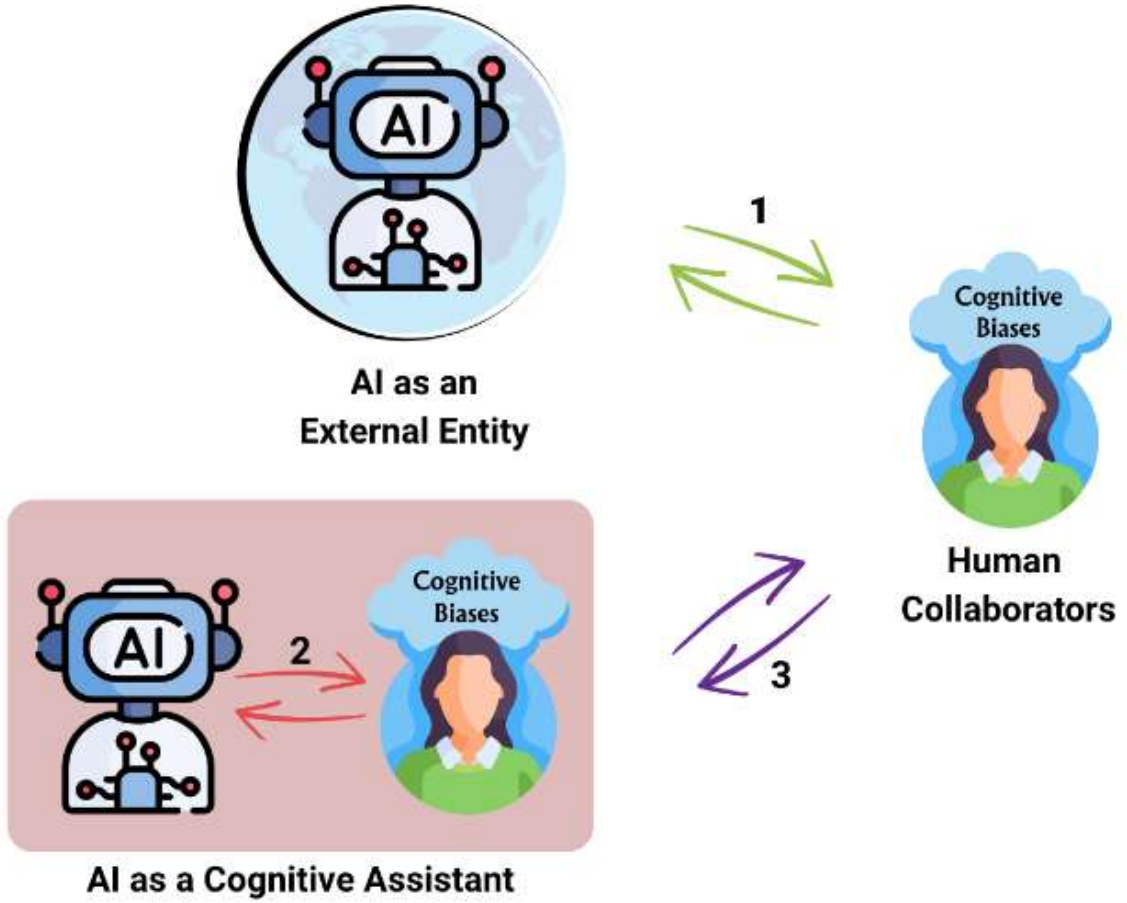
Doctors


MEDICAL EDUCATION
SEP 2020
Believing in Overcoming Cognitive Biases
Tiffany S. Doherty, PhD and Aaron E. Carroll, MD, MS

- Confirmation Bias
- Anchoring Bias
- Affect Heuristic
- Outcomes Bias

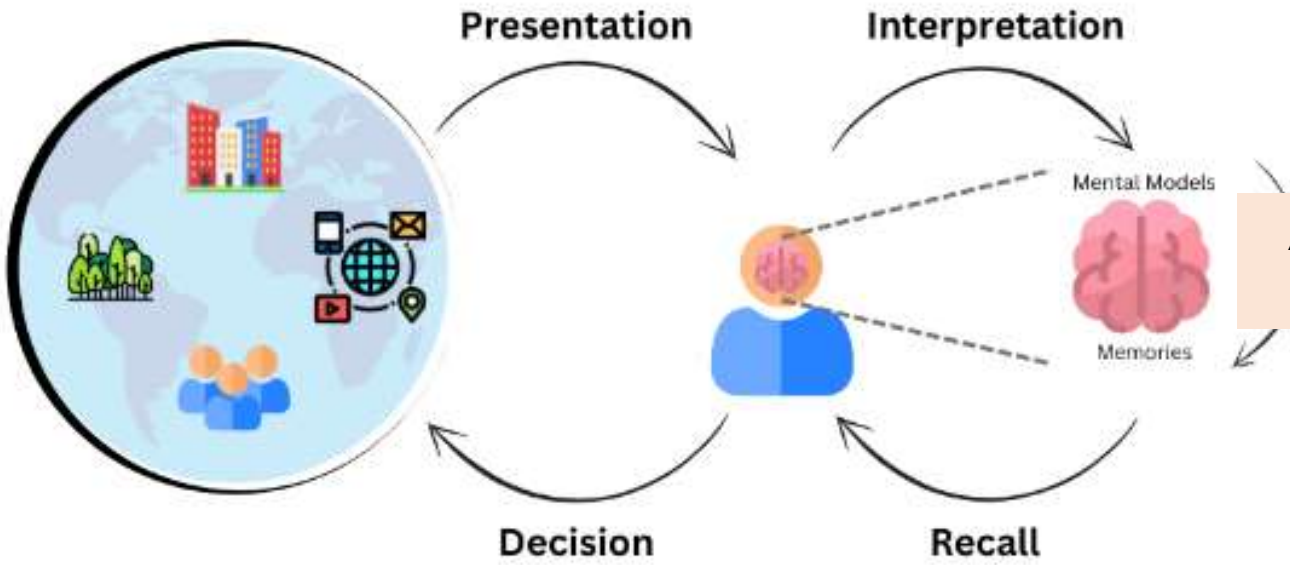
Cognitive Biases and AI Systems

- As we increasingly interact with AI systems and as AI systems make important decisions about us, it is of paramount importance to study the relationship between **our cognitive biases** and **AI algorithms**



Cognitive Biases and AI Systems

- To do so, we need to **classify biases** into categories along the human **decision-making cycle** to aid research in the **interaction** between human cognitive bias and AI systems
- Select biases that
  - Have the **most empirical, scientific** support
  - Are the **most relevant** for the design of AI systems



Under review IEEE Internet Computing Special Issue on Humans Meet AI;
Follow up of "BIASeD: Bringing Irrationality into Automated System Design" presented at AAAI Fall Symposium 2022 on Thinking Fast and Slow and Other Cognitive Theories in AI.

The Attractiveness Halo Effect

- *"What is beautiful is good"* → Attractive individuals are perceived as more intelligent, trustworthy, competent, hirable, less likely to be criminals, etc….
- Appearance biases our decisions even when they should not in domains such as hiring, sentencing, fund raising / investing, credit granting….

Efran, Michael G. "The effect of physical appearance on the judgment of guilt, interpersonal attraction, and severity of recommended punishment in a simulated jury task." *Journal of Research in Personality* 8.1 (1974): 45-54.

Attractiveness and Beauty Filters: Research Questions

RQ1 – Do beauty filters reduce diversity?

RQ2 – Do beauty filters impact recognizability?

> **"OpenFilter: A Framework to democratize Research Access to Social Media AR filters"**
> Piera Riccio, Bill Psomas, Francesco Galati, Francisco Escolano, Thomas Hofmann, Nuria Oliver, NeurIPS 2022 – Datasets and Benchmark track

RQ3 – Are there racial biases encoded in the beauty filters?

> **"Mirror, Mirror on the Wall, Who Is the Whitest of All? Racial Biases in Social Media Beauty Filters**
> P Riccio, J Colin, S Ogolla, N Oliver, Social Media+ Society 10 (2), 20563051241239295"

RQ4 – What is the interplay between beauty filters and cognitive biases?

> ***"What is beautiful is still good",*** *Gulati et al.* Royal Society Open Science, 2024

RQ5 – Do MLLMs and T2I models exhibit an attractiveness bias?

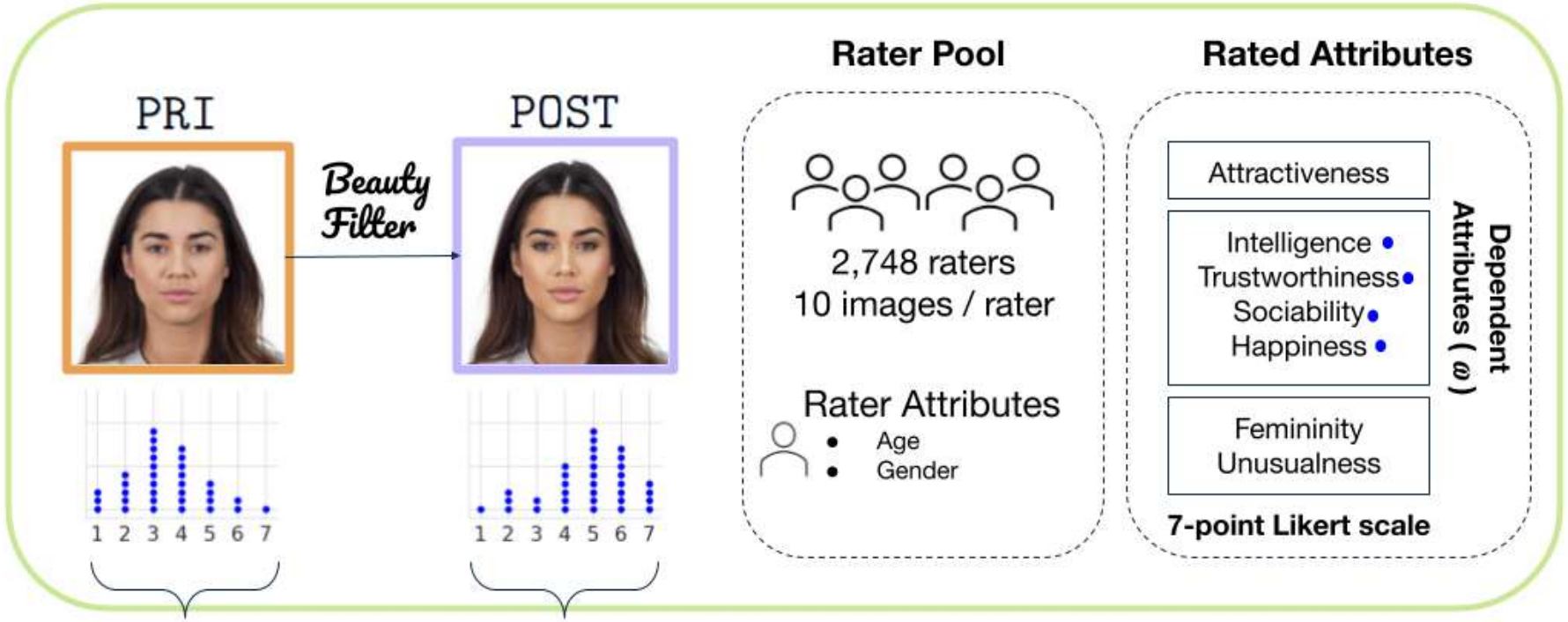
> ***"Beauty and the Bias: Exploring the impact of attractiveness on MLLMs",*** *Gulati et al.* AIES, 2025
>
> ***"The Aesthetics of Harm: Algorithmic Lookism in Generative AI and its Systemic Propagation",*** *Doh et al.* submitted 2025

