



ECOLE
POLYTECHNIQUE
DE BRUXELLES



POLYTECH
MONS

Faces in the Machine:
Algorithmic Vision, Social Codes,
and the Making of Digital Identity

Thesis presented by Miriam Doh

in fulfilment of the requirements of the PhD Degree in Engineering Sciences and Technology (Doctorat en Sciences de l'Ingénieur et Technologie)
Academic year 2025-2026

Supervisor: Prof. Dr. Hugues Bersini

Supervisor: Prof. Dr. Bernard Gosselin

Co-supervisor: Dr. Matei Mancias

IRIDIA Lab – Artificial Intelligence research laboratory
of the Computer & Decision Engineering department (ULB)

ISIA Lab – Information, Signal and Artificial Intelligence (UMons)

Thesis jury :

Prof. Dr. Sidi Mahmoudi (Université de Mons, Chair)
Prof. Dr. Olivier Debeir (Université Libre de Bruxelles)
Prof. Dr. Noa Garcia (Osaka University)
Dr. Hinda Haned (Owls & Arrows)



Abstract

English

As face-analysis systems permeate daily life—from security checkpoints to social-media beauty filters—their opacity gives rise to pressing questions of fairness, accountability, and transparency. In response, this dissertation brings together insights from artificial cognition, feminist theories, and media studies to develop an interdisciplinary methodology, aiming to clarify algorithmic behavior and expose the broader socio-political implications of computational vision. The methodology is applied through a set of interrelated case studies spanning face verification, augmented-reality beauty filters, generative AI for synthetic faces, gender classification, and the use of race in machine-learning fairness frameworks.

From a technical perspective, the work introduces concept-based explainable AI methods tailored to face-verification systems. These methods combine targeted perturbations with semantic mappings of facial landmarks, then use large language models to translate the outputs into explanations that align with human cognitive processes. The resulting explanations are designed to be clear and relevant to people from different backgrounds. Their effectiveness is evaluated quantitatively, and they are implemented into interactive prototypes to ensure accessibility across demographic groups. In the context of augmented reality filters, the dissertation builds a transparency-driven “Disclaimer Block” that exposes transformation parameters and audits how gender classification is embedded in visual aesthetics. Using controlled generation pipelines, the research isolates attractiveness as a variable to investigate how it affects downstream gender classification accuracy, revealing systematic biases. Finally, it critiques the use of rigid racial taxonomies in machine-learning fairness frameworks, proposing context-sensitive, non-essentialist alternatives informed by mixed-race positionality and situated European legal norms.

Taken together, these empirical studies challenge the idea that technical fixes alone can ensure trustworthy AI, advocating instead for a reframing of how algorithmic systems are designed, explained, and governed. The result is a set of sociotechnical interventions that aim to render facial analysis systems more transparent, equitable, and critically aware of the visual politics they encode.

Français

Alors que les systèmes d'analyse faciale se répandent dans la vie quotidienne — des contrôles de sécurité aux filtres de beauté sur les réseaux sociaux — leur opacité soulève des questions pressantes en matière d'équité, de responsabilité et de transparence. En réponse, cette thèse mobilise des apports issus de la cognition artificielle, des théories féministes et des media studies pour élaborer une méthodologie interdisciplinaire visant à clarifier le fonctionnement algorithmique et à mettre en lumière les implications socio-politiques plus larges de la vision computationnelle. Cette méthodologie est appliquée à travers un ensemble d'études de cas interconnectées portant sur la vérification faciale, les filtres de beauté en réalité augmentée, l'IA générative pour les visages synthétiques, la classification de genre et l'usage de la notion de race dans les cadres d'équité en apprentissage automatique.

D'un point de vue technique, ce travail introduit des méthodes d'explicabilité par concepts adaptées aux systèmes de vérification faciale. Ces méthodes combinent des perturbations ciblées avec des cartographies sémantiques de points de repère faciaux, puis utilisent de grands modèles de langage pour traduire les résultats en explications alignées sur les processus cognitifs humains. Les explications produites sont conçues pour être claires et pertinentes pour des publics aux profils variés. Leur efficacité est évaluée quantitativement et elles sont intégrées dans des prototypes interactifs afin d'en assurer l'accessibilité au sein de différents groupes démographiques. Dans le contexte des filtres de beauté en réalité augmentée, la thèse met en place un dispositif de transparence — le « Disclaimer Block » — qui révèle les paramètres de transformation et examine la manière dont la classification de genre est intégrée aux esthétiques visuelles. En utilisant des chaînes de génération contrôlées, la recherche isole l'attractivité comme variable pour étudier son effet sur la précision de la classification de genre en aval, révélant des biais systématiques.

Enfin, elle critique l'utilisation de taxonomies raciales rigides dans les cadres d'équité en apprentissage automatique, et propose des alternatives contextualisées et non essentialistes, éclairées par la positionnalité métisse et par les normes juridiques européennes. Pris ensemble, ces travaux empiriques remettent en cause l'idée que des ajustements purement techniques puissent, à eux seuls, garantir la fiabilité de l'IA. Ils plaident plutôt pour repenser la manière dont les systèmes algorithmiques sont conçus, expliqués et gouvernés. Le résultat est un ensemble d'interventions sociotechniques visant à rendre les systèmes d'analyse faciale plus transparents, plus équitables et plus conscients des politiques visuelles qu'ils véhiculent.

This work was supported by the ARIAC project (No. 2010235), funded by the Service Public de Wallonie (SPW Recherche).



ESTOU
MUDANDO
COISAS
QUE EU
NÃO POSSO
MAIS ACEITAR

Figure 1: Fragment of the quote “*I am no longer accepting the things I cannot change. I am changing the things I cannot accept.*” by Angela Davis, painted on the wall of the cultural center Bombu Mininu, Mindelo, São Vicente (Cape Verde). Photograph taken by the author in 2025.

Acknowledgements

“Ma il coraggio era anche quello. Era la consapevolezza che l’insuccesso fosse comunque il frutto di un tentativo. Che talvolta è meglio perdersi sulla strada di un viaggio impossibile che non partire mai.” [1]¹

Ten years ago, I posted this quote by Giorgio Faletti on my Instagram page. I had just started my engineering studies in Trento, coming from a humanities high school with a shaky grasp of mathematics. I remember sharing that quote to give myself courage in a moment of doubt and discouragement. A part of me truly believed I wouldn’t make it—that I’d eventually give up, overwhelmed by formulas and the deeply masculine environment I had entered. But maybe that post was a way to materialize another version of myself—the one who has always believed, who has always been resilient. To make visible (even digitally, already under the spell of *if you don’t post it, it didn’t happen*) the strength I was trying to hold on to.

Over the course of this PhD, I have considered quitting an unreasonable number of times—possibly enough to qualify as a hobby. I’ve cried more than I ever expected to during a PhD, felt like an impostor on a near—professional level, and faced a few “bonus challenges” far beyond academia: losing centimeters of intestine to Crohn’s disease (five-month pause included), switching my thesis topic after a year, and nearly losing vision in one eye. And yet, somehow, here I am—writing these acknowledgments instead of my resignation letter.

This journey has been anything but linear, and it would have been impossible without the people who walked it with me.

To my supervisors—Hugues and Bernard, thank you for allowing me to begin this PhD journey and for supporting me throughout with the freedom I needed—to lose myself, find myself, transform, and grow.

To Matei—my co-supervisor, who has seen me cry, get excited, feel defeated, and still always listened and offered encouragement or guidance. Thank you, sincerely.

To my jury—for engaging with my work and offering thoughtful feedback. I am grateful for the time and attention you dedicated to this process.

To my two labs—IRIDIA and ISIA Labs: over the years I’ve been a bit of a wanderer, coming and going between trips, but every time I returned, you welcomed me like home.

To Alpha and Kevin: I probably wouldn’t have started a PhD without you two. Alpha, I met you ten years ago in the corridors of the University of Trento, and I’ve always felt safe sharing my dreams, projects—and insecurities—with you. You introduced me to Kevin, who then helped make this journey possible, encouraging me to apply and even guiding me through my Master’s thesis. You’ve both been like academic brothers—listening, challenging me when needed, and always being there. I’ll be forever grateful.

To Nuria—thank you for believing in my ideas and for giving me the courage to submit papers to conferences I never thought I’d have the strength to aim for. From the very first day you welcomed me in Alicante, you’ve made me feel that I could do this—that my ideas were worth pursuing—and

¹English translation: Courage also meant this: knowing that failure would still be the result of a meaningful attempt. That sometimes it’s better to lose yourself on the path of an impossible journey than never to set out at all.

you’ve backed that up by generously reviewing every single paper I’ve written since. Working with you has been truly inspiring, and your mentorship has shaped me profoundly as a researcher.

To the ELLIS Alicante research group—Piera, Aditya, Adrian, Lucille, Erik, Gergely, Julien, Cristina and Ben—thank you for making me feel truly part of the team during my three-month visit, and for all the kindness, generosity, and laughter you shared with me.

To my new lab (MLG group)—thank you for welcoming me so warmly into this new postdoctoral adventure, for being incredibly supportive and genuinely friendly from day one. Thank you for the coffee breaks, for listening to me through the final stretch of this academic odyssey, and for making everything feel lighter with your energy and encouragement. I’m truly happy to be here.

Speaking of labs—I must thank my dear labmates, **Nathan and Feten**. Sharing this journey with you has meant so much. We’ve talked for hours, supported and confided in each other, and most of all, laughed until we cried. I’ll always treasure those nights in Lisbon. And to Nathan in particular—thank you for all the train rides to Mons, for the time we spent together outside of work just talking and laughing, and for bravely trying every questionable restaurant in Mons with me.

To my ARIAC colleagues, Gianmarco and Jérémy—thank you for listening to my endless complaints and low moments, and for making projects and workshops lighter and fun—without a doubt. A special thank you to Gianmarco, whose words of encouragement have lifted me more times than I can count—and for making this thesis lighter also in a very literal sense, by helping me survive L^AT_EX compilation issues!

To Piera—for supporting me always—both in research and in life. I’ll never forget reading your papers before even knowing you and being struck by the power of your work. Then I got to meet you, work with you, laugh and cry with you, eat insanely stretchy mozzarella in the streets of Osaka, and listen to each other through everything. You’ve been, in many ways, the voice in my ear telling me to take the leap—to throw myself into things, just like in Marco Castello’s song, “*che puoi buttarti senza avere paura, che puoi fregartene di sembrare scema, di fare l’orgogliosa, che puoi lasciarti andare*”²—words you’ve made real for me, time and again. I cherish our friendship deeply.

To Corinna—my most frequent co-author—thank you for walking this journey alongside me. Working with you has been enriching on every level. Your knowledge is boundless, and our discussions made each paper stronger while teaching me so much. Your perspective deepened my work, and it was amazing to see how often we looked at the world through the same critical lens. You showed me how stepping outside of disciplinary boxes makes our research richer. Thank you.

To Nassim—my “Belgian” best friend—thank you for your incredible friendship. I’ve always thought of us as yin and yang: you, super chill; me, constantly anxious—but somehow a perfect balance. Thank you for all your very relaxed advice throughout this PhD, for being there to listen when I felt sad or discouraged, and for all the days we spent working (or not working) together. Most of all, thank you for your warmth and care.

To Ben—thank you for all the times you forced me out for pizza or an alcohol-free beer in Alicante, or for evening walks on the beach while I was buried in paper deadlines with absolutely no schedule. First of all—thank you for keeping me sane. But also, those moments allowed us to grow a friendship beyond work, and that was a true gift. Now I can say I have a friend in tiny Tübingen.

To Monique—*minha marida acadêmica*. A vida nunca deixa de trazer surpresas maravilhosas; encontrar mulheres como você é uma delas. À mulher e pesquisadora incrível que você é. Que você possa sempre alcançar as estrelas.

To Caroline—thank you for your trust, for following my ideas, for building things together, and for helping me explore the world of explainability.

To everyone who helped proofread this thesis—Etienne, Piera, Feten, Nathan, Ben, Emanuele, Nassim, Gianmarco and Matei—thank you for your time and your thoughtful feedback. You helped me when I was drowning in too many ideas and had lost the thread—and also my sense of time.

²Marco Castello—*Cicciona* (2021)

To all my friends in Belgium—Federica, Giselle, Yeelen, Giulia, Jasmin, Billy, Mamadou, Yves, Nathan, Nassim, Simon, Yasin, Alex, Elif, Gianluca, Basma, Sara, Camilla, Angelica, Emanuele, Gianmarco and Marco—thank you for these years. Thank you for always making me feel “smart” and telling me how proud you were of me and my work, even when I insisted it wasn’t true. Your words gave me so much strength. Thank you for the Prides, the festivals, the DJ sets, the afternoons in the park, the beers at night and the dinners at home. These moments gave me joy, lightness, and serenity. I love you all.

To my dear friend Carla—Thank you for the unwavering support you’ve shown me ever since we first met in that student dorm full of lost souls. You have always been there—with messages of encouragement, laughter over our fictional saga featuring a certain Johnathan, tarot readings, and an endless stream of MJ references. I will always cherish our late-night calls during my first months in Belgium, both of us utterly exhausted by life yet somehow lifting each other up. Grazie di cuore for being a constant light and for sharing both the madness and the chaos with me.

To all my friends far from me—geographically but never in my heart—Camila, Luca, Alice, Beatrice, Maria, Monique, Sofia, Gabriele, Theresa, Maylis, Loyani, Anna and Roberto: thank you for reminding me that true friendship knows no distance. Your messages, calls, and shared moments—no matter how rare in person—have been little anchors of joy and grounding throughout these years. Even from afar, you’ve made me feel seen, supported, and connected, and that has meant more than words can say.

To my therapist—whose guidance has been invaluable throughout this journey. A PhD is not only an intellectual challenge but also a profound emotional one, where insecurities, stress, and self-doubt constantly resurface. Your support has been invisible to most, but essential to me.

À ma famille en Côte d’Ivoire — Tonton Honoré, Tonton Ben, Tata Valentine, Tata Pauline, Yann, Élodie, Kenza et Fernande — merci de m’avoir accueillie à bras ouverts et de m’avoir fait sentir chez moi. Vous m’avez appelée *Djinsia* sans savoir à quel point cela m’a rendue heureuse. Je reviendrai bientôt.

Ai miei fratelli Ingrid ed Emmon (e alla mia sorella acquisita Francesca)— grazie per questo legame che dura una vita. Grazie per avermi fatto sentire sostenuta e per aver creduto in me, anche quando non lo dicevate ad alta voce. Non potrei chiedere fratelli più divertenti di voi. Grazie per tutte le risate — e anche per i momenti complicati — perché tutto questo fa parte di chi siamo insieme.

Ai miei nipoti Mauro e Lucas — con l’augurio che questa tesi vi ispiri a seguire sempre le vostre passioni, anche quando il percorso sembra ancora tutto da scrivere. Possiate crescere curiosi, coraggiosi e pronti a seguire dove vi porta la vostra curiosità.

To Etienne—“*Eu gosto de você, E gosto de ficar com você, Meu riso é tão feliz contigo, O meu melhor amigo é o meu amor*”³... Etienne, mon amour, meu coração, la mia stella del cielo—all the beautiful things I could say to you in all the languages you speak. I’ve known you practically since the beginning of this journey, and you’ve always been by my side. You’ve been my number-one fan, living through all the emotional phases of this path: despair, joy, fear, and determination. I will always be grateful that our paths crossed, and that I get to share my life with you. Thank you for the love you give me every single day—for the endless laughter, endless hugs, and endless “from 1 to 100?” Because the answer will always be: 100.

E infine, ai miei genitori — **Claudia, Daniela, Philippe e Giambattista.**

Papà Pippi e Daniela — grazie per avermi sempre dato amore e per avermi accolta ogni volta che tornavo in Italia. Grazie per la vostra pazienza, per aver ascoltato le mie storie infinite e per avermi ricordato, ognuno a modo suo, che la casa non è solo un luogo ma le persone che ti sono accanto.

Mamma e Giambi (e Kira) — grazie per tutto il sostegno e l’incoraggiamento che mi avete dato, e per esserci stati davvero in ogni singolo passo di questo dottorato. La vostra presenza e cura sono state una base costante, rendendo questo cammino più possibile di quanto avrei mai potuto immaginare. In particolare, grazie a te, Giambi, per avermi fatto credere che potevo uscire dal “percorso

³Tribalistas—*Velha Infancia* (2004)

previsto” e seguire qualcosa di più ricco, stimolante e nuovo.

Being the child of two heritages, raised by four parents, and both the daughter of an immigrant and an immigrant myself, has undeniably broadened my perspective and taught me to see the world through multiple lenses. For a long time, this in-between identity made me feel like I didn’t fully belong anywhere. Today, I know that belonging is not something you wait to be given; it is something you learn to create. Communities can be built, spaces can be shaped, and the in-between can become a home rather than a fracture.

Learning to inhabit that in-between has also meant learning to create possibilities that did not exist before. Reaching this milestone as the first person in my family—and the first woman—to earn a PhD carries a weight that I feel deeply. I come from women whose lives were often shaped by limited opportunities, by expectations they did not choose, and by roles that left little space for education or self-determination.

To the women of my family—especially **my grandmothers Hélène, Luigina, and the many others whose footsteps came before mine**—who navigated worlds that offered them little choice and whose quiet strength made it imaginable for me to walk a different road: this thesis is also for you.

Contents

Thesis abstract	1
Fundings	4
Acknowledgements	8
1 Introduction: Faces as Social Codes	19
2 Backgrounds and Motivation	26
2.1 <i>Trustworthy AI</i> : From Ethical Principles to Sociotechnical Critique	26
2.2 Interdisciplinary Approaches to Trustworthy AI	28
2.3 Motivations and Research Questions	32
2.3.1 Thesis Contributions:	35
2.4 Research Outputs from this Dissertation	36
2.4.1 Submitted	36
2.4.2 Accepted	36
2.5 Research Outputs Outside this Dissertation	37
3 Reclaiming Sight: Democratizing Vision	40
3.1 Situating Explainable Artificial Intelligence in the <i>Trustworthy Artificial Intelligence (AI)</i> Debate	40
3.2 Rethinking Visual Explainability in Face Verification: Background and Motivation .	42
3.3 Research Questions	44
3.4 Methodology	45
3.4.1 A Cognitively-Inspired Semantic Perturbation Framework	45
3.4.2 Definition of Human-based Semantic Features	46
3.4.3 Concepts Extraction	47
3.4.4 Similarity map	53
3.4.5 Generation of Textual Explanations Using Large Language Models	55
3.5 Metrics	56
3.5.1 Concept Evaluation Experiments	56
3.5.2 User Evaluation and Feedback Methodology	56
3.6 Experiments and Results	57
3.6.1 Semantic Extraction	57
3.6.2 Local Explanation and Sensitivity Analysis	57
3.6.3 User Feedback Evaluation	60
3.7 Conclusion	67
4 Filtering Vision: Challenging Beauty Norms through Transparency	71
4.1 From Transparency to Agency: Extending <i>Trustworthy AI</i>	71
4.2 Motivation: Algorithms, Aesthetics, and Identity	73

4.3	Related work	74
4.4	The Disclaimer Block as a Transparency Proposition	75
4.4.1	Transparency Challenges in Beauty Filters	75
4.4.2	The Disclaimer Block Concept	76
4.4.3	Technical Development: From DB V1 to DB V2	76
4.5	Analysis of the <i>Bold Glamour</i> Beauty Filter	79
4.5.1	Research Questions	80
4.5.2	Dataset for <i>Bold Glamour</i> Analysis	80
4.5.3	Methodology to Characterize Facial Features	81
4.5.4	Results	82
4.5.5	RQ2.1: Does <i>Bold Glamour</i> brighten the faces?	82
4.5.6	RQ2.2: Are the filter transformations dependent on gender and race?	83
4.5.7	RQ2.3: Does <i>Bold Glamour</i> apply a facial feature morphological alignment?	86
4.6	Platform Policies	86
4.6.1	Guidelines vs. Actual Practice	87
4.6.2	Filtering Out The “Ugly”	88
4.7	Discussion and Implications	89
4.8	Conclusion	91
5	<i>Cascading Gazes: Aesthetic Bias Migration in AI Ecosystems</i>	95
5.1	From Aesthetic Governance to Systemic Discrimination	95
5.2	<i>Algorithmic Lookism</i> and Its Systematic Propagation	97
5.3	Research Questions	99
5.4	Dataset Creation	99
5.5	Methodology	101
5.5.1	Feature Extraction	101
5.5.2	RQ3.1.1: <i>Algorithmic lookism</i> Reference Point Analysis	101
5.5.3	RQ3.1.2: Cross-Attribute Correlation Analysis	102
5.5.4	RQ3.2.1 Impact of <i>Algorithmic lookism</i> on Gender Classification Performance	102
5.6	Results	103
5.6.1	RQ3.1: Do synthetic facial images generated by diffusion models exhibit <i>algorithmic lookism</i> , i.e., an implicit correlation between attractiveness and unrelated attributes?	103
5.6.2	RQ3.2: Does <i>algorithmic lookism</i> impact the performance of downstream tasks, particularly gender classification?	106
5.7	Qualitative Analysis of the Visual Content in the Generated Faces	108
5.8	The Multiple Dimensions of Algorithmic Harm	110
5.9	Conclusion	112
6	<i>Who Gets to Define: Mixed-Race Identity as Methodological Critique</i>	115
6.1	Anatomies of Belonging: From McWatt to Machine Learning	115
6.2	The Categorization of Race in ML is a Flawed Premise	117
6.3	Related Work	118
6.4	U.S. centrism vs the European perspective	118
6.5	The Mixed-race Problem in ML	120
6.6	From Reifying to Stereotyping Race in AI	122
6.7	Moving Beyond Race Labels in ML	125
6.7.1	Tabular Data	126
6.7.2	Visual Data	127
6.7.3	Implementation in practice: Context and Participation	128

6.8	Alternative Views	129
6.9	Reflections on Methodological Evolution: Chapter 6 as Post-Empirical Critique . . .	129
6.10	Conclusion	132
7	Conclusion	134
A	Appendix Chapter 3	137
A.1	Table of Facial Areas and Input Values Used for Model Testing	137
A.2	Generated Outputs from LLMs showed in the survey	137
A.2.1	Analysis of Key Instructions and Progressive Removal Tests	139
A.3	Survey Questions	140
B	Appendix Chapter 4	143
B.1	Updates on Governmental Acts	143
B.2	From Faces to Semantic Feature Vectors with FaceGen	143
C	Appendix Chapter 5	145
C.1	Gender Classification Performance as a Function of Different Attributes	145
C.2	NSFW Content Analysis	145
D	Appendix Chapter 6	149
D.1	U.S. census categories	149
D.2	Additional information on datasets	149
D.3	Additional information on testing the Face Race Lightness Illusion in VQA Models .	150
D.4	Phenotypic-based framework	151

List of Acronyms

AI	Artificial Intelligence
XAI	Explainable Artificial Intelligence
STS	Science and Technology Studies
CAV	Concept Activation Vectors
LLM	Large Language Model
CaOc	Class-aware Order Correlation
MAGs	Maximum Activation Groups
EaOC	Embedding-aware Order Correlation
MTCNN	Multi-task Cascaded Convolutional Network
AR	Augmented Reality
DB	Disclaimer Block
ML	Machine Learning
DL	Deep Learning
SD	Stable Diffusion

Chapter 1

Introduction: Faces as Social Codes

“Vision is always a question of the power to see—and perhaps of the violence implicit in our visualizing practices. With whose blood were my eyes crafted?” [3].

Donna Haraway’s provocative question captures the complexity of how we understand sight itself as a political act. When applied to the realm of facial analysis—both human and algorithmic—her words reveal the hidden histories behind our seemingly natural ability to read faces, categorize identities, and make social judgments based on appearance.

Every time an algorithm scans a face to determine gender, predict criminality, or assess emotion, it exercises this power to see. However, the “eyes” of such technologies are not neutral; they carry the legacy of centuries of pseudoscientific thinking that has historically linked facial features to assumptions about character and worth—rationalizations that have been used to sustain hierarchies and exclude those who did not conform to dominant norms.

This dynamic becomes clearer when we consider Ruha Benjamin’s observation about the social power of naming:

“What’s in a name? Your family story, your religion, your nationality, your gender identity, your race and ethnicity? [...] Like a welcome sign inviting people in or a scary mask repelling and pushing them away, this thing that is most *ours* is also out of our hands” [4].

Just as names carry racialized, gendered, and classed meanings that shape how we are perceived and treated, faces function as similarly loaded social codes. Nevertheless, while most people recognize that names are socially constructed, faces are often perceived as reflecting natural, biological truths about identity. This perception lends facial analysis technologies an aura of objectivity and scientific authority that obscures the social biases they reproduce.

The belief that faces reveal essential truths about character and identity has deep historical roots. Physiognomy—the pseudoscientific practice of reading moral worth and social categories from facial features—provided a seemingly scientific foundation for social hierarchies across different domains and historical periods [5, 6, 7, 8, 9].



Figure 1.1: *Planche 37. Les dimensions du nez observées de face.—Identification anthropométrique*, published 1893 [2].

Though such ideas existed in various forms for centuries [10], the 19th and early 20th centuries witnessed their codification into systematic theories that operated across multiple contexts [7, 11]. Theories of criminality claimed that deviant behavior could be detected through skull measurements and facial asymmetries, with figures like Cesare Lombroso arguing that criminals represented biological regressions marked by primitive physical traits [13, 5]. Racial classification systems used skull measurements and facial morphology to construct hierarchical taxonomies that justified imperial domination and slavery [7, 14, 15]. Gender-based physiognomy pathologized women who deviated from social expectations, treating non-conforming behavior and facial features as evidence of biological deficiency [16, 17, 18, 19].

While these explicitly physiognomic theories lost academic credibility by the mid-20th century [20], their underlying logic—that essential truths about character and identity can be read from facial features—has found new expression in AI systems that claim scientific objectivity while perpetuating the same categorical thinking. A 2016 study [21] claimed to detect “criminality” from facial images, yet the algorithm primarily identified facial expressions—effectively criminalizing those who appeared less joyful. Research on AI “gaydar” [22] purported to identify sexual orientation from dating profiles (Figure 1.3a), raising serious ethical concerns about privacy and safety—particularly in contexts where homosexuality is criminalized or socially sanctioned. Similarly, companies like Faception¹ [23] market facial recognition systems that claim to detect “terrorists” and “pedophiles” for security applications. This practice not only lacks scientific validity but also poses severe risks of wrongful profiling, discrimination, and violation of fundamental rights. Even seemingly benign applications like gender classification—still deployed in consumer photo-editing and face-filter apps such as FaceApp²—reduce the complexity of gender expression to binary categories based on superficial visual cues, including facial features or hairstyle.

These applications translate socially constructed categories into quantifiable markers, stabilizing identities that are in reality fluid and contested, and embedding them in algorithmic classifications that both obscure their social origins and reinforce existing hierarchies.

This historical continuum finds a contemporary parallel in what Ruha Benjamin calls the *New Jim Code* [4]. Unlike the explicit discrimination of the Jim Crow laws—state and local statutes that enforced racial segregation and disenfranchised Black Americans in the United States from the late 19th century until the mid-20th century [24]—or 19th-century pseudoscience, the *New Jim Code* operates through the seemingly neutral domain of digital technologies and algorithms. These systems promise to transcend human bias and achieve scientific objectivity, yet as Benjamin observes, “[...] these tech advances are sold as morally superior because they purport to rise above human bias”, while in reality “they could not exist without data produced through histories of exclusion and discrimination” [4].

Benjamin’s insight about data “produced through histories of exclusion and discrimination” be-



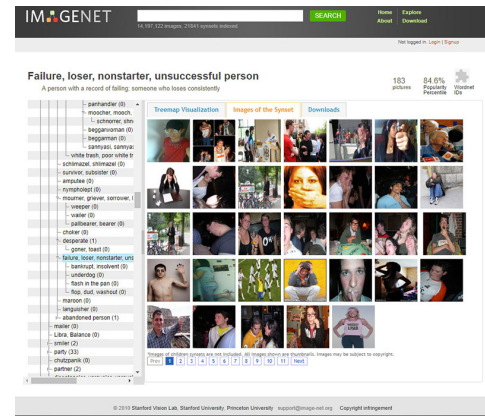
Figure 1.2: Photo illustration by The New York Times; Archive.org (photo plate from *Identification anthropométrique*, published 1893 [2]); Amazon Rekognition (facial landmarks and boundaries) [12].

¹<https://www.faception.com>, Last Access: 20-08-2025.

²<https://www.faceapp.com>, Last Access: 20-08-2025.



(a)



(b)

Figure 1.3: Image (a) Composite faces and average face outlines produced by averaging faces/outlines classified as most likely to be gay or straight. Image from [22]. Image (b) Selection for “person” classes in ImageNet. Image from [25].

comes concrete when we examine how AI datasets handle both *presentation* and *representation*. Presentation concerns the fundamental choices about which images to include for different categories and how to label them—decisions that can embed harmful stereotypes or inappropriate representations directly into training data. Representation, meanwhile, involves questions of inclusion and exclusion: whose faces appear in these datasets, in what proportions, and how different demographic groups are depicted.

This *New Jim Code* becomes tangible when we examine how contemporary AI systems are trained. These dynamics are visible in the datasets that power facial analysis algorithms, which reveal how historical patterns of exclusion and stereotyping become embedded in supposedly neutral technical infrastructure. The presentation problem is exemplified by the ImageNet dataset [26], which was forced to remove part of its training data after it was revealed that labels included sexist, racist, and stereotyped slurs [25] (Figure 1.3b.b). The problems extended beyond labeling to image selection itself. The project included offensive photographs, and in one notable case, images of people in *blackface*³ were labeled as representations of Black individuals—illustrating how racialized bodies have historically been captured through a *white gaze*⁴ that dissects, categorizes, and fixes them in structures of power, rather than allowing for self-representation [29].

Representation issues emerge in the demographic composition of these datasets. Different works [30, 31] showed that many facial analysis datasets are heavily skewed toward lighter-skinned and male subjects, leading to imbalanced performance across demographic groups and perpetuating systemic biases. Such imbalances reveal that AI does not merely reflect the world but actively constructs it by reproducing historical patterns of exclusion and marginalization.

These issues of presentation and representation point to a fundamental question about the nature of visual data itself. Crawford and Paglen address this directly [32], drawing inspiration from René Magritte’s famous painting *The Treachery of Images*—where a precisely painted pipe is accompanied by the text *ceci n’est pas une pipe*⁵ (Figure 1.4). Through this reference, they remind us that images do not describe themselves: their meaning is always constructed, mediated, and situated within social and political contexts. Magritte’s insight proves particularly relevant to facial analysis systems. Just as his painting reveals the gap between visual representation and reality—demonstrating that even the

³Blackface performance, characterized by white actors darkening their skin to portray racist caricatures of African Americans, represents a foundational practice in American entertainment that both reflected and reinforced white supremacist ideologies [27].

⁴Building on Morrison’s literary analysis, the white gaze can be understood as a dominant mode of perception that frames non-white subjects as ‘other,’ using this otherness to construct and maintain white identity and subjectivity [28].

⁵English translation: “This is not a pipe”



Figure 1.4: The Treachery of Images, René Magritte (1929).

most realistic image remains fundamentally a representation rather than the thing itself—Crawford and Paglen expose how AI training datasets perpetuate the illusion of objective categorization. The artist understood that whoever has the power to label an image controls its meaning, a principle that becomes especially troubling when applied to human subjects.

Where physiognomists and contemporary AI researchers assume a direct, measurable relationship between facial features and human characteristics [5], Magritte's work suggests that this relationship is fundamentally unstable. As Crawford and Paglen observe, meaning can be “sculpted by whoever has the power to say what a particular image means” [32]. This insight challenges the notion that systems trained on datasets can objectively categorize people based on their appearance, revealing the deep entanglements between data, power, and social hierarchies, which facial analysis systems inevitably encode—hierarchies that shape not only how identities are labeled but also whose faces are made visible or kept invisible within these technological paradigms.

These insights about the constructed nature of visual meaning and the power dynamics embedded in classification systems directly inform the empirical investigations that follow. If faces function as social codes rather than neutral biological facts, then understanding how AI systems process, classify, and generate faces requires examining not just their technical mechanisms, but the social assumptions encoded within them.

This dissertation situates contemporary facial analysis systems within a broader socio-technical lineage, revealing how the act of seeing through algorithms carries forward long-standing dynamics of categorization, control, and exclusion. These systems are not merely computational tools but infrastructures of perception that shape who is rendered visible, knowable, and classifiable in digital societies. In this sense, facial analysis technologies reflect not only technical design choices but broader political and historical logics. As argued in *Calculating Empires* [33], AI systems are embedded in imperial infrastructures of extraction, classification, and abstraction. From the physiognomic charts of the 19th century to today's neural networks, these systems share a common imperative: to convert human bodies into data, and data into decisions—often without accountability to those being seen. By tracing these entanglements between vision and power, this dissertation challenges the assumption that algorithmic systems can be fair or neutral by design. Instead, it proposes a shift in perspective: from fixing biased outcomes to interrogating the political conditions under which such systems are imagined, built, and deployed. The following chapters trace these entanglements across diverse sites of inquiry—from the interpretability of face verification systems, to the transparency of Augmented Reality (AR) beauty filters, to the aesthetic biases of generative models, and finally to the contested politics of racial categorization in Machine Learning (ML). Together, these cases examine how facial analysis technologies both inherit and reshape longstanding dynamics of power.

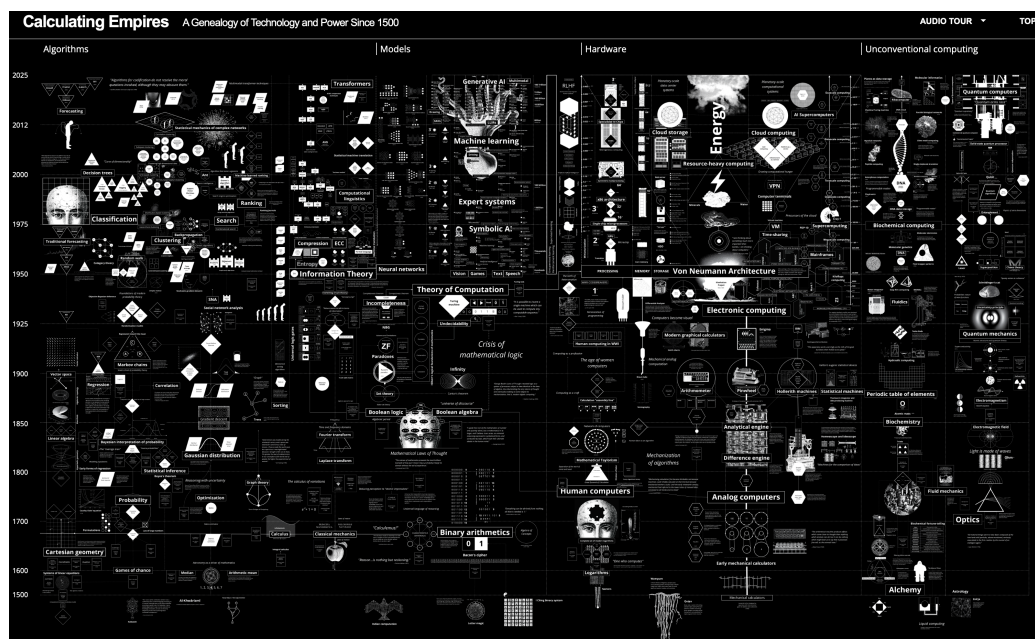


Figure 1.5: Excerpt from “Calculating Empires: A Genealogy of Technology and Power Since 1500,” by Kate Crawford and Vladan Joler. Image excerpted from [33] © 2023 Crawford & Joler.



Chapter 2

Backgrounds and Motivation

This chapter establishes the conceptual and methodological foundations that guide this dissertation’s interdisciplinary investigation of facial analysis systems. It begins by examining the emergence of *Trustworthy AI* as a normative and technical framework for governing artificial intelligence systems, with particular attention to fairness and explainability as key principles. The chapter then traces the evolution from narrow, technical approaches toward more critical and interdisciplinary perspectives that foreground the sociotechnical nature of algorithmic systems, outlining how cognitive and critical approaches offer complementary lenses for understanding AI’s social implications.

Building on this theoretical foundation, the chapter articulates the specific motivations driving this research, situating facial analysis as a particularly revealing case study for understanding how AI systems encode and amplify human biases. It introduces four interconnected research questions that span technical transparency, aesthetic governance, bias propagation, and methodological critique, each addressing different dimensions of how faces are computationally processed while maintaining focus on social implications.

While this chapter provides the overarching conceptual framework, each of the following research chapters develops its domain-specific background concerning the particular question under investigation.

2.1 *Trustworthy AI*: From Ethical Principles to Sociotechnical Critique

The concept of *Trustworthy AI* has emerged as a dominant paradigm for addressing the societal implications of artificial intelligence [34, 35], particularly as these systems expand into socially sensitive domains such as hiring [36, 37], law enforcement [38, 39], education [40], and healthcare [41]. With AI technologies moving from research laboratories into high-stakes applications affecting human lives, calls for ensuring their responsible and ethical deployment have intensified [42, 43]. In response, governments, academic institutions [44], and industry actors [45, 46] have converged around this framework, promoted by the European Commission’s High-Level Expert Group on AI [47] and echoed by organizations like the OECD [48] and UNESCO [49]. Notably, many of the leading research laboratories developing AI technologies are now embedded within multinational corporations, blurring the boundaries between scientific inquiry, product development, and commercial interests [50].

The European Commission’s Ethics Guidelines for *Trustworthy AI* established a foundational framework centered on three pillars [47]: AI systems should be lawful, ethical, and robust. This framework outlines key normative principles, including human agency and oversight, fairness, transparency, accountability, technical robustness, privacy and data governance, diversity, and societal and environmental well-being. Together, these principles aim to ensure that AI systems are not only

compliant with regulations but also aligned with broader ethical and societal values.

Among the principles outlined in the framework, *fairness* and *explainability*¹ are of particular relevance to this dissertation. Within this paradigm, the two are often presented as interconnected technical challenges meant to address the democratic deficits of algorithmic decision-making. The fairness dimension seeks to ensure that AI systems do not perpetuate or amplify social inequalities, while explainability aims to make their operations transparent and comprehensible to affected stakeholders. These principles are typically treated as complementary: explainable systems support the detection of unfair outcomes, while fair systems require a degree of transparency to validate their behavior [51]. The algorithmic fairness domain has largely approached these challenges through mathematical formalization and technical optimization. Foundational work such as Fairness and Machine Learning [52] established comprehensive taxonomies of fairness metrics and bias sources, treating discrimination as a measurable property that could be mitigated through model design. Metrics such as demographic parity, equalized odds, and counterfactual fairness [53, 54, 55] offered a way to detect and constrain disparate outcomes across social groups. In parallel, the field of Explainable Artificial Intelligence (XAI) developed techniques for interpreting the behavior of complex models. These include feature attribution methods like LIME [56] and SHAP [57], as well as taxonomies of interpretability criteria—such as simulatability, decomposability, and algorithmic transparency [58]². This body of work has aimed to address the opacity of black-box systems and make them more accountable to developers, regulators, and end users.

However, this technical orientation began to reveal fundamental limitations that challenged the core assumptions of the Trustworthy AI framework. In the domain of explainability, for example, technical methods for generating explanations have often proven insufficient for achieving meaningful transparency [59]. Empirical studies highlight persistent gaps between what algorithms can reveal and what human users—particularly non-experts—need to understand in order to evaluate and contest automated decisions [60].

Selbst et al.’s influential critique [61] identified the “abstraction trap”—the tendency to treat fairness as a property that can be achieved through technical design while ignoring the social contexts in which these systems operate. Their analysis demonstrated how seemingly neutral choices about problem formulation (*framing trap* - failing to model the entire sociotechnical system), data representation (*portability trap* - assuming solutions transfer across contexts), and evaluation metrics (*formalism trap* - reducing complex social concepts to mathematical measures) embed social assumptions that often reproduce the inequalities they claim to address. Beyond identifying these structural traps, the paper proposes a *sociotechnical perspective* that “shifts from seeking a solution to grappling with different frameworks that provide guidance in identifying, articulating, and responding to fundamental tensions, uncertainties, and conflicts inherent in sociotechnical systems”[61]. Drawing on insights from STS, it calls for approaches that embrace ambiguity, conflict, and contextual specificity. This includes methodological strategies such as heterogeneous engineering that incorporates social actors within system boundaries; attending to user scripts and situated use; recognizing interpretive flexibility and the role of relevant social groups; and critically assessing design choices through the lenses of reinforcement and reactivity. As Gebru and Denton [62] argue, it also raises the fundamental question of *when not to design*—a challenge to the assumption that technological intervention is always desirable.

This sociotechnical turn directly informs the orientation of this dissertation. Rather than treating facial analysis as a purely computational task that can be rendered “fair” through algorithmic refinement, the research investigates how these systems function as infrastructures of social categorization,

¹Explainability is under the Transparency principle.

²In Lipton’s taxonomy [58]: *Simulatability* means the model is small and simple enough for a person to understand it fully and follow how it works step by step; *Decomposability* means each input, parameter, and calculation has a clear and intuitive meaning; *Algorithmic transparency* means the training process itself is well understood and predictable, so we can anticipate how the model will behave on new problems.



Figure 2.1: Visual representation of the background and motivation flow. (author’s illustration using copyright-free Canva elements).

aesthetic governance, and identity formation. The emphasis thus shifts from fairness as a property to be optimized, to fairness as a process—an ongoing negotiation shaped by power, context, and contested values. The following section traces how this interdisciplinary turn has taken shape within AI fairness and explainability research.

2.2 Interdisciplinary Approaches to Trustworthy AI

The limitations of purely technical approaches to Trustworthy AI have prompted researchers to seek interdisciplinary solutions that can address the complex, interconnected challenges of fairness, explainability, and robustness in AI systems [63, 64, 61, 65]. Rather than treating these principles as isolated technical properties to be optimized, emerging scholarship recognizes that meaningful progress requires drawing on methodological traditions from psychology, sociology, critical theory, and other disciplines that have long grappled with questions of cognition, interpretation, power, and social meaning [66, 63, 67, 68, 69, 70, 52, 71].

As outlined in the previous section, Selbst et al.’s critique of the “abstraction trap” [61] has been one of many important contributions to the argument that fairness cannot be separated from the sociotechnical systems in which algorithmic decisions are embedded. Their analysis of how framing, portability, and formalism traps reproduce structural inequalities underscores the need for frameworks that go beyond technical correctness. The sociotechnical turn they advocate—grounded in insights from Science and Technology Studies (STS)—calls for approaches that embrace contextual specificity, participatory design, and critical reflexivity.

This transformation reflects what has been described as a “critical technical awakening” [72]—a profound shift whereby researchers trained in narrow computational paradigms come to recognize the epistemological limitations of viewing the world solely through formal abstraction. Such awakenings involve more than adopting new methodological tools; they represent fundamental transformations in cognitive style, moving from privileging technical formalization to engaging thoughtfully with the normative, political, and social assumptions embedded in algorithmic systems. This reorientation demands that technical practitioners recognize themselves as political actors participating in the construction of social order, requiring an explicit politics grounded in justice, accountability, and lived experience rather than the comfortable fiction of neutral expertise [62, 73, 65].

The concrete implications of this shift can be observed in how fundamental concepts within AI research have been reconceptualized. The very notion of fairness in ML, for instance, has undergone significant transformation as researchers have moved away from the epistemic ideal of “mechanical objectivity” that once dominated statistical science [74]. In its place, a new epistemic regime has emerged—one that embraces “trained judgment,” reflexivity, and the recognition that quantification is always a situated, normative act. Rather than treating fairness metrics as universally applicable technical solutions, researchers increasingly understand their work as requiring partial, politically

informed justifications that confront the epistemic and moral assumptions embedded in algorithmic systems.