Our Study: The Beauty Survey

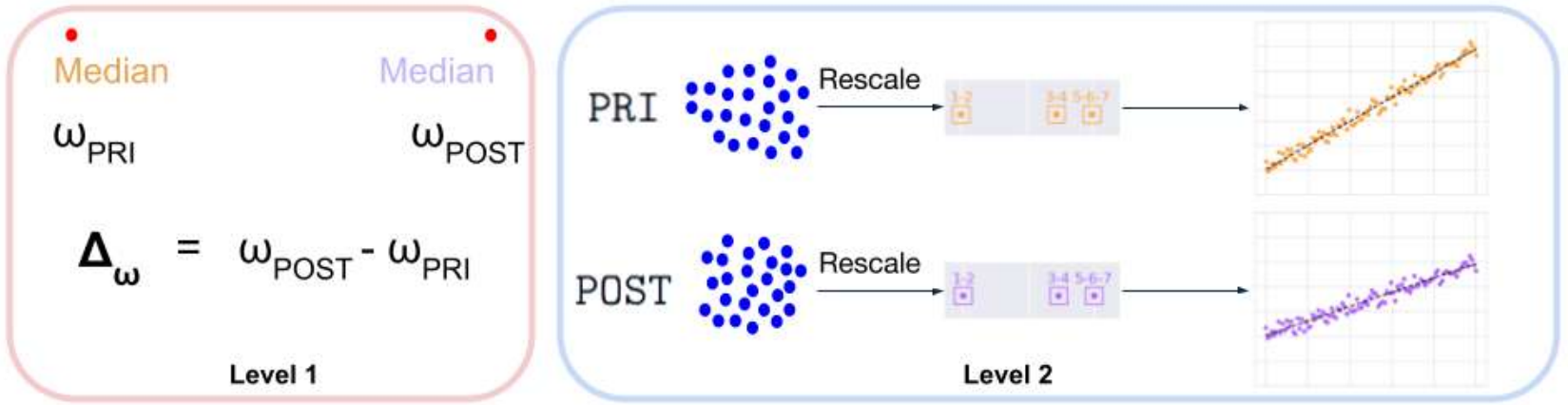
- What happens when we beautify our faces with AI-based beauty filters?
- Images of **462 different individuals** and their beautified counterparts
  - Gender balanced
  - Large diversity and age and ethnicity
  - Neutral expressions, uniform background
- **2748 participants**
  - Everyone rated 10 images
  - 5 with filters, 5 without

Methodology

Data Collection



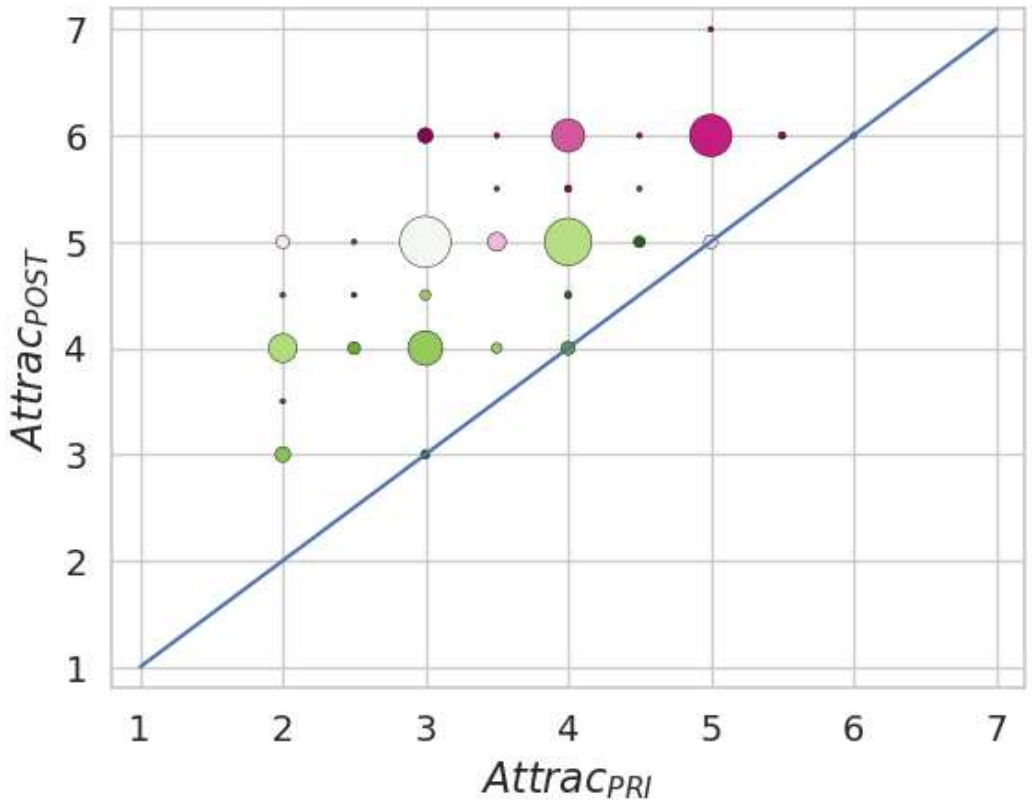
Analysis

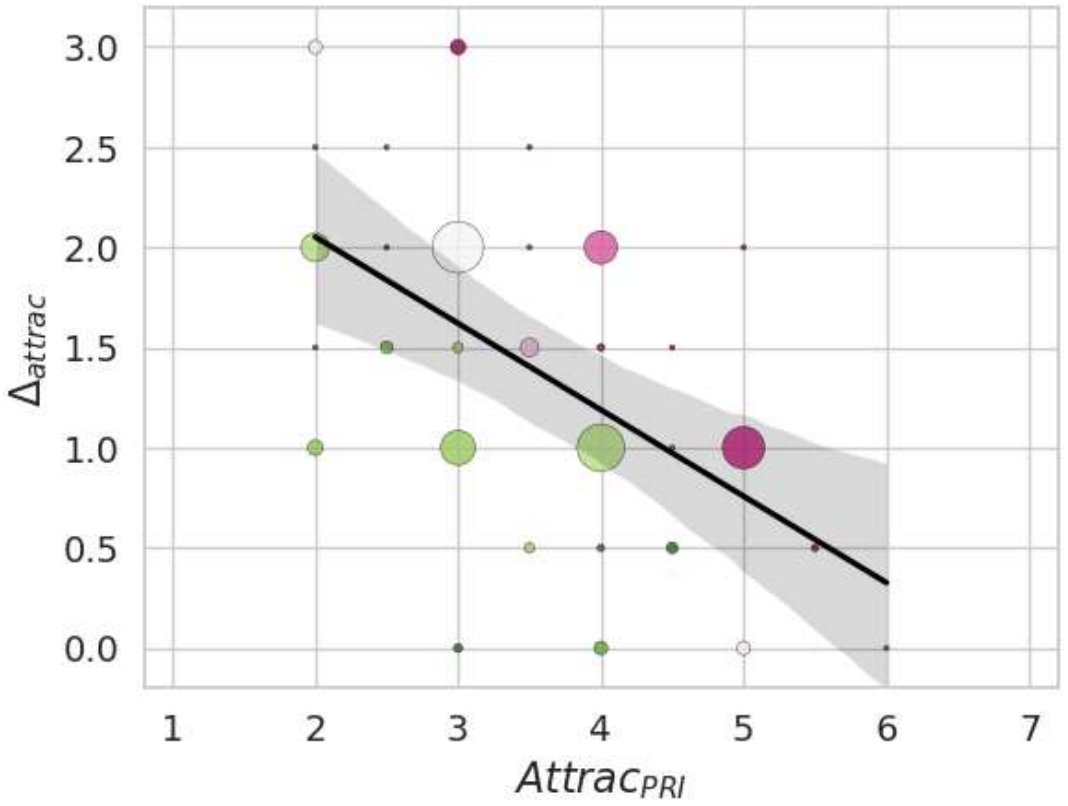


An Example

attractive							
happy							
intelligent							
trustworthy							
sociable							

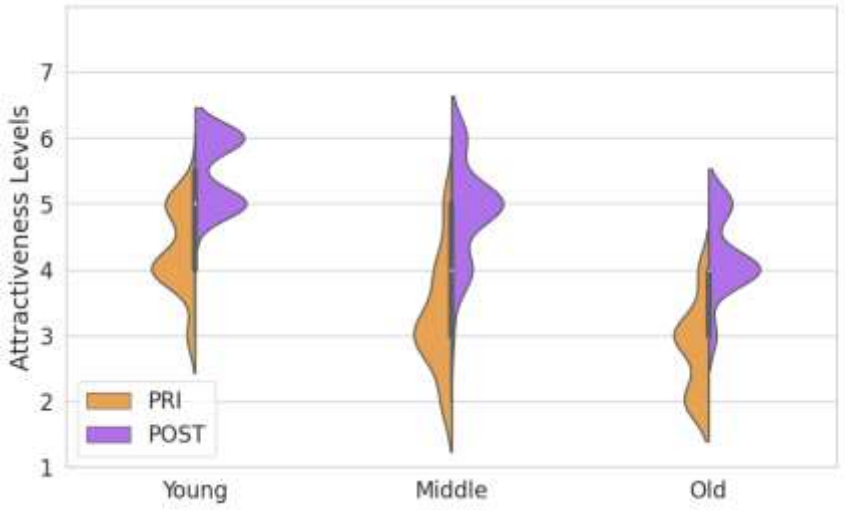
RQ1: Do Beauty Filters work?



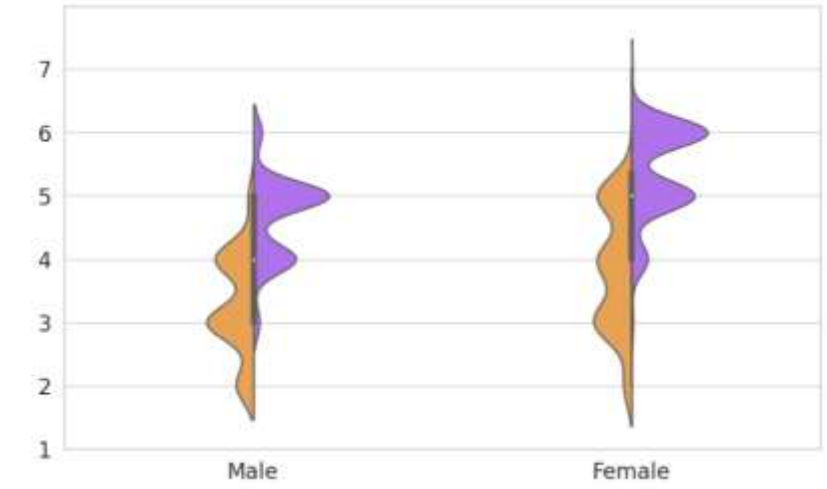


Do Age, Gender and Ethnicity matter?

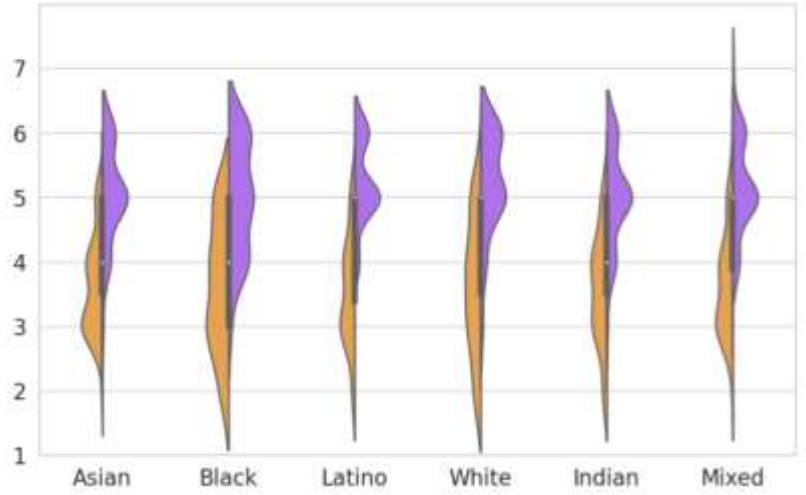
- **Age** matters in perceptions of attractiveness
  - Younger people get higher attractiveness scores



- **Gender** matters in perceptions of attractiveness
  - Females get higher attractiveness scores



- Ethnicity does not!

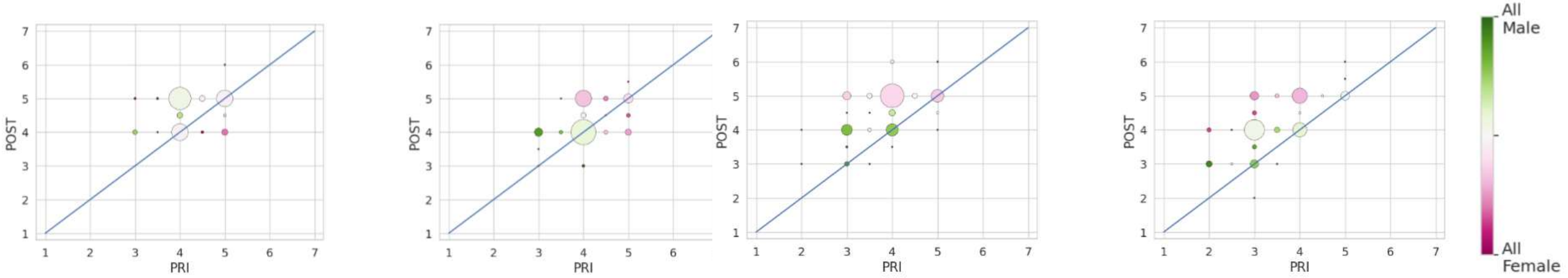


RQ2: Does the attractiveness halo effect exist after beautification?

	Attractiveness	Intelligence	Trustworthiness	Sociability	Happiness
W/n	213.83***	63.15***	33.13***	123.49***	118.57***

Wilcoxon paired rank tests *** *indicates the p-value < 0.001*

YES! Individuals are generally perceived as more attractive, intelligent, trustworthy, sociable and happy after beautification

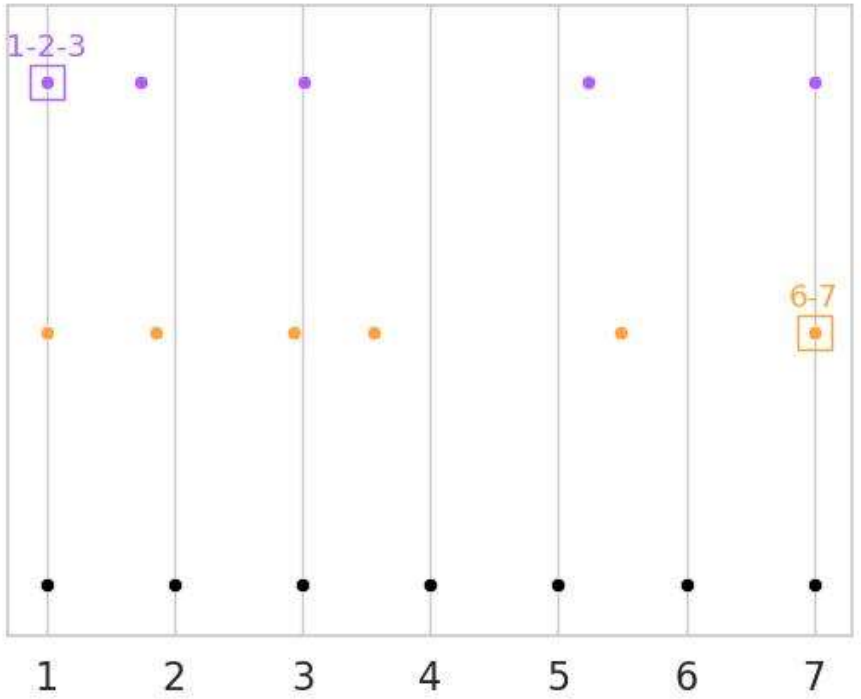


(a) Intelligence (b) Trustworthiness (c) Sociability (d) Happiness

Is beauty in the eye of the beholder?

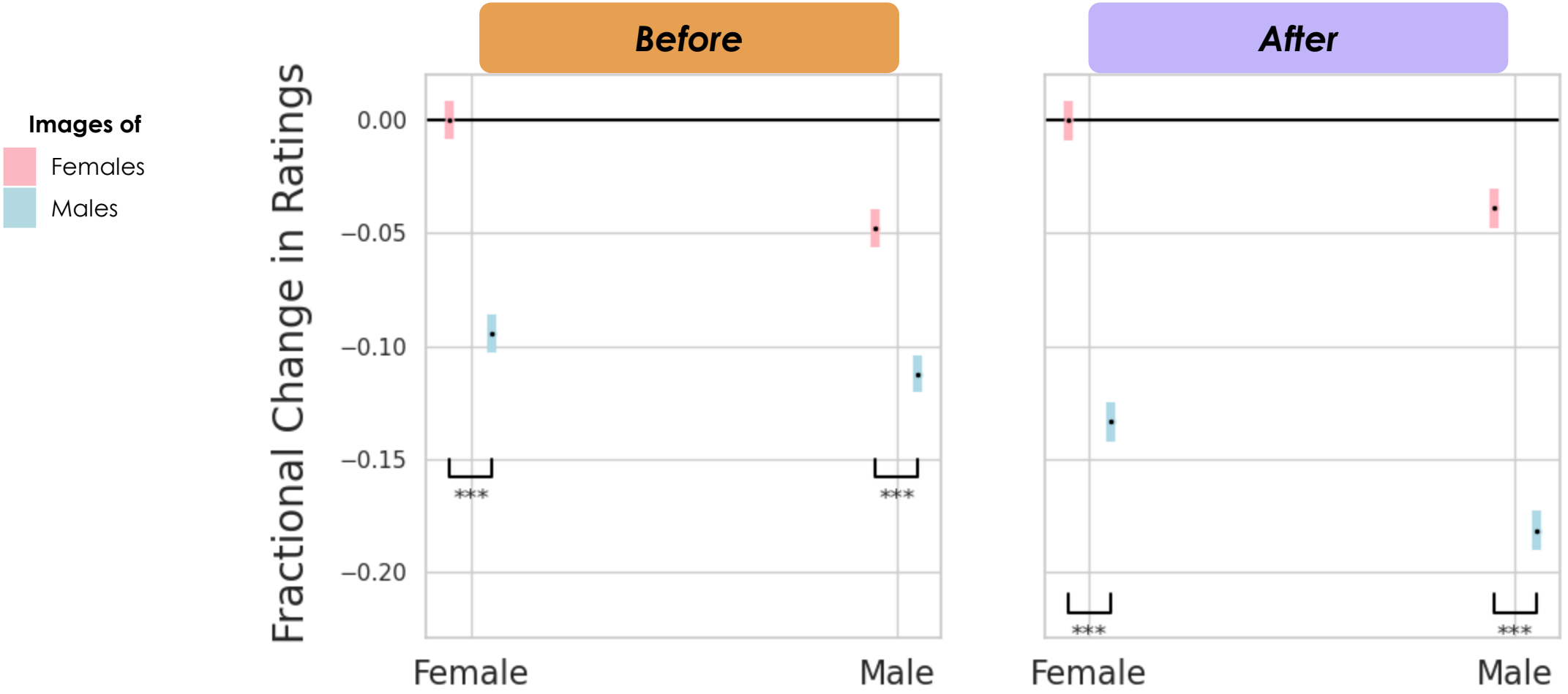
Inclusion of the participant attributes in the models