This interdisciplinary transformation extends beyond fairness research into XAI, where a similar evolution has unfolded from narrow technical concerns toward broader epistemic and social questions. What emerges is a recognition that explainability cannot be conceived as a static property of models, but must be understood as a dynamic, dialogical process involving diverse stakeholders and forms of knowledge [75]. This perspective reveals the inadequacy of current approaches that, despite their technical sophistication, often fail to account for how humans actually process, understand, and use explanations in real-world contexts [76]. The challenge becomes not merely developing more transparent algorithms, but understanding cognitive biases, scaffolding processes, dialogical interaction patterns, and the social pragmatics of explanation—insights that purely technical approaches cannot provide.

This interdisciplinary engagement operates along two complementary dimensions: cognitive approaches, which apply psychological and perceptual methods to understand AI behavior and improve human-AI interaction; and critical approaches, which interrogate the sociopolitical dimensions embedded in algorithmic systems. Cognitive approaches help clarify how AI systems process information, how humans interpret their outputs, and how alignment between artificial and human cognition can improve transparency and trust. Critical approaches, in turn, reveal whose perspectives are centered, whose experiences are marginalized, and how power operates through seemingly neutral design choices. Together, these lenses reframe Trustworthy AI not as a fixed metric to be optimized, but as a dynamic and contested value—shaped by cognition, context, and power.

Cognitive Approaches: Understanding AI Through Human Cognition The first dimension of interdisciplinary engagement draws on cognitive psychology to develop new methods for understanding and evaluating AI systems. This approach emerges from a foundational recognition that both human and machine cognition present analogous epistemic challenges: just as the human mind operates as a “black box” with internal processes hidden mainly from direct observation, AI systems exhibit similar opacity that resists straightforward interpretation [68, 77]. The crucial insight is that cognitive psychologists have developed sophisticated experimental methods over more than 150 years to systematically infer internal processes, cognitive mechanisms, and decision-making patterns from observable behaviors—methodological expertise that proves directly transferable to understanding algorithmic systems [77].

This approach, referred to as *artificial cognition* [77], encompasses two complementary research directions. The first treats AI systems as subjects for psychological investigation, applying experimental methods to understand, evaluate, and explain ML models’ decision-making processes through systematic comparison with human performance [77, 69]. The second focuses on designing AI systems that align with human cognitive processes to facilitate interpretation and coordination between human and artificial decision-making [78, 70].

The experimental approach follows a distinctive four-step framework that moves beyond traditional ML evaluation. Rather than simply measuring performance on benchmarks, researchers document behavioral variations across different conditions, infer causal mechanisms through controlled experimentation, identify boundary conditions where particular behaviors emerge or break down, and systematically manipulate model components to test theoretical understanding [77]. This methodology has revealed surprising insights: Ritter et al.’s investigation of shape bias in deep neural networks [68] demonstrated that architecturally identical models can exhibit qualitatively different cognitive biases despite equivalent classification performance, while Jacob et al.’s systematic adaptation of twelve classic psychology experiments [79] revealed that some human-like perceptual phenomena emerge only after specific training conditions.

Complementing these behavioral investigations, frameworks like Zhang et al.’s perceptual processing model [70] structure AI explanations around human cognitive stages of selection, organiza-

tion, and interpretation, aligning algorithmic processes with familiar patterns of human perception to make AI decision-making more intuitive and interpretable to non-expert users.

Beyond these experimental and alignment approaches, cognitive psychology has also informed fairness research by providing theoretical frameworks for understanding how human cognitive biases systematically propagate through AI systems. Drawing on foundational work on heuristics and biases [80, 81], recent research has demonstrated systematic mappings between the four primary human heuristics—representativeness, availability, anchoring and adjustment, and affect³—and computational biases that emerge throughout the AI lifecycle [82]. This perspective reveals that algorithmic bias is often not merely a technical failure but reflects deeper patterns of human cognition embedded in decisions made during data collection, feature selection, model design, and system deployment.

These cognitive approaches offer significant methodological advantages for AI transparency. Their model-agnostic nature requires only input manipulation and output observation, making them applicable across architectures and domains without internal parameter access. Moreover, they emphasize hypothesis-driven investigation that produces *a priori* explanations rather than *post hoc* interpretations. However, when AI systems operate within complex social contexts, cognitive approaches alone cannot capture the full dynamics at play. Understanding how these systems function as sociotechnical infrastructures—embedded within broader structures of power, institutional logics, and social meaning—requires complementary approaches that can illuminate the political and social dimensions of algorithmic systems [82, 4, 33, 83]. This recognition points toward the necessity of critical approaches that examine not only how systems process information, but how they acquire meaning and exert power within specific social contexts.

Critical Approaches: Interrogating Power and Social Construction While cognitive approaches help us understand how AI systems process information, critical perspectives invite us to examine a more complex terrain: how these systems acquire meaning and exert power within specific social contexts. This body of scholarship fundamentally challenges computational paradigms that reduce algorithmic harms to technical optimization problems, instead drawing from feminist theory, critical race studies, STS, and decolonial frameworks to expose the structural inequalities embedded in technological systems [9, 3, 84].

As mentioned in the introduction, Benjamin provides a foundational critique by merging critical race theory with STS [4]. The work’s concept of the *New Jim Code* powerfully demonstrates how racial bias becomes engineered into technical systems through seemingly objective metrics, particularly in predictive policing and emotion recognition technologies. This work finds resonance with Browne, which historicizes facial recognition within centuries of racialized surveillance, revealing how contemporary algorithms inherit colonial logics of classification and social control [39].

Feminist scholars have similarly transformed our understanding of data practices. D’Ignazio and Klein articulate seven intersectional principles that reframe data science—from rejecting the myth of detached objectivity to making visible the often-exploited labor behind dataset creation [85]. Their perspective resonates with the work of Noble, who demonstrates how commercial search engines reinforce racial and gender stereotypes—for example, by associating search terms like “Black girls” with pornography [86].

The decolonial turn in AI critique has been particularly transformative. Birhane integrates Ubuntu philosophy (“*I am because we are*” [87]) with Black feminist epistemology [88, 89] to challenge the rationalist individualism dominating AI development [63]. Similar concerns are raised in Indigenous AI scholarship, where Lewis et al. examine how extractive data practices in ML violate Indigenous

³In Vakali and Tantalanki’s work [82], these four heuristics are defined as: *representativeness*: the tendency to estimate probability by comparing events to mental prototypes or stereotypes; *availability*: assessing frequency or likelihood based on how easily examples come to mind; *anchoring and adjustment*: the tendency to rely excessively on initial information (the “anchor”) and make few adjustments from that starting point when reaching final decisions; *affect*: making decisions based on emotions and immediate feelings rather than analytical reasoning.

sovereignty and relational ways of knowing [90].

These theoretical frameworks have spurred methodological innovations that reshape how we study and design algorithmic systems. The Gender Shades work [31] revolutionized bias auditing by applying intersectional analysis to facial recognition systems, exposing error rates of up to 34% for darker-skinned women compared to 0.8% for lighter-skinned men. Queer and trans scholars [91, 92, 93] have shown how binary gender classification systems erase non-conforming identities, while [94] reveal how gender recognition tools become instruments of bodily surveillance.

Practically, these critiques have informed participatory approaches to technology design. Costanza-Chock translates critical theory into concrete design methodologies by centering marginalized communities as co-designers rather than passive subjects [95]. Meanwhile, historical-materialist analyses by Crawford and Paglen trace how the ImageNet dataset [26] inherits and amplifies problematic classification logics [32], while Birhane et al. conduct similar investigations of the LAION dataset [96], connecting algorithmic categories to deeper regimes of race, gender, and social domination [97, 98].

Collectively, this scholarship has contributed to a fundamental reconceptualization of algorithmic fairness. While traditional approaches focus on technical metrics such as demographic parity, critical frameworks have shifted attention toward more fundamental questions regarding who gets to define fairness, whose interests are prioritized, and how technological systems redistribute power [4, 85, 95]. This body of work has reframed fairness not as a technical property to be optimized, but as an ongoing sociopolitical negotiation shaped by power relations and social context.

The convergence of cognitive and critical approaches thus points toward a more comprehensive understanding of trustworthy AI—one that recognizes the need for interdisciplinary frameworks capable of addressing both the cognitive dimensions of human-AI interaction and the sociopolitical contexts in which these systems operate. This perspective directly informs the methodological approach adopted in this dissertation.

Integrating Theoretical Frameworks Combining cognitive and critical approaches requires a clear theoretical foundation to ensure methodological coherence. This dissertation draws on three key frameworks that connect technical work with social analysis.

Critical Technical Practice [99] shows how computer scientists can critically examine their work while maintaining technical rigor. Instead of treating social critique as separate from algorithm design, this approach integrates critical reflection directly into technical decisions—questioning how choices about data, models, and metrics reflect particular values and assumptions. While *ethics by design* emphasizes embedding normative principles—such as fairness, transparency, and accountability—into system development from the outset [47], *Critical Technical Practice* complements this stance by interrogating the underlying assumptions, abstractions, and framings that guide technical work, ensuring that ethical commitments are not only implemented but also critically examined at the epistemic level.

Situated Knowledges [3] argues that all knowledge, including algorithmic knowledge, comes from specific positions and perspectives, never from a neutral “view from nowhere”. This insight is essential for facial analysis—the primary empirical focus of this dissertation—where systems claim to measure identity traits objectively actually embed particular cultural viewpoints. Recognizing this situatedness helps reveal how AI systems reflect and amplify their critical perspectives.

Critical algorithm studies and *media platform analysis* [100, 101] provide tools for understanding how algorithmic systems operate within larger digital ecosystems that shape social norms and distribute power. As we will see, facial analysis systems exemplify this dynamic particularly well by being embedded within platform logics that govern visibility, identity, representation, and social interaction.

2.3 Motivations and Research Questions

As discussed in the previous chapter, facial analysis technologies are not novel in their logic, but continue a long history of attempts to extract meaning, value, and social order from the human face. From physiognomy to biometric surveillance, these practices have consistently operated through a regime of visibility that encodes identity into visual markers, often in the service of normative classification and social control. This legacy remains embedded in contemporary algorithmic systems, which inherit and update older forms of seeing and sorting through data-driven infrastructures.

In recent years, facial analysis applications have faced growing criticism due to repeated instances of discriminatory outcomes [102, 103, 104, 105, 31] and controversial applications [106, 107, 108], prompting regulatory responses that recognize the inherent risks of these technologies. The European Union’s AI Act⁴ designates facial recognition and analysis systems as “high-risk” applications⁵ requiring strict oversight and transparency obligations⁶.

However, significant exceptions remain—for law enforcement⁷, military and national security purposes⁸, and research activities⁹, such as the identification of suspects through surveillance footage or the monitoring of public spaces—underscoring the need to examine the assumptions and limitations of these technologies critically.

Meanwhile, the emergence of generative AI has introduced new complications. While standalone facial analysis tools now face regulatory scrutiny, the datasets and computational frameworks that once powered them have been repurposed into generative models capable of synthesizing human faces [109, 110, 111, 112]. These generative tools often perpetuate the same essentialist assumptions about identity and appearance, albeit through different technical means. As a result, even as direct facial classification tools encounter regulatory barriers, the underlying logic of reading identity from facial features persists in contemporary AI systems.

Facial analysis thus provides a particularly revealing case study for understanding how AI systems encode and amplify human biases. It illuminates broader concerns about algorithmic fairness, representation, and the computational construction of identity.

This dissertation approaches these challenges from a deliberately interdisciplinary perspective, building on the emerging field of *artificial cognition* while expanding its scope to account for the multidimensional social nature of facial analysis. *Artificial cognition* [77, 113, 69] applies theories from cognitive psychology to explain and evaluate AI model behavior. This approach is particularly valuable for promoting transparency, as it clarifies when and how models succeed or fail—essentially using the same tools once employed to “open the black box” of human cognition to now interrogate artificial ones.

However, because faces function as complex social codes—carrying meanings related to identity, power, and belonging—this research extends beyond cognitive psychology to incorporate insights from gender studies, critical race theory, and social media studies. By starting from this expanded interdisciplinary foundation and moving toward technical implementation, this approach asks how theories of human perception, identity construction, and social power can inform more equitable and transparent AI systems.

Rather than beginning with algorithmic optimization and later considering social implications, this methodology inverts the traditional computer science research pipeline—using social theory and cognitive psychology as the foundation for technical innovation. This interdisciplinary orientation reflects a conviction that meaningful progress in AI fairness requires not just better algorithms or

⁴Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence, OJ L 2024/1689, (thereafter, “AI Act”), Art. 5(1)(d)-(g)

⁵AI Act, Art. 5(1)(d)-(g)

⁶AI Act, Chapter III

⁷AI Act, Art. 5(g), Art. 5(4)–(7)

⁸AI Act, Art. 2(3)

⁹AI Act, Art. 2(6), (10)

clearer explanations, but fundamentally different ways of framing the problems these algorithms are meant to solve—including critically questioning whether these applications should exist at all.

Given these theoretical foundations and contemporary challenges, this dissertation pursues four interconnected lines of inquiry that move from technical transparency to social resistance, from aesthetic bias to methodological critique. Each question addresses a different aspect of how faces are computationally processed while maintaining focus on the social implications of these technical choices.

To operationalize this interdisciplinary perspective, this research focuses on four specific applications of facial analysis that are among the most widespread and socially relevant today: face verification systems, beauty filters, gender classification, generative AI models, and the use of race categories in facial analysis research itself.

Within the broader Trustworthy AI framework discussed earlier—which encompasses fairness, accountability, robustness, and transparency as interconnected principles—this dissertation focuses specifically on reconceptualizing transparency itself. Rather than treating transparency as a static technical property, the research demonstrates how transparency must be understood as operating across multiple scales of analysis, each requiring distinct approaches and raising different stakes for contestability.

The choice of these five applications is motivated by the fact that each reflects a distinct way in which transparency and contestability are at stake in facial analysis systems, while operationalizing the theoretical frameworks introduced earlier in Section 2.2.

First, face verification addresses opacity at the technical level (Level 1). Here, Critical Technical Practice [99] is central: explainability is not a tool for legitimizing the system, but for exposing its assumptions and enabling contestation. The case is particularly urgent because, despite face analysis being classified as “high-risk” under the EU AI Act, face verification remains legally permitted for law enforcement. The framework developed in Chapter 3 does not aim to improve system accuracy or justify deployment, but rather equips those subjected to automated identity decisions with the technical understanding needed to challenge outcomes that restrict their access to work, services, and rights.

Second, beauty filters foreground transparency in the sociopolitical and platform domain (level 2). Drawing on Situated Knowledges [3], this case shows that these filters are not neutral technical artifacts but culturally positioned technologies that encode and circulate particular ideals of femininity and desirability. At the same time, platform analysis [100, 101] highlights how beauty filters serve the economic logics of social media—governing visibility, virality, and aesthetic labor [114]—while shaping identity performances through opaque mechanisms.

Third, generative AI and gender classifiers exemplify how values and ideologies propagate across systems (level 3). Rather than focusing only on a single model’s transparency, this case shows how aesthetic biases migrate through datasets and architectures, producing “bias cascades” across the AI ecosystem. This resonates with Haraway’s critique of the “view from nowhere,” [3] by demonstrating that supposedly neutral generative models reproduce specific cultural and political viewpoints, and with Gillespie’s insight that algorithms operate within infrastructures of circulation and influence, rather than in isolation [100].

Finally, racial taxonomies bring transparency into the epistemic domain (level 4). Here, the key issue is who has the authority to define categories. Standard demographic analyses in AI fairness often treat race as a fixed technical variable, but by recentring the lived experience of mixed-race individuals, this case destabilizes rigid taxonomies. It reflects Haraway’s insistence on situated perspectives over universal objectivity [3], while embodying Agre’s call for a critical technical practice that questions not only how categories are implemented, but why they exist at all [99]. These cases also illustrate a progression in how the relationship between people and models is conceptualized. In face verification, individuals confront opacity through technical explanations, keeping the focus at the level of model–user interaction. In beauty filters and generative AI, the perspective expands: the encounter

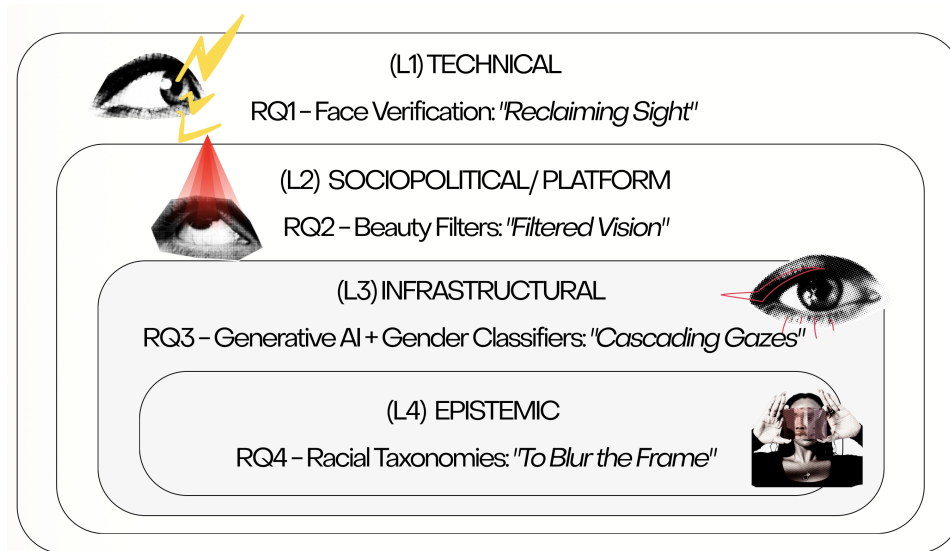


Figure 2.2: Four interconnected dimensions of Trustworthy AI in facial analysis systems, moving from technical transparency to epistemic reconceptualization. Level 1 (Technical) focuses on model explanations enabling contestation. Level 2 (Sociopolitical/Platform) reveals platform governance and aesthetic norms. Level 3 (Infrastructural) traces bias propagation across AI ecosystems. Level 4 (Epistemic) challenges categorical authority itself. Each level represents an expanding conception of how fairness, transparency, and accountability must be understood in facial analysis.

is no longer just between person and model, but is mediated by the broader platform ecosystems and ideological logics that govern visibility, desirability, and cultural circulation. Finally, in the study of racial taxonomies, the focus recenters on the person, but now as an epistemic agent whose lived experience contests and redefines the categories imposed by technical systems.

Building on these motivations and theoretical frameworks, this dissertation articulates four interconnected research questions that operationalize the interdisciplinary perspective outlined above. Each question translates key insights from the preceding sections into concrete empirical inquiries: how algorithmic transparency can be cognitively grounded and socially accessible (RQ1); how aesthetic and gender norms can be challenged through transparency mechanisms (RQ2); how infrastructural aesthetic bias propagates across model stages (RQ3); and how racial categorization can be rethought through the lens of lived experience (RQ4). Specifically:

(RQ1 - L1) *Reclaiming Sight: Democratizing Algorithmic Vision* (Chapter 3) How can Explainable Artificial Intelligence methods be designed to democratize users' understanding of face verification systems across diverse technical backgrounds and age groups? This question investigates how to move beyond technical visualizations (such as heatmaps) and develop semantic explainability frameworks that align with human cognitive processes. The goal is to make these systems more transparent and accessible to non-expert users.

(RQ2 - L2) *Filtered Vision: Challenging Beauty Norms through Transparency* (Chapter 4) How can transparency tools be transformed from passive disclosures into active instruments for challenging the aesthetic and gender norms embedded in beauty filters? This question examines how seemingly playful beauty filters on social media platforms can enforce aesthetic governance through automated gender classification, and how interventions like the Disclaimer Block (DB)—a transparency mechanism described in Chapter 4 Section 4.4—can empower users to critically engage with these technologies.

(RQ3 - L3) *Cascading Gazes: Aesthetic Bias Migration in AI Ecosystems* (Chapter 5) How do generative AI systems create and enforce systematic patterns of unintended harm through aesthetic discrimination that operates as a coherent infrastructure across AI ecosystems? This question explores how generative models embed aesthetic preferences that then propagate to downstream classification tasks, creating “aesthetic bias cascades” that amplify discrimination across AI ecosystems.

(RQ4 - L4) *To Blur the Frame: Mixed-Race Identity as Methodological Critique* (Chapter 6) How can the lived experience of mixed-race identity inform alternative frameworks for studying algorithmic bias that move beyond rigid racial taxonomies? This question addresses the limitations of standard racial categories in algorithmic fairness research and proposes alternative frameworks that foreground lived experience, aiming to challenge the categorical assumptions that often marginalize mixed-race individuals.

These research questions reflect the multifaceted challenges posed by contemporary facial analysis technologies—challenges that span technical, aesthetic, social, and epistemological domains. By focusing on widely used applications, this dissertation situates its analysis within everyday encounters with AI, making visible how systems of classification, representation, and exclusion operate as mutually reinforcing processes within the infrastructures of digital life, where categorizing and depicting are never neutral acts but practices shaped by normative assumptions that actively produce and maintain social boundaries. The methodological orientation—grounded in artificial cognition and enriched by critical social theory and lived experience—seeks not only to diagnose existing harms, but also to reimagine how fairness, transparency, and identity could be conceptualized in AI research. Each question corresponds to a dedicated chapter in this dissertation, offering a distinct but interconnected lens on the broader problem of how faces are computationally seen, classified, and made meaningful.

2.3.1 Thesis Contributions:

This dissertation makes four main contributions at the intersection of computer science, cognitive psychology, and critical theory, each aligned with the research questions presented above:

1. **A cognitively grounded framework for explainable face verification:** introducing a novel approach to semantic interpretability that combines concept-based perturbations, landmark-based mappings, and natural language generation to produce user-aligned, cognitively plausible explanations.
2. **A transparency-driven intervention for AR beauty filters:** developing the Disclaimer Block, a lightweight tool that exposes the parameters and logic of automated beauty transformations, empowering users to interrogate aesthetic and gendered norms embedded in filter design.
3. **A multi-level analysis of aesthetic bias propagation in generative AI:** empirically demonstrating how attractiveness preferences embedded in image generation models like SD migrate to downstream classification tasks, contributing to a broader understanding of “aesthetic bias cascades” in multimodal AI systems.
4. **A methodological critique of racial categorization in AI fairness:** leveraging mixed-race identity as a lens to challenge rigid racial taxonomies, proposing alternative frameworks grounded in lived experience, intersectionality, and situated knowledge.

Collectively, these contributions present a cohesive methodological framework that combines technical advancement with critical analysis, promoting both the scientific and ethical objectives of Trustworthy AI in facial analysis systems.

2.4 Research Outputs from this Dissertation

These research questions have been pursued through a series of empirical studies that bridge theoretical analysis with technical implementation. The following publications represent the core contributions of this dissertation:

2.4.1 Submitted

1. **Doh, M.**, Mazini Rodrigues, C., Boutry, N., Najman, L., Mancas, M., & Gosselin, B. (2025). Found in Translation: Semantic approaches for enhancing AI interpretability in face verification [115].
2. **Doh, M.**, Gulati, A., Canali, C., & Oliver, N. (2025). The Aesthetics of Harm: Algorithmic Lookism in Generative AI and Its Systematic Propagation.
3. Canali, C., & **Doh, M.** (2025). Femininity as Infrastructure in Visual AI and the Work of Normative Governance. **accepted abstract*
4. **Doh, M.**, Riccio, P., Hölting, B., Calderon Lopez, O., Munarini, M., Canali, C., Ogolla, S., & Oliver, N. (2025). Between and Beyond: Designing for Identity Complexity in HCI.

2.4.2 Accepted

1. **Doh, M.**, Mazini Rodrigues, C., Boutry, N., Najman, L., Mancas, M., & Bersini, H. (2023). Bridging Human Concepts and Computer Vision for Explainable Face Verification. In *Proceedings of the BEWARE-23 Joint Workshop at the 22nd International Conference of the Italian Association for Artificial Intelligence (AIxIA 2023)*, Rome, Italy [116].
2. **Doh, M.**, Canali, C., & Karagianni, A. (2024). Pixels of Perfection and Self-Perception: Deconstructing AR Beauty Filters and Their Challenge to Unbiased Body Image. In *Proceedings of the ACM International Conference on Interactive Media Experiences (IMX 2024)*, Stockholm, Sweden [117].
3. **Doh, M.**, & Karagianni, A. (2024). “My kind of woman”: Analysing gender stereotypes in AI through the averageness theory and EU law. In *Proceedings of the Imagining the AI Landscape after the AI Act Workshop (IAIL 2024) at the Conference on Hybrid Human-Artificial Intelligence (HHAI 2024): Hybrid Human-AI Systems for the Social Good*, Malmö, Sweden [118].
4. **Doh, M.**, Gulati, A., & Oliver, N. (2025). Attractive by Design: How The Attractiveness Halo Effect Shapes AI Perception. *Collaborative AI and modeling of Humans (CAIHu 2025) - Bridge program at 39th Annual AAAI Conference on Artificial Intelligence (AAAI 2025)*, Philadelphia, US.
5. **Doh, M.**, Gulati, A., Mancas, M., & Oliver, N. (2025). When algorithms play favorites: Lookism in the generation and perception of faces. In *Proceedings of the European Workshop on Algorithmic Fairness (EWAF 2025)*, Eindhoven, Netherlands [119].
6. **Doh, M.**, Canali, C., & Oliver, N. (2025). What TikTok claims, what Bold Glamour does: A filter’s paradox. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT 2025)*, Athens, Greece [120].

7. **Doh, M.**, Canali, C., & Oliver, N. (2025). Filters of identity: AR beauty and the algorithmic politics of the digital body. presented at *Body Politics: Unpacking Tensions and Future Perspectives for Body-Centric Design Research in HCI Workshop at the ACM Conference on Human Factors in Computing Systems (CHI 2025)*, Yokohama, Japan [121].
8. **Doh, M.**, Hölting, B., Riccio, P., & Oliver, N. (2025). Position: The categorization of race in ML is a flawed premise. In *Proceedings of the Forty-Second International Conference on Machine Learning (ICML 2025)*, Vancouver, Canada [122]. *Spotlight paper

2.5 Research Outputs Outside this Dissertation

1. Stassin, S., Englebert, A., Nanfack, G., Albert, J., Versbraegen, N., Peiffer, G., **Doh, M.**, Riche, N., Frenay, B., & De Vleeschouwer, C. (2023). An experimental investigation into the evaluation of explainability methods for computer vision. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2023)*, Torino, Italy [123].
2. Albert, J., Balfroid, M., **Doh, M.**, Bogaert, J., La Fisca, L., De Vos, L., Renard, B., Stragier, V., & Jean, E. (2023). User preferences for large language model versus template-based explanations of movie recommendations: A pilot study. In *Proceedings of the Dutch-Belgian Workshop on Recommender Systems (DBWSR 2023)*, Antwerp, Belgium [124].
3. Karagianni, A., & **Doh, M.** (2024). A feminist legal analysis of non-consensual sexualized deepfakes: Contextualizing its impact as AI-generated image-based violence under EU law. *Porn Studies* [125].
4. Albert, J., Balfroid, M., **Doh, M.**, Rossello, L., Jean, E. & Frenay, B., & Dumas, B. (2025). Towards Graph-based Tools for Literature Exploration: An Analysis of User Requirements. In *Proceedings of 36th Conference Internationale Francophone sur l'interaction Humain-Machine (IHM 2025)*, Toulouse, France [126].



Chapter 3

Reclaiming Sight: Democratizing Vision

This chapter examines how explainability might help close the gap between machine decision-making and human understanding in the context of face verification systems. Grounded in the paradigm of *artificial cognition*, it introduces a cognitively inspired, concept-based interpretability framework that moves beyond conventional saliency maps. By integrating semantic perturbation techniques with human-centered design and natural language outputs, the chapter proposes a preliminary approach to generating more intuitive and accessible explanations for **face verification**. While exploratory in nature, this work aims to contribute to more interpretable and accountable uses of AI in face analysis.

3.1 Situating Explainable Artificial Intelligence in the *Trustworthy AI* Debate

“Its origin and purpose still a total mystery.” (2001: *A Space Odyssey*) [127]

In *2001: A Space Odyssey* [127], Stanley Kubrick stages the sudden appearance of a smooth, impenetrable black monolith on prehistoric Earth. Its meaning is unknown, yet its presence catalyzes a profound transformation: hominids begin to use tools, reshaping their environment and themselves. The monolith does not explain itself—it acts. Its opacity is not a failure, but part of its power (Figure 3.1).

Artificial intelligence today operates with a similar aura of inscrutability. It influences decisions about healthcare, hiring, education, policing, and more, increasingly structuring our collective futures. Yet, despite its ubiquity, AI often remains opaque—its decisions unreadable, its reasoning inaccessible. As Frank Pasquale notes in [128], the mystery surrounding algorithms produces an imbalance of power. Like Kubrick’s monolith—black and impenetrable—these systems give those who design and deploy them the ability to shape markets, influence opinions, and monitor individuals, while those affected are left unable to understand or contest what is done to them.

This opacity lies at the heart of contemporary debates on *Trustworthy AI*, a paradigm that, as outlined in the previous section, has emerged in response to the growing societal and ethical concerns surrounding automated decision-making. Among the foundational principles articulated by frameworks such as the EU’s Ethics Guidelines for *Trustworthy AI* [47], transparency and explainability occupy a central position. They are not merely technical features to be engineered but normative demands tied to democratic accountability, legal rights, and epistemic justice.

XAI has developed as the technical instantiation of this principle. Initially rooted in early expert systems [129] and now revitalized in the era of Deep Learning (DL) [130], XAI encompasses a broad set of techniques aimed at rendering AI model behavior more intelligible. It includes both post-hoc explanation tools and inherently interpretable models, and has been positioned as a key strategy to mitigate the democratic deficits of algorithmic systems.

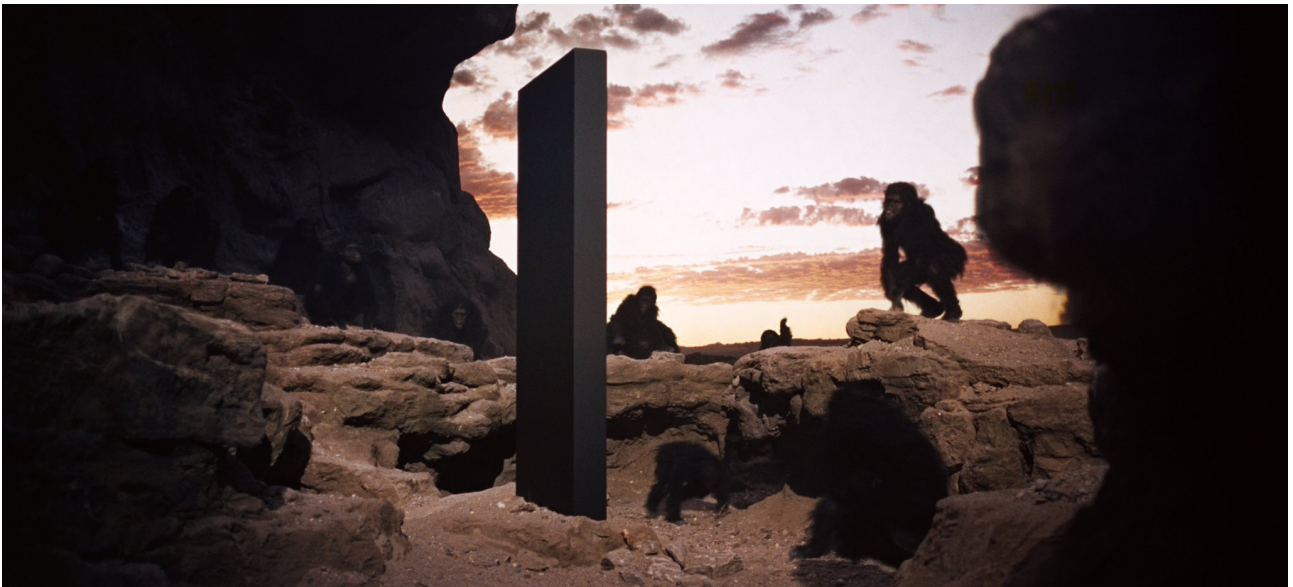


Figure 3.1: Frame from *2001: A Space Odyssey* (dir. Stanley Kubrick, 1968). The monolith as a symbol of opaque and transformative technology [127].

Recent regulatory frameworks reinforce this need [131]. The General Data Protection Regulation (GDPR)¹ [132] and the EU AI Act [133] both affirm individuals’ rights to receive understandable explanations for decisions that significantly affect them. While the legal formulations remain debated—particularly regarding what constitutes a meaningful explanation [134, 135]—the underlying recognition is clear: opacity is not a neutral technical characteristic, but a site of contestation over autonomy, agency, and control.

However, the effectiveness of XAI in fulfilling these demands remains contested. As noted in critiques of the *Trustworthy AI* paradigm, formal explainability tools often fall short of providing the kinds of insight users—especially non-experts—need to make sense of decisions [59, 60, 75]. Many popular techniques [56, 136, 137] produce outputs that are technically valid yet cognitively misaligned [138, 139, 75, 140], requiring inferential leaps that can obscure rather than clarify the model’s logic—risking not only misunderstanding, but misplaced trust in automated decisions. When people cannot understand or contest these decisions, the principles of *Trustworthy AI* and meaningful contestability are undermined. This highlights a crucial distinction: explanations must not only exist, but also be comprehensible. Without this, explainability risks becoming a formal exercise rather than a genuine means of accountability.

This has led to a shift from explainability as a technical fix to explainability as a sociotechnical process. As discussed in the previous section, a sociotechnical critique foregrounds how systems operate not just through algorithms, but through contexts, institutions, and power dynamics. In this view, the goal of XAI is not simply to open the black box, but to enable forms of understanding, contestation, and negotiation that are meaningful to those affected.

This chapter engages with this expanded vision of explainability by situating it within a broader idea of democratizing algorithmic vision. Far from being a purely technical matter, the opacity of AI systems—particularly in face analysis—raises urgent questions about visibility, accountability, and control. Face verification technologies are increasingly embedded in everyday infrastructures, where they can determine access to essential services, employment, and rights. While this chapter focuses on face verification—comparing two facial images to determine whether they belong to the same individual—the technology is closely related to face recognition, which seeks to identify a person by

¹European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with regard to the Processing of Personal Data and on the Free Movement of Such Data (General Data Protection Regulation). Official Journal of the European Union, L119, 4 May 2016, pp. 1–88.

matching them against a larger database. Both tasks rely on similar underlying architectures and raise comparable concerns about bias, opacity, and contestability.

The harms of opaque face analysis systems are well-documented. In 2019, for instance, a federal study confirmed the racial bias of many commercial facial recognition systems, revealing disproportionate error rates for non-white and female faces [141]. Similar issues have emerged in the private sector: Indian Uber drivers have been locked out of their accounts due to failed facial recognition checks [104], and Uber Eats has settled legal complaints for racialized misidentification of workers in the UK [142]. In the United States, several high-profile wrongful arrests [143, 144, 145, 146] have exposed the disproportionate harms of police use of facial recognition, where false matches have led to detention, prosecution, and lasting social and economic consequences.

These cases highlight how facial recognition systems increasingly shape access to work, services, and rights. In this context, explainability becomes essential—not only to improve technical transparency, but to ensure that affected individuals can understand and respond to the outcomes these systems produce. As one wrongfully accused man put it: “I knew I was innocent, so how do I beat a machine?” [146]. His words capture the central problem: without meaningful ways to challenge automated decisions, individuals are left powerless in the face of inscrutable systems. In the case of face verification, this calls for moving beyond abstract saliency maps and toward cognitively aligned, semantically structured explanations—ones that reflect how people perceive and interpret faces. By aligning model outputs with human reasoning, such explanations can help restore a degree of intelligibility and trust in automated identity decisions.

The aim is to shift from explanations that are merely *available* to those that are truly *understandable*—and, ultimately, to support more equitable and accountable forms of algorithmic vision.

This research does not aim to improve face verification systems or justify their deployment. Rather, it adopts Critical Technical Practice [99] to interrogate a foundational assumption: that technical explanations—such as saliency maps or feature attributions—are actually interpretable to those affected by automated decisions. The framework developed in this chapter prioritizes cognitive alignment over technical sophistication, translating algorithmic decisions into forms that reflect how people actually perceive and reason about faces. This enables individuals subjected to automated identity decisions to understand why a system rejected them and to challenge outcomes that restrict access to work, services, or rights. Explainability here functions as a tool for contestation, not legitimation.

3.2 Rethinking Visual Explainability in Face Verification: Background and Motivation

XAI in computer vision has primarily focused on classification tasks, where the goal is to assign labels to single inputs. A wide range of interpretability methods has been developed, including model-specific techniques such as Class Activation Mapping (CAM) [147, 148] and Layer-wise Relevance Propagation (LRP) [149], as well as model-agnostic approaches like LIME [56] and RISE [150].

Yet, the mere generation of explanations does not guarantee meaningful understanding. The real challenge lies in ensuring that these outputs are truly interpretable, that is, translatable into concepts users can grasp. As Renftle et al. [134] argue “one of the challenges is *Translation*”: turning the technical attributes used by models to discriminate between data points into human-understandable features. This translation lies at the heart of interpretability.

Despite their popularity, attribution methods—techniques that assign importance scores to input features based on their contribution to a model’s output—often fall short of this goal. Most rely on pixel-level saliency—highlighting image regions deemed relevant by the model (examples in Figure 3.2). While these visualizations are common, they are not always intuitive [75, 151, 152, 153], especially for non-experts, as they require users to infer high-level meaning from low-level visual cues [154]. As a result, users may misinterpret the explanation or place misplaced trust in it, especially

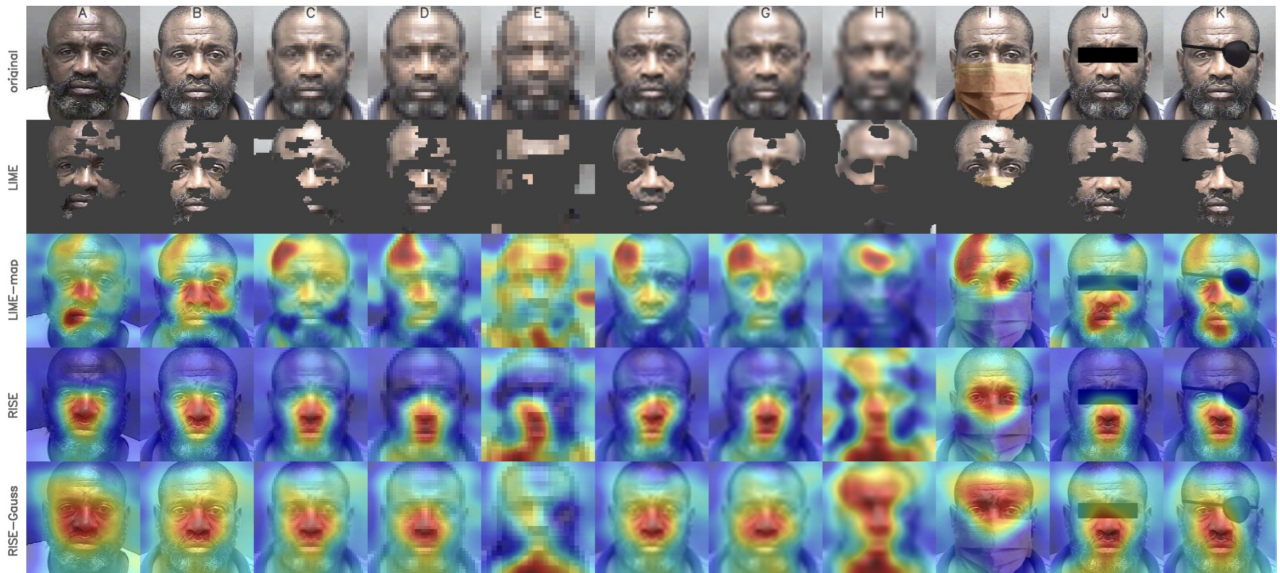


Figure 3.2: Example of traditional visual explanation methods (LIME and RISE) applied to face verification, where image A is compared to A, B, C .. K. Image adapted from [155].

when the output lacks semantic structure [139].

To mitigate this burden, researchers have proposed concept-based explanation techniques. Some focus on global interpretability—mapping internal model representations to human concepts across entire classes. Notable examples include Prototypes [156] and Concept Activation Vectors (CAV) [157], which help visualize how models associate inputs with specific attributes. However, these methods often rely on curated concept datasets, which are costly to produce and prone to bias, especially in specialized domains [75, 158].

A promising alternative lies in unsupervised approaches that discover human-interpretable features directly from data. Clustering techniques [159], for instance, can reveal coherent visual patterns without requiring predefined labels. Similarly, Apicella et al. [139] propose hierarchical segmentation combined with autoencoders to generate explanations at multiple levels of abstraction, adaptable to diverse user needs.

Despite this progress, explainability in face analysis remains underdeveloped. This field is central to highly sensitive applications such as surveillance, emotion detection, demographic inference, and biometric authentication [160, 161, 162, 163]. Yet, most XAI methods have been studied in object classification scenarios, limiting their applicability to facial tasks.

Within this context, face verification introduces unique challenges. Unlike classification, which assigns labels to individual inputs, verification compares two facial images to assess identity similarity—typically through a similarity score. Explaining such decisions requires understanding not just salient features in one image, but also the relationship between two inputs. Systems used in passport control, phone unlocking, or secure authentication rely on these relational inferences, yet standard XAI methods are ill-equipped to capture them.

To address this gap, this chapter introduces an approach that combines mid-level spatial explanations with concept-based reasoning. Specifically, we propose using superpixels—visually coherent image segments—explicitly linked to semantically meaningful, user-definable concepts. This integration of spatial structure with human-centered semantics enables personalized explanations that are both coherent and cognitively accessible.

By drawing on insights from cognitive psychology, our approach contributes to what has been termed XAI 2.0 [75]—a paradigm advocating interdisciplinary, user-aware, and empirically grounded explainability methods.

This work builds upon and extends existing XAI approaches in several key ways. While perturbation-

based methods like LIME and occlusion-based explanations [155] have been applied to face analysis, they typically rely on arbitrary image patches (superpixels or Gaussian masks) rather than semantically meaningful regions. Concept-based methods like TCAV [157] and Prototypes [156] offer semantic structure but require curated datasets and focus primarily on classification rather than similarity-based tasks.

Our framework addresses these gaps by: (1) adapting three complementary XAI techniques (KernelSHAP (local), LIME (context-aware), and MAGE (globally structured)) specifically for face verification rather than classification; (2) replacing arbitrary perturbations with user-defined semantic regions grounded in cognitive science research on facial perception [164, 165, 166]; (3) introducing a global-local aggregation strategy that combines per-image explanations into dataset-level importance rankings; and (4) integrating LLM-based transcription to bridge technical outputs and lay understanding. This positions our work within the XAI 2.0 paradigm [75], which emphasizes user-centered, interdisciplinary approaches that prioritize interpretability for affected individuals over purely technical validation metrics. Unlike approaches that focus primarily on technical validation metrics, our framework emphasizes empirical assessment of whether explanations are comprehensible to non-expert users, treating interpretability as a sociotechnical question that requires both technical rigor and user-centered evaluation.

The remainder of this chapter focuses on the case of face verification, where we evaluate this strategy both quantitatively and through user feedback, responding to the growing demand for practical, user-tested XAI systems.

3.3 Research Questions²

This chapter investigates how cognitively aligned explanations can improve interpretability in face verification systems. The following questions guide the inquiry:

RQ1.1: How can a concept-based, cognitively inspired framework enhance the interpretability of similarity-based decisions in face verification? Traditional explainability methods in computer vision rely on pixel-level saliency maps that often lack semantic structure and cognitive alignment. There is a need to explore whether concept-based explanations grounded in human-interpretable facial regions (eyes, nose, mouth) can provide more intuitive and accessible explanations for face verification decisions compared to conventional attribution methods.

RQ1.2: How do local, context-aware, and globally structured explanation strategies perform when adapted to a semantic perturbation framework for interpreting face verification models? While most XAI methods focus on classification tasks, face verification presents unique challenges as it involves comparing two images to assess identity similarity. There is limited research on how different explanation strategies—specifically KernelSHAP (local), LIME (context-aware), and MAGE (globally structured)—perform when adapted to semantic perturbation approaches for interpreting relational similarity decisions in face verification systems.

RQ1.3: How do users from diverse backgrounds interpret and assess semantic explanations generated by the proposed framework? Given that explanations must be meaningful to those affected by algorithmic decisions, there is a critical need to understand how users with different technical backgrounds perceive and evaluate concept-based explanations. This includes assessing user preferences for different levels of semantic granularity, visualization methods, and the integration of natural language transcriptions generated by large language models.

²Under the main research question (RQ1) *Reclaiming Sight: Democratizing Algorithmic Vision* presented in Chapter

3.4 Methodology

3.4.1 A Cognitively-Inspired Semantic Perturbation Framework

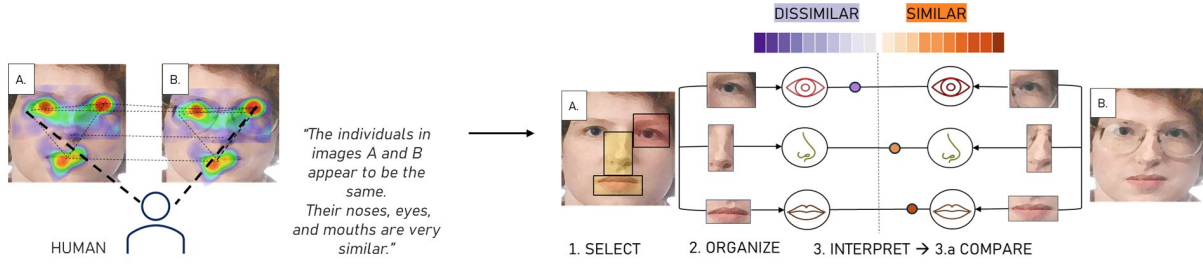


Figure 3.3: On the left, the process of human face recognition is illustrated, highlighting how attention is directed to specific facial regions. On the right, an adaptation of the XAI Perceptual Processing Framework—originally proposed by Zhang et al. [167]—is presented, tailored to face verification and inspired by human strategies for visual processing.

In the broader debate around transparency and intelligibility in artificial intelligence, understanding how decisions are made requires more than technical access to inner workings. It requires an alignment with how humans perceive, reason, and interpret [167]. This section introduces a cognitively informed methodology for interpreting face verification systems, grounded in the premise that explanations must be meaningful not only to engineers but also to those affected by algorithmic decisions.

Cognitive psychology provides foundational insights into how individuals perceive, organize, and interpret visual stimuli—particularly in the context of facial recognition. The human perceptual process typically unfolds in three stages: selection, organization, and interpretation [168]. When recognizing faces, attention is often drawn to specific regions such as the eyes and nose [164, 165, 166]. These salient areas are then grouped into meaningful concepts, enabling comparison and categorization. This comparative mechanism may operate against a memorized internal representation or between two simultaneously perceived faces. A familiar example of this process is captured in expressions such as “You look like your mother—you have the same eyes,” where perceived similarity is conveyed through a conceptually defined facial feature (see Figure 3.3).

In line with the paradigm of artificial cognition [77, 113, 69]—which seeks to understand and explain machine behavior through cognitive psychology—this chapter adopts a methodological framework designed to bridge the gap between machine-generated reasoning and human interpretability. Artificial cognition is conceptualized here as a two-way translation process between machine and human cognition, viewed as distinct yet interacting systems. The aim is not to replicate human reasoning, but to establish an intelligible protocol through which humans can meaningfully interpret decisions made by artificial agents.

From this perspective, conventional saliency maps in computer vision are critically assessed for their misalignment with human perceptual strategies. Human attention in face processing is guided both top-down—through knowledge of typical facial structures—and bottom-up—by sensitivity to atypical or distinctive features. In contrast, AI saliency maps often emphasize low-level textures or patterns that carry little semantic relevance for human observers.

To address this misalignment, the cognitively inspired framework proposed in [167] has been adapted to develop a task-specific methodology for face verification. The core idea is to interpret a face verification system in a model-agnostic manner, using a semantic perturbation approach grounded in similarity scoring.

Let \mathcal{I}_A and \mathcal{I}_B denote two facial images, and let Ξ be a trained face verification model. The model computes embeddings $Activ(\mathcal{I}_A; \Xi)$, $Activ(\mathcal{I}_B; \Xi) \in \mathbb{R}^d$, and their similarity is measured by cosine

similarity:

$$S_{AB} = \frac{\text{Activ}(\mathcal{I}_A; \Xi) \cdot \text{Activ}(\mathcal{I}_B; \Xi)}{\|\text{Activ}(\mathcal{I}_A; \Xi)\|_2 \|\text{Activ}(\mathcal{I}_B; \Xi)\|_2} \quad (3.1)$$

Let $\mathcal{C} = \{c_1, \dots, c_K\}$ denote a set of semantic concepts corresponding to meaningful facial regions (e.g., eyes, nose, mouth). The method aims to identify which concepts contribute to perceived similarity by selectively masking specific regions. For each concept $c_i \in \mathcal{C}$, perturbed versions $\mathcal{I}_A^{(i)}$ and $\mathcal{I}_B^{(i)}$ are generated by masking that region, and a new similarity score $S_{A^{(i)}B^{(i)}}$ is computed. A decrease in the score suggests a positive contribution of the masked concept to overall similarity ($\Delta S_i \geq 0$), while an increase implies a negative contribution ($\Delta S_i < 0$).

Crucially, this masking process is applied only to globally relevant semantic concepts—those identified as conceptually meaningful by the model—rather than across all possible facial segments. This constraint ensures that the analysis focuses on the regions the system relies on most when making its verification decision.

To operationalize this cognitively grounded explanation pipeline, a four-phase framework has been developed and is illustrated in Figure 3.4. The process begins with the semantic formalization of facial features, followed by the extraction of concept-based regions using three interpretability strategies (KernelSHAP [136], LIME [56], and MAGE [169]). These regions form the basis for constructing similarity-based explanations through targeted perturbation. A transcription layer, powered by large language models (Large Language Model (LLM)s), then generates natural-language explanations from the spatial impact of each concept. Finally, the process incorporates an initial phase of user evaluation to assess the clarity and perceptual validity of the generated explanations.

Each phase of the framework is described in detail in the following sections: 1) *semantic features* definition, 2) *concepts extraction*, 3) *similarity map*, and 4) *transcription of semantic contributions using LLMs*.

Taken together, this framework constitutes a communication interface between algorithmic vision and human cognitive expectations.

3.4.2 Definition of Human-based Semantic Features

To bridge the gap between model-generated explanations and human interpretability, facial regions are translated into semantically meaningful concepts using landmarks extracted from Mediapipe [170], an open-source framework for face detection and facial landmark estimation. Based on these landmarks, polygons were defined to segment the face into distinct semantic areas, including the background—intended here as the region outside the facial contour.

For this purpose, a custom annotation script was developed to manually define polygons over Mediapipe’s 468 facial landmarks. The script allowed the user to select landmark points on an example face, snapping each point to the closest landmark to ensure precision, and to combine them into closed polygons. Each polygon was then assigned a semantic label corresponding to a facial region. This procedure was repeated iteratively until the complete set of semantic areas was defined. This semantic segmentation allows for customizable masks based on cognitively relevant facial regions. Three human-based semantic sets were defined to reflect different levels of granularity: SET_0, SET_1, and SET_2 (Figure 3.5).

- SET_0 defines 13 broad features, where areas such as the eyes, nose, and mouth are represented as unified regions without distinguishing between left and right sides.
- SET_1 also contains 13 regions but introduces a bilateral structure, distinguishing symmetrical features like eyes and lips across the left and right sides.
- SET_2 increases granularity to 30 features, offering a finer subdivision of key facial components.

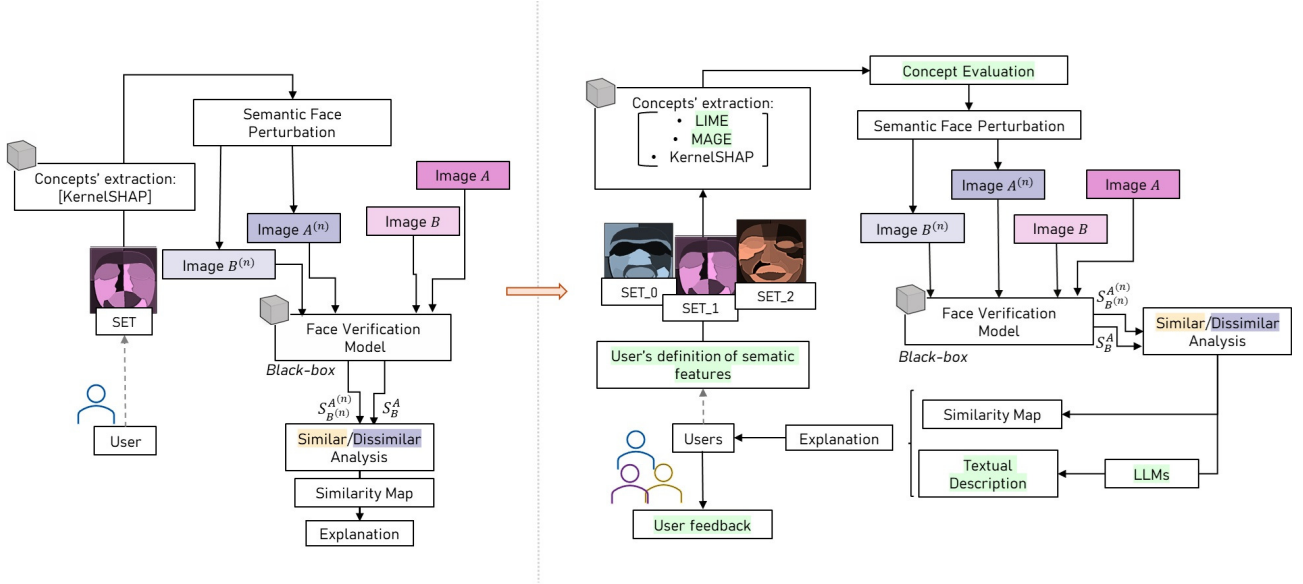


Figure 3.4: This figure illustrates the methodological evolution that took place throughout of the PhD project for this chapter. On the left, the initial framework is shown, outlining the core idea of concept-based perturbation for face verification. On the right, the finalized version is presented, incorporating several extensions: the evaluation of three alternative concept extraction methods (KernelSHAP, MAGE, LIME), the introduction of multiple human-based semantic features to assess variability, the addition of a transcription layer using large language models (LLMs), and the integration of user feedback. New components are highlighted in green.

The three sets were designed hierarchically, with increasing granularity, in order to capture different levels of abstraction. The definition of these sets was not arbitrary: it was inspired by findings from cognitive science, which show that humans organize facial information hierarchically, prioritizing some regions (*e.g.*, eyes and mouth) over others (*e.g.*, hair) in judgments [171].

This hierarchical structuring serves two objectives. First, it enables an evaluation of the method’s sensitivity to different semantic definitions and levels of granularity. Second, it avoids imposing a singular or static interpretation of facial regions, acknowledging the challenge of collecting large-scale user-defined semantic sets for each image or individual.

The landmark-based segmentation includes only the face area, excluding external regions such as hair or ears. This choice aligns with prior findings showing that certain areas—like ears—have minimal influence on similarity judgments compared to core facial regions such as the eyes [171] (Figure 3.6). Moreover, standard face verification pipelines commonly include preprocessing steps that isolate the face region, further justifying this focus.

Figure 3.5 illustrates the process of creating these semantic features. In step 1, Mediapipe’s landmarks are projected onto an example face. Then, users define semantic areas by selecting specific landmarks on Mediapipe’s facemesh. Step 2 shows the three sets of human-based concepts with varying levels of granularity that we hypothesized for this study.

By allowing algorithmic decisions to be explained in terms of user-relevant semantic areas, this approach shifts the explanation process from low-level pixel (individual pixels) attribution to more comprehensible, mid-level (aggregates of pixels) conceptual reasoning. It offers a structure for evaluating how localized facial features contribute to perceived similarity within face verification tasks.

3.4.3 Concepts Extraction

Human cognition intuitively segments images into distinct semantic entities, such as eyes, nose, and mouth. Machines do not have this innate ability; the identification and differentiation of such concepts

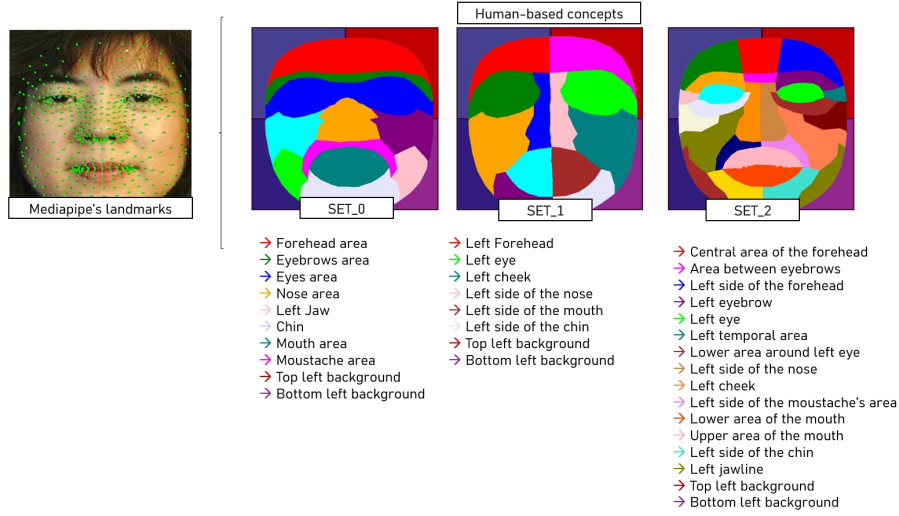


Figure 3.5: The process of creating human-based semantic features. (1) Mediapipe's landmarks projected onto an example face. Using Mediapipe's facemesh, users can define semantic areas by selecting specific landmarks, as shown in the example. (2) Three sets of human-based concepts with varying granularity (SET_0, SET_1, SET_2) created from these user-defined areas. SET_0 and SET_1 have 13 features each, while SET_2 has 30 features.

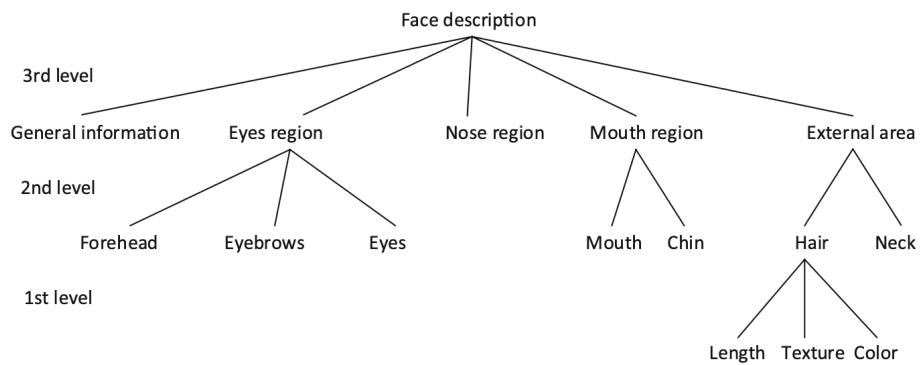


Figure 3.6: Hierarchical structure of facial information used in human face description, as proposed by Karczmarek et al. [171]. The diagram illustrates three levels of abstraction, ranging from low-level features (*e.g.* forehead, eyes, mouth) to more abstract groupings (*e.g.* eyes region, external area), reflecting how humans organize and process facial information.

must be acquired through the training process. A significant challenge arises when the concepts learned by machines do not match those understood by humans, complicating the interpretation of machine decision processes.

Traditionally, XAI methods have focused on generating case-specific explanations to clarify the reasoning behind a model’s decision for a given input. However, this local perspective provides only partial insight. To move toward more generalizable interpretability, the present framework includes a procedure for aggregating multiple local explanations into a global ranking of semantically defined facial regions—ordered by their relative importance to the model’s similarity judgments.

To this end, three interpretability techniques are compared to obtain the model-based relative importance: KernelSHAP [136], LIME [56], and MAGE [169]. While KernelSHAP provides locally grounded explanations based on Shapley value theory, LIME introduces a context-aware surrogate modeling approach, and MAGE offers a globally structured analysis by identifying coherent concept groups from internal activation patterns. Each method is adapted to operate over user-defined semantic masks. This comparison enables an evaluation of whether globally oriented methods provide more informative or consistent explanations than their local counterparts in the context of facial similarity assessment. Moreover, the selection of interpretability methods was guided by two main criteria. First, all chosen techniques are *model-agnostic*, ensuring their applicability to the embedding-based face verification task without requiring class logits or network-specific gradients. Second, these methods are easily adaptable to an input perturbation paradigm, allowing their integration into the proposed semantic-based framework, where explanations are derived from masking human-defined facial concepts rather than individual pixels.

MAGE and Embedding-aware Order Correlation (EaOC): The MAGE technique proposed by Rodrigues et al. [169] is adapted to the present context. The method leverages the last convolutional layer of the network to identify groups of activation patterns that correspond to high-level concepts. To evaluate the importance of such concepts, we use an adapted version of the *Class-aware Occlusion* (CaOc) metric, reformulated here as *Embedding-aware Order Correlation* (EaOC).

Let $\mathcal{D} = \{\mathcal{I}_1, \dots, \mathcal{I}_N\}$ denote a dataset of N face images, and let Ξ be a trained embedding model. Let $\mathcal{C} = \{c_1, \dots, c_K\}$ represent the set of semantic concepts (e.g., eyes, nose, mouth). For each image \mathcal{I}_l , its embedding is denoted by $\text{Activ}(\mathcal{I}_l; \Xi) \in \mathbb{R}^d$.

To characterize the output space, we compute the Euclidean norm of each embedding vector:

$$OS_{\text{Activ}}(\mathcal{D}) = [\|\text{Activ}(\mathcal{I}_1; \Xi)\|_2, \dots, \|\text{Activ}(\mathcal{I}_N; \Xi)\|_2] \in \mathbb{R}^N. \quad (3.2)$$

From these norm values, we define a permutation $\sigma_{\text{orig}} : \{1, \dots, N\} \rightarrow \{1, \dots, N\}$ that orders images in decreasing order of their embedding norms:

$$\|\text{Activ}(\mathcal{I}_{\sigma_{\text{orig}}(1)}; \Xi)\|_2 \geq \|\text{Activ}(\mathcal{I}_{\sigma_{\text{orig}}(2)}; \Xi)\|_2 \geq \dots \geq \|\text{Activ}(\mathcal{I}_{\sigma_{\text{orig}}(N)}; \Xi)\|_2. \quad (3.3)$$

We denote by $\text{rank}_{\sigma_{\text{orig}}}(\mathcal{I}_j)$ the position (rank) of image \mathcal{I}_j in this ordering, i.e., the unique index $r \in \{1, \dots, N\}$ such that $\sigma_{\text{orig}}(r) = j$.

For a given image \mathcal{I}_j and a concept $c_i \in \mathcal{C}$, we denote by $\mathcal{I}_j^{(i)}$ the occluded version of \mathcal{I}_j where concept c_i has been masked out. Replacing \mathcal{I}_j with $\mathcal{I}_j^{(i)}$ in the dataset yields a perturbed dataset:

$$\mathcal{D}_j^{(i)} = \mathcal{D} \setminus \{\mathcal{I}_j\} \cup \{\mathcal{I}_j^{(i)}\}. \quad (3.4)$$

We compute the embedding norms for this perturbed dataset and obtain a new ordering $\sigma_{\text{occ}}^{(j,i)} :$

$\{1, \dots, N\} \rightarrow \{1, \dots, N\}$ defined analogously to Equation 3.3, but using $\mathcal{D}_j^{(i)}$ instead of \mathcal{D} . The EaOC score measures the change in rank of \mathcal{I}_j caused by occluding concept c_i :

$$EaOC(\mathcal{I}_j, i) = |\text{rank}_{\sigma_{orig}}(\mathcal{I}_j) - \text{rank}_{\sigma_{occ}^{(j,i)}}(\mathcal{I}_j^{(i)})|. \quad (3.5)$$

Intuitively, if occluding concept c_i significantly alters the embedding of \mathcal{I}_j , its position in the ranking will shift substantially, producing a high *EaOC* value. Conversely, concepts with little influence will produce only minor rank variations. Figure 3.7 provides a visual illustration of this process.

In this work, MAGE is further adapted to operate on human-defined semantic regions rather than arbitrary image patches. This bridges the network’s internal activation patterns with human-interpretable concepts, allowing us to identify which facial regions contribute most strongly to the model’s embedding space organization.

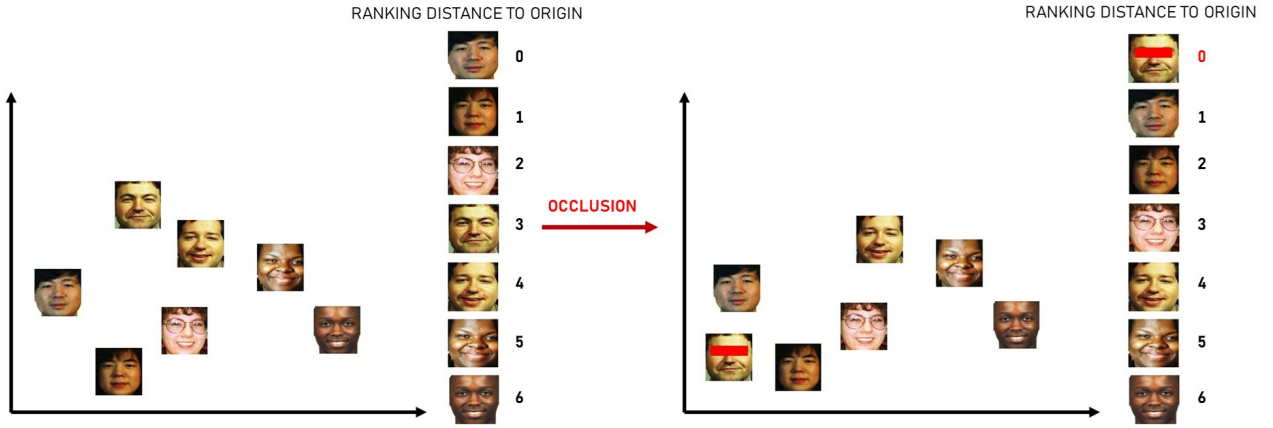


Figure 3.7: Illustrative example of EaOC behavior under occlusion. Each image represents its corresponding embedding obtained from a trained model. Initially, images are ranked based on their distance to the origin in embedding space. After occlusion, the ranking is recomputed; significant occlusions may alter the image’s position in the ranking.

LIME: In addition to the adapted method described above, Local Interpretable Model-Agnostic Explanations (LIME) [56] is employed as a context-aware technique. Although originally designed as a local method, its use of a surrogate model to approximate a local partition of the embedding space introduces a degree of spatial awareness.

The method builds a local surrogate model to approximate the behavior of the black-box embedding function $Activ(\cdot; \Xi)$ around a given input image \mathcal{I}_j .

Formally, let $\mathcal{D} = \{\mathcal{I}_1, \dots, \mathcal{I}_N\}$ denote the dataset, and let $\mathcal{C} = \{c_1, \dots, c_K\}$ represent the set of semantic concepts. For image \mathcal{I}_j , LIME generates a set of perturbed versions $\{\mathcal{I}_j^{(S)}\}$, where each subset $S \subseteq \mathcal{C}$ corresponds to a binary masking of some concepts.

Since the original LIME formulation targets class-specific outputs, the surrogate model is adapted here to approximate the ℓ_1 norm of each image’s embedding instead:

$$y^{(S)} = \|Activ(\mathcal{I}_j^{(S)}; \Xi)\|_1. \quad (3.6)$$

The choice of the ℓ_1 norm, rather than the ℓ_2 norm used elsewhere in the framework, provides a simpler aggregation measure that maintains sensitivity to component-wise changes in the embedding space while reducing computational overhead in the surrogate fitting process.

These input–output pairs $(\mathcal{I}_j^{(S)}, y^{(S)})$ are then used to fit a simple interpretable surrogate model, typically a sparse linear regressor:

$$y^{(S)} \approx \beta_0 + \sum_{i=1}^K \beta_i \cdot z_i^{(S)}, \quad (3.7)$$

where $z_i^{(S)} \in \{0, 1\}$ indicates whether concept c_i is present (1) or occluded (0) in the perturbed image $\mathcal{I}_j^{(S)}$.

In the original LIME formulation, perturbed samples are weighted by a proximity kernel to emphasize local fidelity. In our adaptation, we retain this weighting mechanism, where the weight $\pi_{\mathcal{I}_j}(\mathcal{I}_j^{(S)})$ decreases with the distance between the original and perturbed instances in the semantic feature space. The coefficients β_i serve as local importance scores, estimating how strongly each concept c_i influences the embedding norm of \mathcal{I}_j in its neighborhood.



Figure 3.8: Examples of two images’ human-semantic part importance scores computed using KernelSHAP [136], as presented in the previous study [172] for two models CasiaNet [173] in (a) and (c), and VGGfaces2 [174] in (b) and (d). Green regions indicate higher importance according to the Shapley scores.

KernelSHAP: This method combines the interpretable components of LIME [56] with Shapley values [175] from cooperative game theory, which quantify the marginal contribution of each input feature to the final model output (Figure 3.8).

Let $\mathcal{D} = \{\mathcal{I}_1, \dots, \mathcal{I}_N\}$ be a dataset of face images, and let Ξ be a trained embedding model. For each image \mathcal{I}_l , its embedding is denoted by $Activ(\mathcal{I}_l; \Xi) \in \mathbb{R}^d$. The set of semantic concepts (e.g., eyes, nose, mouth) is denoted by $\mathcal{C} = \{c_1, \dots, c_K\}$.

For a given image \mathcal{I}_j , KernelSHAP estimates how much the presence or absence of each concept c_i contributes to each dimension of the embedding $Activ(\mathcal{I}_j; \Xi)$.

Formally, for each concept c_i and embedding dimension $m \in \{1, \dots, d\}$, the Shapley value is defined as:

$$\phi_{i,m}(\mathcal{I}_j) = \sum_{S \subseteq \mathcal{C} \setminus \{c_i\}} \frac{|S|! (K - |S| - 1)!}{K!} \left[f_m(S \cup \{c_i\}; \mathcal{I}_j) - f_m(S; \mathcal{I}_j) \right], \quad (3.8)$$

where $f_m(S; \mathcal{I}_j)$ denotes the value of embedding dimension m for image \mathcal{I}_j when only the subset of concepts S is present (all other concepts are occluded).

In practice, KernelSHAP approximates these Shapley values rather than computing them exhaustively, as the latter would be computationally intractable for large K . The approximation is achieved by sampling a subset of coalitions S and fitting a weighted linear model, similar to the surrogate approach used in LIME.

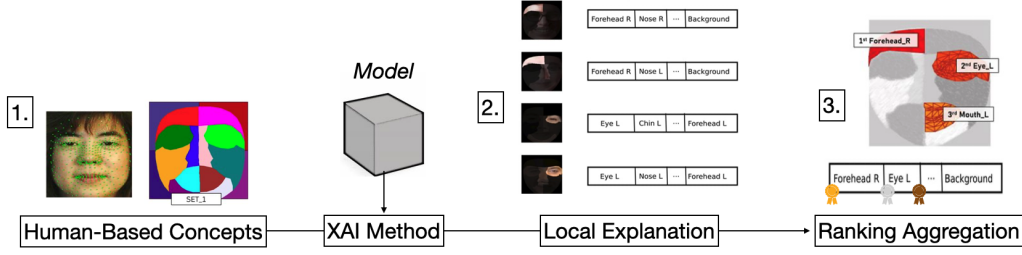


Figure 3.9: Methodology to extract globally important concepts. Human-segmented regions are used to compute explanations for all images (1) via XAI methods such as LIME, KernelSHAP, and MAGE. Each image’s regions are then ordered by importance according to the chosen method (2). Finally, the individual rankings are aggregated to derive a global ranking of facial segments (3).

To obtain a single importance score for each concept c_i on image \mathcal{I}_j , we aggregate across all embedding dimensions:

$$I(\mathcal{I}_j, c_i) = \sum_{m=1}^d |\phi_{i,m}(\mathcal{I}_j)|. \quad (3.9)$$

Unlike classification tasks where the sign of $\phi_{i,m}$ may indicate a positive or negative influence on a class probability, in similarity-based verification tasks both increases and decreases in embedding activations are informative. The absolute values are therefore used, so that $I(\mathcal{I}_j, c_i)$ reflects the overall sensitivity of the embedding to concept c_i for image \mathcal{I}_j . Figure 3.8 shows examples of importance scores for human-defined facial regions.

Concepts aggregation

After applying an XAI method to a single image $\mathcal{I}_j \in \mathcal{D}$, we obtain importance scores $I(\mathcal{I}_j, c_i)$ for each semantic concept $c_i \in \mathcal{C}$ (as shown in Figure 3.9 (1)). These scores represent the contribution of each facial segment to the model’s decision for that specific image. While these local explanations are informative, they reflect only the behavior of the model on individual instances.

The objective of concept aggregation is to identify facial segments that are consistently important across the dataset—namely, segments that influence the model’s behavior in the majority of cases. This is achieved by aggregating multiple image-level explanations into a unified global importance ranking.

Per-image ranking. For each image $\mathcal{I}_j \in \mathcal{D}$, the importance scores $\{I(\mathcal{I}_j, c_1), \dots, I(\mathcal{I}_j, c_K)\}$ are used to produce a ranking $\rho_j : \mathcal{C} \rightarrow \{1, \dots, K\}$, where $\rho_j(c_i)$ denotes the rank position of concept c_i for image \mathcal{I}_j (Figure 3.9 (2)). A rank of 1 indicates the most important concept, while K indicates the least important.

Global ranking via BORDA count. Given the set of per-image rankings $\{\rho_1, \dots, \rho_N\}$ for all N images in \mathcal{D} , we aggregate them using the BORDA count [176]. For each concept c_i , the BORDA score is computed as:

$$B(c_i) = \sum_{j=1}^N (K - \rho_j(c_i) + 1), \quad (3.10)$$

where concepts with lower (better) ranks across many images receive higher BORDA scores. The global ranking $\rho_{global} : \mathcal{C} \rightarrow \{1, \dots, K\}$ is then obtained by sorting concepts in decreasing order of their BORDA scores (Figure 3.9 (3)).

Adaptation for MAGE. A different aggregation strategy is adopted for MAGE, since the method decomposes the network into multiple concept clusters based on activation patterns. For each image \mathcal{I}_j , BORDA count is first applied to combine rankings from different network layers, producing a unified per-image ranking ρ_j . These image-level rankings are then aggregated across the dataset using Equation 3.10 to obtain the global importance ordering.

Weighting for visualization. Rather than restricting the visualization to only the top-ranked concepts, the current framework incorporates each segment’s global rank position as a weight $g_i = K - \rho_{global}(c_i) + 1$ in the similarity map computation (see Algorithm 1). This weighting modulates the local explanations, enabling a more nuanced and consistent visualization across samples that reflects both local impact and global relevance.

Together, these methods—adapted to operate over human-based semantic sets—enable the identification and ranking of facial regions offering a practical trade-off between methodological rigor and cognitive alignment.

3.4.4 Similarity map

To generate local explanations of similarity decisions, a modified version of the single removal algorithm $S0$ —originally introduced with Gaussian masks [155]—is employed. This method was subsequently adapted to work with semantic masks defined over human-interpretable facial regions. In the present framework, the algorithm is further refined by incorporating the global importance of each facial region, as derived from the *Concepts Aggregation* step (Section 3.4.3), into the perturbation weighting mechanism of Algorithm 1.

Let $\mathcal{C} = \{c_1, \dots, c_K\}$ denote the set of semantic concepts, and let $\rho_{global} : \mathcal{C} \rightarrow \{1, \dots, K\}$ be the global ranking obtained from concept aggregation. For each concept c_i , we define a global importance weight as:

$$g_i = K - \rho_{global}(c_i) + 1, \quad (3.11)$$

where g_i is highest for the most important concepts (rank 1) and lowest for the least important concepts (rank K).

Given two facial images \mathcal{I}_A and \mathcal{I}_B , the face verification model computes embeddings $Activ(\mathcal{I}_A; \Xi)$ and $Activ(\mathcal{I}_B; \Xi)$, and their similarity is measured by the cosine similarity:

$$S_{AB} = \frac{Activ(\mathcal{I}_A; \Xi) \cdot Activ(\mathcal{I}_B; \Xi)}{\|Activ(\mathcal{I}_A; \Xi)\|_2 \|Activ(\mathcal{I}_B; \Xi)\|_2}. \quad (3.12)$$

For each semantic concept $c_i \in \mathcal{C}$, let $\mathcal{I}_A^{(i)}$ and $\mathcal{I}_B^{(i)}$ denote the occluded versions of the two images where concept c_i has been masked out. The perturbed similarity score is computed as:

$$S_{A^{(i)}B^{(i)}} = \frac{Activ(\mathcal{I}_A^{(i)}; \Xi) \cdot Activ(\mathcal{I}_B^{(i)}; \Xi)}{\|Activ(\mathcal{I}_A^{(i)}; \Xi)\|_2 \|Activ(\mathcal{I}_B^{(i)}; \Xi)\|_2}. \quad (3.13)$$

The similarity difference for concept c_i is then:

$$\Delta S_i = g_i \cdot (S_{AB} - S_{A^{(i)}B^{(i)}}), \quad (3.14)$$

where the global weight g_i modulates the contribution based on the concept’s importance across the dataset.

The basic idea of this algorithm is to perturb facial images to understand the impact of specific facial regions on the similarity score between two images. The main steps are outlined in Algorithm 1.

where $M_A^{(i)}$ and $M_B^{(i)}$ denote the binary masks corresponding to concept c_i in images \mathcal{I}_A and \mathcal{I}_B , respectively.

Algorithm 1 Single Removal - S0 with Global Weighting

```
1: Input: Images  $\mathcal{I}_A, \mathcal{I}_B$ ; semantic concepts  $\mathcal{C} = \{c_1, \dots, c_K\}$ ; global weights  $\{g_1, \dots, g_K\}$ 
2: Output: Similarity maps  $S0_A$  and  $S0_B$ 
3:
4: Initialize contribution maps  $H0_A(c_i) \leftarrow 0$  and  $H0_B(c_i) \leftarrow 0$  for all  $c_i \in \mathcal{C}$ 
5: for each semantic concept  $c_i \in \mathcal{C}$  do
6:   Compute  $\Delta S_i = g_i \cdot (S_{AB} - S_{A^{(i)}B^{(i)}})$  ▷ Equation 3.14
7:   Compute pair-wise weight  $W_i^{AB} = \frac{\text{Area}(c_i \text{ in } \mathcal{I}_A) \cdot \text{Area}(c_i \text{ in } \mathcal{I}_B)}{\sum_{j=1}^K \text{Area}(c_j \text{ in } \mathcal{I}_A) \cdot \text{Area}(c_j \text{ in } \mathcal{I}_B)}$ 
8:   Calculate contribution  $C_i = \Delta S_i \cdot W_i^{AB}$ 
9:   Update  $H0_A(c_i) \leftarrow C_i$  and  $H0_B(c_i) \leftarrow C_i$ 
10: end for
11:
12: Normalize contributions separately for positive and negative values:
13: for each semantic concept  $c_i \in \mathcal{C}$  do
14:   if  $H0_A(c_i) \geq 0$  then
15:      $H0_A^+(c_i) = \frac{H0_A(c_i)}{\sum_{c_j: H0_A(c_j) \geq 0} |H0_A(c_j)|}$ 
16:      $H0_A^-(c_i) = 0$ 
17:   else
18:      $H0_A^+(c_i) = 0$ 
19:      $H0_A^-(c_i) = \frac{H0_A(c_i)}{\sum_{c_j: H0_A(c_j) < 0} |H0_A(c_j)|}$ 
20:   end if
21:   Repeat for  $H0_B^+(c_i)$  and  $H0_B^-(c_i)$ 
22: end for
23:
24: Generate similarity maps:
25:  $S0_A = \sum_{i=1}^K (H0_A^+(c_i) + H0_A^-(c_i)) \cdot M_A^{(i)}$ 
26:  $S0_B = \sum_{i=1}^K (H0_B^+(c_i) + H0_B^-(c_i)) \cdot M_B^{(i)}$ 
```

The algorithm begins by initializing the contribution maps to zero for both images. For each semantic concept c_i , the similarity difference ΔS_i is calculated between the original similarity score S_{AB} and the perturbed similarity score $S_{A^{(i)}B^{(i)}}$ obtained after masking concept c_i in both images. The global importance weight g_i (Equation 3.11) modulates this difference to reflect the concept's relevance across the dataset.

The contribution C_i is determined by multiplying ΔS_i by a pair-wise weight W_i^{AB} , which accounts for the relative spatial extent of concept c_i in both images. This ensures that larger regions do not dominate the explanation purely due to their size.

Once all contributions are calculated, they are normalized separately for positive and negative values. Positive contributions indicate that the presence of a concept increases similarity between the images, while negative contributions indicate that the concept introduces dissimilarity. This separation enhances interpretability by clearly distinguishing between features that align versus features that diverge.

Finally, the overall similarity maps $S0_A$ and $S0_B$ are obtained by summing the normalized contributions weighted by their corresponding semantic masks $M_A^{(i)}$ and $M_B^{(i)}$.

While the overall procedure builds on the Single Removal (S0) algorithm introduced by Mery et al. [155], our implementation departs from the original in several key respects. We retain the core principle of estimating regional contributions by occluding image parts and comparing the resulting similarity scores, but introduce five main contributions: (1) replacing Gaussian patches with human-

interpretable semantic masks derived from the set \mathcal{C} ; (2) integrating a global weighting factor g_i (Equation 3.11), derived from the Concepts Aggregation step, to modulate the impact of each region based on its importance across the dataset; (3) introducing a pair-wise weighting term W_i^{AB} to balance the relative size and contribution of regions across both images; (4) extending the computation of contribution maps to both images in the verification pair, rather than only the probe; and (5) separating the normalization of positive and negative contributions to improve interpretability.

The resulting similarity maps serve as a bridge between algorithmic reasoning and human perception, providing a visual basis for the natural-language explanations and user evaluation discussed in the following sections.

3.4.5 Generation of Textual Explanations Using Large Language Models

To enhance the accessibility of visual and tabular outputs for non-expert users, an additional layer of explanation is introduced through the generation of natural language descriptions using large language models (LLMs). This process translates the numerical contributions of each semantic region and the global similarity score into concise and human-readable explanations.

Three small-scale LLMs available on Hugging Face were selected for this task: CodeLlama-7B [177]³, Zephyr-7B [178]⁴, and Beagle14-7B⁵. These models were chosen for their capacity to produce coherent outputs efficiently and without the need for additional training. All experiments were conducted within the LMstudio environment⁶. All models were downloaded via LMstudio and executed locally, without relying on external APIs or remote servers.

CodeLlama-7B, developed by Meta, is optimized for code-related tasks but is also capable of generating structured technical prose. Zephyr-7B specializes in natural language understanding across general domains, while Beagle14-7B is fine-tuned for instruction following and task-oriented reasoning, providing precise and context-aware responses.

Each model received a structured prompt designed to contextualize the explanation task. The prompt included the cosine similarity score [`cosine_similarity_percentage`] and a table of regional similarity contributions [`contributions_table`]. It clarified how positive and negative values indicate similarity or dissimilarity, and referred to a corresponding color map where orange denotes similarity and purple denotes dissimilarity, with color intensity reflecting the degree of alignment or divergence.

Importantly, no facial images were ever provided to the LLMs: the inputs consisted exclusively of numerical similarity scores and regional contribution tables produced by the algorithm described earlier. In this sense, the role of the LLMs was limited to transcribing structured numerical outputs into natural language, rather than processing biometric data directly.

The exact prompt is reported below:

Prompt:

“Context: A face verification system assigns a cosine similarity score between two images. In this instance, the cosine similarity is [`cosine_similarity_percentage`] (a percentage from 0 to 100%). From the model’s knowledge, several main human-understandable concepts are extracted; these concepts are used to explain the model’s output (cosine similarity). These concepts are associated with a similar/dissimilar score. Specifically, when a value is positive or equal to zero (≥ 0), the model perceives these areas in the two images as similar. Conversely, they are seen as dissimilar when the value is negative (example: -0.5): [`contributions_table`]. Given that a color map is displayed where shades of purple

³<https://ollama.com/library/codellama:7b>, Last Access: 20-08-2025.

⁴<https://huggingface.co/HuggingFaceH4/zephyr-7b-beta>, Last Access: 20-08-2025.

⁵<https://huggingface.co/mlabonne/Beagle14-7B>, Last Access: 20-08-2025.

⁶<https://lmstudio.ai>, Last Access: 20-08-2025.

indicate dissimilarity and shades of orange indicate similarity, with color intensity proportional to the magnitude of the similarity or dissimilarity, provide a simple explanation of why the cosine similarity between the two images is [\[cosine_similarity_percentage\]](#).
No long explanation”

To ensure clarity and consistency across models, the prompt included precise formatting instructions—such as “*a percentage from 0 to 100%*”, “*example: $\geq 0/-0.5$* ”, and “*No long explanation*”—which constrained the style and structure of the generated outputs.

The use of LLMs in this context supports the automatic generation of textual explanations that complement visual outputs, adapting their tone and clarity depending on the model used. A detailed discussion of the generated responses and prompt design rationale is provided in Appendix A.1.

3.5 Metrics

3.5.1 Concept Evaluation Experiments

Inspired by the metrics proposed by Bommer et al. [179], three aspects of the model’s concept-based explanations are evaluated: (1) *faithfulness*, (2) *sensitivity*, and (3) *randomization*. Faithfulness assesses the extent to which an explanation reflects the actual behavior of the model. Sensitivity measures how responsive the model’s output is to changes in the input. Randomization evaluates how the output is affected when random alterations are introduced, providing a robustness baseline.

Occlusion of top concepts: This experiment addresses faithfulness and sensitivity. After identifying the globally important concepts through the aggregation procedure described in Section 3.4.3, concepts are sequentially occluded, beginning with the most important and proceeding to the least important. A reliable explanation method is expected to produce a steep initial increase in the difference from the original output—corresponding to the removal of high-importance concepts—followed by a plateau as lower-ranked concepts are occluded.

Randomized occlusion: To provide a comparative baseline, the same occlusion procedure is repeated using a random ordering of concepts. The resulting curve is expected to reflect a lower-bound behavior in terms of impact, against which structured XAI methods can be assessed.

3.5.2 User Evaluation and Feedback Methodology

A preliminary user study was conducted to explore the subjective perception of the system’s explanations and the overall framework. Since several design choices were made under the assumption that they would enhance interpretability, collecting direct feedback from users was essential to assess the effectiveness of these decisions.

The survey focused on evaluating the proposed semantic segmentations and the perceived clarity and usefulness of the additional information provided alongside the similarity map—such as contribution tables and textual descriptions generated by LLMs. Participants were also asked to compare the semantic-based explanations with traditional methods (LIME with superpixels), in terms of clarity and user-friendliness.

The goal was to engage participants from a variety of disciplinary backgrounds and gather diverse perspectives on interpretability. The preliminary survey was disseminated through the `r/SampleSize` subreddit, which includes over 222,000 members who voluntarily participate in surveys. It was also shared via mailing lists from AI, law, and social science laboratories affiliated with the authors’ institutions: Université Libre de Bruxelles, Université de Mons, Vrije Universiteit Brussel (VUB), and Université Gustave Eiffel.