New scales for attractiveness using OSM's

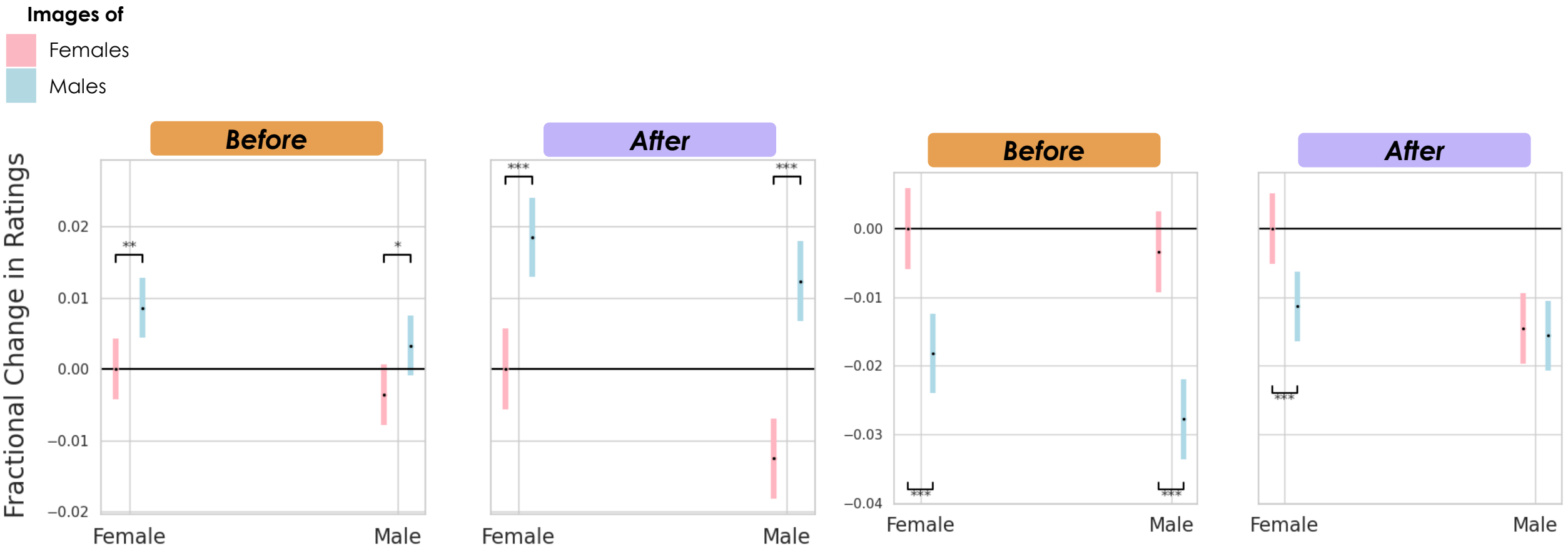


$$\omega = \beta_0 + \beta_1 \cdot Attrac_I + \beta_2 \cdot Gender_I + \beta_3 \cdot Age_I + \beta_4 \cdot Gender_R + \beta_5 \cdot Age_R + \beta_6 \cdot Gender_I \cdot Gender_R + \beta_7 \cdot Age_I \cdot Age_R + RandEff_{Rater}$$

Is beauty in the eye of the beholder?



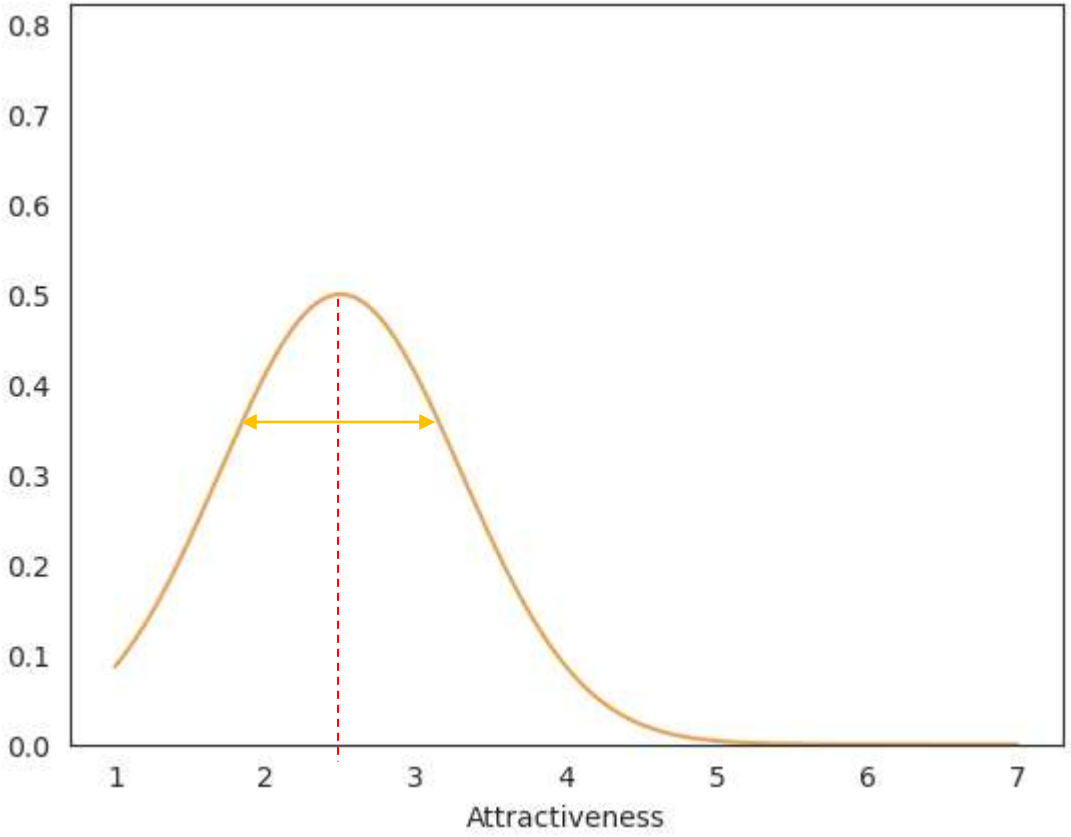
Are other attributes also in the eye of the beholder?



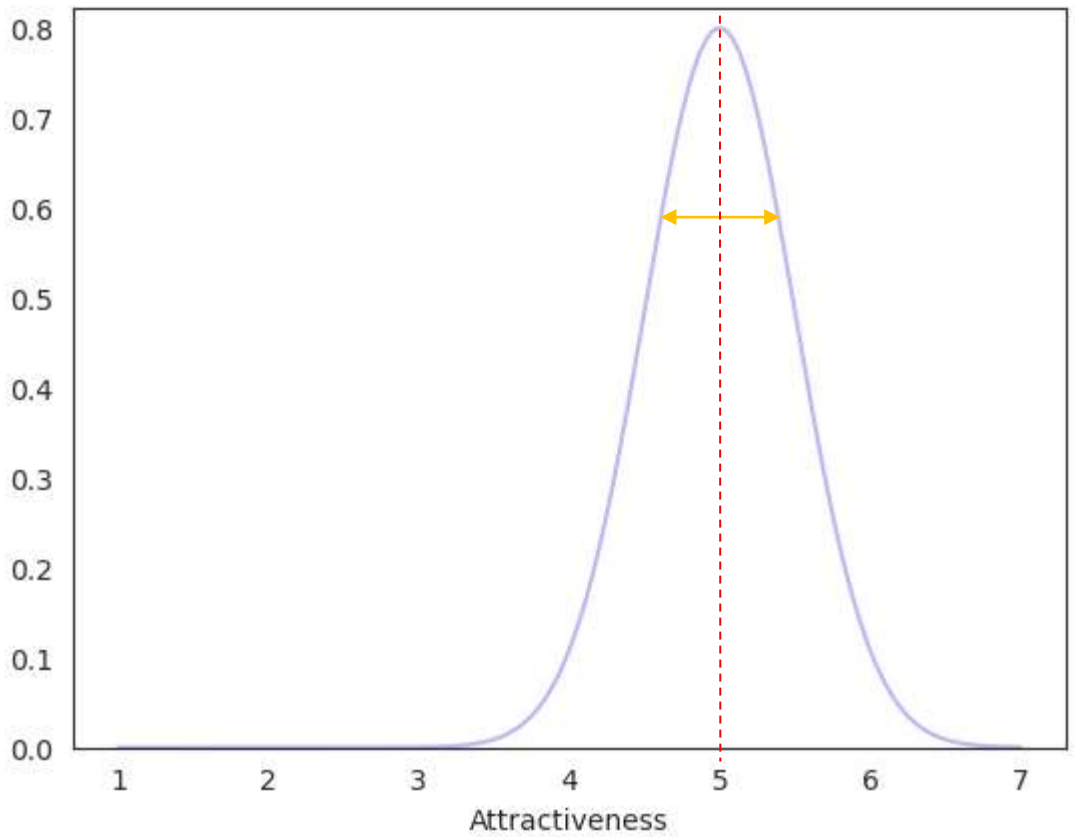
Intelligence

Trustworthiness

Could beauty filters mitigate the halo effect?