3.6 Experiments and Results

The experimental evaluation was conducted using the color FERET dataset [180]. We selected the Color FERET database because its structure aligns particularly well with the aims of this chapter. The dataset was acquired in semi-controlled conditions and provides multiple images of the same subject across different sessions, poses, and expressions. This makes it especially suitable for a verification setting, where the task is to decide whether two images represent the same person. The controlled variability allows us to probe model behaviour under small, interpretable changes, and to test whether the explanations produced by our pipeline can meaningfully capture how the system responds to such shifts. Moreover, because FERET simulates the kind of standardized capture protocols often found in regulated contexts such as border control or access management, it provides a realistic proxy for scenarios in which explainability might directly support user understanding. At the same time, we acknowledge that FERET is a demographically limited dataset [181]. For this reason, the experiments reported here should not be read as evidence about fairness across demographic groups, but rather as a methodological evaluation of explanation quality under controlled conditions.

All facial images were preprocessed using the Multi-task Cascaded Convolutional Network (MTCNN) method [182] to ensure consistent cropping and resizing to a fixed dimension of 256×256 pixels. The proposed interpretability techniques were assessed on pre-trained FaceNet [183] models that had been trained on two widely used face recognition datasets: Casia-WebFace [173] and VGGFace2 [174]. Specifically, the experiments employed the pre-trained *InceptionResnetV1* implementation provided by the `facenet_pytorch` library, without any additional training or fine-tuning. All analyses were carried out in a Python environment with GPU support, as the objective of this study is to evaluate interpretability rather than model optimization.

3.6.1 Semantic Extraction

These experiments assess the extent to which the top-ranked human-based semantic sets influence the final face representation and the face verification task. Regions corresponding to these sets are successively occluded, from the most to the least important, and the variation from the original output is measured. Three different human-based semantic sets (SET_0, SET_1, SET_2) are evaluated. For the face representation comparison, the Euclidean distance between the original and occluded 512-dimensional vectors is computed for 750 images. The results for models trained on Casia-WebFace and VGGFace2 are presented in Figure 3.10.

For the face verification task, the Euclidean distance between the similarity scores of 350 image pairs is computed before and after occlusion. The corresponding results for models trained on Casia-WebFace and VGGFace2 are shown in Figure 3.11.

Four main observations can be drawn from the results. First, all methods outperform the random baseline at specific stages of the occlusion process. Second, LIME consistently achieves the best performance across all models and semantic sets in this type of occlusion-based evaluation. Third, due to its decomposition-based design—partitioning the network into clusters of conceptually coherent components—MAGE achieves strong results only after the occlusion of multiple concepts; this leads to an initially weaker performance, which can, however, surpass that of SHAP after several occlusions in certain settings. Finally, SHAP exhibits average performance overall, as its local explanation nature is sensitive to generalization: since the images used for evaluation differ from those used during concept extraction, effectiveness is contingent on the generalization capacity of the network.

3.6.2 Local Explanation and Sensitivity Analysis

In this section, a representative output of the local explanation framework is presented. Figure 3.12 (left) shows the explanation generated for a pair of images (A and B) using the VGG-Face model,

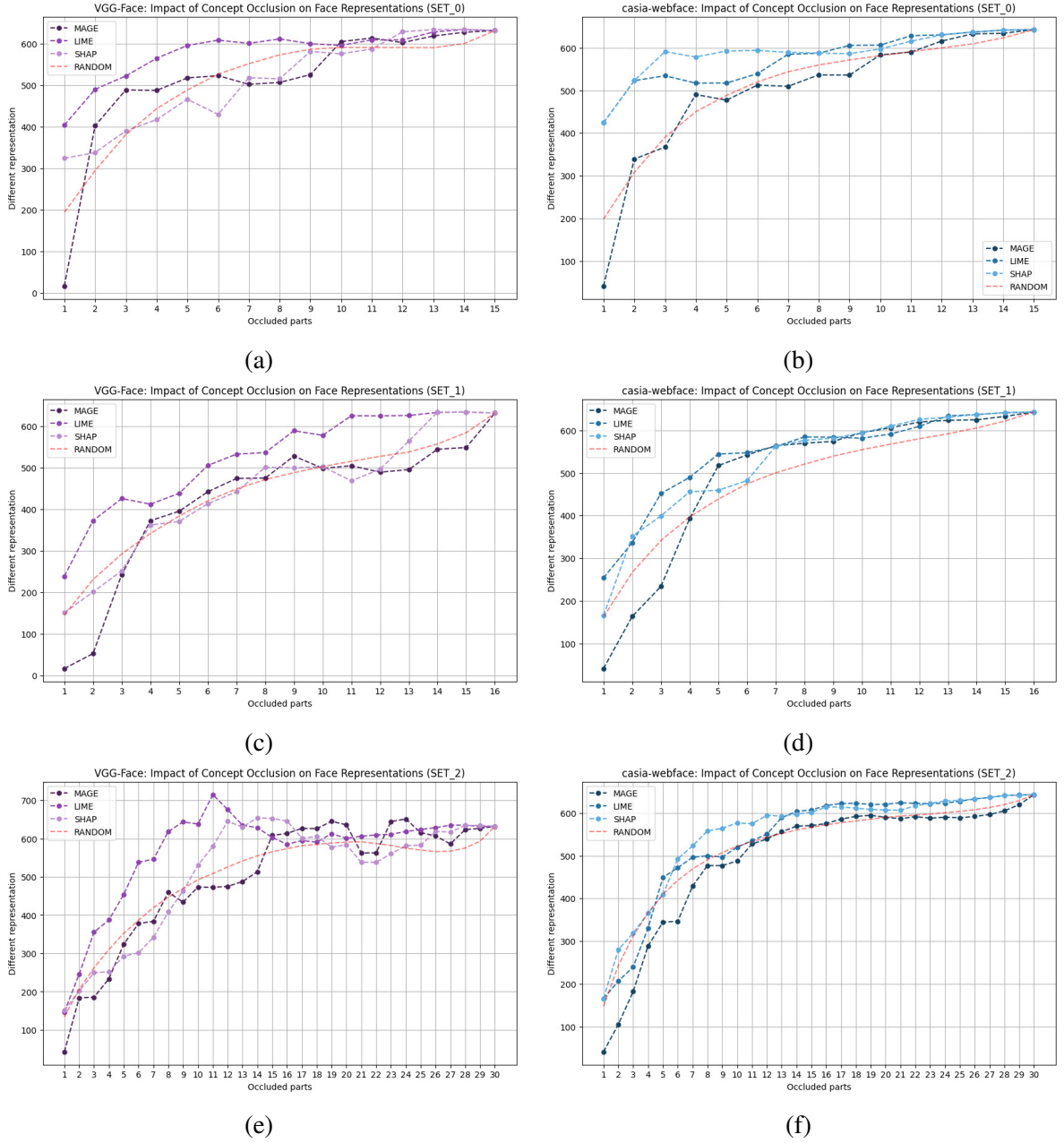


Figure 3.10: Difference in face representations resulting from successive occlusion of concepts—ranked from most to least important—across three semantic sets, for the VGGFace (a, c, e) and Casia (b, d, f) models. The x-axis denotes the number of occluded parts, while the y-axis reports the Euclidean distance between the original and occluded face representations. Three XAI techniques (MAGE, LIME, and SHAP) are compared, alongside a random occlusion baseline. The random approach provides an average benchmark, with higher values indicating better performance.

which assigned a similarity score of $S_{AB} = 0.94$. The figure includes two similarity maps indicating semantically similar (orange scale) and dissimilar (purple scale) regions, along with a contribution table displaying the corresponding semantic areas and the magnitude of their contribution to similarity or dissimilarity. This example is based on the semantic segmentation defined in SET_2, the most granular among the hypothesized sets. Due to space limitations, only the similarity maps and contribution table are shown—both resulting from the single removal algorithm described earlier. Textual explanations generated by large language models are discussed in the user feedback section.

On the right side of Figure 3.12, additional similarity maps are provided for multiple image pairs, illustrating different face verification outcomes. These examples include both genuine pairs (e.g., two

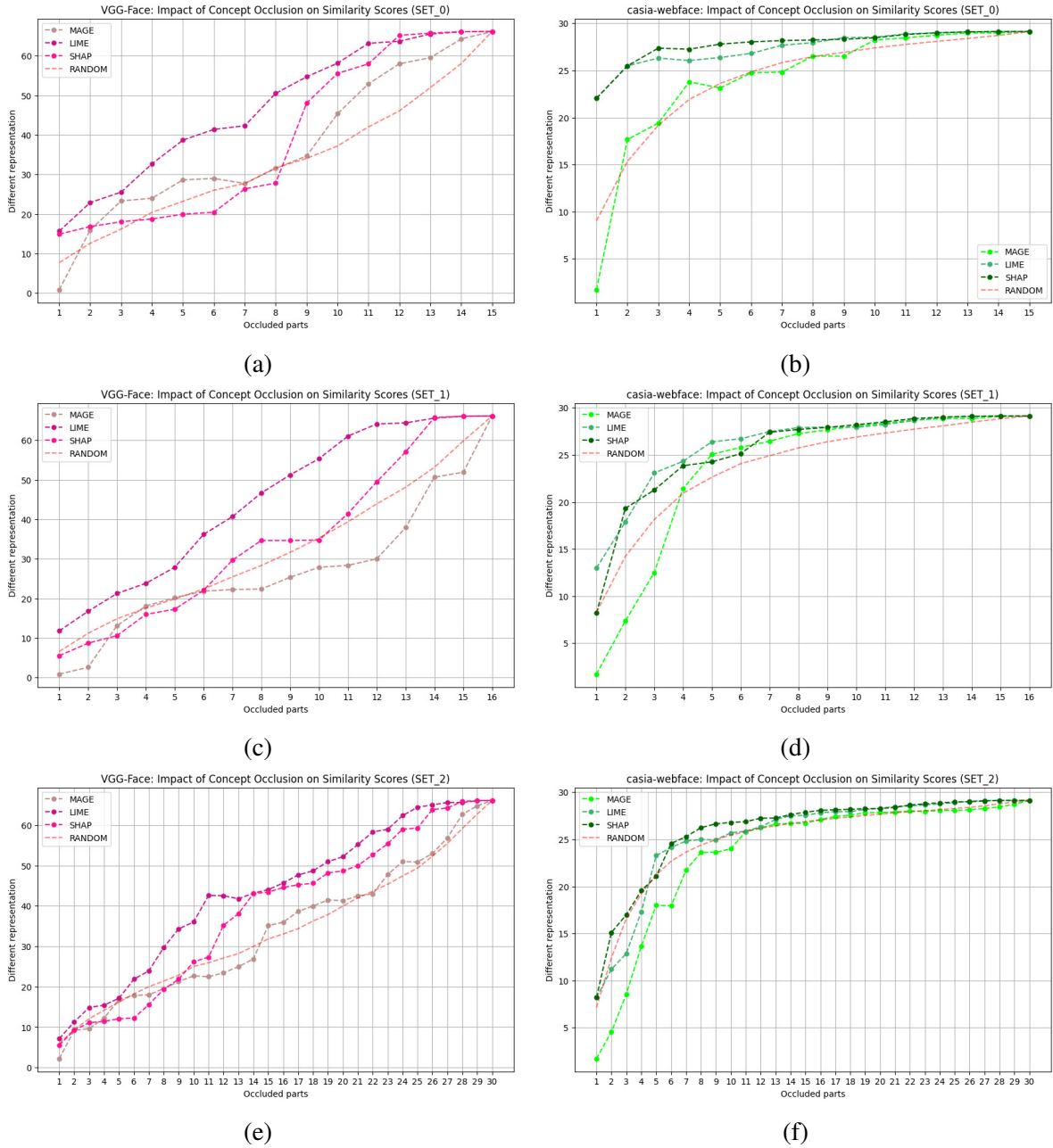


Figure 3.11: Difference in similarity scores resulting from successive occlusion of concepts—ranked from most to least important—across three semantic sets, for the VGGFace (a, c, e) and Casia (b, d, f) models. The x-axis denotes the number of occluded parts, while the y-axis shows the distance between the original and occluded similarity scores. Results are reported for three XAI techniques (MAGE, LIME, and SHAP), along with a random baseline. Higher values suggest better concept selection performance.

images of the same subject) and impostor pairs (e.g., images of different individuals). In the first case, the images depict the same person with a minor change—closed eyes in one of the images. The similarity map accurately identifies the eyes as dissimilar, while other facial regions remain similar. In the second case, involving distinct individuals, the maps indicate global dissimilarity. The third example presents a more nuanced case: although the individuals differ in certain features (e.g., the left side of the nose and the lower lip), the model identifies similarity in the upper face, resulting in a moderate similarity score of 0.47.

In all cases, the ten most influential semantic features are highlighted, based on their local impact (in absolute value) weighted by their global importance. This selection strategy ensures that the

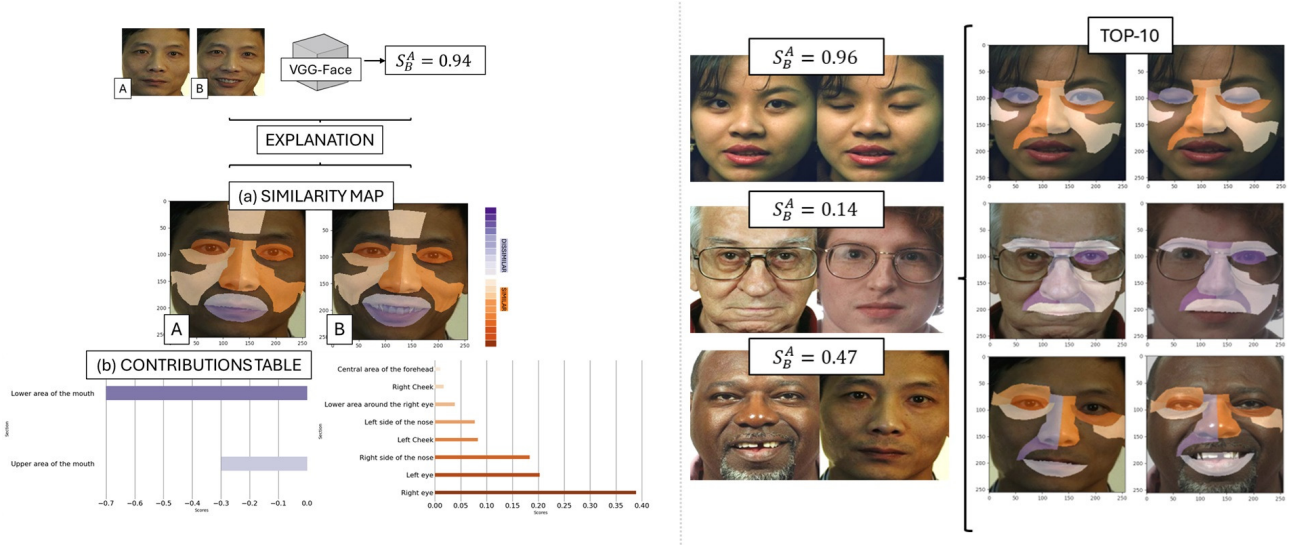


Figure 3.12: (Left) Example of local explanation for VGG-Face model output ($S_{AB} = 0.94$) showing similarity maps and a contribution table for semantic features from SET_2. (Right) Similarity maps for various cases, highlighting the top 10 features with the greatest local impact, balanced by their global importance.

most locally relevant regions are emphasized, while maintaining alignment with the model’s overall behavior.

A sensitivity analysis was conducted using a “Cut-and-Paste” experiment to assess the robustness of the similarity maps, as shown in Figure 3.13. In this setup, specific facial regions within an image were modified to evaluate how the system responds to both synthetic and occlusive alterations. Starting from a true match (A–B), modifications were applied to the original image to create variants: in image C, the eyebrows were removed; in image F, a fake mouth was added; and in image G, a facial mask occlusion was introduced. For each case, the ten most influential features were visualized, with the exception of the mask occlusion case, where the top 25 features were shown to assess whether the eyes remained identifiable as similar.

The resulting similarity maps demonstrated high sensitivity to both structural and appearance-based changes. Notably, when dissimilarities were detected in regions belonging to the unaltered reference image (A), this was interpreted as an effect of the model’s holistic processing of facial structure. Even local modifications may alter the global embedding and shift the network’s perception of the entire face. This behavior is consistent with findings from cognitive studies, such as the Thatcher effect [184], and further supported by Jacob et al. [185], which show that facial recognition—whether human or machine—is influenced by configural relationships, reinforcing the hypothesis that face models internalize a holistic understanding of facial identity.

3.6.3 User Feedback Evaluation

This section presents the findings from a preliminary user survey conducted to assess the interpretability and perceived effectiveness of the proposed framework. The survey aimed to gather user perceptions on three core elements: (1) participants’ background, (2) their evaluation of the semantic-based framework, and (3) their comparison of the framework with a traditional visualization method. Although the comparison question between semantic and traditional (LIME with superpixels) explanations was presented at the beginning of the questionnaire to minimize bias, in reporting the results we opted for a different order. Specifically, we first describe participants’ backgrounds, then their evaluation of the semantic framework, and finally their comparative preferences between the two approaches. This structure was chosen to provide a more coherent narrative flow.

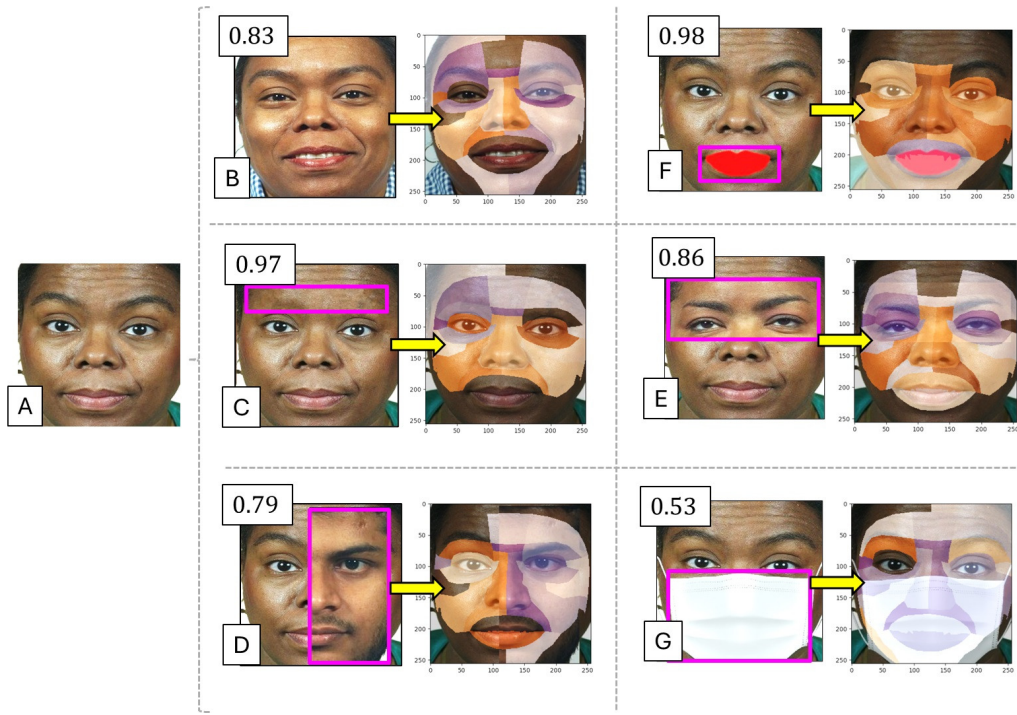


Figure 3.13: Example of local explanations for the VGG-Face model with their respective similarity scores. (A) Original image. (B) True positive image. (C) Image with eyebrows removed. (D-E) Images with parts from another individual’s face (D: eyes and forehead, E: half face). (F) Image with a fake mouth added. (G) Image with a mask occlusion. The similarity maps indicate semantically similar (orange) and dissimilar (purple) areas, with the top 20 most important features highlighted for cases B-F, and the top 25 features for case G.

Survey design. The questionnaire was structured into multiple formats: closed-choice questions (e.g., education, professional background), likert scales from 1 to 5 (e.g., clarity, satisfaction, usefulness), and open-ended questions (e.g., preferences for semantic sets or explanation styles). This mixed design enabled both quantitative comparisons across participants and qualitative insights from free-text responses. An overview of the survey questions is provided in Appendix A.3.

Ethical considerations. The survey was conducted via Google Forms without collecting personally identifiable information. Participants were not asked to provide names, email addresses, or login credentials, and IP addresses were not stored. The only metadata recorded consisted of the submission date and time, used solely for organizational purposes. All responses were analyzed in aggregate form, and no attempt was made to trace individual answers. Since the study did not involve sensitive personal data, biometric information, or vulnerable populations, and participation was entirely voluntary, formal institutional approval was not required under the applicable institutional and European research ethics guidelines.

The survey remained open for one week and gathered responses from 61 participants.

Demographics and Background:

To assess the interpretability of the generated explanations, a survey was conducted via Google Forms, as introduced in Section 3.5.2. The distribution of participants’ professional backgrounds is visualized in Figure 3.14.

The aim was to obtain feedback from a diverse audience representing different disciplinary perspectives. Participants were asked to indicate their professional background, which was subsequently categorized into two primary groups: technical and non-technical. The technical group, comprising 41% of the respondents, included individuals with expertise in fields such as technology or engineer-

ing. The remaining 59% formed the non-technical group, which encompassed participants from a broad range of other domains, yielding a relatively balanced distribution across the two categories.

Additional demographic information was collected to contextualize the responses. Most participants were young adults, with 72% falling within the 25–34 age range. Other age brackets were represented as follows: 18–24 (10%), 35–44 (7%), 45–54 (2%), 55–64 (5%), 65 and above (2%), and under 18 (2%). The predominance of respondents in the 25–34 age group may reflect a sampling bias linked to the authors’ own academic and professional networks.

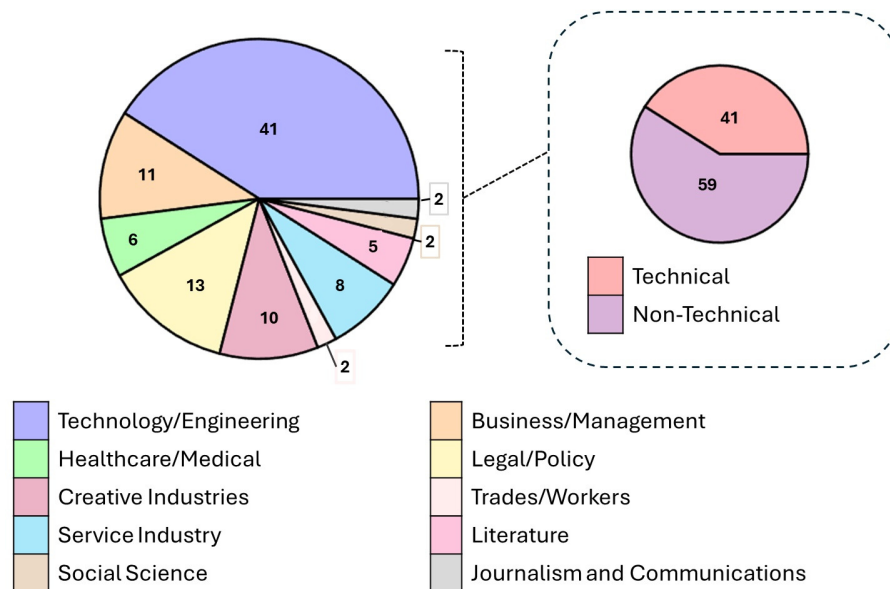


Figure 3.14: Distribution of participants’ professional backgrounds categorized into technical (41%) and non-technical (59%) roles. The technical group includes fields such as technology, engineering, and related disciplines, while the non-technical group encompasses diverse areas, including health-care, creative industries, business, and more.

Regarding educational attainment, most respondents reported high levels of education: 55% held a Master’s degree, 16% had completed secondary education or equivalent, 14% held a Bachelor’s degree, and 5% had completed only primary education.

Participants also reported varied levels of familiarity with artificial intelligence: 36% described their understanding as basic, 20% considered themselves very knowledgeable, 17% reported moderate knowledge, 15% indicated no familiarity with AI, and 12% identified as experts.

Understanding of XAI similarly varied: 32% of respondents had a basic understanding of the concept, while 28% had never heard of it. An additional 18% had heard the term without knowing its meaning, 15% felt able to explain it to others, and 7% reported advanced knowledge with practical experience.

To support a preliminary analysis, survey responses were grouped according to participants’ technical or non-technical backgrounds. This categorization was used as a proxy to evaluate differences in interpretability perception, enabling a simplified yet meaningful comparison. While more granular demographic analysis would have been possible, such detail was beyond the scope of this initial study.

Framework and Explanation Evaluation

Participants were asked to assess several components of the proposed framework, including the type of facial semantics, the visualization method, and the inclusion of textual transcriptions.

Table 3.1: Overview of semantic evaluations based on four properties: Completeness, Clarity, Precision, and Simplicity. Each semantic set (SET_0, SET_1, SET_2) is evaluated by users, with percentages indicating the preference for each property. SET_0 is primarily chosen for its simplicity, SET_1 for its balanced approach, and SET_2 for its completeness and precision. Numbers are expressed in (%).

	Completeness	Clarity	Precision	Simplicity
SET_0	22.22	33.33	22.22	77.78
SET_1	28.57	28.57	14.19	28.57
SET_2	62.50	18.75	68.75	6.25

Semantic Preferences. Because collecting personalized semantic definitions for each user would have been computationally intensive, the evaluation was conducted using three predefined sets of facial regions. Participants compared these sets based on four commonly mentioned properties:

- *Completeness*: the extent to which a semantic set captures relevant facial features in a detailed and exhaustive way.
- *Clarity*: the ease with which a semantic set can be understood and interpreted.
- *Precision*: the degree of specificity in segmenting facial components.
- *Simplicity*: the balance between informativeness and cognitive load.

As shown in Tab. 3.1, semantic SET_2 emerged as the most preferred (50%), while SET_1 and SET_0 received 25% and 26% of the votes, respectively. These preferences offer early insights into how users weigh detail and usability in facial segmentation schemes.

As shown in Table 1, SET_0 was primarily chosen for its simplicity. One participant noted, “I prefer SET_0 as it is simpler and more flexible.” SET_2 was appreciated for its completeness and precision. A user commented, “SET_2 because the granularity is higher, so I can have higher accuracy in space.” SET_1 was perceived as the most balanced, with one respondent stating, “I chose SET_1 because it feels closer to me when I evaluate similarities in facial features.”

Differences in preferences emerged based on participants’ backgrounds:

- **Technical background:** Preferences were relatively evenly distributed, with 35% choosing SET_2, 30% SET_1, and 26% SET_0. Participants in this group often emphasized precision and completeness. For instance, one technical user remarked, “I chose SET_0 because the separation of the human face into its basic parts is sufficient to study the similarity.”
- **Non-technical background:** A stronger preference for SET_2 emerged among this group (61%), followed by SET_1 (22%) and SET_0 (11%). Respondents in this group tended to value clarity and perceived accuracy. One participant commented, “SET_2: It contains more details, making it seem more accurate.”

These responses suggest that while SET_2 is generally favored, the reasons behind this preference vary: technical users tend to value the structural rigor of the sets, while non-technical users are more drawn to ease of understanding and perceived completeness.

Clarity of the Visualization: Most participants rated the explanations provided by the framework as either fairly clear or extremely clear. Among respondents with a technical background, approximately 76% found the explanations clear, 19% were neutral, and only 5% considered them somewhat unclear. A similar trend was observed among non-technical participants: 74% found the explanations clear,

18% were neutral, and 9% perceived them as somewhat unclear. Notably, none of the respondents in either group rated the explanations as completely unclear.

Participants with technical backgrounds generally expressed higher levels of satisfaction. More than three-quarters reported being fairly satisfied, while around 10% indicated they were very satisfied. The remaining 14% were neutral, and none expressed dissatisfaction. In contrast, the responses from non-technical participants were more diverse: 35% reported being fairly satisfied, 24% were very satisfied, about one-third remained neutral, and approximately 9% expressed mild dissatisfaction. No participants reported being completely dissatisfied.

When asked about the clarity of similarity values presented in the accompanying table, all technical participants reported a positive experience. Specifically, 86% described the values as fairly or extremely clear, and the remaining 14% were neutral. This unanimity underscores the table’s effectiveness for technically proficient users.

Among non-technical participants, most also found the similarity values clear, although approximately 6% reported difficulty in understanding them. For example, one respondent commented: “I am not very familiar with tables. I understand, for example, that the lower area around the right eye is very dissimilar and the right side of the nose is similar, but I cannot understand how you got the 64% value.” This suggests that while many participants could interpret the output, some struggled with the numerical presentation of similarity values.

The perceived usefulness of the table was also explored. Among technical participants, 95% found it useful, with 5% neutral and none reporting it as not useful. Among non-technical users, 74% considered the table helpful, 18% were neutral, and 9% found it not useful. These results suggest a slightly lower level of utility perceived by non-technical users, compared to their technical counterparts.

Participants who found aspects of the similarity table unclear often suggested improvements. Technical respondents proposed visual enhancements to improve usability. One participant noted, “I find the table clear, but it would be more visually pleasant with some modifications.” Another suggested an interactive format: “It would be interesting to visualize similarity values only by clicking on the area of interest without having the table displayed directly.”

Feedback from non-technical participants highlighted broader challenges, such as difficulty interpreting percentages and limitations when viewing the table on mobile devices. One respondent observed: “It is an image not optimized for smartphones. Considering that most people do this on mobile, every image should be optimized for mobile.” These comments point to the need for greater accessibility and responsive design.

Overall, the explanations were perceived as clear by the majority of participants, with technical users reporting slightly higher levels of satisfaction. Both groups found the table valuable for interpreting the model’s output, although technical participants were more unanimous in their positive assessment. The findings also suggest a relationship between general satisfaction with the explanations and the perceived clarity of the similarity values. While technical participants focused on usability improvements, non-technical users highlighted comprehension and accessibility issues, emphasizing the importance of designing explanations that accommodate varying levels of expertise.

Textual Transcription: As mentioned in Section 3.4.5, automatic textual transcriptions of similarity values were generated using Codellama Instruct-7B, Zephyr-7B-beta, and Beagle14-7B. User preferences are summarized in Table 3.2. Each model produced stylistically distinct explanations of the face verification output, based on a standardized prompt designed for consistency. Due to space limitations, the complete text outputs for each model are provided in Appendix A.1.

Codellama Instruct-7B produces detailed and practical explanations, emphasizing specific facial features to illustrate the similarities and dissimilarities between two images. For instance, it notes: “Specifically, it (*the model*) is seeing differences in the right eye, left eye, upper area of the mouth, central area of the forehead, and right cheek.” It also incorporates relatable analogies to aid understanding, such as: “For instance, if you have a photo of your grandmother from one year ago and

Table 3.2: Comparison of preferences for model transcription styles between participants with technical (T) and non-technical (NT) backgrounds. The table illustrates the percentage of users who preferred each transcription style: Beagle, Zephyr, and Codellama. Beagle is the most preferred style among both technical (42.86%) and non-technical (44.12%). Numbers are expressed in (%).

	Beagle	Zephyr	Codellama
T	42.86	23.81	32.35
NT	44.12	23.53	31.30

another photo of your grandmother now, the model will see her nose as dissimilar while seeing some features (like the eyes) as similar.” This accessible and example-driven style enhances comprehension, particularly for users with non-technical backgrounds.

Zephyr-7B-beta offers balanced and concise explanations that clarify how different facial regions contribute to the overall similarity score. For example, it states: “Areas with dissimilarities tend to lower the overall cosine similarity score, while areas with similarities tend to increase it.” The model deliberately avoids technical jargon, ensuring that the explanation remains accessible and informative for both technical and non-technical audiences.

Beagle14-7B adopts a structured and detailed explanation style, making use of visual aids such as color-coded maps to represent varying levels of similarity and dissimilarity. For example, it explains: “The color map will display shades of orange where those similarities are found (i.e., higher scores), while shades of purple indicate differences or dissimilarities (lower scores).” Unlike other models, it provides comprehensive textual descriptions without assigning specific numerical values to each area. Statements like “The ‘Left side of the nose’ and ‘Lower area around the right eye,’ for example, have lower similarity values indicating that they were seen as less alike in both images” help convey nuanced insights. The model concludes by summarizing that “the final cosine similarity score combines the contributions made by each facial area.” This structured presentation, combined with visual references, significantly improves clarity and user comprehension.

User preferences varied depending on the clarity, conciseness, and structural features of each transcription style. Codellama instruct-7B was appreciated for its practical examples and straightforward language, making technical concepts more accessible—particularly for participants with non-technical backgrounds.

Zephyr-7B-beta received positive feedback for its balanced and concise explanations. Its ability to communicate essential insights without unnecessary complexity made it suitable for quick understanding across both technical and non-technical groups.

Beagle14-7B emerged as the most preferred option in this preliminary evaluation. Its structured narrative and references to visual elements (such as color-coded regions) were considered helpful for interpreting model behavior. Technical users highlighted its clarity in describing the evaluation process, while non-technical participants found the visual guidance particularly intuitive.

Overall, although the user sample was limited and informal, this preliminary feedback suggests that Beagle14-7B’s structured and visually supported explanations were most effective in aiding interpretability. These early observations reinforce the importance of accessible and well-organized outputs when communicating AI decisions, especially to diverse audiences.

Proposed Method vs. Traditional Approach

As part of the evaluation, participants were also asked to indicate their preference between the semantic explanation method proposed in our framework and a traditional approach. For the traditional method, we used LIME in its conventional visual representation—namely, superpixels combined with heatmaps, where red highlights more important areas and blue indicates less relevant ones (Figure 3.15.a). A brief explanation accompanied the image to clarify the color coding and the role

of superpixels.

In this preliminary assessment, the majority of users expressed a preference for the semantic approach (78%), while only 12% favored the traditional method. The remaining participants did not express a clear preference. Based on user feedback, the reasons behind these preferences can be grouped into four main categories, with technical and non-technical distributions summarized in Table 3.3. Firstly, participants appreciated the level of detail and clarity provided by the semantic

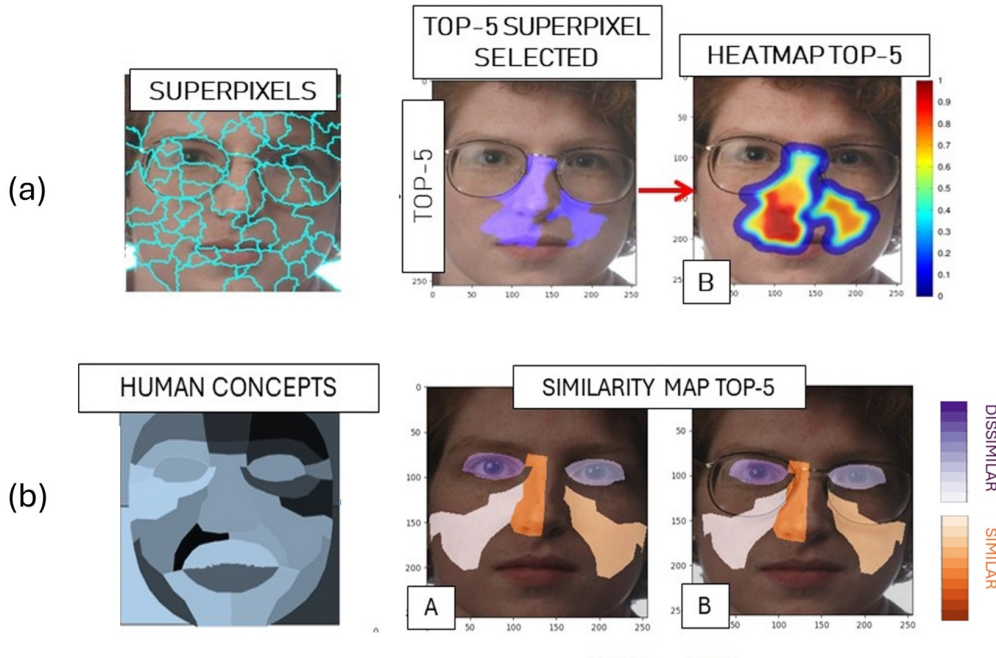


Figure 3.15: (a) Traditional method represented by LIME using superpixels and standard heatmap’s colours to represent the explanation (b) Proposed method using human concepts with Similar and dissimilar colours to represent the explanation. Both methods are represented with the top 5 most important concepts.

approach. Several respondents noted that it “provides more details” and “gives a clearer picture of the metrics.” Others highlighted that it allowed for “better understanding of specific features.”

Secondly, the semantic method was perceived as more aligned with how individuals naturally compare faces. One participant remarked that it is “closer to the way I reason when comparing two faces,” while another commented that “the semantic approach seems clearer since the different parts of the face are easier to identify and compare.”

Thirdly, participants described the semantic explanations as more logical and meaningful. Some stated that the method “is linked to how humans attribute meaning” and that it “explains similar areas better.”

Finally, the semantic approach was valued for its clearer visual structure. Users pointed out that “the location of importance is clearer” and that it “breaks down the analyzed areas and scores them,” making the explanation more interpretable.

Conversely, the small group of participants who preferred the traditional method cited reasons such as simplicity and familiarity. Some described it as “easier to understand quickly” or “more traditional and intuitive.” One participant stated that “the traditional method is clearer than the semantic one.”

A few respondents also perceived the traditional method as offering more direct or accurate visual insight into model behavior. For instance, one user explained: “I chose the traditional method rather than the semantic approach. It’s more relevant. I don’t care about trying to “translate” an AI’s methods into human-friendly terms if it’s less accurate as to what it’s actually doing.” Others highlighted the

Table 3.3: Comparison of Technical (T) and Non-Technical (TN) Background Preferences for XAI method visualization. Numbers are expressed in (%).

	Traditional	Semantic
T	2.40	90.48
NT	29.41	70.59

strong visual impact of the heatmaps, noting their immediate and intuitive appeal.

In summary, the semantic method was generally favored for its richer descriptive power, alignment with human reasoning, and improved visual clarity. The traditional method, meanwhile, was appreciated for its simplicity, familiarity, and straightforward visual cues. Preferences appear to reflect a trade-off between interpretability and immediacy—between cognitive accessibility and visual simplicity.

While the findings are preliminary and based on a modest sample, they suggest that semantic-based explanations may offer a valuable complement to traditional visualizations in face verification systems. As machine learning models grow in complexity, especially in sensitive domains such as face analysis, developing explanation techniques that align with human interpretive processes becomes increasingly important. This study adds to existing work by proposing a combined global-local framework that incorporates semantic concepts and large language models to support more accessible, human-centered AI explanations. Early user feedback indicates a promising preference for the semantic approach—particularly among both technical (over 90%) and non-technical (over 70%) respondents—pointing toward a need for further research into explanation methods that bridge computational reasoning and human understanding.

3.7 Conclusion

This study introduced a preliminary framework aimed at enhancing explainability in AI-based face verification systems through the integration of human-centric semantic approaches. The proposed method combines local and global explanations grounded in user-defined semantic facial features, and explores the potential of contextual algorithms (such as LIME and MAGE) and LLMs to support interpretability.

Several extensions were made beyond earlier iterations of this framework. These include a more refined strategy for combining global and local explanations, the adaptation of existing XAI techniques for semantic aggregation, the use of occlusion experiments to quantify concept relevance, and the generation of textual descriptions through LLMs to support accessibility. A small-scale user study was also conducted to gather feedback on the perceived clarity and utility of the proposed method.

Results from the semantic occlusion experiments suggest that global-aware methods such as LIME may be more stable in identifying relevant semantic face regions, whereas local explanation methods—while useful for analyzing individual examples—may lack robustness when generalized to broader model behavior.

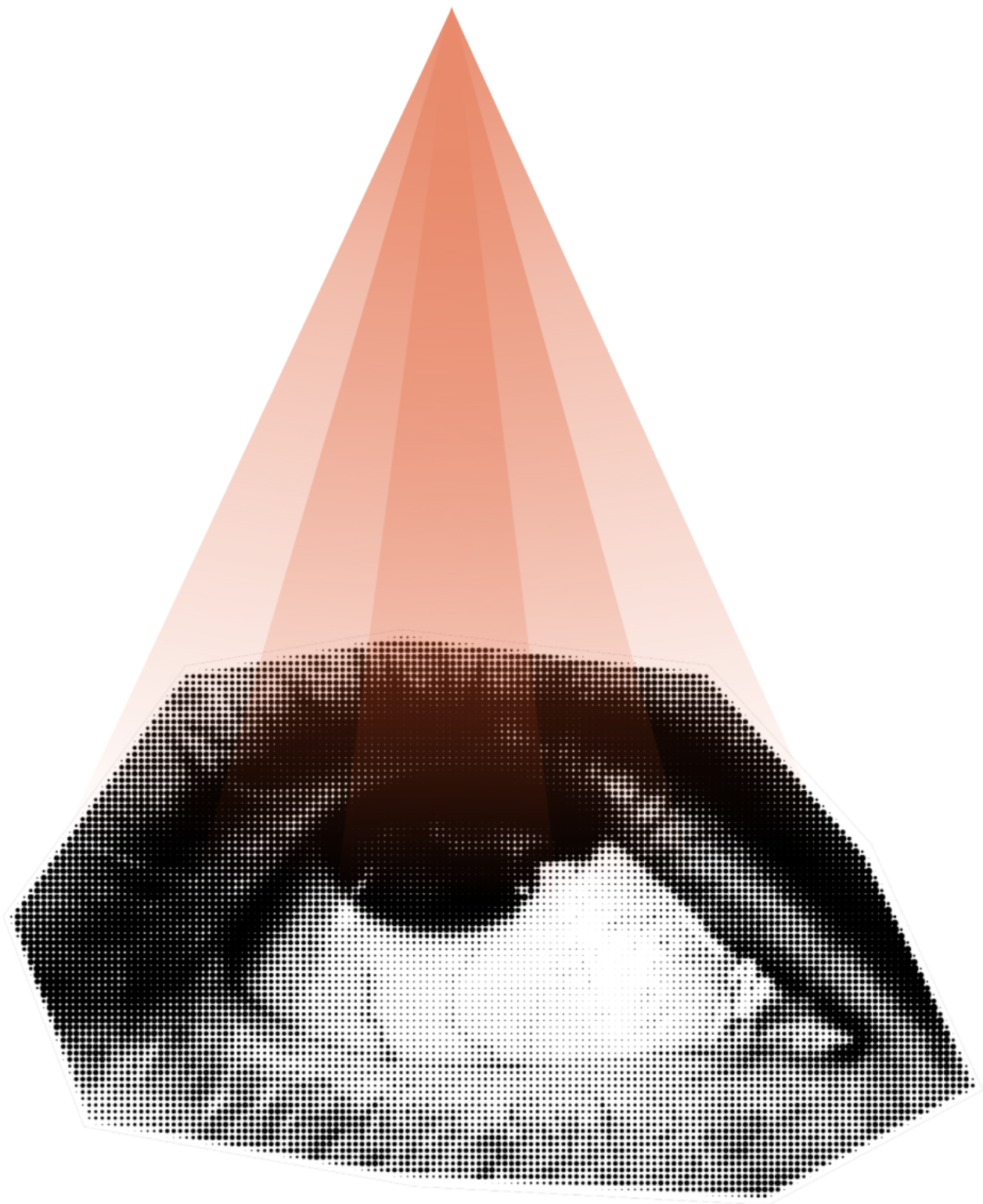
The user study, while limited in scope, yielded encouraging feedback. Participants generally found the semantic approach more intuitive and meaningful than traditional visualizations such as superpixel-based heatmaps. However, responses also revealed areas for improvement. Some participants expressed uncertainty regarding the granularity of semantic concepts or the clarity of the tabular and textual outputs, particularly on mobile devices or among non-technical users.

These findings should be interpreted with caution. The sample size of 61 respondents, while sufficient to gather exploratory feedback, does not permit generalizable conclusions. Moreover, the demographic composition of the sample—potentially skewed by shared academic or professional networks—may introduce bias in preferences and perceptions.