Beauty Filter



** Graphs are illustrative and not based on real data from the study*

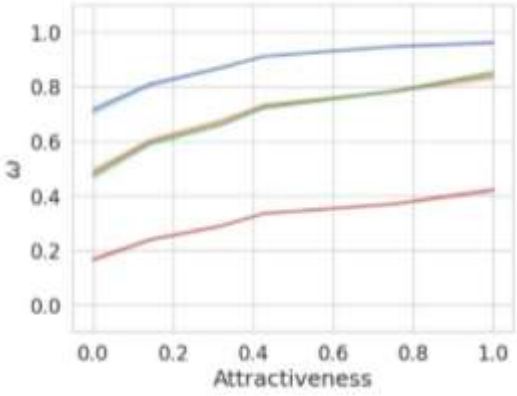
Saturation of the halo effect

$$\omega = \beta_0 + \beta_1 \cdot Attrac + \epsilon$$

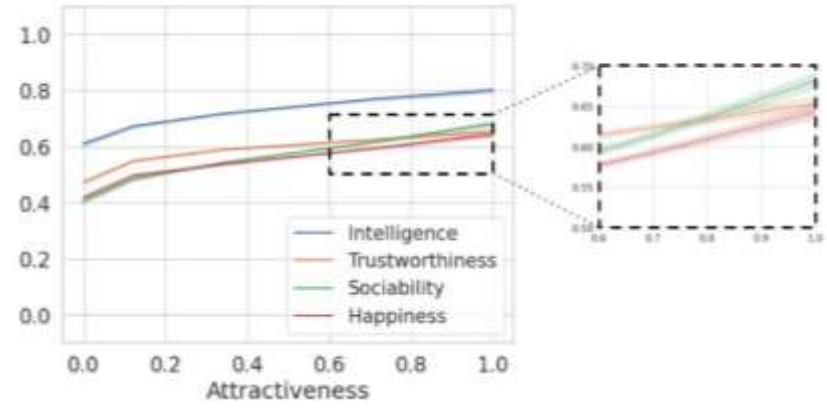
Dependent Attribute (ω)	Before			After		
	β_0	β_1	R^2	β_0	β_1	R^2
Intelligence	3.18***	0.30***	0.327	4.11***	0.12***	0.036
Trustworthiness	3.34***	0.20***	0.181	3.50***	0.17***	0.069
Sociability	2.56***	0.39***	0.363	2.78***	0.38***	0.321
Happiness	2.08***	0.39***	0.261	2.47***	0.35***	0.186

*** represents $p < 0.001$

Before



After



- **Reduced** correlation between attractiveness and (1) intelligence and (2) trustworthiness after beautification i.e., applying a beauty filter helps mitigate the effect
- **Saturation** of the effect in intelligence and trustworthiness
- This phenomenon does not occur with sociability and happiness
- Indicates that **beauty filters could be used to mitigate the attractiveness halo** effect as hypothesized, but not always

Summary of Findings

- Beauty filters **increase perceptions of attractiveness** for almost everyone
- The **same individuals** are perceived as more **intelligent, trustworthy, sociable** and **happier** after **beautification**
- **Age** and **gender** matter in perceptions of attractiveness, **ethnicity** does not!
- Beauty filters **partially mitigate** the attractiveness halo effect
- Beauty filters exacerbate **female gender stereotypes**
- There is a need for **transparency** and **ethical guidelines** surrounding the use of beauty filters in the digital world, including the metaverse

"What is Beautiful is Still Good: The Attractiveness Halo Effect in the era of Beauty Filters", Gulati et al, to Royal Society Open Science Journal, Nov 2024, ***Top 5% world***

ROYAL SOCIETY OPEN SCIENCE

royalsocietypublishing.org/journal/rsos

Research

Cite this article: Gulati A, Martínez-García M, Fernández D, Lozano MA, Lepri B, Oliver N. 2024 What is beautiful is still good: the attractiveness halo effect in the era of beauty filters. *R. Soc. Open Sci.* **11**: 240882.
https://doi.org/10.1098/rsos.240882

Received: 3 June 2024
Accepted: 30 September 2024

Subject Category:
Computer science and artificial intelligence

Subject Areas:
artificial intelligence, human-computer interaction, psychology

Keywords:
cognitive biases, attractiveness halo effect, beauty filters, artificial intelligence, gender stereotypes

Author for correspondence:
Aditya Gulati
e-mail: aditya@ellisalicante.org

THE ROYAL SOCIETY PUBLISHING

What is beautiful is still good: the attractiveness halo effect in the era of beauty filters

Aditya Gulati[1,2], Marina Martínez-Garcia[3], Daniel Fernández[4], Miguel Angel Lozano[2], Bruno Lepri[5] and Nuria Oliver[1]

[1]ELLIS Alicante, Alicante, Spain
[2]University of Alicante, Alicante, Spain
[3]Universitat Jaume I de Castellon, Castellon, Spain
[4]Universitat Politècnica de Catalunya - BarcelonaTech, Barcelona, Spain
[5]Fondazione Bruno Kessler, Trento, Italy

AG, 0000-0002-0356-2987; MM-G, 0000-0002-2228-4396; DF, 0000-0003-0012-2094; MAL, 0000-0002-4757-5587; BL, 0000-0003-1275-2333; NO, 0000-0001-5985-691X

The impact of cognitive biases on decision-making in the digital world remains under-explored despite its well-documented effects in physical contexts. This paper addresses this gap by investigating the attractiveness halo effect using AI-based beauty filters. We conduct a large-scale online user study involving 2748 participants who rated facial images from a diverse set of 462 distinct individuals in two conditions: original and attractive after applying a beauty filter. Our study reveals that the *same* individuals receive statistically significantly higher ratings of attractiveness and other traits, such as intelligence and trustworthiness, in the attractive condition. We also study the impact of age, gender and ethnicity and identify a weakening of the halo effect in the beautified condition, resolving conflicting findings from the literature and suggesting that filters could mitigate this cognitive bias. Finally, our findings raise ethical concerns regarding the use of beauty filters.

1. Introduction

Beauty matters, even when we know that physical attractiveness is not correlated with other measurable traits, such as intelligence [1–3]. In fact, decades of research in several disciplines—

About this Attention Score

In the top 5% of all research outputs scored by Altmetric

MORE...

Mentioned by

- **41** news outlets
- **2** blogs
- **26** X users
- **1** peer review site
- **2** Facebook pages
- **1** Wikipedia page
- **8** Bluesky users

watson
Halo-Effekt: Filter auf Social Media sind problematisch
watson, 04 Dec 2024
Beautyfilter lassen uns tatsächlich attraktiver erscheinen – das gilt besonders für Frauen.

DAGENS NYHETER
Så lurar skönhetsfiltren din hjärna
Dagens Nyheter, 30 Nov 2024
Tiktok planerar att förbjuda så kallade skönhetsfilter. Men hur lurar dessa filter vår hjärna och blir vi verkligen vackrare av...

Heute
Beautyfilter – Frauen gelten als weniger intelligent
Heute, 30 Nov 2024
Laut einer Studie wirken "gefilterte" Gesichter auf Fotos attraktiver. Frauen steigen punkto Intelligenz jedoch schlechter aus...

EL DIA
Los filtros de belleza hacen que las mujeres sean percibidas como menos listas que los hombres, según un estudio
El Dia, 28 Nov 2024
Los investigadores pidieron a 2.748 personas de entre 18 y 88 años que evaluaran con un "juicio rápido" cientos de imágenes de...

Diario de Mallorca
Los filtros de belleza hacen que las mujeres sean percibidas como menos listas que los hombres, según un estudio
Diario de Mallorca, 28 Nov 2024
Los investigadores pidieron a 2.748 personas de entre 18 y 88 años que evaluaran con un "juicio rápido" cientos de imágenes de...

Levante
Los filtros de belleza hacen que las mujeres sean percibidas como menos listas que los hombres, según un estudio
Levante, 28 Nov 2024
Los investigadores pidieron a 2.748 personas de entre 18 y 88 años que evaluaran con un "juicio rápido" cientos de imágenes de...

elPeriódico
Los filtros de belleza hacen que las mujeres parezcan menos listas que los hombres, según un estudio
El Periódico - ES, 28 Nov 2024
TikTok restringirá el acceso a filtros de belleza entre los menores de 18 años Cerca del 90% de las mujeres de entre 18 y 30 años...

the guardian
News story from The Guardian on Thursday 28 November 2024
The Guardian, 28 Nov 2024

BUG
Percepcija privlačnosti u eri filtera ljepote
Bug Online , 27 Nov 2024
Ljepota je važna, čak i kada znamo da fizička privlačnost nije u korelaciji s drugim mjerljivim osobinama, poput inteligencije.

DIE WELT
Soziale Medien: Fotofilter machen nicht nur schöner, sondern auch intelligenter und vertrauenswürdiger - WELT
Die Welt, 27 Nov 2024
In sozialen Netzwerken wollen Menschen attraktiv wirken und bearbeiten ihre Fotos mit Schönheitsfiltern.

From Research to Regulation

ellis ALICANTE UNIT

TikTok and Meta's ban on beauty filters is due any day now. But is it too late?

THE ECONOMIC TIMES | News

English Edition | Today's ePaper

Home BUDGET'25 ETPrime Markets Market Data News Industry Politics Wealth MF Tech Careers Opinion NRI Panache Videos

India Web Stories Economy Politics Newsblogs Elections Defence International More

No more TikTok, Instagram & Facebook for kids under 16 in Australia as govt set to ban social media

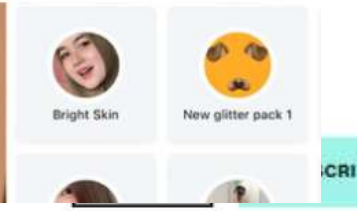
Reuters - Last Updated: Nov 07, 2024, 08:00:00 AM IST

ALL NEWS | SOCIAL MEDIA | ENTERTAINMENT | TECHNOLOGY

Meta Removes Third-Party AR Filters on its Platforms Starting 14 January

By **Oulayvanh Sisounonth** *January 9, 2025*

Share



Review

POLICY

Why Meta is getting sued over its beauty filters

A new sweeping lawsuit takes aim at various Meta features that allegedly endanger children and their privacy. It could have a big impact on child online safety.