Several limitations should be acknowledged. First, the evaluation was performed primarily on the FERET dataset, which, despite being widely used in face verification research, lacks demographic diversity. This restricts the generalizability of the results and prevents the analysis of potential performance differences across demographic groups. Second, the user study was limited to a modest sample of 61 participants, with possible biases introduced by recruitment through academic networks and online communities. Third, the survey design did not include a baseline or control condition, which would have provided a reference point for interpreting satisfaction levels. Without such a baseline, positive feedback may partly reflect the novelty of structured explanations rather than their actual effectiveness. Finally, while three small-scale language models were used to generate textual explanations, the study did not systematically compare their performance across user groups or contexts. Future research should address these issues by incorporating more diverse datasets, larger and more balanced user samples, explicit baseline conditions, and systematic evaluations of textual explanation strategies.

The motivation behind this work echoes the broader concerns introduced at the beginning of this thesis. Like Kubrick’s monolith in *2001: A Space Odyssey*, AI systems often appear as opaque and alien artifacts—present and influential, yet ultimately unintelligible. By attempting to bridge this gap through concept-based, human-centered explanations, this work aligns with a growing recognition that transparency is not only a technical desideratum, but also a democratic imperative. If users are to trust and critically engage with AI, explanations must resonate with how people actually reason and perceive.

Overall, this study offers a step toward more human-aligned forms of AI interpretability in face analysis. While preliminary in scope, the approach holds potential for supporting more accessible and meaningful explanations, provided that future efforts continue to engage with users and reflect critically on the limits and assumptions embedded in both models and methods.



Chapter 4

Filtering Vision: Challenging Beauty Norms through Transparency

This chapter explores how transparency tools can evolve from passive mechanisms of disclosure into active instruments for questioning the aesthetic governance embedded in *beauty filters*. Moving beyond narrow notions of technical interpretability, it examines how **AI-powered augmented reality (AR) filters** function as systems of bodily and identity control under the guise of playful self-expression. Focusing on TikTok’s *Bold Glamour* filter, the chapter introduces the Disclaimer Block—a framework for quantifying and visualizing the transformations enacted by such filters. Drawing on computer vision and feminist media studies, the analysis situates the *Bold Glamour* filter within a broader ecosystem of algorithmic aesthetics. By exposing systematic gender and racial biases that contradict platform claims of inclusivity, it highlights how transparency can reframe users’ understanding of these systems and their role in shaping digital self-representation.

4.1 From Transparency to Agency: Extending *Trustworthy AI*

“Bring a different version of yourself to life”

– Message displayed when opening the *Bold Glamour* filter.

In 1990, French performance artist ORLAN began a series of surgical operations titled *The Reincarnation of Saint ORLAN*, systematically altering her face to incorporate features from classical paintings of feminine beauty [187]. ORLAN’s radical intervention exposed the violent mechanics of aesthetic transformation and the social forces that govern feminine appearance, making visible the pain and coercion inherent in beauty standards. Her work demonstrated how the body is never simply personal but always political—shaped by systems that define whose bodies matter and how they should appear.

Three decades later, ORLAN’s prophetic critique finds new resonance in the ubiquitous beauty filters that populate social media platforms. Contemporary AR filters like TikTok’s *Bold Glamour* operate within the same political terrain, but with crucial differences. Where ORLAN’s scalpel made visible the violence embedded in aesthetic transformation, digital filters promise seamless, painless modification—maintaining the illusion of autonomy while obscuring the systems of control that continue to govern aesthetic choice.

Unlike cosmetic surgery, which requires a conscious and often costly decision, beauty filters project algorithmic ideals directly onto the face—freely, instantly, and invisibly. This accessibility makes them not only more pervasive, but also more insidious: the filter’s aesthetic logic is inscribed on the user’s image in real time, blurring the line between self-perception and machine-mediated ideal. The psychological effects are significant. Recent studies have shown that repeated exposure



Figure 4.1: Screenshot of ORLAN performing *OMNIPRESENCE*, November 21, 1993, in New York. © MYRIAPODUS Films & ORLAN. Image from [186].

to filtered appearances can fuel body dissatisfaction and even prompt individuals to seek cosmetic surgery in an attempt to match their digital selves [188, 189].

In the same way as ORLAN’s performances, tools that reveal how beauty filters work algorithmically can transform controlling technologies into opportunities for critical analysis. This is where transparency becomes a form of agency—a way to resist the black box logic of social media platforms and reclaim critical awareness of how digital technologies shape embodied experience.

As established in the background chapter, the limitations of purely technical approaches to *Trustworthy AI* have led researchers to seek more comprehensive frameworks that address the sociotechnical nature of algorithmic systems—particularly when these systems are embedded in broader infrastructures like social media platforms.

This chapter extends the exploration of transparency from technical necessity to political discourse. Where Chapter 3 addressed face verification systems and the need to render opaque classification decisions interpretable to users, the present chapter considers how transparency tools can confront the aesthetic governance embedded in beauty filters that masquerade as entertainment.

Beauty filters pose a unique challenge to *Trustworthy AI* frameworks because they operate at the intersection of entertainment, self-expression, and algorithmic regulation. Marketed as harmless enhancements, they embed curated aesthetic norms beneath the rhetoric of personal choice and creativity. Their appeal lies precisely in their invisibility: users experience their effects not as external impositions, but as reflections of their own preferences—when in fact, these preferences are subtly shaped by platform-driven ideals.

This blurring of autonomy and control reveals the sociotechnical complexity of contemporary AI systems. Unlike face verification, where explainability serves regulatory and technical aims, beauty filters circulate in consumer environments where the stakes are cultural, psychological, and commercial. They do not merely process facial features—they participate in constructing the very standards by which faces are evaluated and made acceptable.

In this context, explainability becomes inseparable from governance: who defines beauty, and how are these values encoded in machine vision systems? What appears as individual customization is often the result of invisible algorithmic defaults aligned with normative ideals of gender, race, and

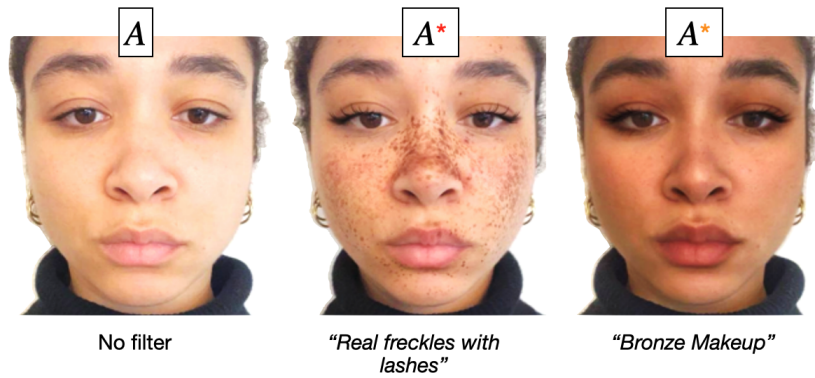


Figure 4.2: Examples of TikTok beauty filters “*Real fleckers with lashes*” and “*Bronze Makeup*” compared with the original face. Images by the author.

desirability.

The movement from transparency to agency thus marks a shift from technical opacity to cultural opacity—from understanding what machines see to interrogating how they shape what it means to be seen. The following analysis of TikTok’s beauty filter and its *Bold Glamour* filter reveals how such systems encode intersectional biases while presenting themselves as neutral tools for playful self-expression. By making visible the aesthetic rules that shape filtered appearances, transparency empowers users to reclaim agency over how they present themselves.

4.2 Motivation: Algorithms, Aesthetics, and Identity

Beauty filters represent a subset of AI-powered AR filters explicitly designed to apply predefined beauty standards, typically smoothing the skin and modifying features such as lips, nose, eyes, eye-lids, and cheeks to conform with socially constructed ideals of attractiveness (two examples in Figure 4.2). They offer a critical site for understanding how algorithmic systems increasingly mediate self-perception, visibility, and identity formation [190, 191, 192, 193]. While often presented as tools for playful self-expression, such filters operate at the intersection of computational vision, aesthetic norms, and platform economies. Their ubiquity, technical opacity, and affective reach make them uniquely positioned to illuminate the sociotechnical mechanisms through which bodies are governed and aesthetic ideals enforced [194, 195, 196].

Social media platforms like TikTok [197], Instagram [198], and YouTube [199] have radically reshaped how individuals construct and perform identity in digital environments [190, 191]. This transformation has been driven by substantial investments in technologies designed to curate user experience through image-based interaction—technologies often developed in alignment with monetization strategies and market logics [200]. As Hearn notes, these visually-oriented platforms are increasingly the primary sites for identity-making online, where bodily appearance is simultaneously expressed and commodified [201].

Within these environments, algorithmic infrastructures play a central role in shaping what is seen—and what is seen as desirable. From personalized advertising to content classification, AI-based systems increasingly embed sexual, racial, and gender biases in the core logics of digital interaction and visibility [86, 202, 203, 204]. Beauty filters extend these logics into the domain of the face and body, transforming them into sites of normative reinforcement. Their seemingly innocuous or empowering interfaces obscure the deeper mechanisms of “soft governance” they enact: reinforcing ideals under the guise of user choice, safety, or fun [114].

TikTok’s *Bold Glamour* filter exemplifies these dynamics. Capable of hyperrealistic, real-time modification of the user’s face, it reflects a new frontier in AR-based facial analysis systems. While

recent policy changes¹ signal growing concern over their societal impact, empirical investigations remain limited, especially within TikTok’s high-frequency, short-form, algorithmically curated content ecosystem.

Studying beauty filters is therefore not merely a matter of analyzing interface design or aesthetic trends. It is an opportunity to interrogate how algorithmic systems construct, constrain, and normalize embodied identities. These tools allow researchers to examine: (1) how algorithmic aesthetics are operationalized in everyday interaction; (2) how normative values become encoded and enacted through vision; (3) how notions of identity, beauty, and desirability are regulated under computational infrastructures; and (4) how users may resist or reinterpret these norms through critical interventions [195, 196, 117].

As visual interfaces that touch on selfhood, affect, and representation, beauty filters sit at the intersection of computer vision, critical media studies, and body politics. They offer a powerful lens for analyzing how power is exercised not just through classification or prediction, but through seeing itself—what this chapter refers to as *filtered vision*—and for developing strategies of resistance, transparency, and more accountable forms of algorithmic design.

4.3 Related work

AI-based AR filters are automated photo-editing tools that leverage computer vision algorithms to apply customized effects to images or videos of faces in real time. Since their introduction on social media in 2015, these filters have evolved significantly in accuracy and popularity. Millions of users now engage with AR filters on visual-based platforms which also offer filter creation tools such as Meta Spark AR² and TikTok’s Effect House³. AR filters serve various purposes, ranging from playful photomontages to brand-specific applications that allow users to virtually try on products like makeup or accessories. Within this wide spectrum, beauty filters have become particularly prominent, as they are among the most widely used and debated categories due to their role in reinforcing or challenging aesthetic norms.

In the past few years, although still limited by the platforms’ non-disclosure practices, growing academic and public scrutiny has advanced the general understanding of the biases embodied and perpetuated by AR beauty filters. Previous work has addressed racial bias and colorism [206], and the filters’ ability to reinforce idealized and stereotypical beauty canons that affect self-perception [207, 208, 209, 210], leading to heightened body self-surveillance based on Western-aligned standards. Recent work by Riccio et al.[196] presents an in-depth examination of how beauty filters perpetuate racial biases, employing explainable AI techniques and publicly available datasets of beautified faces [195]. A complementary strand of research takes a more ethnographic and phenomenological approach and has explored how users experience and are impacted by these filters on a personal level through small ethnographic studies. For example, Rosalind Gill conducted in 2021 a survey for the City, University of London, where 175 UK-based young women (aged 18–30) were interviewed. The findings showed that around 90% of participants had used AR beauty filters, with 48% using them at least once a week [211, 212]. The study further highlighted that participants most commonly used these filters to even or alter skin tone, whiten teeth, enlarge eyes, plump lips, narrow noses, reshape jaws, and reduce weight (the so-called “skinny-filter”). Additionally, the survey reported that 94% of participants felt “pressure to look a certain way on social media” ([211]), with nearly 80% stating that social media negatively impacts their self-perception, and 60% reporting feelings of depression due to these aesthetic expectations. The psychological and social impact of beauty filters, especially on young females—a demographic group traditionally affected by prescriptive beauty

¹Such as Instagram’s 2025 ban on beauty filters and TikTok’s 2024 restriction for under-18 users [205].

²<https://spark.meta.com/>, Last Access: 20-08-2025.

³<https://effecthouse.tiktok.com/>, Last Access: 20-08-2025.

standards [213, 214]—has been further studied by several authors [215, 209, 216].

Building on this body of work, further scrutiny of beauty filters remains necessary. Drawing on Gerrard and Thornham’s framework of social media “sexist assemblages” to understand how digital governance perpetuates normative gender roles through both human and mechanical elements [217], this analysis similarly employs Deleuzian theory to conceptualize beauty filters as collective assemblages that articulate territoriality and identity, matching “those social forms capable of generating them and using them” [218, 219, 220]. Accordingly, it becomes crucial to critically examine their opaque modes of creation, deployment, and governance in order to unpack the nature and extent of the intersectional biases embedded within their design. This approach also calls for sustained engagement with the complex sociotechnical and political heritage informing the technologies under study [221, 118]. To address these challenges, the present study adopts a multifaceted methodology that integrates socio-cultural analysis with technical perspectives, examining how platform governance and algorithmic design contribute to the reproduction and circulation of gender biases. As a case study, the analysis focuses on TikTok’s *Bold Glamour* beauty filter, released in early 2023 and widely adopted across the platform.

4.4 The Disclaimer Block as a Transparency Proposition

4.4.1 Transparency Challenges in Beauty Filters

In this section, the investigation into the use of beauty filters is detailed, stimulated by the analysis of demographic trends observed on TikTok. The research commenced with selecting ten filters for their popularity—measured in terms of the frequency with which users saved these filters—and for how regularly they appeared in the videos viewed on the platform. For each filter, a screening based on the gender distribution of 150 randomly selected users was conducted. The screening considered the binary conception of gender (female/male) and relied on the users’ self-identification⁴. This examination highlighted a greater prevalence of filter use among female users, as reported in Table 4.1. Female users’ dominance of beauty filter usage aligns with the initial hypothesis concerning the influence of gender stereotypes in the choice and application of filters.

Furthermore, we believe that beauty filter names inherently select their target audience from the outset, primarily orienting towards a female audience. The names assigned to filters, being frequently associated with positive descriptors such as “Pure eyes”, “Cutie”, or “Prettiest,” insinuate a certain aesthetic desirability linked to a specific facial type and ideal of beauty inherited from Western-aligned standards of feminine attractiveness. In other words, such filters also imply what a user should look like to fit the given canon and gain appreciation and popularity within spaces where platforms capitalise on users’ aesthetic labour [114]. This tendency not only reflects but potentially reinforces such expectations and cultural norms associated with femininity, affecting the perception of filters based on gender stereotypes. During the research, two main areas of interest for further investigation were identified:

- *Beyond Simple Aesthetic Adjustments:* Filters significantly modify facial features, such as enlarging lips or reshaping the nose, distorting fundamental aspects of the face.
- *Nominalism of Filters and Presentation Autonomy:*
 - *Positive Descriptor Influence:* As mentioned above, filter names with positive descriptors, suggesting certain aesthetic traits, question the self-representation autonomy of female users and the reinforcement of restrictive beauty standards.

⁴Gender was determined based on the pronouns indicated in each user’s bio.

Table 4.1: Analysis of the gender distribution percentage among users who utilised this filter, based on a sample size of 150 users. ‘N-time Posted’ refers to the number of videos that have been publicly posted.

Filter name	N-time Posted	Female (%)	Male (%)
“Princess makeup”	3.4 M	97.33	2.66
“Real freckles with lashes”	390.9 M	96.66	3.33
“Light make up”	285.6 M	98.66	1.33
“Glamour audience”	196.4 M	68.66	31.33
“Posh”	182.9 M	52.66	47.33
“Pure eyes”	89.3 M	97.33	2.66
“Hazel Eyes”	187.6 M	65.33	34.66
“Cutie”	47.8 M	96.66	3.33
“Prettiest”	62.2 M	98.66	1.33
“Bronze Make up”	5.4 M	94.00	6.00

- *Name Ambiguity*: The poor correspondence between the filters’ names and their actual effect on the face raises concerns regarding transparency and the ethics of self-representation, questioning the implications of this practice for personal awareness and acceptance.

In investigating beauty filters, particular attention was paid to the dynamics of “Nominalism of Filters and Presentation Autonomy”, focusing specifically on the ambiguity side. This aspect raised significant concerns about transparency and ethics in self-representation, highlighting how filters often function as “black boxes” with unclear mechanisms for users. Users see other users’ related videos and the creator’s account when using a filter but lack disclaimers about the filter’s effects.

4.4.2 The Disclaimer Block Concept

As explained above, in light of the transparency requirements imposed not only by the AI Act for limited-risk AI systems such as AR filters, but also by the EU Digital Services Act (DSA)⁵, we propose the introduction of a “disclaimer” option (Figures 4.3 and 4.5). This option would allow users to visualise the modifications applied by a beauty filter, comparing pre- and post-filter faces, and thereby make more informed choices about their digital self-representation.

At the same time, large-scale quantitative filter analysis on TikTok presents challenges: users cannot upload arbitrary images to apply filters, and access is limited to a curated set of predefined effects. These constraints motivated the design of a dedicated methodological framework—the Disclaimer Block (DB)—to capture and visualise filter-induced changes.

4.4.3 Technical Development: From DB V1 to DB V2

The Disclaimer Block was developed in two iterations. The first version (V1) was conceived as an exploratory prototype, aimed at testing the feasibility of detecting and visualising filter-induced modifications. While it provided initial insights, V1 relied on video-based acquisition and frame averaging, which reduced noise but also blurred finer changes and introduced artefacts.

Building on this proof of concept, the second version (V2) was designed to overcome these limitations. V2 introduces a more rigorous acquisition pipeline (single high-quality frames with controlled

⁵Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services (Digital Services Act) and amending Directive 2000/31/EC, OJ L 277, 27.10.2022, p. 1–102.

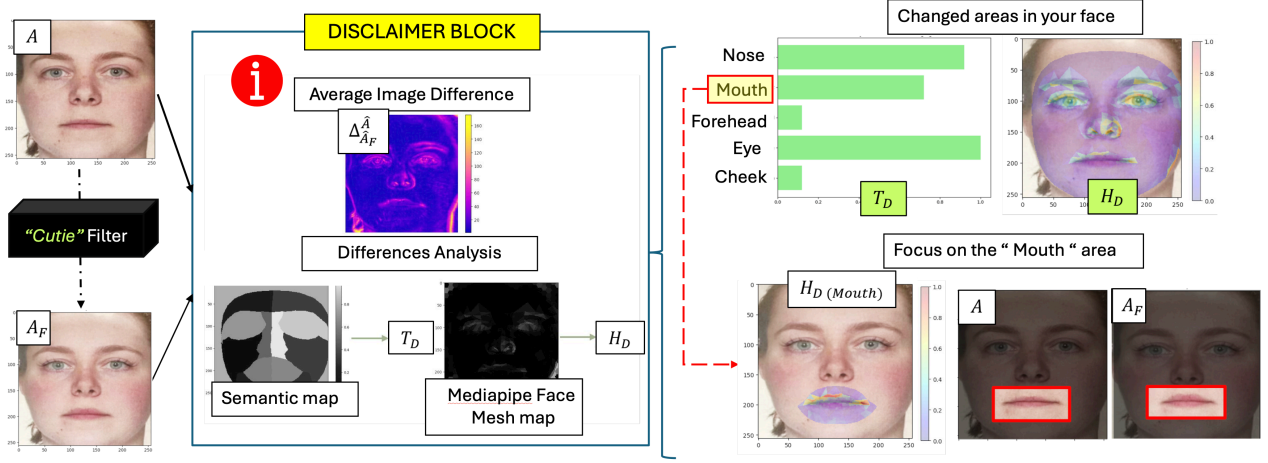


Figure 4.3: Hypothetical *Disclaimer* block. The difference image $\Delta_{A_F}^A$ is created from the unfiltered image A and the post-filter image A_F . A semantic analysis uses the semantic masks defined in Chapter 3 Section 3.4.2 (SET_1) and the Mediapipe face mesh [170]. The output includes a table T_D of the top N semantic areas that have changed and to what extent (in this case $N = 5$), as well as a heatmap H_D to show where the major differences are located spatially. A further possibility would be to have a focus on certain areas. In the example, this focus is on the “mouth”. Sample image from the Face Research Lab London set [222].

lighting), advanced alignment techniques (SIFT-based homography), and finer-grained semantic segmentation (30 masks instead of 13). These enhancements make it possible to move from coarse, region-level approximations to pixel-level mappings of filter effects, while still providing semantically interpretable summaries.

Figures 4.3 and 4.5 illustrate this progression: from the initial heatmap-and-table outputs of V1 to the more precise $IDif_A$ and semantically enriched $S(IDif_A)$ of V2.

Below, we outline the key enhancements introduced in V2 and how they address the limitations of V1.

Disclaimer Block Acquisition Pipeline (V1 vs V2) In Disclaimer Block V1, the data acquisition process relied on recording short videos for each subject, first unfiltered and then filtered. From these, multiple frames were extracted and averaged across corresponding positions (*e.g.*, frame 1 pre-filter vs. frame 1 post-filter). While this averaging reduced noise caused by lighting or movement, it also smoothed out finer, localized modifications and occasionally introduced artefacts.

To overcome these issues, V2 shifted to directly comparing single high-quality frames of unfiltered and filtered images. This avoided the loss of detail observed in V1 and resulted in a cleaner, more accurate dataset. To ensure consistency across subjects, V2 employed a controlled setup with a phone

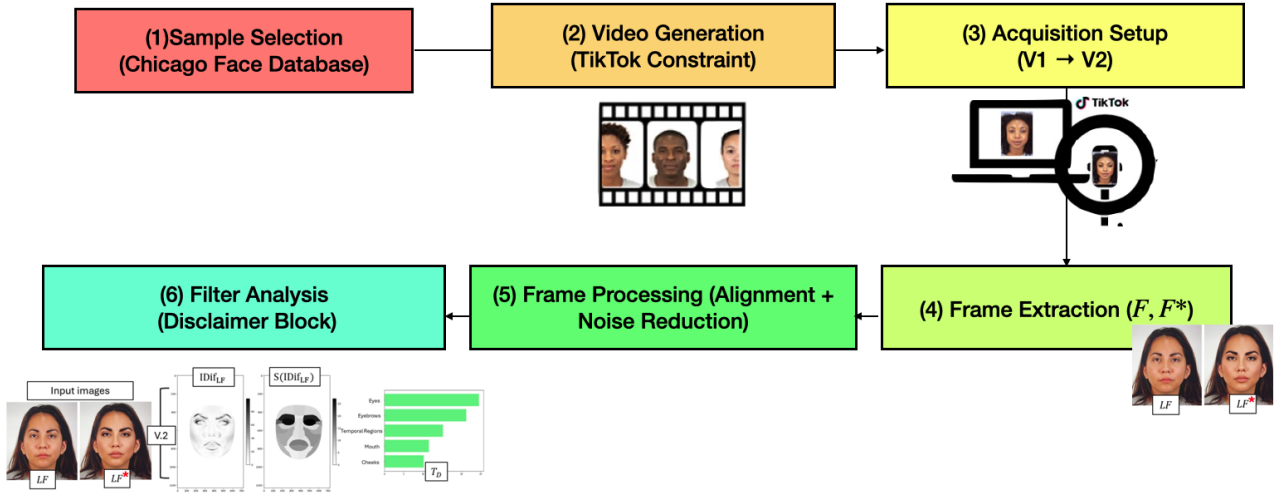


Figure 4.4: Pipeline of the Disclaimer Block (DB) V2: from sample selection and video acquisition to frame extraction, processing, and filter analysis with semantic and pixel-level outputs.

mounted on a selfie ring light, stabilizing lighting and positioning during collection. By relying on static, high-quality images instead of averaged frames, V2 eliminated unnecessary artefacts and achieved greater precision.

From H_D to $IDif_A$ Because V1 data was noisy and often misaligned, the image processing pipeline aggregated changes within coarse facial regions defined by Mediapipe tessellation. The output was a heatmap (H_D), which provided a visually intuitive but imprecise representation, smoothing over many localized changes.

With the cleaner dataset in V2, we could adopt a more precise alignment technique using Scale-Invariant Feature Transform (SIFT) [223]. SIFT matches keypoints between filtered and unfiltered images, enabling homography-based correction of shifts, rotations, or distortions. A Gaussian blur was also applied to minimize residual noise. These improvements made it possible to move from region-based aggregation to pixel-by-pixel comparisons, producing $IDif_A$: a detailed footprint of the filter’s impact at the pixel level. Unlike H_D , $IDif_A$ maps changes directly, without approximation, across the entire face.

Semantic Analysis Improvements: T_D and $S(IDif_A)$ In V1, semantic analysis complemented H_D through T_D , a table summarizing average filter-induced changes across 13 macro areas of the face (SET_1, see Chapter 3). While this facilitated interpretability, its coarse segmentation obscured finer variations.

In V2, we refined this approach by using 30 semantic masks defined in Chapter 3 Section 3.4.2 (SET_2) derived from Mediapipe landmarks, allowing for more granular segmentation of facial regions. For example, cheeks, previously treated as a single region, were divided into multiple sub-regions, capturing subtler modifications. Importantly, these masks are now directly integrated into $IDif_A$, producing $S(IDif_A)$ —a semantic map that overlays pixel-level differences with meaningful facial regions. This dual representation bridges the gap between detailed spatial differences and human-interpretable areas, significantly enhancing the explanatory power of the analysis.

While V1 of the Disclaimer Block emerged as a critical design idea, V2 consolidates it into a methodological instrument capable of systematically detecting and measuring how beauty filters alter faces across gender and race. Having established this refined analytical framework, the following section demonstrates its application through an examination of TikTok’s *Bold Glamour*, a filter that exemplifies the sophisticated capabilities and concerns associated with modern AR beauty technologies.

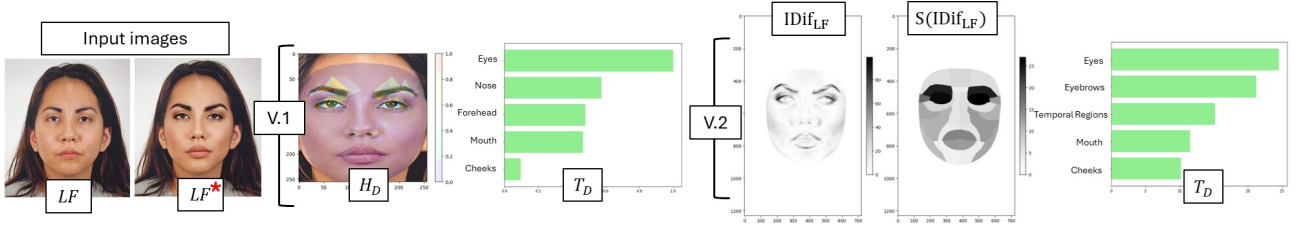


Figure 4.5: Left: Original (LF) and beautified (LF^*) faces. Middle: Output of the Disclaimer Block V.1 framework, showing a heatmap (H_D) and semantic analysis (T_D) of the facial changes. Right: Output of the Disclaimer Block V.2 framework proposed in this chapter. It uses $IDif_{LF}$ for more precise noise reduction and $S(IDif_{LF})$ for detailed regional analysis, enhancing the accuracy and granularity of the assessment of the filter on facial features (T_D). Example of a Latina = L Female = F.

Scope and usage of acquisition pipelines It is important to clarify the scope of these pipelines. The V1 acquisition process was only an exploratory prototype, used to test feasibility and motivate the refinements of V2. All subsequent analyses of the *Bold Glamour* filter rely exclusively on the V2 pipeline (Figure 4.4). When we refer to the *Dataset for Bold Glamour Analysis* in Section 4.5, we specifically mean the set of pre-/post-filter frames constructed with DB V2 from Chicago Face Database images.

4.5 Analysis of the *Bold Glamour* Beauty Filter



Figure 4.6: Examples of the *Bold Glamour* filter applied to a female and a male sample (A = Asian, L = Latino/a) and genders (F = Female, M = Male). In each pair, the first image represents the original, unfiltered face (e.g., LF), while the second image shows the face after the filter has been applied (e.g., LF^*), illustrating the aesthetic changes made by the filter. Original samples from Chicago Face Database [224].

The introduction of AR filters on Snapchat in 2015 revolutionized digital self-representation by incorporating 3D elements the platform branded as “Lenses” [225]. AR filter technology has since advanced, with platforms like Instagram⁶ and TikTok enabling user-generated filters and utilizing sophisticated AI to create real-time, seamless transformations, with beauty filters—enhancing appearances by reshaping and adjusting facial features—becoming particularly popular.

An emblematic example is TikTok’s *Bold Glamour*, developed in 2023. Renowned for its unprecedented hyper-realism and seamlessness [229], it quickly went viral [230] and has since been featured

⁶Following criticism and legal actions [226], Meta announced that, starting in January 2025, it will discontinue AR beauty filters across its platforms. This includes the removal of all third-party face filters and AR effects on Facebook, Instagram, and Messenger, along with the shutdown of APIs enabling users to create filters via Meta Spark AR [227, 228].

in around 250 million videos. The filter’s most notable characteristic is its exceptional stability during use—maintaining its transformative effects even as users move their faces, gesture, or obstruct the camera with their hands. This seamless tracking capability not only fueled its widespread adoption but also amplifies concerns about digital deception, as the filter’s sophisticated operation can make it difficult for viewers to detect when augmentation is being applied [231]. Moreover, there are growing suspicions among TikTok’s users that the filter employs gender classification algorithms to customize its transformations, potentially reinforcing binary gender norms and excluding non-conforming identities⁷. By obscuring the distinction between natural and algorithmically enhanced appearances, *Bold Glamour* exemplifies how beauty filters can subtly influence perceptions of attractiveness, self-worth, and identity [193, 232, 233, 234, 235], while remaining opaque in their design, deployment, and adoption (as discussed in Section 4.4).

This blurring of digital and physical selves is reinforced by social media platforms, which profit from and promote normative representations of femininity. TikTok exemplifies this dynamic, leading the global beauty e-commerce sector: in 2023 alone, the platform generated an estimated Gross Merchandise Value (GMV) of 2.5 billion USD from beauty sales, with over 370 million beauty and personal care products sold via TikTok Shops [236].

Section 4.4 introduced the improved Disclaimer Block (V2) as a framework for analyzing the transformations applied by beauty filters. We now apply this framework to TikTok’s *Bold Glamour*, using a diverse dataset of faces to systematically examine how the filter modifies appearance across gender and race. The following subsection formulates the specific research questions that guide this analysis.

4.5.1 Research Questions⁸

This section presents the methodology and results of analyzing the impact of the *Bold Glamour* filter on a diverse set of faces. Our analyses are driven by the following research questions:

RQ2.1: Does *Bold Glamour* brighten the faces? Previous work has reported a potential brightening of the faces due to the application of Instagram’s beauty filters [196], yet there is a lack of conclusive evidence in this regard.

RQ2.2: Are the filter transformations dependent on gender and race? While the *Bold Glamour* filter claims to personalize its effects based on the user’s face shape and features, there is a lack of research on the role that gender and race play in the face modifications applied by the filter.

RQ2.3: Does *Bold Glamour* apply a facial feature morphological alignment? Recent research has suggested the existence of a *white* racial bias in Instagram’s filters [196]. Yet, there is a lack of detailed, quantitative research on the transformations applied by the filter and its potential morphological alignment.

4.5.2 Dataset for *Bold Glamour* Analysis

Inspired by the data collection techniques described by Mirabet-Herranz et al. [237], we selected 208 face images from the Chicago Face Database [224]: 26 images per race⁹ and gender category, across two genders [M = Male, F = Female] and four race categories [W = White, B = Black, A = Asian, L = Latino], which will be used as abbreviations throughout the chapter.

⁷<https://tinyurl.com/mucxp82d>, Last Access: 20-08-2025.

⁸Under the main research question (RQ2) *Filtering Vision: Challenging Beauty Norms through Transparency* presented in Chapter 2 Section 2.3.

⁹The term *race* is used as in standard ML datasets, acknowledging it as a social construct distinct from *ethnicity* [238, 239]. This study does not seek to promote or reify racial categories in AI, but rather to critically examine how AI systems encode and propagate biases linked to socially constructed categories such as race. This problem is discussed in Chapter 6.

The Chicago Face Database was selected because it provides standardized, neutral images with consistent pose, clothing, and background. This level of control minimized the influence of external factors such as lighting variation, background clutter, or facial expressions, allowing us to isolate the transformations performed by the filter more clearly. An additional advantage is the use of self-defined demographic labels, which are particularly valuable for analyses involving gender and race, reducing the risk of relying on potentially unreliable inferred labels.

One-minute videos were generated from the images to adapt to TikTok’s constraints, each featuring multiple face images. This process resulted in 16 videos capturing both unfiltered and filtered facial data. The videos were then displayed in front of a phone mounted on a selfie ring light to ensure consistent lighting and stable conditions. Each face was first recorded without filters and then with the *Bold Glamour* filter applied to them. This setup minimized variability in lighting and positioning across recordings, allowing accurate comparisons between unfiltered (F) and filtered (F^*) images.

After video capture, frames corresponding to F and F^* were extracted and reassociated with their respective gender and racial labels per the Chicago Face Database metadata. The extracted frames were then processed and refined, preparing them for the analysis described next.

Importantly, the entire dataset construction for the *Bold Glamour* analysis relied exclusively on the V2 acquisition pipeline described in the previous Section 4.4, ensuring consistency and avoiding the artefacts associated with the earlier V1 setup. At the same time, using images with neutral backgrounds introduces a potential limitation: while it reduces noise and supports precise comparisons, it does not fully capture the variability of real-world contexts where faces appear against diverse and dynamic settings. This gap should be considered when interpreting the generalizability of our findings.

4.5.3 Methodology to Characterize Facial Features

To automatically characterize the facial features, we adapted and expanded the Disclaimer Block or DB framework introduced by [117], which is designed to bridge the gap between the technological opacity of beauty filters and user awareness. This revised approach, called Disclaimer Block V.2 (DB V.2) integrates enhanced imaging techniques to provide a more detailed analysis of the changes introduced by the beauty filter. Figure 4.5 illustrates both the original (V.1) and refined (V.2) DB frameworks applied to the original LF and beautified LF^* images shown on the left.

From facial features to semantic feature vectors Facial features are transformed into semantic vectors using FaceGen [240], which models facial morphology via PCA on 273 diverse 3D scans, yielding reproducible high-dimensional shape vectors that capture subtle structural variations. As a result, we transform the face into quantifiable vectors, $\bar{\mathbf{F}}^{(g,r)} \in \mathbb{R}^n$ and $\bar{\mathbf{F}}^{(g,r)*} \in \mathbb{R}^n$, where $\bar{\mathbf{F}}$ and $\bar{\mathbf{F}}^*$ denote the vectors corresponding to the original and beautified faces, respectively, and the superscript (g, r) denotes the gender and race group. These vectors enable a detailed comparative analysis, providing a structured way to quantify the aesthetic changes imposed by the filter, as illustrated in Figure 4.7. For further methodological details on how semantic feature vectors are derived from facial features with FaceGen, refer to Appendix B.2.

Average face per gender and race The semantic feature vectors are computed from the average face within each racial and gender group to analyze general trends and patterns without relying on individual-level data, addressing practical and methodological limitations. Processing individual images is time-intensive and resource-demanding, particularly with manual facial analysis software like FaceGen, which requires manual processing of each image. Averaging provides a manageable and scalable method to analyze the broader impact of the filter while highlighting systemic biases and overarching trends that the filter may impose on distinct racial and gender groups. This approach is supported by Owens et al.[241], who demonstrated that face averaging effectively reveals systemic

biases in automated systems. By revealing patterns such as biases in skin tone or facial shape, average faces offer insights beyond individual-level analysis, making them a robust tool for assessing group-level effects in line with the study’s objectives. Furthermore, TikTok’s architecture makes large-scale data extraction challenging. Working with average representations circumvents these technical barriers, enabling research while respecting the platform’s limitations. By focusing on average faces, our study examines the broader sociotechnical implications of the filter while maintaining logistical efficiency.

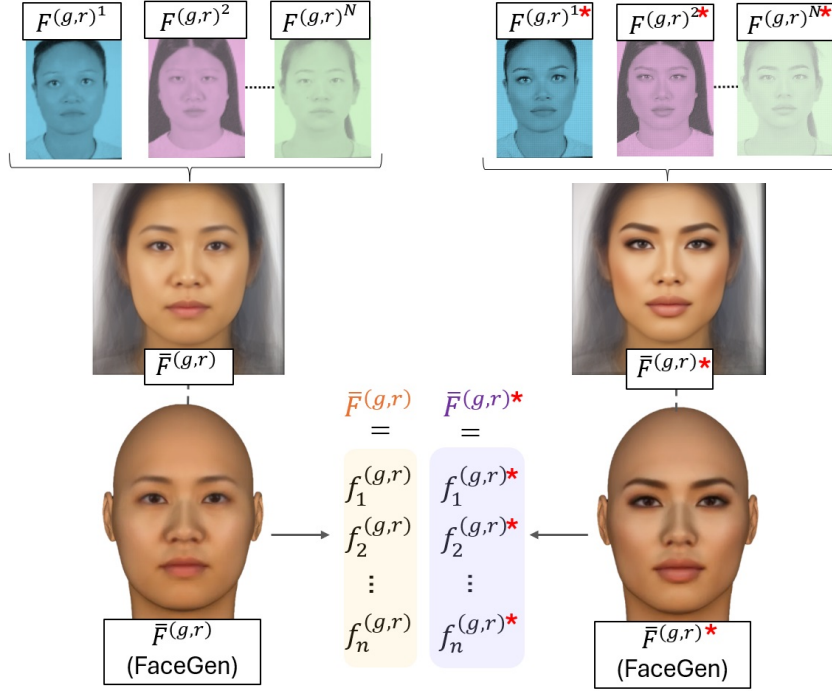


Figure 4.7: Creation of 2D average faces for groups categorized by race and gender, shown here for Asian females. Both original (\mathbf{F}) and beautified (\mathbf{F}^*) averages were used to derive 3D reconstructions with FaceGen, yielding vector representations.

Note that our approach does not provide an absolute measure of facial feature changes. Instead, it allows us to compare the relative impact of the filter across different groups by analyzing how each group’s average facial features change after the filter’s application. Previous studies have addressed the challenge of interpreting changes without an absolute reference by establishing reference baselines, such as using blurred images to quantify information loss [196]. Similarly, our analysis focuses on ranking facial features based on the degree of change observed in the average faces, providing a relative understanding of the filter’s impact.

4.5.4 Results

4.5.5 RQ2.1: Does *Bold Glamour* brighten the faces?

We assess brightness modifications from TikTok’s *Bold Glamour* filter to detect potential skin tone bias. Following Riccio et al. [196], we also convert RGB values to the HSV space and extract the Value component, but extend their approach by applying it to a dataset with self-reported gender and race labels and by analyzing subgroup-specific differences. For each image, ΔB is defined as the mean brightness difference between the filtered and original versions. Figure 4.8 presents the distributions of ΔB by gender and race.

Because ΔB deviates from normality, we apply Wilcoxon signed-rank tests [242] to each gender–race subgroup. All male subgroups exhibit significant brightness decreases: Asian (AM, $p =$

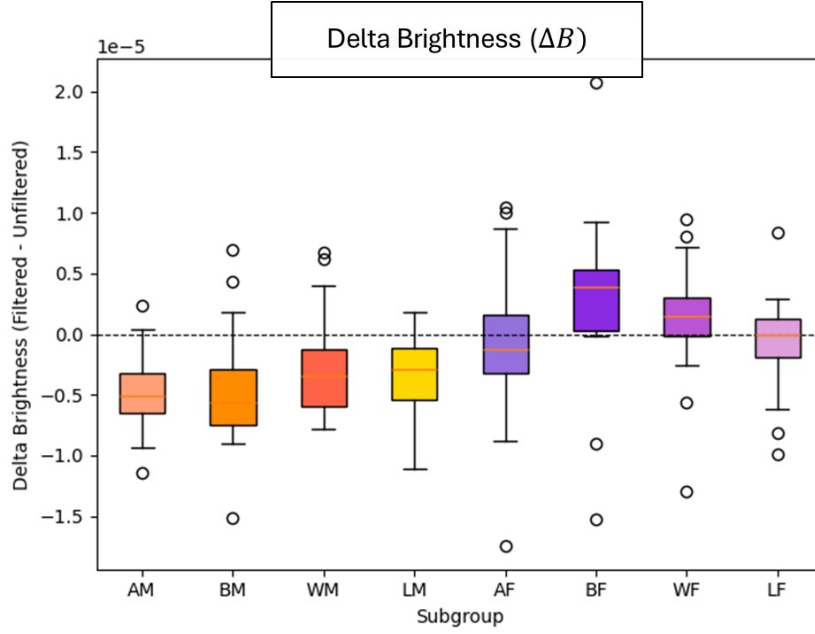


Figure 4.8: Boxplots illustrating variations in ΔB (Delta Brightness) across subgroups. Values below the zero line represent a decrease in brightness, while values above indicate an increase.

7.45×10^{-7}), Black (BM, $p = 6.032 \times 10^{-5}$), Latino (LM, $p = 1.11 \times 10^{-5}$) and White (WM, $p = 6.69 \times 10^{-3}$). Among females, Black (BF, $p = 0.0304$) and White (WF, $p = 0.0516$) faces show significant brightness increases, while Asian (AF) and Latina (LF) faces do not.

Given the small samples in each gender–race subgroup, we also aggregate all 208 images by gender and then compare mean and median ΔB using Welch’s t-test ($t = -6.409$, $p = 1.50 \times 10^{-9}$) and the Mann–Whitney U test ($U = 2168$, $p = 1.436 \times 10^{-11}$). Male faces consistently darken (mean $\Delta B = -3.88 \times 10^{-6}$), and female faces consistently lighten (mean $\Delta B = 4.82 \times 10^{-7}$), confirming a robust gender-dependent brightness bias.

These results indicate that *Bold Glamour* tends to darken male faces and lighten female faces across all races. This consistent, gender-dependent pattern points to systematic differences in how the filter processes faces. While the underlying cause cannot be established from our analysis, the observed trend raises important questions about potential algorithmic design factors and their implications for user representation and identity.

4.5.6 RQ2.2: Are the filter transformations dependent on gender and race?

Male vs Female Modifications A close examination of $IDif_A$ and $S(IDif_A)$ reveals that the *Bold Glamour* filter adapts noticeably based on two distinct transformation styles that correlate with perceived gender cues in the face. Because TikTok’s API does not expose any explicit gender labels, we inferred the presence of these styles by observing two consistent patterns—“Feminine Output” ($F - (IDif_A)$) and “Masculine Output” ($M - (IDif_A)$)—applied across the sample (see Figure 4.9).

In the case of faces that receive the feminine-typed output, the filter transforms the lips, cheeks, and eyes, adding makeup to the eyelids and enhancing cheekbones with blush (Figure 4.9 (a)). Conversely, faces that receive the masculine-typed output experience a vertical blush or shadow effect down the face, in contrast with the diagonal blush along the cheekbones in the feminine output. This vertical shading highlights a more angular, square facial structure, particularly by enhancing the jawline, reinforcing a masculine aesthetic with a firmer jawline contour. In both styles, the filter increases the eyebrows’ volume and intensity (Figure 4.9 (b)).