Outline

- Biases in social media
  - Beauty filters
- Cognitive biases
  - In humans in the context of AI
  - In ML algorithms

Attractiveness and Beauty Filters: Research Questions

RQ1 – Do beauty filters reduce diversity?

RQ2 – Do beauty filters impact recognizability?

"OpenFilter: A Framework to democratize Research Access to Social Media AR filters"
Piera Riccio, Bill Psomas, Francesco Galati, Francisco Escolano, Thomas Hofmann, Nuria Oliver, NeurIPS 2022 – Datasets and Benchmark track

RQ3 – Are there racial biases encoded in the beauty filters?

"Mirror, Mirror on the Wall, Who Is the Whitest of All? Racial Biases in Social Media Beauty Filters
P Riccio, J Colin, S Ogolla, N Oliver, Social Media+ Society 10 (2), 20563051241239295"

RQ4 – What is the interplay between beauty filters and cognitive biases?

"What is beautiful is still good", *Gulati et al.* Royal Society Open Science, 2024

RQ5 – Do MLLMs and T2I models exhibit an attractiveness bias?

"Beauty and the Bias: Exploring the impact of attractiveness on MLLMs", *Gulati et al.* AIES, 2025

"The Aesthetics of Harm: Algorithmic Lookism in Generative AI and its Systemic Propagation",
Doh et al. submitted 2025

Lookism: an overlooked bias in Computer Vision

Lookism: Preferential treatment of individuals based on their physical appearance.

Algorithmic Lookism: Lookism present in AI-based algorithms

Discriminative

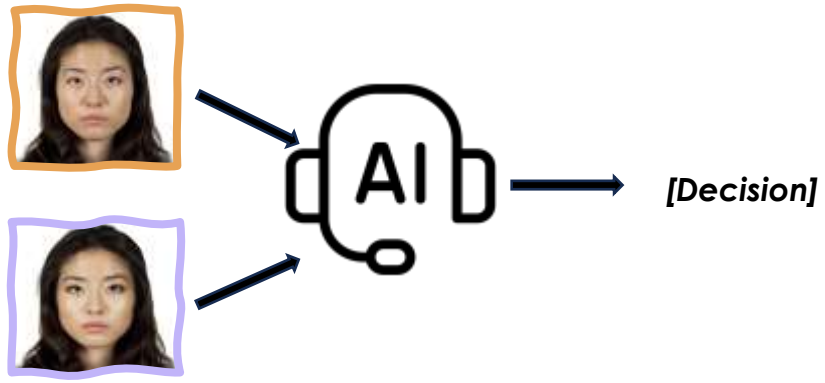
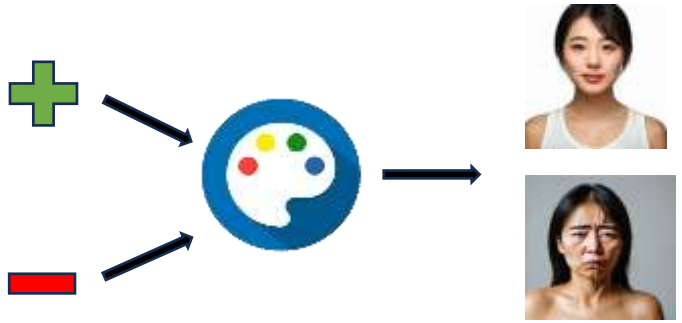


Image or multimodal-based decision-making systems might be impacted by lookism, which can lead to unfair treatment and reinforcement of societal biases

Generative



Lookism might exist in image generation and multimodal generative systems, leading to representational biases in the content they create

"Lookism, the overlooked bias in computer vision", Gulati, Lepri and Oliver, to appear at ECCV 2024, workshop on Fairness and ethics towards transparent AI: facing the chalLenge through model Debiasing (FAILED)

Lookism: an overlooked bias in Computer Vision

Impact: The oversight of lookism can result in systemic disadvantages and discrimination for individuals who do not conform to prevailing aesthetic norms, affecting their opportunities and how they are perceived and judged by automated systems.

Gulati, A., Lepri, B., & **Oliver, N.** (2022). Lookism: the overlooked bias in computer vision. ECCV'24, workshop on Fairness and ethics towards transparent AI: facing the chalLenge through model Debiasing (FAILED)

Discriminative: Are MLMM systems impacted by attractiveness?

RQ1: Does **appearance matter** for MLLM decisions?

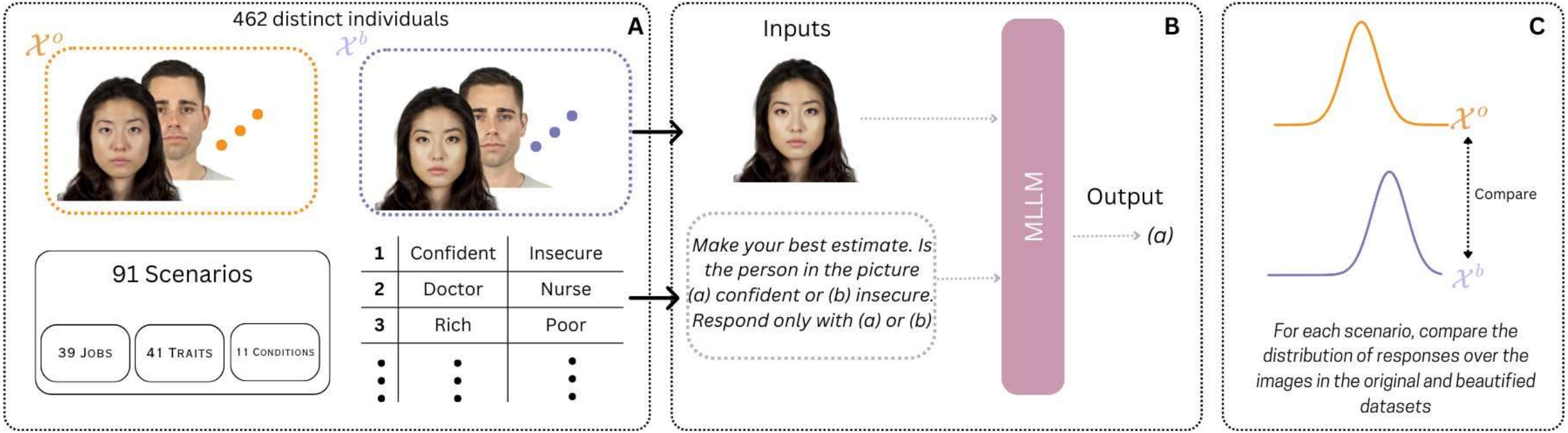
- Are there **statistically significant differences** in the **distribution** of the responses given to faces depending on their attractiveness?

RQ2: Do MLLMs exhibit an **attractiveness halo effect**?

- Is attractiveness associated with more **positive traits** also in MLLMs?

Discriminative: Lookism in MLLMs

Using A-HEAD (the **A**ttractiveness **H**alo **E**ffect **A**ttribution **D**ataset) we evaluate **7 large open-source** MLLMs on **91 unique scenarios** related to **jobs, traits** and **conditions**



"Beauty and the Bias: Exploring the Impact of Attractiveness on Multimodal Large Language Models", *to appear at the AAAI conference on AI, Ethics and Society,* https://doi.org/10.48550/arXiv.2504.16104

Discriminative: Lookism in MLLMs

We observe consistent tendencies of an **attractiveness bias (83.8%)** and the **attractiveness halo effect (92.6%)** in MLLMs

First evidence reported in the literature. In addition, we reveal intersectional gender (76.5%), age (69.2%) and race (62.3%) biases.

	Total (91)	Jobs [■]			Traits [■]		Conditions [■]		
		Gender (19)	Race (12)	Attractiveness (8)	Sentiment (33)	Other (8)	Geography (4)	Wealth (5)	Other (2)
Gemma	89.0%	78.9%	**91.7%**	**100.0%**	93.9%	75.0%	**100.0%**	**100.0%**	50.0%
Phi3.5	79.1%	**84.2%**	83.3%	75.0%	84.8%	50.0%	**100.0%**	60.0%	50.0%
DeepSeek	**90.1%**	**84.2%**	**91.7%**	**100.0%**	93.9%	**100.0%**	75.0%	60.0%	**100.0%**
Molmo	80.2%	68.4%	66.7%	87.5%	90.9%	62.5%	**100.0%**	**100.0%**	50.0%
Qwen2	81.3%	68.4%	66.7%	**100.0%**	87.9%	62.5%	**100.0%**	**100.0%**	**100.0%**
Pixtral	86.8%	68.4%	83.3%	**100.0%**	**100.0%**	75.0%	75.0%	**100.0%**	50.0%
LLaVA 1.5	80.2%	63.2%	75.0%	75.0%	97.0%	75.0%	75.0%	80.0%	50.0%
Average	83.8%	73.7%	79.8%	91.1%	92.6%	71.4%	89.3%	85.7%	64.3%

"Beauty and the Bias: Exploring the Impact of Attractiveness on Multimodal Large Language Models", *to appear at the AAAI conference on AI, Ethics and Society,* https://doi.org/10.48550/arXiv.2504.16104

Generative: When Algorithms Play Favorites

RQ1: Do synthetic faces generated by diffusion models exhibit algorithmic lookism, i.e. an implicit correlation between attractiveness and unrelated attributes?