To quantify the observed gender differences, we compute the difference vector $D^{(g)}$ for each

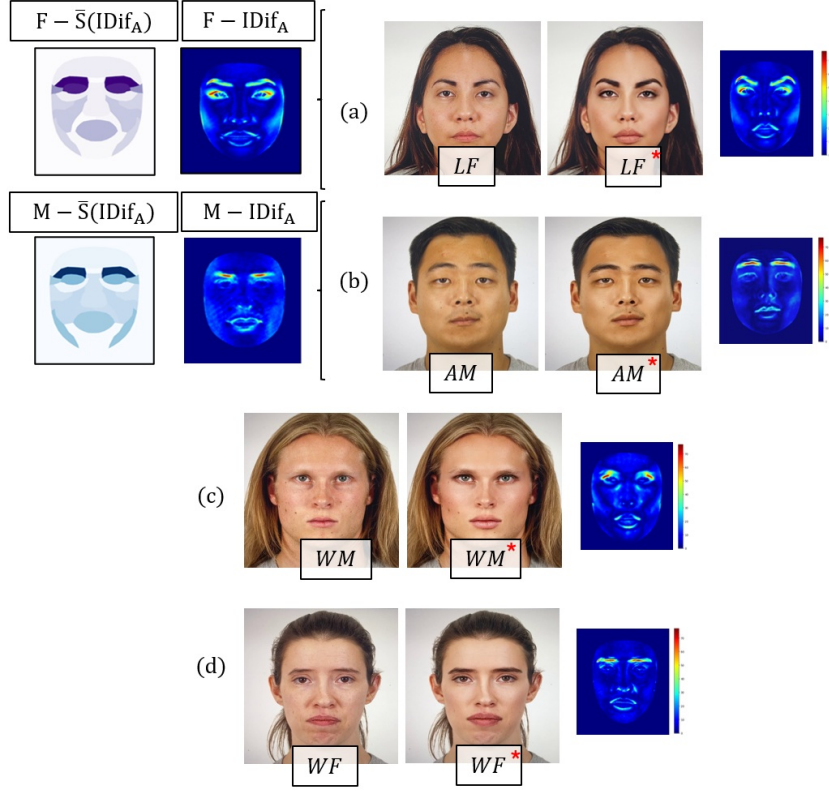


Figure 4.9: Gender differences in the transformations the filter applies based on an implicit gender classification of the input face. (a) Females (F): When the filter classifies the face as a female, its output enhances lips, cheeks, and eyes with makeup; (b) Males (M): When the filter classifies the face as male, it focuses on structural changes without makeup. Unfortunately, the filter might misclassify the input face, as illustrated in (c) and (d), where a man has applied the female transformation (c), and a woman has applied the male transformation (d).

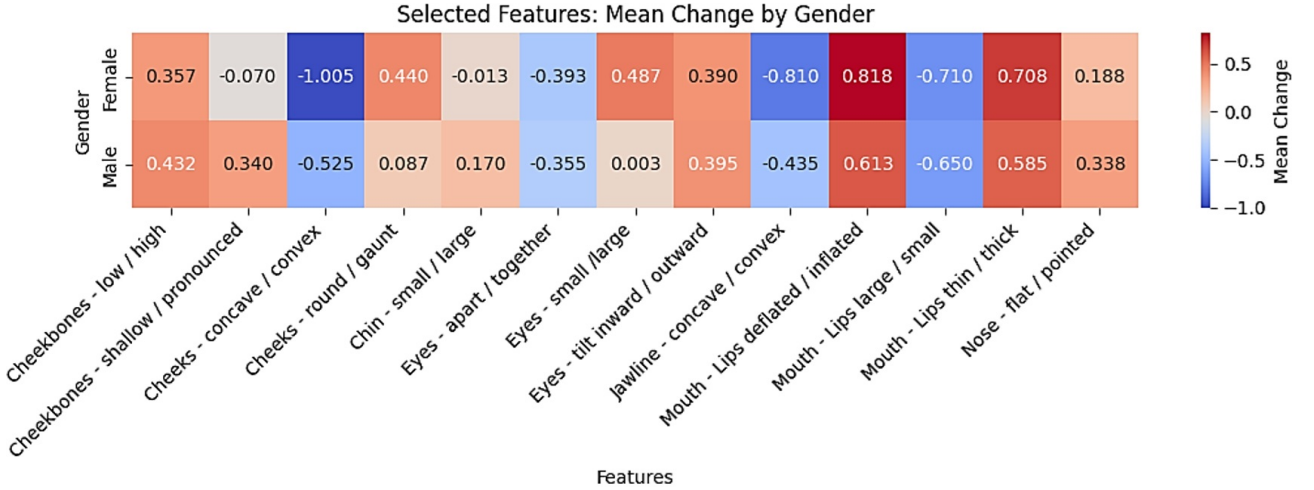


Figure 4.10: Average gender-specific modifications of facial features. The color represents the direction (red for increases and blue for reductions) and the intensity and magnitude of the changes. Observe how the filter's impact on different facial features varies depending on whether the output aligns with the feminine- or masculine-typed style.

gender, where each component $d_i^{(g)}$ represents the change in the i -th facial feature and is defined as

$$d_i^{(g)} = f_i^{(g)*} - f_i^{(g)} \quad (4.1)$$

Here, $f_i^{(g)}$ and $f_i^{(g)*}$ are the components of the previously described feature vectors \bar{F} and \bar{F}^* , representing the pre- and post-filter values, respectively. We select the 13 most commonly modified features, namely areas related to the size and shape of the eyes, cheekbones, chin, jawline, lips, and nose, as shown in Figure 4.10.

The quantitative analysis suggests distinct gender-specific transformations, with women experiencing overall more pronounced changes, as reflected by an average absolute change of 0.491 compared to 0.379 for men. Lips are the most transformed feature across genders, with volumizing effects (+0.818 for women and +0.613 for men). Additionally, women tend to show stronger softening and contouring of the cheeks (−1.005), along with slight enlargement and inward tilting of the eyes (+0.39) and a softened jawline (−0.810), creating a more delicate appearance. In contrast, men’s transformations are comparatively more oriented toward angularity and structure, with enhancements in the jawline (−0.435) and cheekbones (+0.433), reinforcing traditionally masculine features.

Overall, these patterns appear consistent with conventional beauty standards—emphasizing softness and symmetry for women versus structure and definition for men. Importantly, these observations are based on the filter’s output, not on explicit design documentation, and should therefore be interpreted as emergent tendencies rather than confirmed implementation choices. Furthermore, the gender classification is performed automatically by the filter, without the knowledge of the user, which also yields undesirable results in some failure cases, as illustrated in Figure 4.9 (c) and (d).

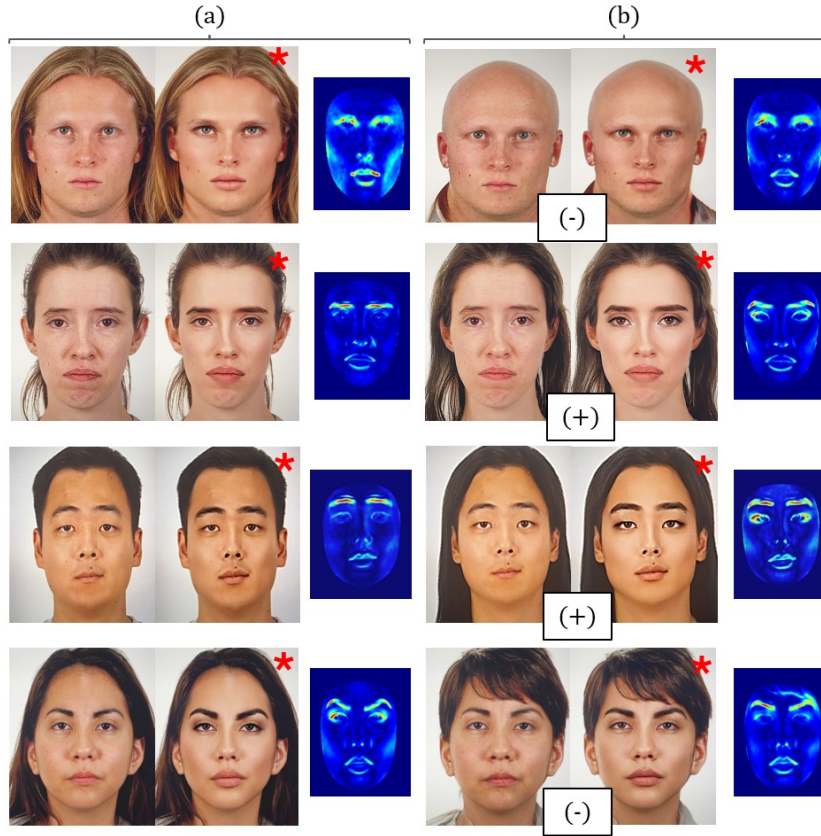


Figure 4.11: Illustrative examples of how changes in hair length may alter the filter’s output transformations. In (a), original images with corresponding filter outputs are shown. In (b), modifications to hair length are applied: The “(+)” and “(-)” symbols indicate extensions or reductions in hair length. The * represents the beautified version of the image.

Sensitivity to facial features As previously described, the filter sometimes applies transformations that align more with those typically observed in the opposite gender group. Our analysis suggests that these mismatched transformations disproportionately affect female subjects. Specifically, 8 out of 26

Black Female images showed this effect (30.76%), compared to 3 out of 26 White Females (11.53%) and 2 out of 26 Latina Females (7.69%). Among male subjects, only 1 out of 26 White Men (3.84%) was affected.

To investigate the factors contributing to these mismatches, we employed an input perturbation approach using FaceApp¹⁰, which allows for modifications in facial features such as hairstyles or facial expressions. While computationally expensive, this approach provided key insights into how the filter perceives faces. Hair length emerged as a recurring factor: many of the affected women had their hair tied back, while the only affected man had longer hair. Figure 4.11 illustrates the impact of hair length on the filter’s transformation.

As shown in Figure 4.11.a and 4.11.b, lengthening (+) or shortening (−) the hair of these cases altered the type of transformation applied in 64.29% of instances. Specifically, 5 Black Female images, 1 White Female image, and 1 Latina Female image remained unchanged after adjustments, while the only affected male was instead modified in line with the opposite pattern. Conversely, altering hair length in originally unambiguous cases had the opposite effect: lengthening the hair of men led the filter to apply transformations typically associated with women, while shortening the hair of women produced transformations usually associated with men.

Taken together, these observations suggest that hairstyle might shapes the type of transformations produced by the filter. This is consistent with prior findings by Albiero et al. [243], who emphasized the role of hairstyle in gender perception and its impact on facial recognition systems.

4.5.7 RQ2.3: Does *Bold Glamour* apply a facial feature morphological alignment?

Next, we turn our attention to RQ2.3 and measure the impact of the filter on aligning facial features across different race-based subgroups. We address this question by studying how the filter affects the distances between facial characteristics of various subgroups, assessing whether there is a trend towards aligning the facial features with those of a specific pre-filter group. For each gender, we compute

$$\Delta d_{(i,j)} = \left\| \overline{F}^{(i)*} - \overline{F}^{(j)} \right\| - \left\| \overline{F}^{(i)} - \overline{F}^{(j)} \right\| \quad (4.2)$$

where $\overline{F}^{(i)}$ and $\overline{F}^{(i)*}$ correspond to the facial features of a specific subgroup (denoted by the i and j subscripts) before and after (*) the application of the filter, as illustrated in Figure 4.12.a). This measure quantifies the change in Euclidean distance between the feature vectors of subgroup i post-filter application and subgroup j pre-filter application.

The heatmaps in Figure 4.12.b depict the $\Delta d_{(i,j)}$ for both female (purple) and male (blue) faces per subgroup. Among female images, a trend emerges where the features of White (WF), Latina (LF), and Asian (AF) participants shift closer to those of Black Females (BF), *i.e.* the $\Delta d_{(i,j)}$ are smallest when $j = \text{Black}$. In male images, the features of White (WM) and Asian (AM) participants show the smallest distances towards Latino Males (LM). Meanwhile, Black (BM) and Latino (LM) groups exhibit mutual proximity, suggesting a reciprocal post-filter alignment.

These insights could serve as a foundation for further studies. Note that the feature vectors only consider facial features and do not have any information regarding skin color.

4.6 Platform Policies

Following the analysis of the *Bold Glamour* filter, we briefly present next the platform policies and reported guidelines for filter creation to shed light on their alignment with the actual behavior of the filter.

¹⁰<https://www.faceapp.com>, Last Access: 20-08-2025.

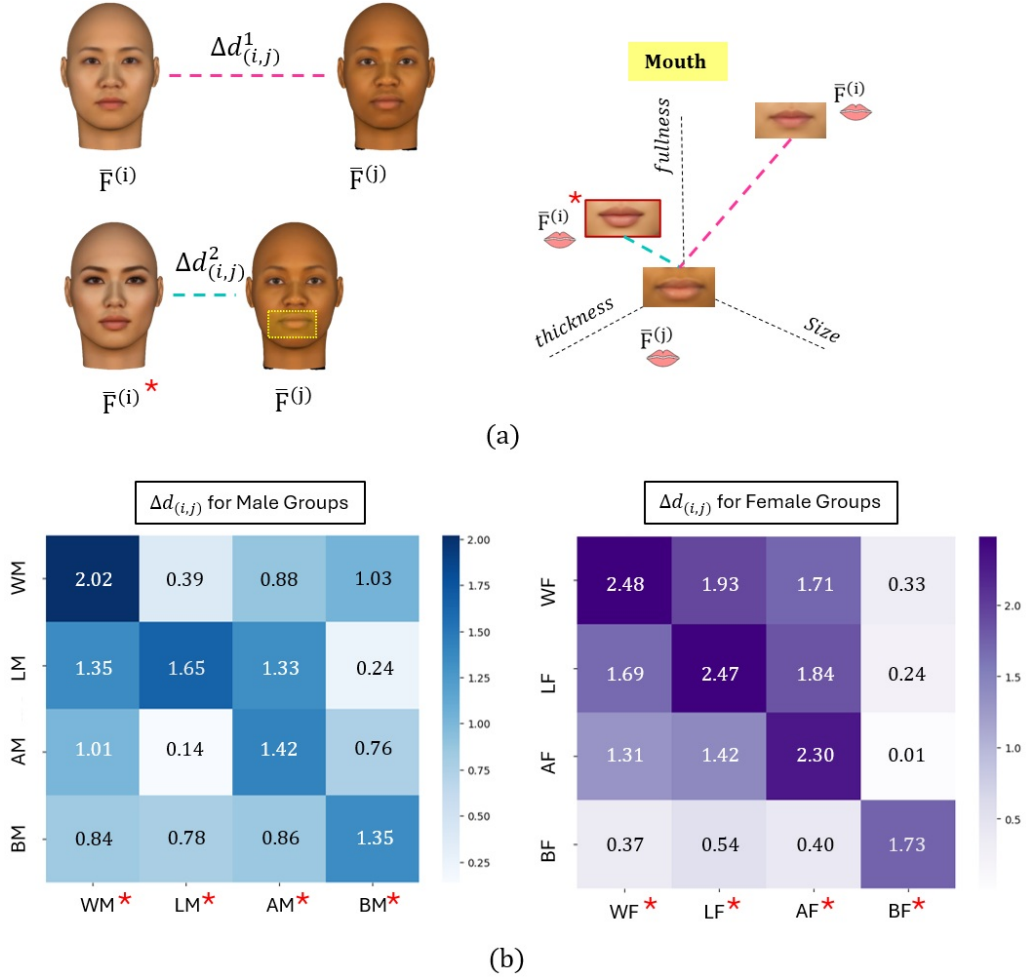


Figure 4.12: (a) Illustration of the metric $\Delta d_{(i,j)}$, which quantifies changes in Euclidean distance between post-filter features of group i and pre-filter features of group j , relative to the pre-filter distance. The mouth diagram provides an example of how alignment is measured through changes in specific features. (b) Heatmaps visualizing $\Delta d_{(i,j)}$ for females (purple) and males (blue). Lighter colors correspond to smaller relative changes, while darker colors indicate larger changes, highlighting numerical patterns of facial feature alignment observed after filter application.

4.6.1 Guidelines vs. Actual Practice

TikTok’s community guidelines, which apply to everyone and everything on their platform, emphasize its focus on creating a welcoming, safe, and entertaining experience [197]. Furthermore, TikTok’s website provides best practices for creating TikTok effects, including a focus on

Diversity and Inclusivity: “When making effects, ensure it is inclusive of a variety of skin tones, hairstyles, facial features, body shapes, accessibility levels, and other differences. Avoid effects that reinforce negative or discriminatory stereotypes relating to gender, sexual orientation, age, ethnicity or disability”; and **Positivity:** “TikTok is a place for authentic, joyful, and uplifting content. Think about creating effects that empower creators to express themselves, explore their self-identity, and share their creativity in uniquely TikTok ways. Effects should promote a positive self-image and avoid reinforcing narrow and unattainable beauty ideals. For example, don’t create effects that make users look thinner or which imply women must wear makeup – or that men can’t” [244]

These guidelines starkly contrast with the actual transformations applied by the *Bold Glamour* filter, as revealed in our analyses, which *seem to apply makeup-like effects more often to female subjects*

in our dataset, for example, while favoring structural changes in male subjects. Such disparities reflect broader systemic patterns in the social media ecosystem, where filters have become pervasive tools that actively shape user experiences and perceptions. In fact, platforms like TikTok and Instagram not only host filters but also incentivize their creation through specialized tools—such as Meta’s Spark AR and TikTok’s Effect House—and financial rewards. ByteDance, for instance, launched the TikTok Effect Creator Rewards program in 2023, dedicating \$6 million to reward creators for viral effects and filters [230]. This investment underscores the market value of AR filters and highlights the pressures these technologies impose on users to conform to prescriptive and lucrative beauty standards.

In this context, the users’ faces and bodies—particularly of female users—are transformed into malleable assets, acting as sites of aesthetic labor that bolster platforms’ profitability. This labor is deeply gendered, as noted by [114], who highlight the emphasis on “female attractiveness” and the pressures to adhere to specific appearance norms, often at significant personal cost. Such dynamics illustrate how visual appearance, mediated by rapidly advancing digital technologies, is increasingly entangled with social media’s commodification of identity. As described by Winch [245], “managing the body is the means by which women acquire and display their cultural capita.” This process reinforces surveillance and normative commodification, subjecting feminine users to persistent pressures to align with idealized standards. The *Bold Glamour* filter exemplifies this trend, showcasing the intersection of technology, market-driven aesthetics, and gendered expectations in the digital age. These transformations reflect biases embedded in algorithmic design and the broader socio-economic structures perpetuating them.

4.6.2 Filtering Out The “Ugly”

TikTok is not unfamiliar with ambiguous techniques for aesthetic curation. Despite what the aforementioned policies claim, and as evidenced by both TikTok’s investments into viral AR filters and our technical investigation into the *Bold Glamour* filter, TikTok actively produces and circulates beauty filters that reinforce “narrow and unattainable beauty standards” in direct contradiction with its own Best Practices for creating AR filters [244]. Moreover, our technical analysis of *Bold Glamour* reveals intersectional biases that appear systemic, emerging from the platform’s infrastructure. Beauty filters like *Bold Glamour* are shaped and circulated within the confines of the platform’s governance, a complex structure of moderation of the content that users produce and consume. Within this framework, aesthetic curation—a practice often imbued with intersectional bias—plays a critical role. Thus, a broader overview of the platform’s (not always transparent) governance techniques is essential to fully understanding the phenomenon of beauty filters.

The results of our analysis on the *Bold Glamour* filter align closely with broader patterns of aesthetic curation and bias on social platforms. TikTok’s emphasis on beauty standards is evident in its filters and algorithmic moderation practices, which actively curate content to favor certain aesthetic ideals. A notable example of this occurred in March 2020, when The Intercept reported that TikTok had internal pressures on content moderation teams to suppress posts from users deemed “too ugly, poor, or disabled” for the platform [246]. Moderators were instructed to filter content for TikTok’s influential For You feed, which most users encounter when opening the app. The criteria for content removal were explicitly biased, listing reasons such as “abnormal body shape”, “ugly facial looks”, “too many wrinkles”, and other “low quality” traits [246]. Moreover, videos filmed in “shabby and dilapidated” environments, such as slums or rural fields, were systematically hidden, while videos showcasing “rural beautiful natural scenery” were exempt from these restrictions.

These moderation practices are deeply intertwined with the effects of AR beauty filters, like *Bold Glamour*, which emphasize specific idealized beauty standards. For instance, the increase in brightness observed for Black and White Female faces in our analysis reflects a trend toward lighter, more “polished” appearances—aligning with the platform’s historical suppression of certain content. As these filters and moderation policies show, platforms are shaping beauty ideals not only through the

tools they provide users but also through the algorithmic curation of content that actively excludes diverse or non-conforming appearances, bodies, and identities ([203, 247, 248, 249, 250, 251]). TikTok’s response to these revelations—acknowledging the primary goal of preventing bullying but dismissing the policies as outdated—suggests that these biases are deeply ingrained and may persist subtly. Our findings, particularly in the context of racial and gender differences, further shed light on how these platforms’ design choices perpetuate exclusionary beauty standards and contribute to the systemic marginalization of certain appearances, reinforcing the pressures placed on users to conform to a specific, profitable image. Thus, beauty filters are not only tools of self-expression but also instruments of broader market-driven agendas that align with established norms in platform economies.

4.7 Discussion and Implications

Our findings underscore critical implications regarding the intersection of beauty filters, social media platforms, and systemic biases. These implications extend across individual, societal, and platform-level dimensions, highlighting the broader consequences of these technologies.

Disclaimer Block, a Tool for Transparency We present the Disclaimer Block which enables us to analyze the transformations performed by beauty filters. This tool provides a detailed breakdown of the modifications applied, offering insights beyond the available minimal information, such as the filter’s name and creator. If widely adopted, the Disclaimer Block could become an integral feature of social media platforms, either as a real-time tool accessible to users or as a mandatory documentation process for filter designers working with platforms like TikTok’s Effect House. Alternatively, the Disclaimer Block could be part of an external platform where users would upload pre- and post-filter images to generate a visual “footprint” of the filter’s effects. Such an interactive analysis could include quantitative data on the magnitude of changes, a heatmap illustrating affected areas, and text-based explanations generated by advanced language models [252, 253, 254]. By clearly communicating semantic and aesthetic alterations, this system would make the transformations more accessible and understandable, empowering users to evaluate how these digital tools shape their self-representation critically.

Implementing tools like the Disclaimer Block represents a crucial step toward promoting transparency and accountability in the design and application of beauty filters. By providing users with detailed insights into the specific changes performed by these filters, platforms could help mitigate the psychological and social pressures tied to unrealistic beauty standards. This transparency can shift the focus from passive consumption of idealized aesthetics to active, informed engagement with digital self-representation. Furthermore, such tools could encourage filter creators to reflect on the societal impacts of their designs, fostering a more ethical approach to digital beauty. Ultimately, widespread adoption of these strategies could help counteract the reinforcement of exclusionary norms and encourage the development of more inclusive and empowering digital environments.

Automatic Gender Inference Our analyses suggest that the *Bold Glamour* filter adapts its transformations in ways that appear correlated with perceived gendered traits, often aligning with traditional gender norms. For female faces in our dataset, the filter enhanced lips, cheeks, and eyes, often adding makeup-like effects to emphasize softness and symmetry. By contrast, male faces more often displayed structural modifications such as sharper jawlines or emphasized angularity, generally without makeup-like effects. We also observed cases where these transformations diverged from face-expected categories, particularly among Black female subjects, who were disproportionately affected. These findings echo concerns raised by studies like Boulamwini and Gebru’s *Gender Shades* [31], highlighting dataset and design biases. While we cannot claim to demonstrate explicit gender classification by the filter, the correlation between facial traits (such as hair length) and diver-

gent transformation patterns points to reductive mechanisms that may unintentionally enforce narrow norms and overlook diverse identities.

Racial and Gender Dependencies Our analyses indicate patterns of racial and gender differences in how the filter applies enhancements. For instance, in our dataset, the filter tended to increase brightness for Black and White females, suggesting a skin-lightening effect that aligns with historical biases favoring lighter skin tones [196]. These transformations reinforce exclusionary beauty standards rooted in Westernized and heteronormative ideals, marginalizing those who do not conform, with potential negative impacts on self-perception and mental health, especially among underrepresented groups. Our study also reveals racial disparities in facial modifications. For women, the filter adjusted features of White, Latina, and Asian individuals to resemble traits associated with Black females, such as fuller lips. However, Black females experienced minimal changes, indicating selective aesthetic idealization. Among men, the filter made features of White and Asian males resemble those of Latino males, while Black and Latino males showed the least alteration post-filter. These findings suggest a dual process: the filter selectively emphasizes traits associated with Black individuals, such as fuller lips while upholding lighter skin tones as the dominant ideal. This mirrors broader beauty trends, including cosmetic surgery, where traits from non-white groups are selectively embraced, but Eurocentric standards, especially lighter skin, remain prioritized [255].

Commodification of Identity and Aesthetic Labor AI-based AR beauty filters blur the boundary between organic and digital self-representation, functioning as algorithmically codified simulations of identity [256]. They render gender performances instantaneous and computationally structured, aligning with Butler’s conceptualization of gender performativity [257], while serving as “normative discursive strategies” that influence user agency [258]. The commodification of identity, driven by the platform’s economic priorities and the incentivization of viral content, positions beauty filters as tools for aesthetic and glamour labor, disproportionately pressuring feminine users to conform to appearance-based norms [259, 114, 260]. This dynamic intertwines visual appearance with social and economic capital, intensifying beauty surveillance [114] and encouraging bodily transformations that enhance platforms’ market appeal. Despite claims that filter use is voluntary, normative governance subtly shapes user decisions through unacknowledged mechanisms, embedding harmful beauty standards that are neither transparent nor easily recognized [261]. These filters thus perpetuate exclusionary norms while serving the platforms’ economic interests.

Beauty Filters as Technologies of Gender Our analyses suggest that the beautification parameters of *Bold Glamour* may reinforce discriminatory gender constructs, echoing [262] and resonating with the concept of “technologies of gender” [263]. Drawing on Judith Butler’s framework, our study emphasizes that gender norms are performatively¹¹ enacted and inscribed on the body through external technologies and narratives, shaping and enforcing idealized body standards [257]. Beauty filters like *Bold Glamour* amplify these dynamics by providing immediacy and hyper-realism into the construction of gendered performance, fostering a personalized yet standardized beauty ideal that imposes biased aesthetic norms on users in opaque and potentially harmful ways to their mental and physical well-being. While promoting creativity, they also reveal an ambiguity in the governance of feminized bodies and identities on Western platforms, where these identities are treated both as valuable commodities and as subjects of normative control. The ambiguous and opaque governance of beauty filters, combined with their biases, limit their users’ ability to critically engage with these technologies

¹¹Gender, according to Judith Butler, is understood as a socially constructed phenomenon that cannot be separated from the “cultural intersection” that both “produce and maintain” it [257]. Butler argues that gender is not something one is, but repeated and ever-changing performances that align with societal norms and expectations. These performances, embedded in cultural, social, and historical contexts, collectively constitute gender identity, highlighting its fluid and performative nature rather than a fixed or inherent characteristic.

and understand their implications. Platforms like TikTok curate and monetize feminine representations, reinforcing heteronormative beauty standards. Addressing these issues requires a multifaceted approach that considers the platforms’ context and governing models. Our analyses reveal and help mitigate the hidden biases embedded in these technologies, debunking corporate self-affirming narratives and critically engaging with their products.

Ethical and Regulatory Considerations Beauty filters reflect broader cultural and technological shifts, where the boundary between physical and digital selves is increasingly blurred. This phenomenon compels users to navigate their identities through the lens of algorithmically mediated aesthetics. The privileging of certain appearances over others exacerbates social inequalities and shapes cultural norms in deeply inequitable and often invisible ways. Our research raises ethical concerns about the responsibilities of platforms in mitigating harm caused by biased filters. While these tools are marketed as empowering and creative, they can contribute to adverse psychological outcomes, such as body dissatisfaction and diminished self-worth. This calls for stricter regulatory oversight of platform practices, including transparency in algorithmic processes, explicit ethical guidelines for filter development, and mechanisms to counteract racial and gender bias.

4.8 Conclusion

This chapter demonstrated how transparency tools can evolve from passive disclosure into active instruments for challenging aesthetic governance. Like ORLAN’s performances that exposed the violence embedded in beauty standards, the enhanced Disclaimer Block makes visible the algorithmic operations that beauty filters conceal—transforming sites of control into spaces of critique.

Our analysis of TikTok’s *Bold Glamour* filter suggests systematic gender- and race-related patterns that stand in tension with the platform’s stated commitments to diversity and inclusion. In our dataset, the filter tended to darken male faces while lightening female faces, and it appeared to adapt transformations differently across genders, with effects that particularly disadvantaged Black women. These transformations occur invisibly, maintaining the illusion of user autonomy while exercising sophisticated forms of aesthetic control.

Where ORLAN’s scalpel made visible the coercion inherent in beauty standards, digital filters promise seamless modification while concealing even more pervasive infrastructures of governance. As Ananny and Crawford [264] point out, seeing does not necessarily mean knowing. Opacity in algorithmic systems is structural rather than just technical, arising from the distributed nature of data, models, and platform economies. What matters, therefore, is not full transparency—which is neither possible nor sufficient—but the creation of sociotechnical conditions for accountability.

In this chapter, the Disclaimer Block is not presented as a purely technical fix but as part of a broader sociotechnical analysis that combines disclosure, bias detection, and policy critique. By situating transparency within these wider dynamics, the chapter reframes it as relational and contextual—a means of enabling contestation, critique, and agency rather than an unattainable ideal of total visibility.

While our investigation focused on a single beauty filter and relied on average faces per demographic group due to data collection constraints—thereby limiting individual variation and generalizability—it provides a foundational contribution to understanding how aesthetic algorithms operate. Moreover, while we used binary gender labels and four racial categories for methodological consistency, we recognize these as socially constructed categories that do not capture the full spectrum of non-binary, multiracial, and other underrepresented identities. Despite these methodological constraints, our work demonstrates how transparency can become a form of resistance against algorithmic aesthetic control, opening pathways for more inclusive and critical approaches to digital beauty technologies. In this context, “resistance” does not mean simply exposing how a filter works, but actively

unsettling the normative assumptions embedded in its design. By making the operations of the filter visible through the Disclaimer Block, users and researchers can contest the gendered and racialized ideals that the system enforces. This aligns with the critical frameworks introduced in Chapter 2.3, where transparency was framed not as a neutral technical property but as a sociotechnical practice. Resistance here is directed toward the opaque governance of platforms and the market-driven logics that incentivize exclusionary beauty standards. The intended outcome is to redistribute interpretive power—shifting part of the authority from platforms and developers back to users and critical communities, enabling them to scrutinize, reinterpret, and challenge how their identities are algorithmically constructed.



Chapter 5

Cascading Gazes: Aesthetic Bias Migration in AI Ecosystems

Warning: This chapter includes content that reflects stereotypical associations between physical appearance and socially constructed attributes. Such associations emerge from biases embedded in generative AI systems—not from empirical truths or the author’s views.

This chapter investigates how aesthetic biases migrate across AI ecosystems, examining the systematic propagation of discriminatory norms through interconnected technological infrastructures. It explores how **text-to-image generative models** operate as sites of *algorithmic lookism*—the systematic association of physical attractiveness with positive attributes—constructing social value hierarchies that determine digital visibility. Through empirical analysis of 26,400 synthetically generated faces, the chapter reveals how generative biases cascade into downstream *classification systems*, creating a coherent normative infrastructure that governs algorithmic legibility. The investigation demonstrates how aesthetic discrimination operates as structural mechanisms that institutionalize exclusion, particularly affecting racialized women who experience both representational invisibility and computational misrecognition. By tracing connections between generative aesthetics and classification performance, it exposes how AI systems collectively participate in a *politics of (in)visibility*—determining who can be seen and how they must appear to be algorithmically recognized.

5.1 From Aesthetic Governance to Systemic Discrimination

The previous chapter’s analysis of TikTok’s *Bold Glamour* filter demonstrated how digital beauty technologies embed normative policies that reinforce gender binaries through algorithmically mediated aesthetic standards (Chapter 4). It has thus been documented how technology does not merely reflect identities, but actively constructs them. Beauty filters operate as instruments of body politics through what can be defined as “aesthetic governance”: the parameters within which digital identity is permitted, tolerated, and recognized are established. However, this governance does not operate in isolation. The artistic work “Past Life Grid” (2021) by Ben Cullen Williams and Isamaya Ffrench [265] reveals with clarity the logic underlying these processes. Images in the training set of a generative system are intentionally deformed by the artists, with these alterations subsequently observed to propagate and become normalized in the output of newly generated images. Deliberate action becomes new generative reality; intervention in the dataset is transformed into visual norms that are reproduced, recreated, and amplified. A crucial mechanism is thus exposed by the work: the politics that are included—or excluded—in training systems become the normatives that systems reproduce and impose. This artistic insight leads to the heart of the present chapter. While *Bold Glamour* demonstrated how a single filter codifies discriminatory aesthetic norms, a more pervasive

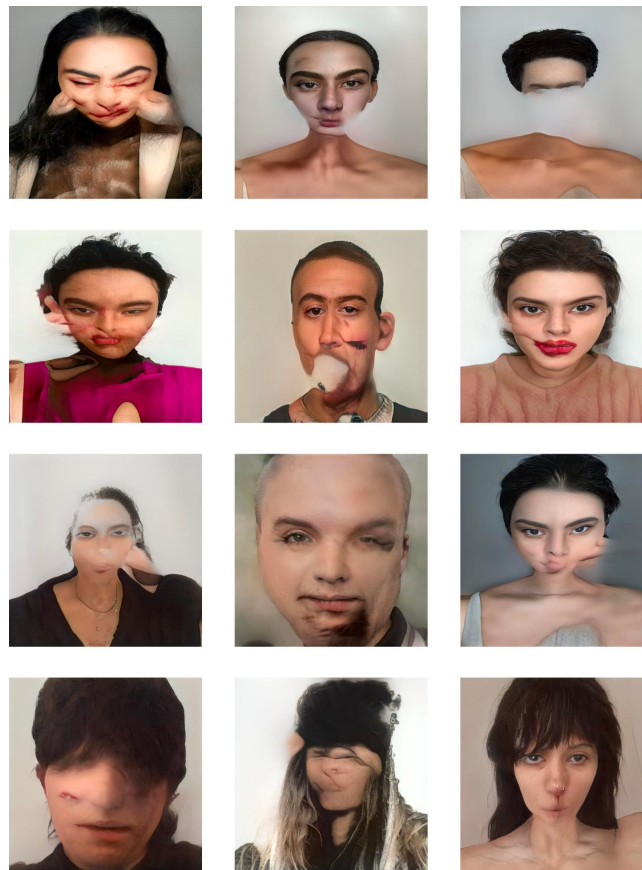


Figure 5.1: “Past Life Grid”, Ben Cullen Williams and Isamaya Ffrench, (2021) © Ben Cullen Williams and Isamaya Ffrench [265].

phenomenon must now be confronted: the technological ecosystem linked to AI—systems that feed on data containing specific patterns and ideologies—repeats, transmits, and reinforces these same patterns across different applications, whether generative AI or classification systems.

In this chapter, we will see that text-to-image generative models do not merely create images: social value hierarchies are created through the systematic association of physical attractiveness with positive attributes such as intelligence, trustworthiness, and happiness—despite these traits having no empirical correlation with physical appearance. This *algorithmic lookism* generates a double harm: first, certain types of bodies and identities are rendered invisible through representation, then this invisibility is punished through recognition.

Gender classifiers—such as those operating behind Bold Glamour—correctly recognize only certain aesthetics while systematically failing on others. Those who do not conform to algorithmic standards of attractiveness are not only misrepresented but also misrecognized, creating systematic barriers to digital access and participation.

A systematic pattern thus emerges where visibility is never given, but always conditioned by conforming to algorithmically mediated aesthetic normatives. Generative and classification technologies reproduce and amplify a singular vision already present in reality, systematically obscuring those who do not fit within these normative politics. As in “Past Life Grid,” the deformed faces inserted in training become the only new digital identities, reproduced and amplified throughout the entire system.

From the aesthetic control of beauty filters to the identity governance of generative systems, the present chapter traces the connections of this discriminatory ecosystem, demonstrating how AI does not simply reflect biases, but institutionalizes them through technological infrastructures that determine who can be seen—and how they must appear to be recognized.



Figure 5.2: Examples of generated images for the promptot <Front photo of a [attribute] White woman> with Stable Diffusion 3.5 Medium.

5.2 Algorithmic Lookism and Its Systematic Propagation

Generative Artificial Intelligence (AI) systems are increasingly shaping the content that we consume online [266, 125, 266, 267], creating new forms of unintended harms that extend far beyond traditional notions of algorithmic bias [268, 269]. While the rapid adoption of text-to-image (T2I) models promises democratized content creation, these systems simultaneously generate systematic disadvantages for marginalized communities through mechanisms that are often invisible to both users and developers [97]. The growing need to detect, quantify and mitigate potential biases [270] has revealed that generative AI systems do not merely reflect existing societal prejudices but amplify and institutionalize them through *synthetic normativity* [271], *i.e.*, the algorithmic construction of aesthetic and social norms that determine who becomes visible and valued in AI-generated content.

Significant work in the literature has focused on gender [272, 273, 274], racial [275, 276, 277] and age [30, 278] biases in computer vision and image generation models, including T2I models, which have been found to exhibit demographic biases, with whiteness and masculinity overrepresented in the generated images [279, 280]. Furthermore, these biases extend beyond demographic traits, affecting object selection, clothing, spatial representations [281] and subtler forms of discrimination, such as those identified by Kumar et al. [282], including *representative bias*, the disproportionate privileging of the perspectives or experiences of particular identity groups, and *affinity bias*, the systematic preference for particular narratives or viewpoints. While examined initially in the context of language models, these dynamics can also emerge in generative image systems, as illustrated by Qadri et al. [283], who document how visual outputs encode skewed demographic portrayals and culturally biased aesthetics, even when prompts are demographically neutral. Another form of discrimination, *algorithmic lookism*—the preferential treatment of individuals based on their physical appearance—represents a particularly insidious form of unintended harm [284]. Rooted in beauty standards and cognitive biases [285, 286, 287, 288, 221], algorithmic lookism creates cascading disadvantages that affect not only representation, but also the behavior of downstream automated decision-making systems.

Here, the term “cascading” emphasizes how data infused with specific aesthetic patterns and ideologies is repeatedly transmitted across the AI pipeline (Figure 5.3). Generative models amplify these patterns by producing vast amounts of synthetic images that both circulate online and serve as training material in data augmentation, thereby reinforcing existing norms. Classification systems then legitimize such patterns by rendering them “recognizable,” which in computational terms means visible and intelligible—while identities that do not conform risk invisibility or misrecognition.

Following this idea, the unintended harms caused by *algorithmic lookism* operate through three

interconnected mechanisms. First, T2I models systematically render certain appearances invisible while privileging others, creating what can be understood as a “*politics of (in)visibility*”—where Eurocentric aesthetic norms determine whose faces are deemed worthy of representation. More troubling still, this extends beyond mere invisibility to active negative association, where non-conforming appearances become linked with undesirable traits or characteristics in the model’s representational space. Second, these appearance-based biases reflect market-driven optimization processes that reconfigure human identity into profitable data under neoliberal rationality, where visibility depends on conformity to algorithmically-determined standards of attractiveness. Third, the aesthetic biases encoded in generative models propagate to image analysis systems used for a variety of purposes, including identity verification, gender classification, security, and access control, creating systematic barriers for individuals who do not conform to algorithmic beauty standards.

In this chapter, we examine how *algorithmic lookism* [284] leads to unintended harms through the systematic correlation of facial attractiveness with positive attributes in images generated using Stable Diffusion (SD) 2.1 and 3.5 Medium. Specifically, we assess four attributes—namely happiness, sociability, trustworthiness, and intelligence—originally operationalized by Gulati et al. [221] in a study with human subjects and grounded in a long-standing body of social-psychological literature [285, 289, 286, 290, 291, 292, 293], which has consistently demonstrated that physically attractive individuals are *perceived* to be more sociable, trustworthy, intelligent and happier, despite limited empirical evidence for such correlations. By adopting the same attribute definitions utilized by Gulati et al., our study facilitates direct comparability between results obtained from synthetic facial images and the large-scale, high-quality human data reported in their work, enhancing the potential for future comparative research.

Furthermore, we evaluate the impact of *algorithmic lookism* on computer vision-based gender classification models not merely as a downstream application, but as evidence of a pervasive aesthetic normativity that operates consistently across AI systems. Our analysis of three gender classification models applied to synthetically generated faces reveals how the same aesthetic biases encoded during generation are systematically recognized and rewarded by classification systems trained on entirely different datasets. This convergence demonstrates that we are not observing isolated bias incidents, but rather a coherent normative infrastructure that determines digital legibility across the AI ecosystem. Building on work by Doh et al. [118], we treat differential classification performance as diagnostic evidence that aesthetic standards function as systematic filtering mechanisms determining who becomes *algorithmically readable* in digital identity systems.

Our study yields three key findings that shed light on the unintended harms of current generative AI systems: (1) T2I generative models show strong signs of *algorithmic lookism*—associating attractiveness with positive attributes, even though attractiveness is not a good predictor of such attributes—revealing how AI systems perpetuate and amplify existing forms of social hierarchy while simultaneously constructing novel mechanisms of aesthetic-based stratification; (2) computer vision-based gender classification models are also impacted by *algorithmic lookism*, with faces generated using negative attribute terms being misclassified more often than their positive counterparts. This phenomenon causes differential accuracy patterns that systematically disadvantage marginalized groups in identity verification processes; and (3) the impact of *algorithmic lookism* is nonuniform for different gender and racial groups with Asian and Black women being disproportionately impacted across multiple dimensions, namely lower classification accuracy, stronger aesthetic-behavioral correlations, and systematic exclusion from normative representations.

These findings demonstrate that T2I systems generate additional layers of harm through the automatic association of physical attractiveness with positive attributes, a phenomenon that operates independently of explicit demographic categories yet systematically disadvantages the same marginalized groups. The demonstrated unintended harms of generative AI extend far beyond isolated instances of bias, revealing systematic patterns of digital discrimination that mirror and amplify existing structures of inequality. As these systems become increasingly integrated into everyday digital interac-

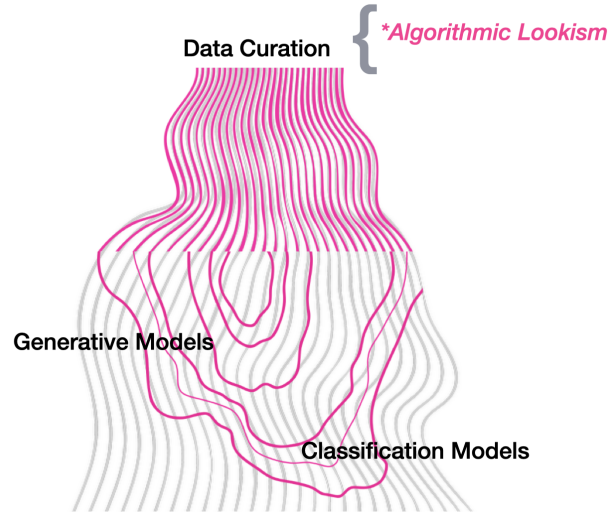


Figure 5.3: From data curation to generation to classification: a schematic of the cascade in which aesthetic correlations learned upstream are reproduced and reinforced downstream.

tions, understanding and addressing these synthetic forms of normativity becomes critical to ensure an equitable participation in an increasingly automated society.

5.3 Research Questions¹

The main research question addressed in this chapter is as follows:

Main RQ3: How do generative AI systems create and enforce systematic patterns of unintended harm through aesthetic discrimination that operates as a coherent infrastructure across AI ecosystems?

This overarching question is articulated in the following two research questions:

RQ3.1: Do synthetic facial images generated by diffusion models exhibit algorithmic lookism, i.e., an implicit correlation between attractiveness and unrelated attributes? To tackle this research question and drawing from the literature of this cognitive bias in humans [221], we study the existence of a systematic association between four positive attributes (happiness, trustworthiness, intelligence and sociability) and the model’s operationalization of facial attractiveness (RQ3.1.1). We also assess whether the strength of this correlation can be measured quantitatively and study it on different demographic groups (RQ3.1.2).