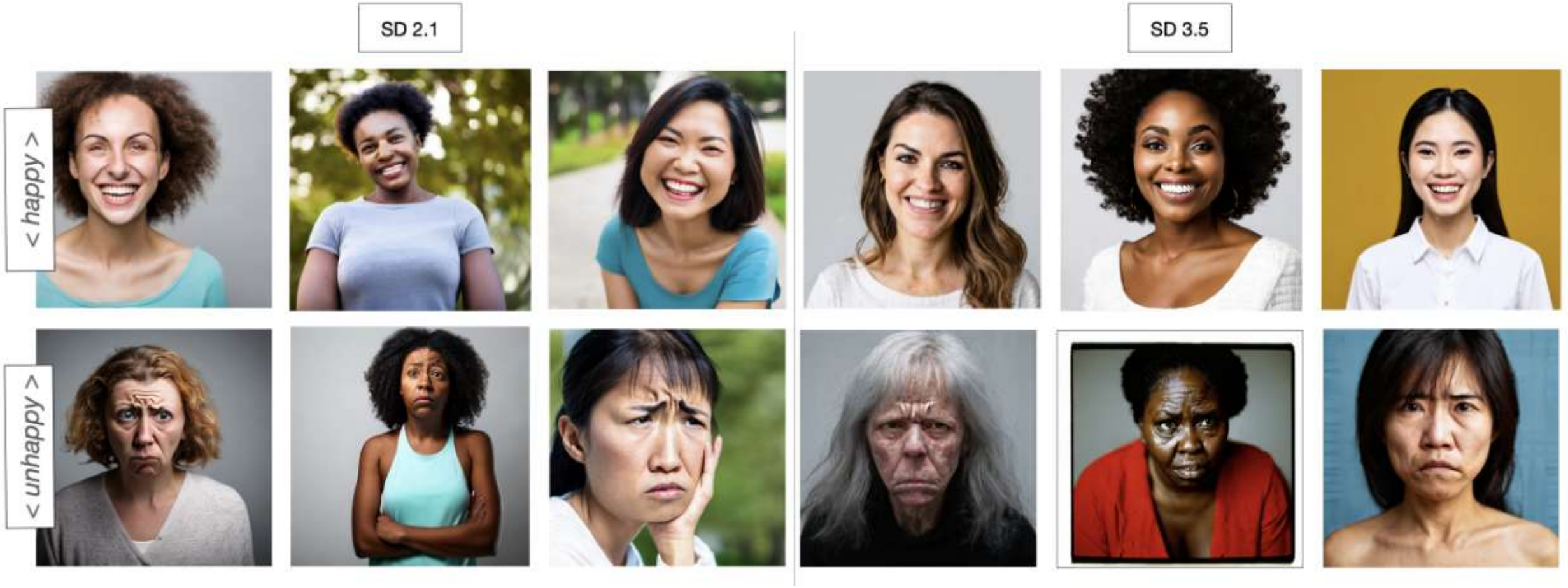
RQ2: Does algorithmic lookism impact the performance of downstream tasks, particularly gender classification?

Generative: When Algorithms Play Favorites

24,600 images generated across **5 attributes** (attractiveness, happiness, intelligence, sociability and trustworthiness) with **3 race categories** (Asian, White, Black) and **2 genders** (Woman, Man) using **Stable Diffusion 2.1** and **Stable Diffusion 3.5**

"When Algorithms Play Favorites: Lookism in the Generation and Perception of Faces", *Doh, et al*. Extended abstract at the Fourth European Workshop on Algorithmic Fairness (EWAF) 2025, Submitted to IEEE TIST 2025

When Algorithms Play Favorites

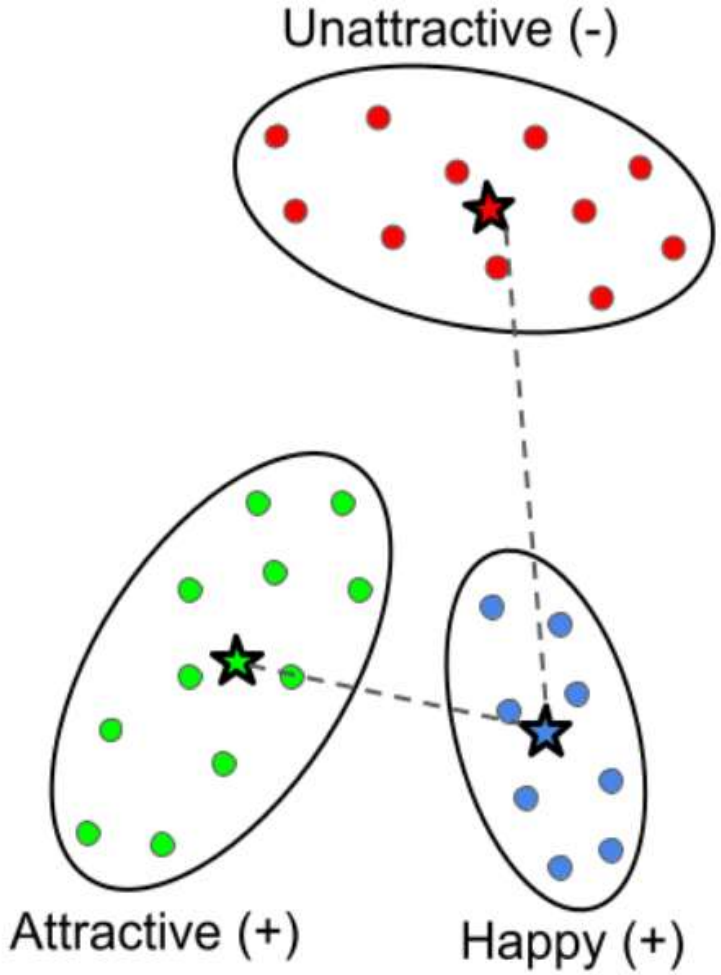


"When Algorithms Play Favorites: Lookism in the Generation and Perception of Faces", *Doh, et al*. Extended abstract at the Fourth European Workshop on Algorithmic Fairness (EWAF) 2025, Submitted to IEEE TIST 2025

Generative: When Algorithms Play Favorites

- Using CLIP embeddings of the images, we compute the **distances between the centroids** of the images generated with the attractive/unattractive descriptor to those with other positive/negative trait pairs

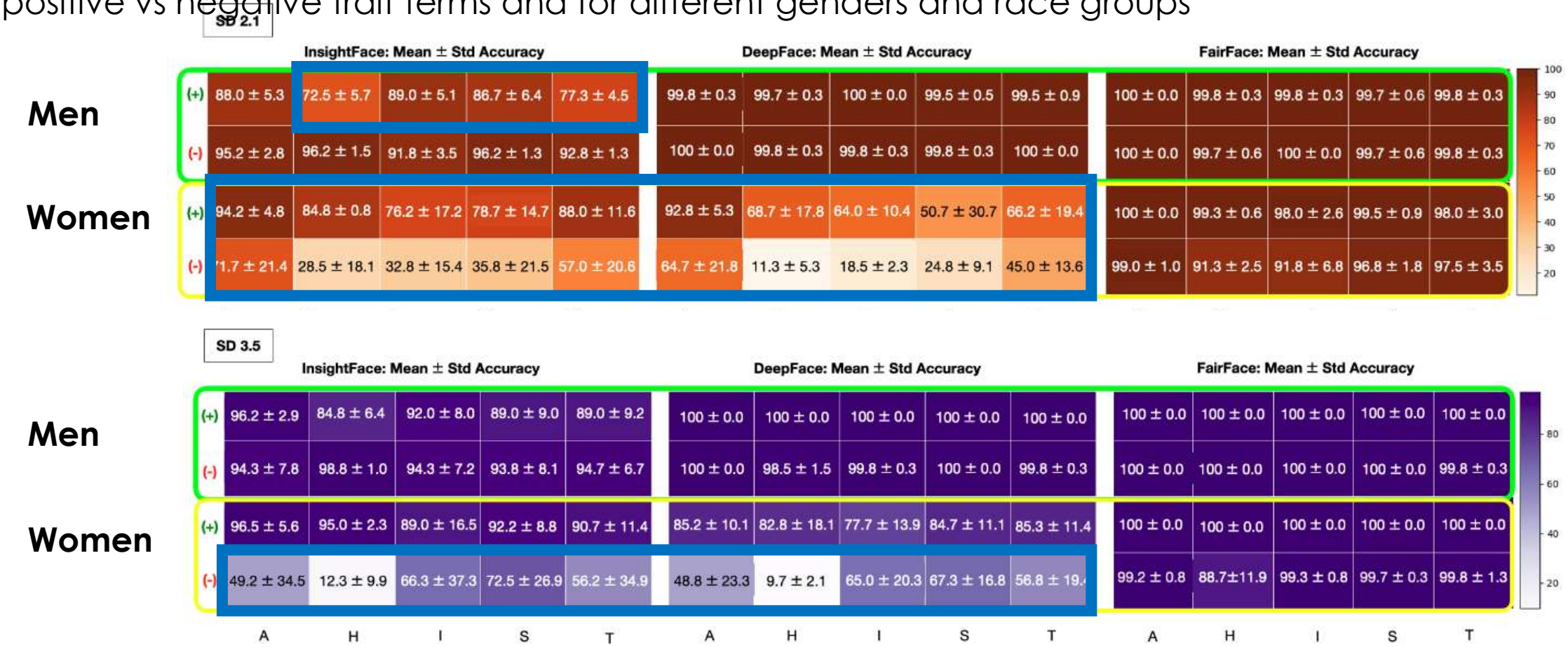
- Images generated with **positive trait terms** are closer to **attractive images** and vice versa



"When Algorithms Play Favorites: Lookism in the Generation and Perception of Faces", *Doh, Gulati,* Mancas and Oliver. Extended abstract at the Fourth European Workshop on Algorithmic Fairness (EWAF)

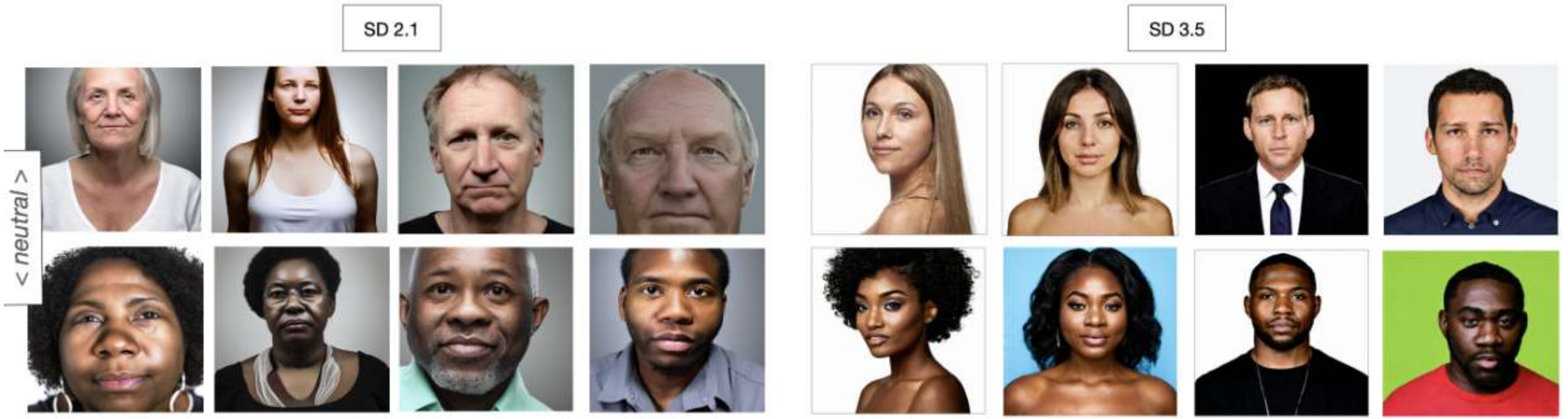
Generative: Impact on Downstream Apps

- We evaluate the gender classification performance of **3 classifiers** (InsightFace, DeepFace and FairFace)
- **Significant difference** in the performance of depending on whether the images were generated with positive vs negative trait terms and for different genders and race groups



"When Algorithms Play Favorites: Lookism in the Generation and Perception of Faces", *Doh, et al*. Extended abstract at the Fourth European Workshop on Algorithmic Fairness (EWAF) 2025, Submitted to IEEE TIST 2025

Are T2I models getting better or worse?



- Images generated with **SD3.5** have **higher resolution and photorealism**, but they tend to depict **younger** and **more attractive and homogeneous** individuals than those generated with SD2.1.

"When Algorithms Play Favorites: Lookism in the Generation and Perception of Faces", *Doh, et al*. Extended abstract at the Fourth European Workshop on Algorithmic Fairness (EWAF) 2025, Submitted to IEEE TIST 2025

Biases are everywhere...

- We are embedding our own biases in social media filters and T2I generative AI models
- While there is a lot of research on algorithmic biases regarding gender and race, humans have tens of cognitive biases, which could be exacerbated in human-AI systems
- Machine learning models, including MLLMs and T2I models, biases consistent with our cognitive biases with an impact on downstream tasks and intersectional effects with age, gender and race
- Our results call for additional research on the interplay between cognitive biases and ML algorithms and for the need for broader definitions of biases in the design, development and evaluation of ML systems

Relevant publications

Riccio, P., Psomas, B., Galati, F., Escolano, F., Hofmann, T., & **Oliver, N.** (2022). OpenFilter: A Framework to Democratize Research Access to Social Media AR Filters. *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track* .

Riccio, P., & **Oliver, N.** (2022). Racial Bias in the Beautyverse. *European Conference on Computer Vision 2022*. Workshop on CV4Metaverse

Riccio, P., Oliver, J. L., **Escolano, F.**, & **Oliver, N.** (2022). Algorithmic Censorship of Art: A Proposed Research Agenda *13th International Conference on Computational Creativity*.

Riccio, P, Colin, J., Ogolla, S. & **Oliver, N.**, "Mirror, Mirror on the Wall, Who Is the Whitest of All? Racial Biases in Social Media Beauty Filters, Social Media+ Society 10 (2), 205630512412392 95"

Riccio, P, Curto, G., Hofmann, T.& **Oliver, N.**, An Art-centric perspective on AI-based content moderation of nudity, ECCV'24, workshop proceedings

Doh, M., Canali, C., & **Oliver, N.** (2025). What TikTok Claims, what Bold Glamour Does: a Filter´s Paradox. *ACM Conference on Fairness, Accountability and Transparency, FAccT 2025*.

Open Filter and Datasets: https://ellisalicante.org/datasets/OpenFilter

Relevant publications

Gulati, A., Lozano, M. A., Lepri, B., & **Oliver, N.** (2022). BIASeD: Bringing Irrationality into Automated System Design. Thinking Fast and Slow and Other Cognitive Theories in AI, AAAI Fall Symposium 2022 .

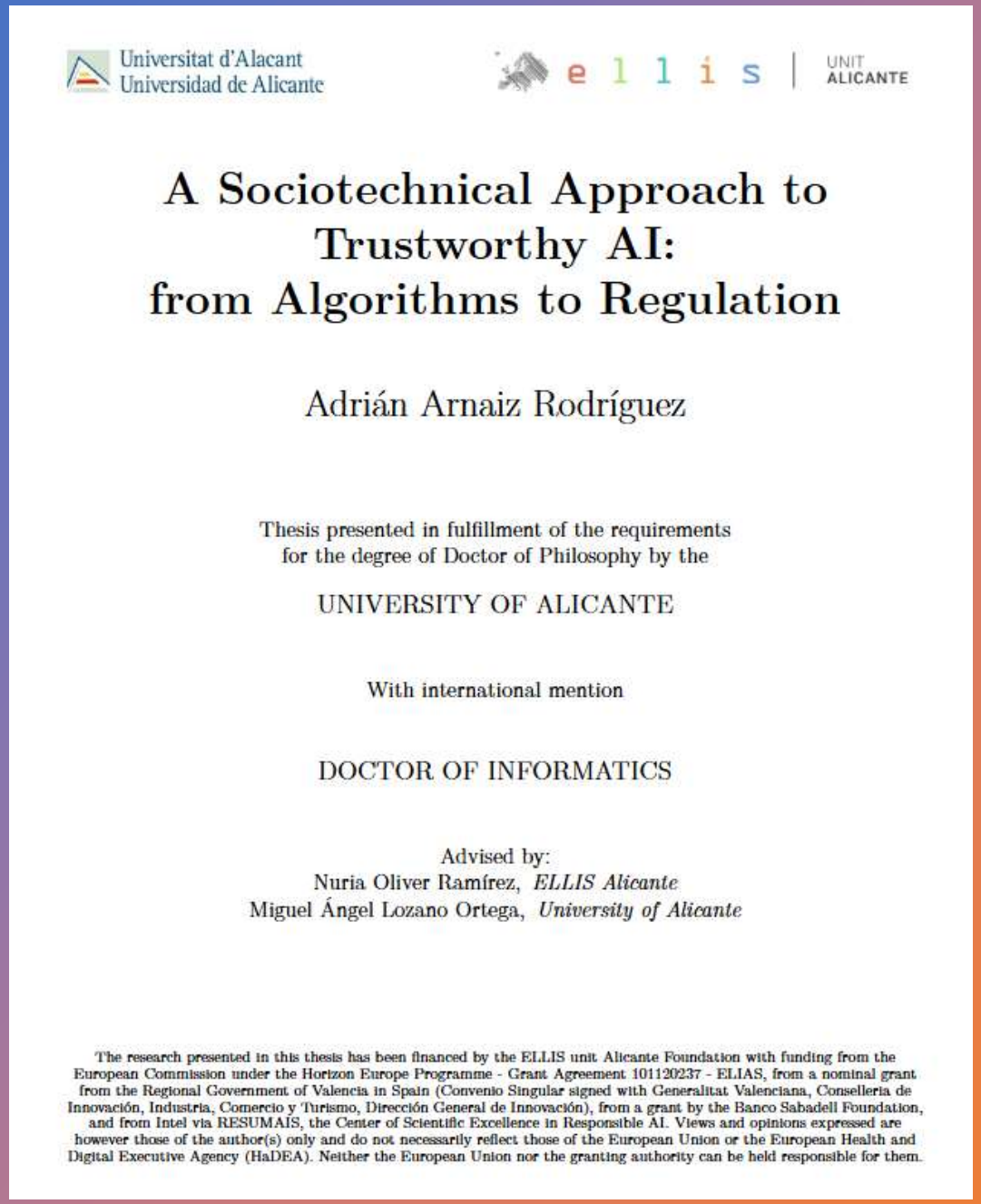
Gulati, A. *et al., "What is beautiful is still good",* Royal Society Open Science, 2024

Gulati, A., Lepri, B., & **Oliver, N.** (2022). Lookism: the overlooked bias in computer vision. ECCV'24, workshop on Fairness and ethics towards transparent AI: facing the chalLenge through model Debiasing (FAILED)

Gulati, A., et al (2025). Beauty and the Bias: Exploring the Impact of Attractiveness on Multimodal Large Language Models, AAAI AIES, 2025

Doh, M., Höltgen, B., **Riccio, P.**, & **Oliver, N.** (2025). Position: The Categorization of Race in ML is a Flawed Premise. *ICML'25: International Conference on Machine Learning*.

Doh, M., **Gulati, A.**, Mancas, M., & **Oliver, N.** (2025). When Algorithms Play Favorites: Lookism in the Generation and Perception of Faces. *European Workshop on Algorithmic Fairness*.



Universitat d'Alacant
Universidad de Alicante

ellis | UNIT ALICANTE

A Sociotechnical Approach to Trustworthy AI: from Algorithms to Regulation

Adrián Arnaiz Rodríguez

Thesis presented in fulfillment of the requirements
for the degree of Doctor of Philosophy by the

UNIVERSITY OF ALICANTE

With international mention

DOCTOR OF INFORMATICS

Advised by:
Nuria Oliver Ramírez, *ELLIS Alicante*
Miguel Ángel Lozano Ortega, *University of Alicante*

The research presented in this thesis has been financed by the ELLIS unit Alicante Foundation with funding from the European Commission under the Horizon Europe Programme - Grant Agreement 101120237 - ELIAS, from a nominal grant from the Regional Government of Valencia in Spain (Convenio Singular signed with Generalitat Valenciana, Conselleria de Innovación, Industria, Comercio y Turismo, Dirección General de Innovación), from a grant by the Banco Sabadell Foundation, and from Intel via RESUMAIS, the Center of Scientific Excellence in Responsible AI. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them.

Thesis Defense on **September 26th!**

International Committee Members:

Prof Isabel Valera, Univ Saarbrucken

Dr. Ciro Cattuto, ISI Foundation

Thank you!

The Institute of Human(ity)-centric Artificial Intelligence

Nuria Oliver, PhD

nuria@ellisalicante.org

ELLIS Alicante Co-founder and Director

ELLIS Co-founder and Vice-president

This work is funded by:

GENERALITAT VALENCIANA
Conselleria de Innovación, Universidades, Ciencia y Sociedad Digital

EUROPEAN UNION

Sabadell Fundación

ELIAS

intel.