RQ3.2: Does algorithmic lookism impact the performance of downstream tasks, particularly gender classification? To address the research question, we evaluate whether the performance of gender classification models changes depending on the attributes and demographic factors of the synthetic images.

5.4 Dataset Creation

Two different datasets, each comprising 13,200 face images, were generated by prompting two widely used, open-source T2I diffusion-based models—Stable Diffusion 2.1 [294] and Stable Diffusion 3.5 Medium [295]—adopting the same protocol. In both datasets, images vary by gender (woman, man),

¹Under the main research question (RQ3) *Cascading Gazes: Aesthetic Bias Migration in AI Ecosystems* presented in Chapter 2 Section 2.3.



Figure 5.4: Examples of the generated faces with Stable Diffusion 2.1 and Stable Diffusion 3.5 Medium with positive (+) and negative (-) variations for three attributes (A = Attractiveness, H = Happiness, and I = Intelligence) together with the neutral faces (N = Neutral). Yellow (♀) and green (♂) correspond to images of females and males, respectively. Light Blue (♂) borders highlight the faces corresponding to the positive Attractiveness trait.

race² (Asian, Black, White), and five attribute pairs associated with lookism in humans—attractiveness halo effect—[221]: attractive vs. unattractive, intelligent vs. unintelligent, trustworthy vs. untrustworthy, sociable vs. unsociable, and happy vs. unhappy. Prompts followed the format: *<Front photo of a [attribute] [race] [gender]>* with 200 images generated per attribute–race–gender triplet. Additionally, a “Neutral” set (200 images per race–gender combination) was generated without any attribute descriptors in the prompt, to serve as a baseline to measure each of the model’s default tendencies.

To isolate facial features from potential confounding factors, all generated images were cropped to focus exclusively on the facial region, removing clothing, backgrounds, and other contextual elements that could introduce additional biases into the analysis. This pre-processing ensures that all subsequent analyses examine aesthetic discrimination based solely on facial characteristics rather than external visual cues. Figure 5.4 shows samples of positive and negative attribute terms across all six gender–race categories. It should be noted that facial attractiveness is neither defined nor measured in this study—instead, the focus is placed on analyzing how T2I models associate attractiveness—and its absence—with other positive and negative attributes.

Training Differences Between Stable Diffusion 2.1 and 3.5 Medium

While both image sets in our study were generated using the same prompting protocol, the underlying diffusion models differ substantially in architecture, scale, and training pipeline. Stable Diffusion 2.1 [294] was trained on filtered subsets of the LAION-5B dataset [296], a large-scale corpus of image-text pairs scraped from the public web. The dataset was preprocessed using automated filters to remove content deemed unsafe or aesthetically low-quality. Nonetheless, prior studies have pointed out that large-scale web-scraped datasets like LAION-5B often contain pervasive misogyny, explicit content, and stereotypical representations, which can propagate harmful associations along lines of gender, race, and perceived physical appearance [297, 97, 98].

The training data used for Stable Diffusion 3.5 Medium, by contrast, has not been publicly disclosed. However, available documentation highlights significant changes in data curation practices. These include the application of NSFW classifiers, aesthetic ranking systems, semantic deduplication techniques, and the use of synthetic captions generated with vision-language models such as CogVLM [295]. While the source data may remain similar in origin—likely large-scale web-scraped

²The term *race* is used as in standard ML datasets, acknowledging it as a social construct distinct from *ethnicity* [238, 239]. This study does not seek to promote or reify racial categories in AI, but rather to critically examine how AI systems encode and propagate biases linked to socially constructed categories such as race. This problem is discussed in Chapter 6.

corpora—the enhanced preprocessing pipeline alters how visual concepts are represented and associated in the model’s latent space.

These differences in training pipelines are not merely technical. They affect how each model encodes relationships between facial characteristics, perceived attractiveness, race, gender, and other high-level attributes. Our analysis examines potential divergences between the outputs of SD 2.1 and SD 3.5 Medium in light of these underlying differences, with particular attention to their implications for demographic representation and aesthetic normativity. While our quantitative analyses (Sections 5.6.1–5.6.2) measure these differences through embedding space calculations, our qualitative visual assessment (Section 5.7) complements them by examining observable aesthetic patterns in the generated imagery.

5.5 Methodology

5.5.1 Feature Extraction

To quantify the relationship between attractiveness and other attributes, we extracted features from the generated images using two complementary embedding spaces: CLIP [298], which provides general-purpose visual–semantic representations learned from diverse image–text pairs, and ArcFace [299, 300], which offers face-specific embeddings trained on facial recognition tasks. This combination ensures that our findings are not dependent on a single representational framework, leveraging both general contextual cues and specialized facial features. By using both, we can validate the robustness of our bias measurements across different embedding spaces and mitigate concerns that our results might be artifacts of a particular learned representation.

5.5.2 RQ3.1.1: *Algorithmic lookism* Reference Point Analysis

This analysis addresses our primary research question by examining whether T2I models systematically associate positive attributes with attractiveness and negative attributes with unattractiveness. We developed a distributional approach that considers all pairwise distances between samples to capture the complete distributional characteristics of attribute associations.

Let the sets be defined as:

Gender (G) {man, woman}

Race (R) {White, Black, Asian}

Attributes (A) {happy, unhappy; intelligent, unintelligent; neutral;
sociable, unsociable; trustworthy, untrustworthy; }

Note that, although *attractive* and *unattractive* were generated in the same way as the other attributes, they are not included in A here because they serve as the reference categories in the distance calculations described below.

Each **demographic group** is defined as a pair $g = (\text{gender}, \text{race}) \in G \times R$. Thus, for each demographic group $g = (\text{gender}, \text{race})$ and attribute $a \in A$, we compute two distance distributions:

$$\mathcal{D}_g^{(a \rightarrow \text{attractive})} = \left\{ \|\mathbf{e}_i^{(a)} - \mathbf{e}_j^{(\text{attractive})}\|_2 : i = 1, \dots, N_a, j = 1, \dots, N_{\text{att}} \right\} \quad (5.1)$$

$$\mathcal{D}_g^{(a \rightarrow \text{unattractive})} = \left\{ \|\mathbf{e}_i^{(a)} - \mathbf{e}_k^{(\text{unattractive})}\|_2 : i = 1, \dots, N_a, k = 1, \dots, N_{\text{unatt}} \right\} \quad (5.2)$$

where $\mathbf{e}_i^{(a)}$ represents the i -th embedding vector for attribute a , and N_a , N_{att} , N_{unatt} denote the number of samples for attribute a , attractive, and unattractive categories, respectively.

We operationalize *algorithmic lookism* via a distributional measure (DLM), defined as the difference between the expected values of the two distance distributions:

$$L_g^{(a)} = \mathbb{E}[\mathcal{D}_g^{(a \rightarrow \text{unattractive})}] - \mathbb{E}[\mathcal{D}_g^{(a \rightarrow \text{attractive})}] \quad (5.3)$$

A positive DLM value $L_g^{(a)} > 0$ indicates that attribute a is systematically closer to “attractive” samples, while a negative value $L_g^{(a)} < 0$ suggests systematic proximity to “unattractive” samples. Therefore, under the *algorithmic lookism* hypothesis, positive attributes are expected to yield positive DLM values and negative attributes to yield negative ones.

For each attribute–demographic combination, we perform statistical comparisons between the two distance distributions. We first assess distributional assumptions using the Shapiro–Wilk test for normality and Levene’s test for homogeneity of variances, then apply the most appropriate statistical test (independent t-test, Welch’s t-test, or Mann–Whitney U test). Effect sizes are quantified using Cohen’s d to assess the practical significance of observed differences.

5.5.3 RQ3.1.2: Cross-Attribute Correlation Analysis

This analysis addresses RQ1.2 by examining how strongly attractiveness and positive attributes are correlated at the representational level in the learned embedding space.

For each attribute with positive/negative polarity (*i.e.*, all attributes but “neutral”), let a^+ and a^- denote the positive (*e.g.*, trustworthy) and negative (*e.g.*, untrustworthy) version of the attribute, respectively.

We define the Correlation Lookism Strength (C) per demographic group pair $g = (\text{gender}, \text{race}), \in G \times R$, as:

$$\Delta C^{(g)} = \mathbb{E}[\rho(\mathbf{e}_i^{(\text{attractive})}, \mathbf{e}_j^{(a^+)})] - \mathbb{E}[\rho(\mathbf{e}_i^{(\text{attractive})}, \mathbf{e}_k^{(a^-)})] \quad (5.4)$$

where $\rho(\mathbf{e}_i^{(\text{attractive})}, \mathbf{e}_j^{(a^+)})$ represents the correlations between the individual embedding vectors of attractive individuals and the embedding vectors of positive attributes, and analogously for negative attributes.

We compute all pairwise correlations between individual attractive samples and all samples created with positive/negative attributes, then test whether the mean correlation difference is significantly greater than zero using the Mann–Whitney U test.

Under the *algorithmic lookism* hypothesis, we expect $\Delta C^{(g)} > 0$ across demographic groups, indicating stronger correlations between attractive and positive attributes than between attractive and negative attributes.

5.5.4 RQ3.2.1 Impact of *Algorithmic lookism* on Gender Classification Performance

This analysis addresses RQ2.1 by investigating whether *algorithmic lookism* in T2I models propagates to downstream classification tasks. Specifically, we examine whether gender classifiers exhibit differential performance on faces generated with positive versus negative attribute prompts, and whether systematic misclassification patterns emerge based on attractiveness levels.

We evaluate gender classification performance using three widely-adopted face analysis models: InsightFace [301], DeepFace [299], and FairFace [30]. For each demographic group pair $g = (\text{gender}, \text{race}) \in G \times R$ and attribute $a \in A$, we compute the gender classification accuracy as:

$$\text{Accuracy}_{g,a} = \frac{1}{N_{g,a}} \sum_{i=1}^{N_{g,a}} \mathbb{I}[\hat{y}_i = y_i] \times 100 \quad (5.5)$$

where $N_{g,a}$ is the total number of images for demographic group g and attribute a , \hat{y}_i is the predicted gender label for image i , y_i is the true gender label specified in the generation prompt, and

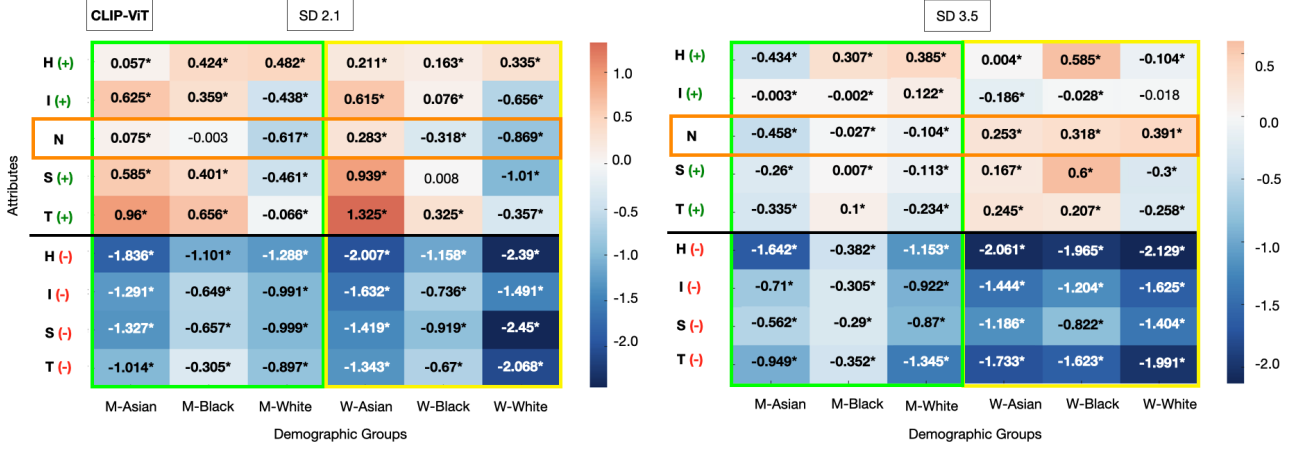


Figure 5.5: Distributional measure (DLM) scores across demographic groups in images generated with Stable Diffusion 2.1 (left) and 3.5 (right), using CLIP embeddings. Negative values (blue) indicate closeness to unattractive faces; positive values (red) to attractive ones. An asterisk (*) indicates statistical significance ($p < 0.05$). The neutral trait is highlighted with a orange box. The results for men/women are highlighted with a green/yellow box, respectively. A = Attractiveness, H = Happiness, I = Intelligence, S = Sociability, T = Trustworthiness, N = Neutral.

$\mathbb{I}[\cdot]$ denotes the indicator function. A classification is considered correct when the predicted gender matches the gender specified in the generation prompt.

We compare classification performance across attributes to assess whether faces generated with positive versus negative attributes exhibit systematic differences in gender recognition accuracy. Misclassification rates are analyzed across all gender–race combinations to identify potential biases that disproportionately affect certain demographic groups based on the attractiveness bias encoded in the generated faces.

This analysis aims to determine whether *algorithmic lookism* in T2I models creates systematic vulnerabilities in automated gender recognition systems, potentially leading to the differential treatment of individuals based on their attractiveness levels embedded in facial representations.

5.6 Results

In this section, we provide a summary of the results of the previously described analyses addressing the proposed research questions.

5.6.1 RQ3.1: Do synthetic facial images generated by diffusion models exhibit *algorithmic lookism*, i.e., an implicit correlation between attractiveness and unrelated attributes?

RQ3.1.1: *Algorithmic lookism* Reference Point Analysis

Our findings consistently indicate the presence of *algorithmic lookism* both in Stable Diffusion 2.1 and 3.5. Using both ArcFace and CLIP embeddings to ensure robustness across representational spaces, we observe that faces generated with positive attributes tend to cluster closer to the “attractive” reference, and faces corresponding to negative attributes are positioned closer to the “unattractive” reference in both embedding spaces. This phenomenon is more pronounced in the CLIP embedding space and affects women significantly more than men.

Neutral faces, those generated without any explicit attribute in the prompt, offer insight into the aesthetic defaults embedded in generative models. In SD 2.1, neutral faces are mostly positioned

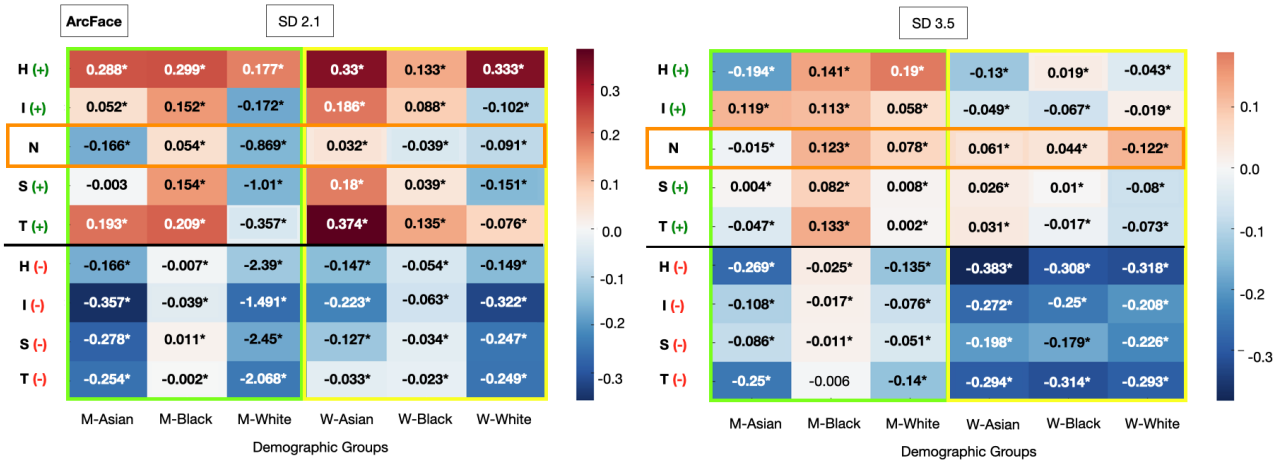


Figure 5.6: Distributional measure (DLM) scores across demographic groups in images generated with Stable Diffusion 2.1 (left) and 3.5 (right), using ArcFace embeddings. Negative values (blue) indicate closeness to unattractive faces; positive values (red) to attractive ones. An asterisk (*) indicates statistical significance ($p < 0.05$). The neutral trait is highlighted with a orange box. The results for men/women are highlighted with a green/yellow box, respectively. A = Attractiveness, H = Happiness, I = Intelligence, S = Sociability, T = Trustworthiness, N = Neutral.

closer to the unattractive cluster, with limited demographic differentiation. Interestingly, this baseline shifts in SD 3.5: neutral faces are closer to the attractive cluster. In the CLIP embedding space, this phenomenon is gender-dependent: while neutral female faces are closer to attractive exemplars, male faces remain near the unattractive pole. This suggests that a form of default beautification is structurally encoded for women, even in the absence of specific attributes, suggesting a gendered aesthetic logic. This pattern reflects a growing gender asymmetry in newer models: femininity is increasingly subject to aesthetic optimization, while masculinity is comparatively ignored or neutralized. The rising polarization in CLIP highlights how so-called neutral representations are anything but neutral: they encode normative assumptions that tether the female identity to beauty, reinforcing gendered expectations even in the absence of explicit prompts.

An intersectional analysis reveals that faces of Asian and Black women are systematically subject to stronger polarizations: they exhibit the largest differences in the distributional lookism scores the between positive and corresponding negative attributes and are also disadvantaged in neutral conditions. In contrast, faces of both White men and White women show weaker-and in some cases reversed-distributional lookism scores, with positive attributes in some cases aligning with unattractiveness. These findings indicate not only the existence of persistent *algorithmic lookism*, but also a demographic asymmetry in how social value is visually encoded: virtue is more tightly coupled with attractiveness for racialized and gendered identities, while White subjects are comparatively less constrained by this aesthetic-moral mapping.

A preliminary visual inspection of SD 2.1 outputs suggests that, for White subjects, positive attributes such as intelligent and sociable are depicted with greater variation in age and appearance than attractive, whereas for Asian and Black subjects, positive attributes tend to converge toward the same youthful and polished aesthetic associated with attractive (Figure 5.7). This pattern may point to an under-representation of more diverse appearances for these groups in the training data (Full qualitative analysis in Section 5.7).

RQ3.1.2: Cross-Trait Correlation Analysis and Neutral Defaults

In the case of SD 2.1, CLIP embeddings reveal the strongest aesthetic polarization for faces of Asian women (+0.079), followed by faces of Black women (+0.049) and Asian men (+0.055). The effect



Figure 5.7: Visual observation of SD 2.1 outputs for intelligent, sociable, attractive, and unattractive across gender–race groups. While White subjects (left columns within each gender) display greater variation in age and appearance for positive attributes (intelligent, sociable) compared to attractive, Asian and Black subjects tend to show a more young-looking and stylized appearance for positive attributes, closely resembling attractive.

is markedly weaker in the case of faces of White women (+0.018) and White men (+0.032), suggesting a partial decoupling between positive attributes and facial attractiveness among White subjects. ArcFace embeddings confirm this trend, though with lower magnitudes: faces of Asian women again show the highest correlation (+0.040), while faces of White men exhibit no significant difference (-0.005 , $p = 0.99$).

Neutral faces further expose the aesthetic defaults embedded in the generative space. In the CLIP embedding space, only faces of White men align with attractiveness in the absence of descriptive prompts (-0.03), whereas all other groups—particularly faces of racialized women—are subtly displaced toward unattractiveness (faces of White women = $+0.07$, Black women = $+0.11$). In the ArcFace embedding space, all groups exhibit positive distances (*i.e.*, closer to “unattractive” references), but faces of White men again register the lowest value ($+0.32$) and faces of Asian women the highest ($+0.51$). These findings indicate that attractiveness is not only unevenly mapped to attributes, but also structurally encoded into the model’s aesthetic defaults.

Regarding SD 3.5, the existence of *algorithmic lookism* persists, albeit with slightly reduced magnitudes. In the CLIP embedding space, the trait-based difference is the largest for faces of Black women ($+0.067$), Asian women ($+0.059$), and White women ($+0.045$). Male groups show weaker associations: faces of White men score $+0.051$, while faces of Asian men display minimal effect ($+0.009$). ArcFace confirms this demographic gradient: Black ($+0.047$) and Asian women ($+0.045$) again show the strongest trait-attractiveness alignment, while male groups remain less affected.

Neutral faces in SD 3.5 continue to reflect gendered and racialized defaults. In CLIP, female faces are systematically pulled towards attractiveness (White = -0.21 , Asian = -0.24), while in the case of faces of White men there is a shift in the opposite direction ($+0.34$), suggesting an intensification of feminized aesthetic encoding. A subtler version of this trend is also present in the ArcFace embedding space: faces of females remain more closely aligned with attractiveness than faces of males (*e.g.*, faces of White women = $+0.10$, Black women = $+0.16$, Asian women = $+0.12$), while faces of White men are closer to the aesthetic center ($+0.06$) and faces of Asian men slightly above ($+0.17$).

Taken together, these findings confirm that both T2I models encode structural lookism: facial attractiveness systematically co-occurs with positive attributes, particularly in the case of faces of women—and most acutely for faces of racialized women.

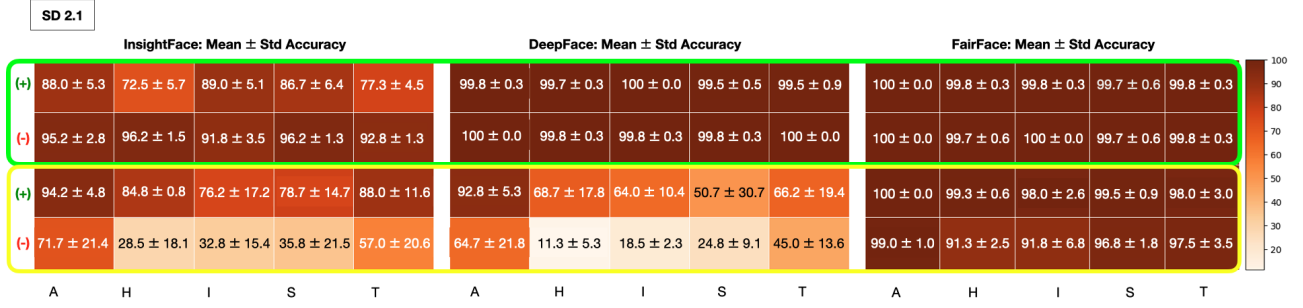


Figure 5.8: Heatmaps of gender classification accuracy (Mean \pm Std) for InsightFace, DeepFace, and FairFace for **SD 2.1**. A = Attractiveness, H = Happiness, I = Intelligence, S = Sociability, T = Trustworthiness. Performance corresponding to female (Women)/male (Men) faces is highlighted in Yellow \bullet / Green \bullet , respectively. Accuracies in classifying neutral faces (%): InsightFace: (W) 80.9 ± 10.7 (M) 87.5 ± 1.5 , DeepFace: (W) 51.1 ± 9.0 (M) 99.6 ± 0.1 , FairFace: (W) 96.8 ± 1.4 (M) 98.9 ± 0.4 .

5.6.2 RQ3.2: Does *algorithmic lookism* impact the performance of downstream tasks, particularly gender classification?

Gender Classification Performance as a Function of Different Attributes

Figure 5.8 and Figure 5.9 report classification performance by model and gender, with race categories aggregated for ease of interpretation in the main text. The disaggregated results by race are provided in Appendix C.1. As shown, performance varies significantly depending on the attributes and the gender of the generated image. The three gender classification algorithms perform the best on the faces of males generated both with SD 2.1 and SD 3.5 for all attractiveness values and attributes. DeepFace achieves near-perfect accuracy ($\geq 98.5\%$) on faces of males across all attributes in both generated datasets, while FairFace maintains consistently high performance, ranging from 99.7% to 100%, across all male categories. Even InsightFace, which shows the greatest variability, maintains relatively stable and competitive performance on faces of males ranging from 72.5%-96.2% on SD 2.1-generated images and improving to 84.8%-98.8% on SD 3.5-generated images.

The stability of classification performance on faces of males persists across racial categories, confirming that algorithmic bias disproportionately affects female classification regardless of racial identity or generative model version (see Appendix C.1).

Conversely, the classification performance of the three algorithms on female faces demonstrates considerably different patterns, with large variations depending on the attributes and exhibiting a systematic deterioration in specific model-attribute combinations on the two generated datasets. On SD 2.1-generated images, the performance of the classification algorithms on female faces with negative attributes shows a substantial drop: the classification accuracy on "unhappy" female faces is only 28.5% in InsightFace and 11.3% in DeepFace, while the classification accuracy of "happy" female faces reaches 84.8% and 68.7%, respectively. Intelligence shows similar patterns, with a classification performance on faces of "unintelligent" women of only 32.8% (InsightFace) and 18.5% (DeepFace), compared to 76.2% and 64.0% for their "intelligent" counterparts. FairFace maintains more stable performance across different attributes of female faces but still exhibits systematic sensitivity to the attributes, with accuracy ranging from 91.3% (for unhappy female faces) to 100% (for attractive female faces).

Classification performance on SD 3.5-generated images reveals a paradoxical pattern of simultaneous improvement and deterioration, depending on the model and the attribute. FairFace demonstrates notable progress toward equity, achieving near-perfect performance on faces created with positive attributes (100%) and substantial improvements on faces corresponding to negative attributes (88.7%-99.7%). DeepFace yields mixed results, with some improvements on faces created with

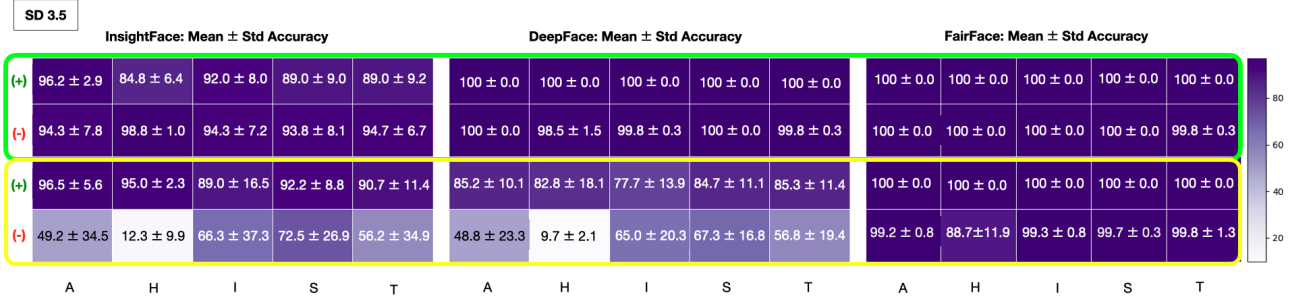


Figure 5.9: Heatmaps of gender classification accuracy (Mean \pm Std) for InsightFace, DeepFace, and FairFace for **SD 3.5**. A = Attractiveness, H = Happiness, I = Intelligence, S = Sociability, T = Trustworthiness. Performance corresponding to female (women)/male (men) faces is highlighted in Yellow \bullet / Green \bullet , respectively. Accuracies in classifying neutral faces (%): InsightFace: (W) 83.0 ± 17.5 (M) 89.8 ± 5.7 , DeepFace: (W) 76.5 ± 14.3 (M) 100.0 ± 0 , FairFace: (W) 99.8 ± 0.2 (M) 100 ± 0 .

certain attributes, like intelligence (65.0% vs 18.5% on SD 2.1-generated images), but persistent poor performance on faces with other attributes, such as happiness (9.7%). Most concerning is InsightFace’s dramatic deterioration on SD 3.5-generated images: its classification performance on “unhappy” and “unattractive” female faces collapses to 12.3% (from 28.5% on SD 2.1-generated images) and 49.2% (from 71.7%), respectively. This deterioration is accompanied by increased variability, with standard deviations reaching 34.5%-37.3%, indicating not only poor but highly unstable performance.

Disaggregated analysis by racial category (detailed in Appendix C.1) reveals that these discriminatory patterns operate through complex intersectional mechanisms on both generated datasets. The drop in performance is the most severe on faces of Black women: InsightFace’s performance for “unhappy” and “unattractive” Black women drops to a catastrophic 1.0% on SD 3.5-generated images (from 9.5% on SD 2.1-generated images) and 10.0% (from 53.0%), respectively. In the case of faces of Asian women, we observe a marked yet less severe deterioration of performance. Interestingly, the performance on faces of White women exhibits more variable patterns.

An analysis of the performance of the gender classification algorithms on neutral faces, *i.e.*, faces generated without any explicit attribute prompts, reveals baseline discriminatory patterns that operate even without explicit attribute cues. On SD 2.1-generated neutral images, significant gender gaps in performance emerge: 6.6 points (W: 80.9% vs M: 87.5%), 48.5 points (W: 51.1% vs M: 99.6%) and 2.1 points (W: 96.8% vs M: 98.9%) for InsightFace, DeepFace and FairFace, respectively. The performance on SD 3.5-generated neutral images shows mixed evolution: FairFace approaches near-perfect equity (W: 99.8% vs M: 100%), DeepFace maintains a substantial performance gap (W: 76.5% vs M: 100%), and InsightFace shows increased variability (W: $83.0 \pm 17.5\%$ vs M: $89.8 \pm 5.7\%$). An intersectional analysis reveals that the gender classification performance on faces of Black women is the lowest in both datasets for InsightFace (39.5% SD 2.1, 58.5% SD 3.5) and DeepFace (32.5% SD 2.1, 58.5% SD 3.5), while FairFace maintains consistently high performance ($\geq 95.5\%$) across all female racial categories. These neutral baselines demonstrate that discriminatory patterns exist independently of the attributes, providing the foundation upon which attribute-based disparities are amplified. Performance levels below 10% in attribute-specific conditions represent what can be termed “algorithmic invisibility”, *i.e.*, the systematic failure to recognize the legitimacy of certain intersectional identity presentations. However, even the neutral face analysis reveals baseline discriminatory patterns, with some groups experiencing significantly lower recognition rates that provide the basis for attribute-based amplification. The systematic nature of these disparities reveals how biases create cascading discrimination effects throughout AI systems. The consistency of bias patterns across independently trained classification models when applied to the same generated datasets suggests the operation of a

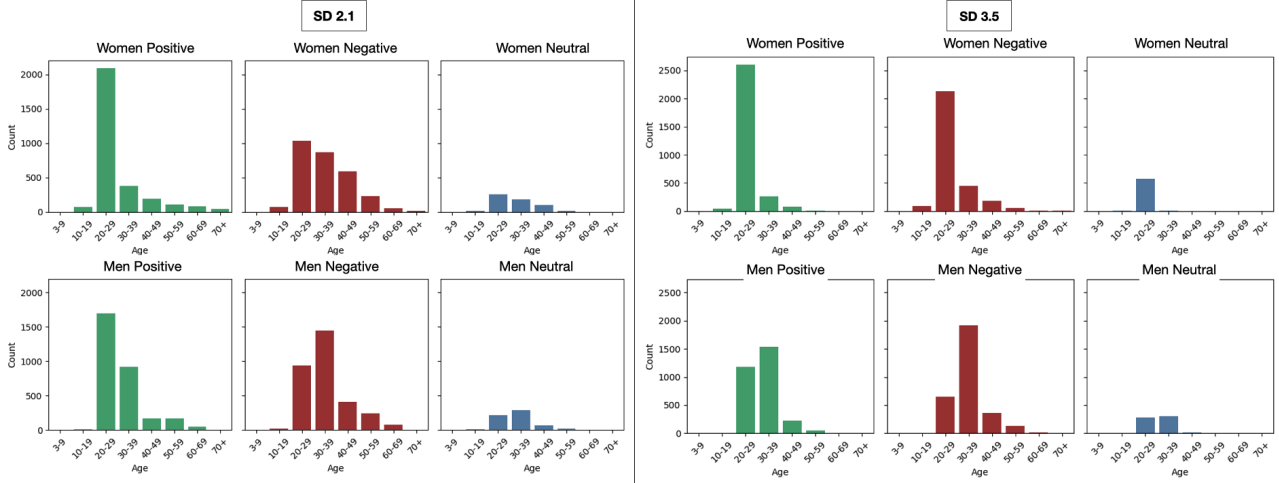


Figure 5.10: Age homogeneity and agism in the generated images: Age distribution of synthetic faces by gender and associated attributes in SD 2.1 and SD 3.5. The images created with positive attributes tend to depict younger individuals than the images created with negative or neutral attributes. Age classification was performed using FairFace. Top/bottom rows correspond to images of women/men, respectively.

coherent normative infrastructure that determines which identity configurations are considered algorithmically legitimate. This infrastructure operates at multiple levels: neutral faces establish baseline hierarchies of recognition, while the attributes amplify these disparities. In the case of faces of females, positive attributes increase the classification performance and negative attributes decrease it, creating systematic barriers to accurate classification. This behavior is reversed in the case of faces of males, where for some attributes of models, the classification performance deteriorates on faces created with positive attributes.

5.7 Qualitative Analysis of the Visual Content in the Generated Faces

Beyond the previously described statistical analyses, we examined the visual content of the generated uncropped images, recognizing that visual representation fundamentally shapes human categorization and judgment. In this section, we report the results of a qualitative assessment of the generated faces. Our analysis reveals systematic patterns that complement the quantitative findings. Most notably, while SD 3.5 produces more convincing and polished images compared to SD 2.1, this apparent technical improvement comes with a troubling trade-off: the newer model demonstrates a pronounced shift toward idealized representations that codify Eurocentric normative aesthetics as the default standard for visibility. This creates a paradox where images become more “realistic” in technical quality while simultaneously becoming less representative of actual human diversity. In particular, we observe the following patterns:

The Progressive Age Homogenization and the Persistence of Algorithmic Agism We analyzed the age distribution of the generated faces using FairFace age classification, as shown in Figure 5.10. The data confirms a systematic shift toward younger appearances in SD 3.5 compared to SD 2.1 for both men and women. Despite SD 2.1’s lower image quality and less photorealistic output, it exhibits greater age diversity across demographic groups. This heterogeneity extends beyond age to include other aesthetic variations, most notably the presence or absence of makeup among women, particularly evident in our neutral baseline generations (Figure 5.11). Furthermore, we observe that positive

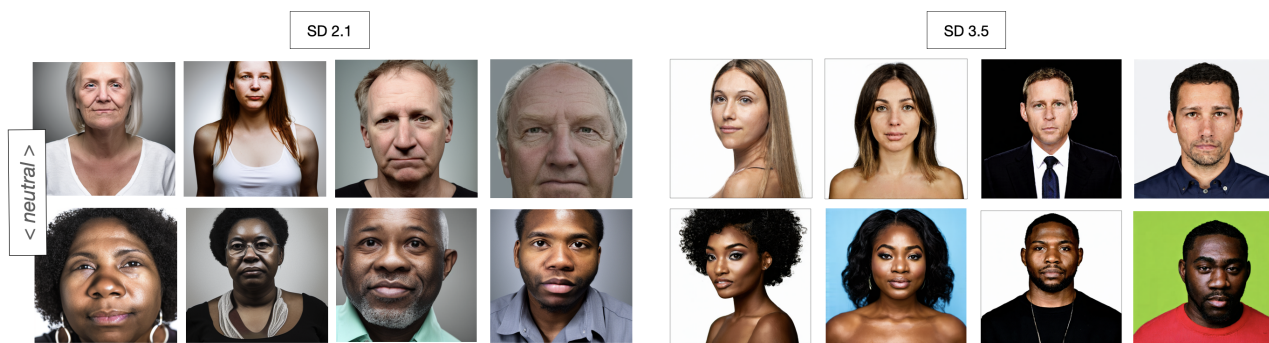


Figure 5.11: Examples of generated faces from SD 2.1 (top) and SD 3.5 (bottom) for the “Neutral” prompt for “White” and “Black” categories, illustrating the shift toward visually younger and more uniform appearances.

attributes generally lead to the creation of images of younger individuals (particularly women) when compared to negative attributes, where there is more diversity in the distribution of ages and the generated faces tend to correspond to older individuals (Figure 5.10, also visible in Figure 5.12). While the official release of SD 3.5 emphasizes its ability to produce “*Diverse Outputs: Creates images representative of the world, not just one type of person, with different skin tones and features, without the need for extensive prompting.*” [302], our findings suggest that, in the specific domain of facial generation, these improvements in prompt adherence and photorealism have been accompanied by a narrowing of aesthetic variation and a reduced representation of age diversity. This trend may be related to changes in the data-curation pipeline introduced in SD 3.5—such as the use of aesthetic ranking systems, semantic deduplication, and NSFW filtering—and potentially also to the inclusion of different or newly sourced images in the training data.

The Happiness-Beauty Conflation One of the most revealing patterns emerges in the representation of attributes—particularly happiness (happy/unhappy). Among all tested attributes, “unhappy” becomes the most visually caricatured, likely because it is the only attribute explicitly tied to an emotional state. In SD 2.1, unhappiness is primarily conveyed through facial expressions: exaggerated frowns, down-turned mouths, and melancholic eyes. In SD 3.5, however, “unhappy” takes on a broader aesthetic meaning: faces are not only sad, but systematically appear older, less groomed, and devoid of makeup—especially among women (see Figure 5.12). This shift suggests that happiness is no longer represented solely as an emotion, but increasingly as an aesthetic ideal. In SD 3.5, to look happy means to look young, polished, and stylized—not merely to smile. In this way, the representation of happiness extends far beyond emotional expression, revealing how affective cues are reinterpreted as markers of social and aesthetic desirability.

The Persistence of Gendered Exposure Across both datasets, women are consistently generated with higher frequency of exposed skin, particularly in the neck, shoulders, and cleavage/décolletage areas. This pattern holds regardless of the specified attribute and persists uniformly from SD 2.1 to SD 3.5, indicating that skin exposure has become a recurring feature of how these models represent femininity (examples are visible in Figure 5.11). This observation is further supported by the analysis of NSFW content presented in the Appendix C.2, which shows that images of women are flagged as NSFW at higher rates than their male counterparts.

Geographic Reductionism in “Asian” Representation The treatment of the “Asian” category exemplifies a form of geographic reductionism that reveals the Western-centric assumptions embedded in these models. Both SD 2.1 and SD 3.5 reproduce exclusively Northeastern Asian traits—essentially

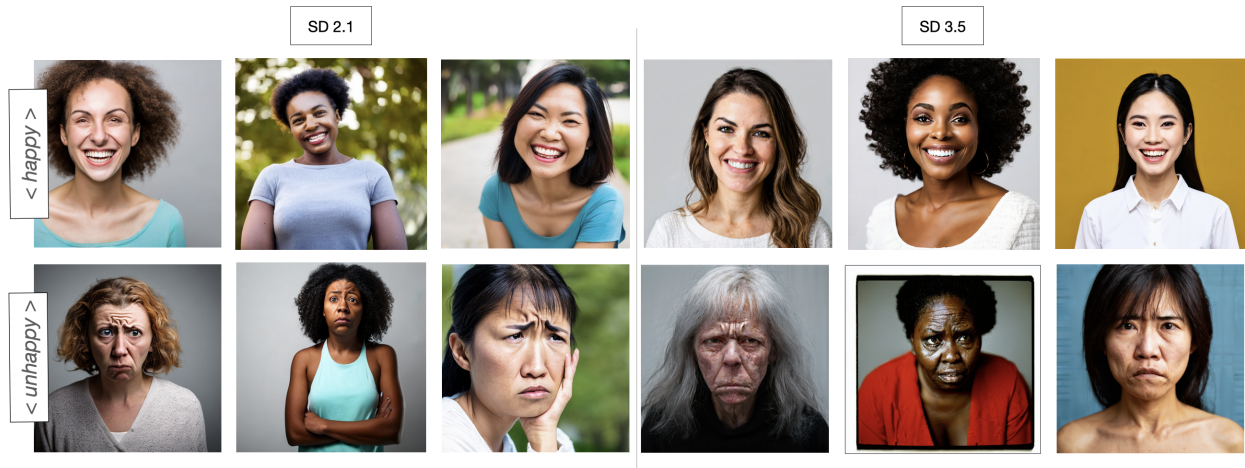


Figure 5.12: Comparison of generated faces for the attributes “happy” (top row) and “unhappy” (bottom row) across Stable Diffusion 2.1 (left) and 3.5 (right) for the female cases.

Chinese, Korean, and Japanese phenotypes—while completely neglecting the vast diversity of the Asian continent. Here, a centralized vision conflates an entire continent with specific regional appearances that reflect hegemonic visual priors likely shaped by Western views, effectively invisibilizing South Asian, Southeast Asian, Central Asian, and Middle Eastern populations.

5.8 The Multiple Dimensions of Algorithmic Harm

From our extensive empirical evaluation, we conclude that diffusion-based T2I models exhibit *algorithmic lookism*. Furthermore, our findings show that the *algorithmic lookism* operates not as an isolated glitch but as a systemic mechanism that discursively shapes both digital representation and computational recognition. This bias reflects and reinforces broader structures of inequality, particularly those shaped by neoliberal rationality, racialized and gendered norms, and platform-driven optimization. Below, we discuss four interrelated implications of our work.

Neoliberal Rationality and the Economics of (In)visibility Our findings reveal systematic lookism across both generative models: faces that are generated with prompts requiring positive attributes (intelligent, trustworthy, happy) consistently align with canonically “attractive” representations, while negative attributes gravitate toward equally canonical “unattractive” ones. This bias intensifies in the case of images of women and becomes most pronounced for Black and Asian women, who experience the strongest aesthetic polarization.

This aesthetic hierarchy cannot be understood outside neoliberal rationality, which, following Wendy Brown [303, 304], operates as a form of *governmentality*³ that not only evaluates, but constitutes and governs subjects through logics of market productivity, appeal, and self-optimization. Here, marketability defines what Sarah Banet-Weiser calls economy of visibility that “increasingly structure not just our mediascapes, but our cultural and economic practices and daily lives” [306]. Under this framework, facial appearance becomes a proxy for individual and social value defined through market demands and capitalization, and visibility itself becomes a commodity. What cannot be optimized

³The term *governmentality*, coined by Michel Foucault in his 1978–79 lectures, refers to “the ensemble formed by institutions, procedures, analyses and reflections, calculations, and tactics” that enable the exercise of power over populations and the subject—where political economy becomes the dominant form of knowledge (Foucault, 1978–79). Derived from the French *gouverner* and *mentalité*, the term underscores how modes of thought structure governance through “the conduct of conduct,” encompassing practices that range from self-governance to the governance of others [305].

aesthetically is gradually rendered invisible, not through deletion, but through erasure by omission, under-representation, and mis-recognition.

From Representation to Recognition: Aesthetic Filtering as Structural Barrier Aesthetic hierarchies embedded in generative models propagate through downstream tasks such as gender classification. Our analysis reveals that female faces that have been generated with negative attributes are systematically misclassified. This demonstrates how aesthetic norms affect not only depiction but algorithmic recognition itself. Aesthetic legibility becomes a precondition for computational recognition. Faces that deviate from normative ideals become less intelligible to classification systems: the less *desirable* the face, the more likely it is to be misread or erased. For marginalized subjects, this mis-recognition compounds representational harms with computational exclusion. As Sarah Banet-Weiser [307] argues, visibility under neoliberalism requires being seen *correctly* through dominant visual codes. Generative systems encode and enforce these norms, creating hierarchies of legibility that downstream systems inherit.

Aesthetic Entrepreneurship and the Privatization of Harm The evolution from SD 2.1 to SD 3.5 exemplifies how technical advancement intensifies aesthetic constraints. While SD 3.5 produces more convincing images, even neutral prompts yield beautified faces, particularly for women. What appears as increased realism is actually a tightening of representational boundaries, where recognition is granted only to those aligning with desirable optimized and profitable ideals. This dynamic resonates with work by Elias et al. [114] on aesthetic labour and its logic of gendered entrepreneurship under which women are encouraged to treat their bodies as ongoing projects of self-transformation and branding. In generative AI contexts, this imperative becomes automated and systems reward conformity to learned attractiveness norms while penalizing deviation. Corporate platforms frame outputs as neutral while those falling outside algorithmic standards become anomalous, less recognizable, misclassified, or invisible. Meanwhile, platforms profit from unpaid data labor and attention economies, transforming aesthetic compliance into computational capital. As Gill [233] argues, apparently progressive media environments often reproduce (and profit from) gendered hierarchies through logics of sexualization, desirability, and objectification. AI systems inherit these logics, encoding them into visual defaults that dictate not only what can be seen, but how it must appear to be recognized correctly. *Algorithmic lookism* reflects not system failure, but success in reproducing dominant ideologies through machine vision—creating infrastructural harm built into how platforms govern digital aesthetics and mediate social legibility.

Orientalist Regimes and Geographic Reductionism Our analysis reveals that both generative models reduce the representation of “Asian” faces to exclusively Northeastern Asian features (Chinese, Korean, and Japanese phenotypes) while completely failing to represent South Asian, Southeast Asian, Central Asian, and Middle Eastern populations. This can be understood as a computationally encoded, aestheticized iteration of a visually reproduced orientalist logic ([308])⁴ rendered through algorithmic selectivity rather than textual discourse, flattening and erasing complexity while reinforcing hierarchies of legibility aligned with Western, neoliberal demands. This geographic reductionism reflects what Quadri et al. [283] term algorithmic Orientalism [308]. Drawing on [309], they show how AI systems perpetuate colonial stereotypes that flatten diverse regions into simplified, Western-legible categories. The “Orient” becomes a singular representation that obscures vast cultural and geographic diversity. Our findings demonstrate the same dynamic: “Asian” identity gets conflated

⁴Through his conception of *Orientalism*, Edward Said outlined how the West has historically constructed ideals of the “Orient” as a homogeneous, exotic, and aestheticized “other” under terms it could recognize and control. This also materialized through the distortion and erasure of complexity of the Asian continent. SD 2.1 and 3.5 in this sense, replicate the epistemological logic of Orientalism through selective visibility, stylization, and erasure.

with East Asian appearance, invisibilizing entire populations across the continent. This represents another form of systematic exclusion embedded in generative AI—one that transforms historical colonial ways of seeing into computational defaults that determine which communities become visible and recognized.

5.9 Conclusion

In this chapter, we are one of the first to investigate the existence of *algorithmic lookism* in text-to-image generative diffusion-based models. By means of an empirical evaluation of 26,400 synthetically generated face images, we find that *algorithmic lookism* is not a sporadic glitch but a structural feature of generative AI systems. An empirical comparison between Stable Diffusion 2.1 and 3.5 reveals that positive traits (*e.g.*, “happy,” “intelligent,” “trustworthy”) are systematically mapped onto normatively attractive faces, while negative traits align with less conventionally attractive ones. These patterns disproportionately affect women—particularly Black and Asian women—who are more strongly aestheticized and more frequently misclassified by downstream gender classifiers. Aesthetic bias thus propagates across the AI pipeline, from generation to recognition.

These harms must be understood within broader systems of power. As Stuart Hall [310] argues, representation is never neutral, but part of a regime of meaning shaped by cultural and ideological forces. Generative AI participates in this regime by encoding normative visual codes that determine which bodies become visible, and how. In this sense, algorithmic lookism is not merely about appearance—it governs legibility, recognition, and social worth.

This dynamic is exacerbated in multimodal systems, where textual and visual biases intertwine. As Birhane [97] warns, combining modalities risks entrenching stagnant stereotypes, especially when visual meaning lacks an explicit vocabulary and is inferred from culturally loaded prompts. In text-to-image generation, “realism” often becomes a proxy for conformity to dominant norms.

At the same time, several limitations must be acknowledged. Our use of “attractive” and “unattractive” prompts presumes a shared standard of beauty without human validation, reducing a culturally variable and multidimensional concept to binary labels. Furthermore, the demographic framing is limited to three racial categories and binary gender, leaving out mixed-race, non-binary, and other marginalized identities whose experiences may diverge from our findings.

Ultimately, as “Past Life Grid” art project demonstrated, these systems do not simply reflect cultural hierarchies, they automate and institutionalize them. What appears as aesthetic optimization is, in fact, a political economy of representation: social norms become default outputs; divergence is penalized through invisibility or mis-recognition. Understanding this requires moving beyond performance metrics to interrogate the politics of visibility embedded in AI infrastructures.

If generative models are to serve public interests, they must be examined not only for what they depict, but for whom they fail to represent and at what cost. We hope that our research will spur a more critical and inclusive approach to dataset curation, model evaluation and deployment practices; an approach that actively addresses neglected representational harms.



Chapter 6

Who Gets to Define: Mixed-Race Identity as Methodological Critique

Warning: This chapter uses the term “race” as it appears in standard ML datasets and research, primarily derived from U.S. contexts [238]. We acknowledge this as a social construct rather than a biological reality, and recognize that different academic and policy contexts may prefer alternative terminologies such as “ethnicity” [239]. We retain “race” terminology to critically examine how these socially constructed categories are encoded in AI systems, without seeking to promote or legitimize such classifications.

This chapter uses mixed-race identity as a methodological critique of fixed **racial categories in ML**. Through analysis of datasets, generative AI outputs, and fairness frameworks, it demonstrates how rigid taxonomies—primarily derived from U.S. census classifications—systematically exclude mixed-race individuals while reifying race as biological rather than social. The investigation reveals how AI systems enforce a logic of external *representation* rather than community-driven *presentation*, rendering complex identities either invisible or illegible. Drawing on standpoint epistemology and the lived experience of categorical in-betweenness, the chapter advocates for methodological sovereignty in AI fairness research—shifting power from those who design categorical systems to communities who must live within them, proposing participatory approaches that capture discrimination without reinforcing racial essentialism.

6.1 Anatomies of Belonging: From McWatt to Machine Learning

“what are you?” [311]

This question, repeatedly directed at writer Tessa McWatt throughout her life and examined in her book *Shame on Me: An Anatomy of Race and Belonging* [311], captures a fundamental tension that extends far beyond personal identity into how we construct knowledge about race (and gender) in Artificial Intelligence. McWatt, with her complex multi-ethnic heritage¹, experienced this interrogation as both a deep intrusion and categorical demand—a requirement to fit within rigid frameworks that could never fully contain her reality. The previous chapters have documented how AI systems systematically determine digital visibility—from beauty filters that enforce aesthetic hierarchies to generative systems that amplify lookism. These technologies operate through what can be understood as *representation*: external systems defining and categorizing identities according to predetermined frameworks. But McWatt’s approach suggests a different possibility: *presentation*—the right of individuals and communities to define themselves on their own terms.

¹McWatt’s heritage is Scottish, English, French, Portuguese, Indian, Amerindian, African and Chinese.

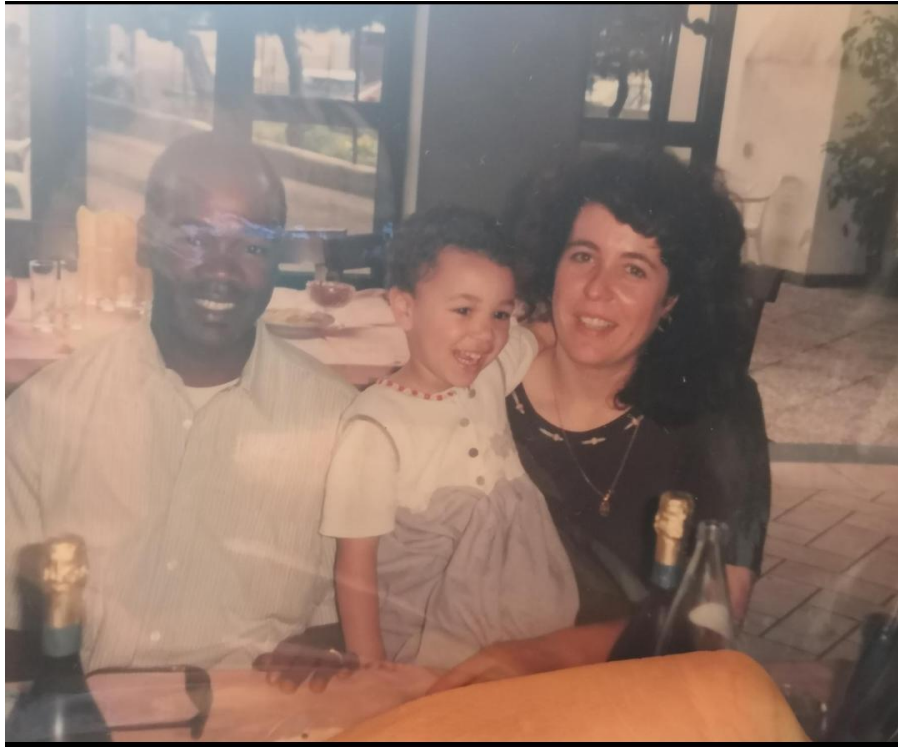


Figure 6.1: *In-betweenness* embodied: a family photo showing the author with her parents (Ivorian and Italian / Black African and White European), revealing the lived reality that exists beyond algorithmic systems’ binary classifications.

In the book, McWatt takes a distinctive approach: she breaks down her body—nose, lips, eyes, skin, bones, blood—and uses each part to reflect on how society assigns racial meaning to physical features. This anatomical framing challenges the idea that identity can be reduced to fixed categories. Her work shows that the question “What are you?” is less about genuine curiosity and more about placing people within existing systems of inclusion and exclusion. More importantly, it demonstrates what happens when individuals reclaim the power to answer that question for themselves.

The question “What are you?” has become computational, embedded in the datasets and fairness frameworks used in AI research. Visual technologies often create what has been described as “multiply exposed” experiences—where some tools fail to recognize Black people, while others render them hypervisible and subject to racial surveillance [4]. For mixed-race individuals, this exposure becomes even more complex: we are often invisible in standard categories, yet hypervisible as exceptions that disrupt the logic of racial classification.

The concept of *scopic vulnerability*—used by Benjamin [4] and informed by Fanon’s [312] analysis of the racializing gaze—describes the experience of being looked at without being truly seen. This framework helps explain how mixed-race individuals are positioned within AI systems, which often misrecognize us, reducing complex identities to rigid racial categories. The result is a double bind: we are rendered hypervisible as algorithmic anomalies, yet erased within the very systems that define “recognizable” identity. Our existence in the interstices of these categories exposes the arbitrariness of the taxonomies themselves.

The standard racial categories used in machine learning research—derived predominantly from U.S. census classifications—render mixed-race individuals either invisible (when forced to choose a single category) or illegible (when subsumed under catch-all labels like “Other” or “Mixed”). This represents the same logic of external representation that governs aesthetic algorithms: someone else deciding who you are and how you should be categorized.

This chapter emerges from a personal discomfort that has shadowed the author’s work in algorithmic fairness—including the research presented elsewhere in this dissertation (Chapter 4 and Chapter

5). Even while engaging in demographic auditing and bias measurement, the author has encountered the inadequacy of the categorical frameworks available to describe her own experience and that of her community. But rather than simply calling for better categories, this chapter asks a more fundamental question: What would it mean to shift from representation to presentation in AI fairness research?

Drawing on theories of standpoint epistemology [88, 313, 3] and participatory AI [314, 315, 316], this chapter uses mixed-race identity as a methodological lens to expose the inadequacy of fixed categorical frameworks. The goal is not inclusion within existing structures, but methodological sovereignty [317, 318]: the right of marginalized communities to determine how they are studied, categorized, and included within technological systems.

This chapter critiques fixed racial categories not merely as technically inadequate, but as fundamentally misaligned with how identities are actually lived and experienced. By centering the voices and experiences of those who have been systematically excluded from current frameworks, it aims to restart the conversation about fairness—shifting power from those who design categorical systems to those who must live within them. As McWatt writes, the body becomes “a battlefield of belonging.” This chapter is about reclaiming that battlefield—about giving back power to communities to define themselves and challenge the epistemological assumptions that shape how AI systems see and categorize human difference.

6.2 The Categorization of Race in ML is a Flawed Premise

“A race is only a sort of average of a large number of individuals; and averages differ from one another much less than individuals. Popular impression exaggerates the differences, accurate measurements reduce them” [319]

The concept of separate human races arose in the 17th and 18th centuries and was used by Westerners to justify slavery despite their Christian faith [320]. This notion persisted into the 20th century and was closely tied to early statistics and eugenics, with pioneers like Galton, Pearson, and Fisher reinforcing the idea of biological racial separability. However, anthropologists have increasingly contested these racial distinctions, recognizing them as social constructs rather than biologically discrete entities [319]. The “one-drop rule” in the U.S. and the Apartheid regime’s classification system in South Africa illustrate how racial categories have been historically fluid and politically motivated [9].

For decades, scholars across disciplines have emphasized that racial classifications are neither genetically discrete [321] nor can they be reliably measured or considered scientifically meaningful [320]. Instead, race is understood as a construct whose meaning is constituted by social arrangements, practices, and intersubjective beliefs [322]. This understanding implies that there is no “correct” racial taxonomy derived from biology, as perceptions of race depend on both phenotypic traits and contextual interpretations.

Despite this knowledge, the Machine Learning (ML) research community has often relied on datasets that label race as a categorical variable, treating it as a “ground-truth” that simplifies its social and historical complexity [323]. Racial labels such as “White,” “Black,” or “Asian” are indeed widely used in image [324, 325, 326, 30] and tabular [327, 328] datasets, often adopting U.S. census-based racial classifications without contextualization, reinforcing North American constructs in global ML research.

By using these categorizations, ML models fail to account for the social, cultural and historical contexts that shape racial identities. Even when these labels are well-intended and used for fairness audits [329, 330, 331], they reflect outdated racial theories. This lack of critique has real-world consequences: from biased hiring algorithms to unequal healthcare diagnostics, these systems risk exacerbating disparities rather than addressing them. To move forward, we contend that the ML community should examine how racial data is collected, labeled and used, avoiding the instrumentalization of ethical concepts, to the point that “bias becomes a form of numerical error to be corrected

with better datasets, and ethics becomes a bureaucratic checklist to be inserted into the production flowchart” [332].

In this chapter, we discuss the challenges and ethical concerns associated with the use of categorical race labels in ML systems. Thus, discussing the benefits for individuals to freely define their identities through racial categories [333, 334] is out of its scope. Instead, we critically analyze ML practices, datasets, and the socio-historical construction of race (Section 6.3), emphasizing the dominance of U.S.-centric frameworks and their limitations from the perspective of Europe (Section 6.4) and the “mixed race” community (Section 6.5). In addition, we address the broader issue of racial reification and stereotyping in ML (Section 6.6). Finally, we propose a research agenda that avoids categorical race labels by integrating context and domain knowledge to more effectively tackle discrimination and social inequality (Section 6.7).

6.3 Related Work

Several studies have criticized the use of fixed racial categories in AI from different perspectives. First, the process of defining racial taxonomies has been questioned, calling for more careful and contextual choices, justification, and documentation [335, 336, 337, 323, 338]. Second, the use of racial categories in assessing fairness and discrimination in ML has been challenged, calling for the use of different categories [339, 340, 341, 342]. In contrast, we argue that methods aimed at measuring discrimination should go beyond using a categorical race variable. Recent work has presented alternatives to categorical race labels both in vision and tabular data. In computer vision, the use of visual cues instead of race has been proposed [31], almost exclusively focusing on skin tone [343, 109, 344, 345]. We argue that other phenotypic features should be considered while avoiding “racial phenotypes” [346, 347], which run the risk of reifying race along phenotypic lines [340, 335]. In the case of tabular data, the most closely related work to ours concerns subgroup fairness [348] and multi-calibration [349]. However, the former approach still relies on predefined race categories while the latter solely aims at general predictive quality without incorporating context or distinguishing between features.

6.4 U.S. centrism vs the European perspective

Race is often invoked as a protected attribute in the context of algorithmic fairness in ML, building on U.S. discrimination doctrines of disparate treatment or impact [36]. This area has been shaped by seminal works, such as the 2016 ProPublica analysis of the COMPAS recidivism prediction algorithm, reporting higher error rates for black than for white inmates [328]. For the past decade, numerous authors have argued for fairness metrics defined in terms of comparisons between socially relevant groups, such as equal error rates or equal outcomes, and ways to optimize for them [350]. In these works, race is typically considered a categorical attribute as any other, often adopting the taxonomy of the U.S. census, as we discuss below. Race has also been extensively used in computer vision, where many datasets with human faces contain race labels that are used to compare error rates across different racial groups [351], provide a measure of dataset diversity [30], or as target variables for models serving a variety of purposes, including “security and defense, surveillance, human computer interface (HCI), biometric-based identification” [352].

Recent studies highlight how fairness research continues to rely on predefined racial classifications without critical engagement. Abdu et al. analyzed 60 ACM FAccT (2018-2020) papers [323], showing that most use racial categories without justification, primarily adopting labels from pre-existing datasets. Extending this analysis, we reviewed 78 ICML and CVPR papers from 2023 and 2024, selecting those with “fair” in the title for ICML and “fair”/“bias” for CVPR, as the latter yielded too few results when searching only for “fair” (Table 6.1).

Among these 78 papers, 45 explicitly discussed group fairness. Within this subset, 29 relied on racial categories as a protected attribute, all adopting rigid racial classifications inherited from existing datasets—confirming the persistence of these taxonomies in fairness research [323]. This reliance suggests a continued adoption of established classification schemes without a critical reassessment of their validity. To further investigate this trend, we examined the most widely used ML datasets based on citation counts from Google Scholar and their recurrence in our chapter analysis. Our findings confirm that these datasets overwhelmingly adopt racial classifications derived from U.S. census categories (Appendix D.1), a pattern also highlighted in [323]. Even when explicit census labels are absent, the underlying racial categories often reflect U.S.-centric taxonomies, reinforcing rigid classification structures. Notably, none of these datasets provide critical reflection of their contextual meaning or applicability beyond the United States [339, 337, 338] (see table 6.2 and Appendix D.2).

Table 6.1: List of race-related datasets used in fairness research, indicating the papers and their respective conferences (CVPR = * and ICML = *). (2023-2024)

Race Dataset	Papers Using the Dataset * = ICML, * = CVPR (2023-2024)
FairFace [30]	[[353], [354], [355], [356]]*
UTKFace [326]	[[354], [357]]*; [[358]]*:
COMPAS [359]	[[360], [361], [362], [363], [364], [365], [366], [367], [368]]*
Census (Adult / Folktables) [369]	[[370], [371], [372], [362],[373],[374],[375]]*
MINNEAPOLIS ²	[[360]] *
FASSEG [376]	[[377]]*
HSLs [378]	[[368]]*
ENEM [379]	[[368]]*
Law School [380]	[[364],[381],[382]] *
Communities & Crime [380]	[[370],[362],[381],[383],[382], [368]] *
Toxic Comments [384]	[[385]] *

Despite the cultural vicinity, the legal and institutional framework in Europe contrasts sharply with the U.S. approach, where racial categories are routinely collected and integrated into governance, research, and AI datasets.

Mainly as a result of post-World War II efforts to combat racialization and discrimination, race is widely rejected as a legitimate classification category in Europe [386, 387, 388]. Also at the legislative level, the European Union has institutionalized its avoidance of racial classification. For instance, the General Data Protection Regulation (GDPR) [389] restricts the collection of racial or ethnic data, permitting it only under narrowly defined conditions, such as explicit consent or when serving a significant public interest. In practice, many European states remain hesitant to engage with race as an analytical category, relying instead on indirect indicators, such as nationality, language spoken at home, or parental country of birth [390, 391, 392, 393, 394]. Indeed, the Court of Justice of the European Union ruled that ‘ethnic origin cannot be determined on the basis of a single criterion’ and judicial decisions on racial or ethnic discrimination need to consider multiple factors [395]. Fur-

thermore, different member states have different standards on which attributes are relevant for racial discrimination [396], highlighting the importance of local context.

As AI systems become increasingly regulated under frameworks such as the EU AI Act [397], with explicit requirements regarding algorithmic fairness and non-discrimination, the direct adoption of U.S.-based frameworks raises significant questions [398, 399]. However, even if the institutional approach to dealing with racial categories is more permissive in the United States, criticism of such taxonomies has not been exclusive to Europe. Even in the U.S., racial beliefs have long been said to “constitute myths about the diversity in the human species and about the abilities and behavior of people homogenized into “racial” categories” [400].

In line with this, we argue that racial categories should be abandoned in ML research whenever possible, incorporating context and domain knowledge instead to better address discrimination. In the following sections, we support this position by identifying two problems which highlight why racial categorization in ML can be both conceptually flawed and harmful. We start by demonstrating that ML frameworks fail to represent mixed-race identities and enforce reductive classifications.

Table 6.2: Race categories utilized in popular Machine Learning datasets. More information in Appendix D.2.

Dataset	#Categories	U.S. Census subset	Note
COLORFERET	4	✓	
MORPH	4	✓	
UTK FACE	5	✓	<i>Indian separated from Asian</i>
FAIRFACE	7	✓	<i>White & Asian subdivided, Indian separated</i>
ADULT / FOLKTABLES	5/9	✓	
COMPAS	6	✓	
COMMUNITIES & CRIME	4	✓	
LSAC LAW SCHOOL	8	×	<i>plus Mexican American & Puerto Rican</i>

6.5 The Mixed-race Problem in ML

“Is belonging something given to you or something you create? [...] I am a kind of middle ground in the current language of belonging: mixed race, middle class, dual nationality. I embody both privilege and oppression.” [311]

Mixed-race individuals, defined as those with parents from different racial backgrounds or who belong to multiple racial groups [401], embody the complexity and fluidity of racial identities.

These individuals navigate multiple cultural and racial contexts, resisting singular classification and exposing the limitations of existing taxonomies [402, 403, 311]. Their dual positioning—simultaneously within and beyond established categories—makes them emblematic of identities that defy categorical frameworks [404]. At the heart of this experience lies the concept of “in-betweenness”, which highlights both the richness of multiplicity and the challenges of fragmented affiliations [405, 406]. This liminal space often amplifies marginalization, reflecting how societal and institutional structures struggle to accommodate fluid identities [403, 404, 407].

A fundamental question emerges: *how should mixed-race identities be classified within categorical race taxonomies in AI?* There are arguably five possible ways to handle mixed-race categories in categorical race taxonomies, which we will discuss as approaches (A) to (E) below, with (A) to (C) being used in practice [323].

- **Approach (A)** assumes that each person only belongs to one race, hence ignoring the reality of mixed-race individuals and invalidating their identity by forcing them to claim only one aspect of it [408, 409]. This approach is largely driven by two factors: (1) the preference for mutually exclusive racial classification schemes, which streamline computational modeling at the expense of identity fluidity, and (2) the push for computational efficiency, where minimizing the number of racial categories simplifies both the analysis and fairness auditing [323, 338]. However, these practical considerations come at the cost of inclusivity and representational accuracy, ultimately reinforcing reductive racial frameworks rather than challenging them. This seems to be the most common approach in ML and especially in computer vision, as exemplified by the FairFace [30] dataset. Marketed as a “fair” dataset, it was designed to mitigate the racial imbalances in existing face datasets by introducing seven racial groups. The authors explicitly frame their motivation as follows: “*Existing public face datasets are strongly biased toward Caucasian faces, and other races (e.g, Latino) are significantly underrepresented.*” This illustrates that under the assumption of rigid racial taxonomies, even well-intended approaches can become highly exclusive.
- **Approach (B)** includes a single “Mixed-race” category, as it is the case of the American Community Survey (ACS) Public Use Microdata Sample (PUMS) data [410], from which the `folktables` dataset [369] is derived.

This choice may seem to be an efficient solution given that, in the U.S. for instance, many mixed-race Black/White individuals identify as Black (and are treated as such) since the early 20th century [411], and hence the mixed-race category only comprised 2.9% of U.S. census respondents as of 2010. In 2020, however, this percentage increased to 10.2%, presumably “largely due to the improvements to the design [and processing]” of the survey [412].³ However, this approach is also highly reductive: Why should, for instance, “Asian-White” be grouped together with “Black-Hispanic”?

We illustrate this limitation with an example in computer vision, using embeddings extracted with CLIP ViT-B/32 [414] from the Chicago Face Database (CFD) and its extension CFD-MR [415, 416]. CFD consists of images of 597 unique individuals with their self-reported race (e.g., Asian, Black, Latino, and White), and CFD-MR includes images of 88 unique individuals who self-reported multi-racial ancestry (Mixed-Race). We conduct an experiment comparing the images of faces in the Asian, Black, White and Mixed-Race groups, ensuring balance across gender and sample size. As shown in Figure 6.2, we observe significant overlap between Mixed-Race samples (shown as magenta dots in the right-hand side of the Figure) and other racial groups.

This blurring of the boundaries between groups in the case of mixed-race individuals further illustrates the limitation of relying on rigid racial taxonomies.

We acknowledge that CFD-MR’s small sample size limits statistical robustness; however, to our knowledge, it is the only publicly available dataset including self-reported mixed-race identity, enabling analysis of mixed-race complexity without imposing external labels. The scarcity of larger, self-reported mixed-race datasets highlights a gap in available data and a shortfall in current collection practices—underscoring the urgent need for more nuanced, self-defined datasets.

- **Approach (C)** subsumes mixed-race under “Other”, as exemplified by [370]. This option not only inherits the limitations of using a single “Mixed-Race” category but even exacerbates them by adding to all possible variations of mixed-race individuals any other individuals that do not fit into the imposed categories.

³This dependence on framing as well as the fact that indicating multiple races was only possible after long negotiations in the 90s [413] exemplifies the volatility of race taxonomies.

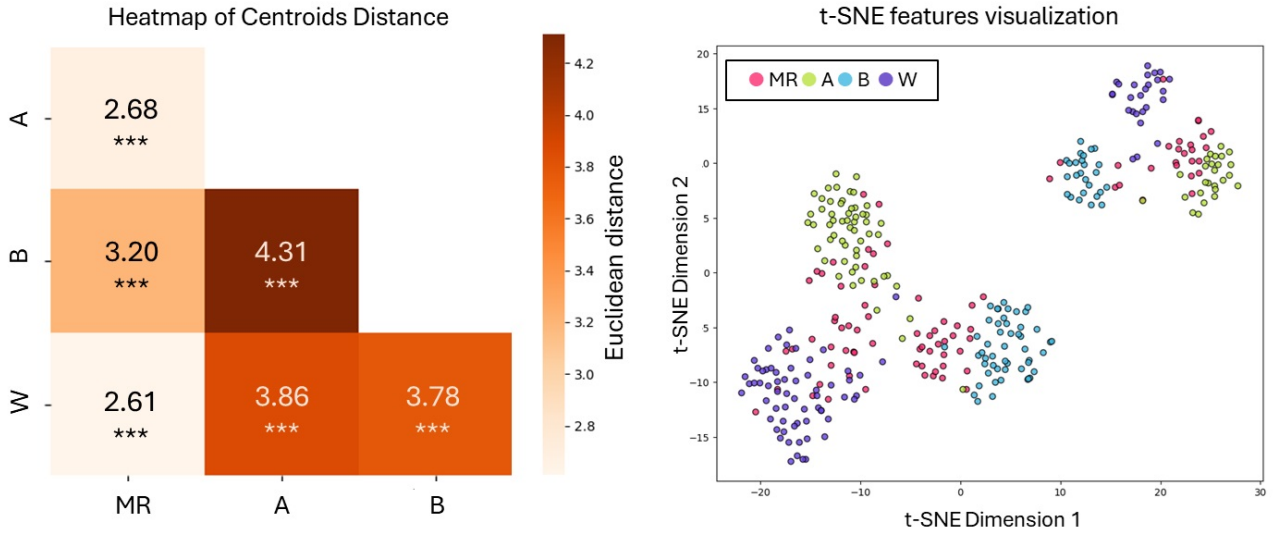


Figure 6.2: Embedding visualization for four racial groups from the Chicago Face Dataset and its extension, CFD-MR: Mixed-Race (MR), Asian (A), Black (B), and White (W). Left: heatmap of the Euclidean distances between group centroids, with significance: *** ($p < 0.001$), ** ($p < 0.01$), * ($p < 0.05$). Right: t-SNE plot of the embeddings of each individual image. Note how Mixed-Race individuals (depicted in magenta) occupy an intermediate space between other groups.

In addition to these three approaches, we describe below two potential approaches that have not been pursued yet in the ML literature.

- **Approach (D)** accounts for all mixed categories separately, *e.g.*, “Black/White”, “Black/Hispanic”, etc. However, this approach does not seem advisable for two reasons. First, it would lead to a combinatorial explosion, as for k categories, this would lead to $2^k - 1$ possible labels. Second, the categorical nature of the variable would mean that, for instance, “Black/Hispanic” would stand in the same relationship to “Black” as to “White”, which is clearly sub-optimal.
- **Approach (E)** proposes to allow individuals to have multiple racial labels instead of being forced into a single category. While this better reflects the reality of mixed-race identities, it is not used in practice due to both data availability issues and the technical complexity of multi-label analysis. For instance, while the U.S. Census initially records race data in this way, it is later simplified into single-label categories in the ACS PUMS database, and hence also `folktables` (approach (B)). Even if such data were available, implementing this approach in ML would require new methods to handle overlapping groups, which most fairness frameworks do not support.

Rather than attempting to develop new technical solutions for such categorical approaches, we argue that race labels should be avoided altogether. We provide further justification for this position in the following section.

6.6 From Reifying to Stereotyping Race in AI

The concept of reification refers to the process of transforming abstract ideas into concrete, seemingly natural entities [417]. In the context of AI, racial reification occurs when socially constructed racial categories are encoded as computer-readable attributes and, thus, appear to be measurable and biologically real.

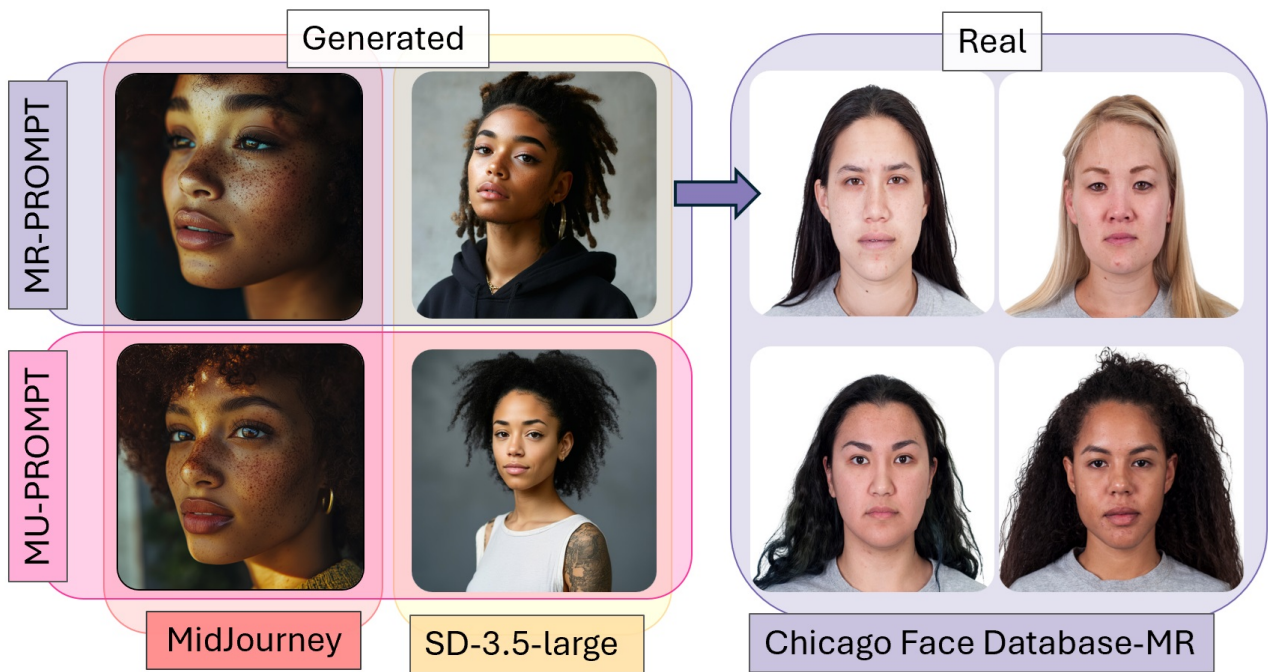


Figure 6.3: Comparison of AI-generated and real-world images of mixed-race individuals. The generated images come from two AI models, MidJourney and Stable Diffusion 3.5 Large, using different prompts: “A mulatto person” (MU-PROMPT) and “a mixed-race person” (MR-PROMPT). The real-world images are sourced from the Chicago Face Database-MR, showcasing greater phenotypic diversity.

This process not only replicates racial classification systems but entrenches race within digital infrastructures. It thereby reinforces the false notion that racial groups are stable, objective, and biologically determined rather than historically contingent and socially constructed [29, 340, 337, 331]. These mechanisms, deeply embedded in machine learning systems, mirror historical physiognomic classification, where racial typologies were used to categorize, rank, and regulate human difference [418, 4, 419, 420].

A key issue in AI’s reification of race lies in how image datasets obtain the labels for racial categories. Many widely used datasets, such as FairFace [30], rely on manual annotation, where individuals classify faces into racial categories without any contextual information. These categorizations, often based on a subjective visual assessment, do not account for the fluidity of racial identity, regional variations, or self-identification, yet they serve as the foundation for how ML models “learn” to recognize or consider race [337]. In addition, these datasets impose a singular, racialized gaze that fixes the representation of marginalized identities into stereotypical, externally imposed frameworks. A clear example of this can be found in ImageNet [26], where the category “Black person” was represented in part by images of individuals performing *blackface*, illustrating how racialized bodies have historically been captured through a *white gaze* that dissects, categorizes, and fixes them in structures of power, rather than allowing for self-representation [29, 335].

While ML -based classification systems reinforce racial categories on a large scale, generative systems for images, videos, and even audio push this further by enabling the creation of content tied to racial classifications [110, 421, 422, 423]. Unlike static datasets which have a finite number of datapoints, Generative AI models can synthesize entirely new datasets, reinforcing and expanding racial stereotypes with no human verification or historical grounding.

To illustrate this phenomenon, we conducted an experiment with Stable Diffusion 3.5 Large [424], Midjourney [425], and DALL·E 3 [426], using the prompts “a mixed-race person” (MR-PROMPT)

and “a mulatto⁴ person” (MU-PROMPT). The inclusion of the term mulatto in this study was intentional, as it historically frames mixed-race identity within a rigid “Black”/“White” binary [427]. We did not include prompts such as “Asian-White,” which lack an equally established cultural reference and could introduce arbitrary assumptions about how these identities should appear.

Because Generative AI systems are trained on datasets reflecting historical and cultural biases, we hypothesized that they might replicate or reinforce such reductive categorizations. Consequently, while our conclusions are necessarily limited to the specific historical stereotypes associated with these terms, this focus ensures that our findings about the replication of pre-existing biases remain clearly interpretable.

To empirically validate this hypothesis, we then generated images using each model under both prompts. Specifically, 30 images per prompt were generated for both Stable Diffusion and Midjourney. Note that we were unable to create images with DALL·E 3 due to prompt moderation on the term “mulatto”. Interestingly, both models produced nearly identical images across these prompts: individuals with medium-brown skin, curly or afro-textured hair, and phenotypic traits associated with a “Black”/“White” racial mix. Figure 6.3, left, shows an example of the generated images for each prompt and model. From a quantitative perspective, the cosine similarity between the embeddings (extracted using CLIP ViT-B/32) of the two sets of images was 0.8190 for Stable Diffusion and 0.7687 for MidJourney. This suggests that both models interpreted the MR-PROMPT and MU-PROMPT similarly, thus creating similar images for both prompts. Conversely, natural images exhibit a much wider range of phenotypic diversity than the images generated with AI models. For example, the images in the CFD-MR [416] showcase greater diversity under the “Mixed Race” category, as illustrated in Figure 6.3. From a quantitative perspective, the mean cosine similarity of embeddings (also extracted using CLIP ViT-B/32) within the CFD-MR dataset is 0.7883, whereas the generated images with the MR-PROMPT exhibit larger internal similarity (0.8694 for Stable Diffusion and 0.8265 for MidJourney). Despite the small sample size, our analysis of generative AI outputs provides key insights into how these models replicate and reinforce existing racial stereotypes.

Given the apparent limitations of existing frameworks to address race in ML, we present next an alternative approach that aims to more effectively account for racial dynamics and mitigate biases in ML systems.

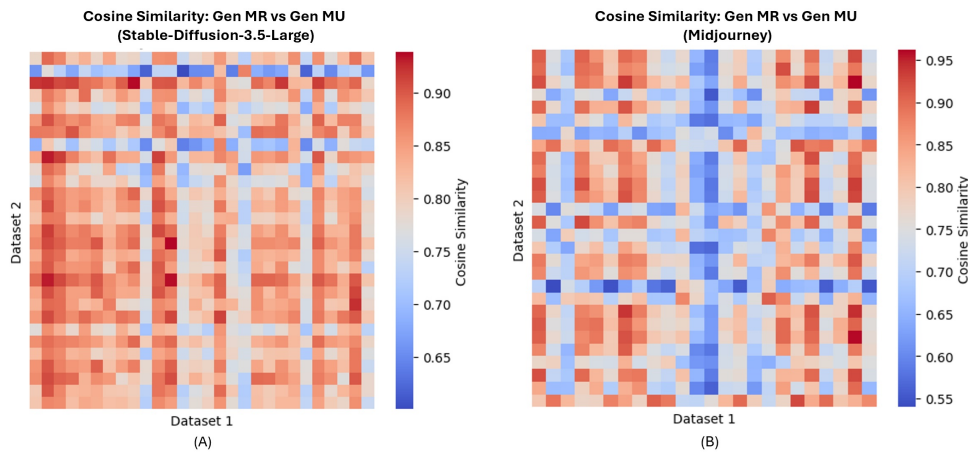


Figure 6.4: Heatmap of cosine similarity between images generated using the prompts “mixed-race person” (MR-PROMPT) and “mulatto person” (MU-PROMPT). On the left (A), results from Stable Diffusion 3.5-Large; on the right (B), results from MidJourney.

⁴**Disclaimer:** The term *mulatto* is considered offensive in contemporary usage. It is used here solely for historical and analytical purposes to critically examine biases in Generative AI representations.

Table 6.3: Cosine similarity distribution between images generated for “a mixed-race person” (MR-PROMPT) and “a mulatto person” (MU-PROMPT) in Stable Diffusion and MidJourney.

Model	Mean	Median	75th Percentile	90th Percentile
Stable Diffusion	0.8190	0.8253	0.8618	0.8849
MidJourney	0.7687	0.7732	0.8228	0.8975

Table 6.4: Cosine similarity statistics for real and AI-generated mixed-race images. The generated images (SD = Stable Diffusion-3.5-Large and MDJ = MidJourney) exhibit higher internal similarity compared to real-world images from CFD-MR, indicating a lower degree of phenotypic diversity.

Dataset	Mean	Median	75th Percentile	90th Percentile
Real Mixed-Race (CFD-MR)	0.7883	0.7911	0.8489	0.8831
Generated Mixed-Race (SD)	0.8694	0.8770	0.9050	0.9272
Generated Mixed-Race (MDJ)	0.8265	0.8388	0.8933	0.9209

6.7 Moving Beyond Race Labels in ML

In response to the concerns raised in this paper and more generally the constructivist understanding of race [320], we propose that ML research should abandon racial taxonomies where possible and instead focus on developing alternative methods that address identity-based disparities without reifying race as a biological category.

The challenge to tackle racial discrimination without fixed racial categories has been addressed in multiple disciplines. There is a growing shift towards fine-grained, context-sensitive approaches that challenge rigid racial classifications in a variety of fields, including biology [428], economics [429], public health [430], social psychology [431], and the law [432]. This perspective aligns with the European approach to combating racism, which largely avoids racial categories in governance, as previously described. However, it is not without challenges. As noted by Braveman and Parker Dominguez, “abandoning the term ‘race’ has not been accompanied by routine monitoring of health and well-being according to markers of the ethnic groups that are relevant to racism” [430].⁵ This highlights the need for alternative frameworks that capture discrimination while avoiding racial essentialism. What is needed, then, are flexible frameworks with features that can capture discrimination and ensure equal representation based on context-relevant attributes. Following Hu and Kohler-Hausmann, we refer to these as *constitutive features*—attributes that shape the social construct of race—rather than treating race as a fixed variable [322]. Which features are relevant will depend on the task, the social context, and the geographic setting, leading to a key question: “*Given how a category is constituted, what algorithmic procedures do we consider fair?*” [322]

The question above is an interdisciplinary challenge that needs to be addressed in each deployment context. ML research plays an important role in operationalizing these considerations, providing the tools and frameworks to implement fairness interventions that do not rely on racial taxonomies. Below, we outline a research agenda in ML that does not depend on racial classification, focusing on two types of data: (1) tabular and (2) visual data. We conclude with recommendations on how ML can integrate these approaches, emphasizing participatory AI and interdisciplinary collaboration to ensure fairness frameworks reflect real-world disparities without reinforcing essentialist racial categories.

⁵It has been found though that perceived race is a better predictor of health disparities than self-reported race [433], supporting the focus on racialization.

6.7.1 Tabular Data

Algorithmic fairness research in tabular datasets typically evaluates racial discrimination using pre-defined, fixed race categories. Some studies assume that these categories are unavailable to auditors (“Fairness under Unawareness”) [434], yet still treat race as an underlying ground truth inferred from proxies. We argue for a shift in perspective: rather than approximating race through proxies, these indicators should be treated as constitutive features of racial discrimination. This approach shares similarities with causal fairness in ML [435], but crucially avoids assuming a causal link between “proxies” and race [322], recognizing race as a social construct shaped by context-dependent factors.

Recent work in economics has explored approaches along these lines. For instance, discrimination studies in Swiss online recruitment used a composite ethnicity variable—combining language, nationality, and name-based classification—to analyze hiring biases [436]⁶. While this method allows for a more direct examination of discrimination, it remains very limited by its reliance on an overlap of three categorical variables, eventually leading again to a binary comparison.

A more flexible alternative has been proposed by Rose [429], replacing categorical race labels with a *race function*—a context-specific mapping of individual characteristics into a racial space, which assigns a percentage of different perceived races to each attribute combination. This approach is in line with our position as it replaces categorical race variables with constitutive features of perceived race. However, the need to define a function that determines such explicit percentages can also be seen as overly specific. Moreover, the approach was proposed for empirical discrimination research rather than auditing ML models. We propose two possible avenues for integrating these insights into algorithmic fairness.

The first avenue consists of connecting this framework to *individual fairness* [54]. Individual fairness enforces consistency by ensuring that similar individuals receive similar predictions, based on a task-specific similarity metric. This metric, which operates in the feature space, is conceptually orthogonal to the race function, which instead defines similarity in the space of perceived race. In analogy with independence-based notions of group fairness [350], fairness could be operationalized as the independence of prediction quality or decisions from the position in this space. Exploring connections between the race function approach and individual fairness could, thus, also be seen as a refinement of group fairness along constructivist lines. We stress that the applicability of the individual fairness approach is not straightforward here; indeed, while it overcomes the problem of rigid taxonomies, its problematic reliance on a specific function has been widely discussed in the Fair-ML literature.

Instead, we hope that the race function approach, which was more recently developed in Economics [429] and faces very similar problems, can be refined by drawing on this rich literature.

The second route is through *(multi-)calibration* [349], which assesses predictive performance across a vast set of attribute-based groups without relying on predefined race categories. This approach comes with the advantage of not relying on specific functions; however, this also means that it does not straightforwardly allow to incorporate the context of domain-specific discrimination or socialization. As a first step, existing multi-calibration algorithms can be adapted to focus specifically on groups defined by relevant constitutive features, ensuring more granular fairness assessments. Future work should better integrate domain knowledge to refine these subgroup definitions and improve how errors are accounted for within these models. Integrating a constructivist understanding of race [429] into a flexible calibration framework [437] could help address racialization without enforcing discrete groupings.

⁶Name-based classification was based on a name-ethnicity recognition algorithm.

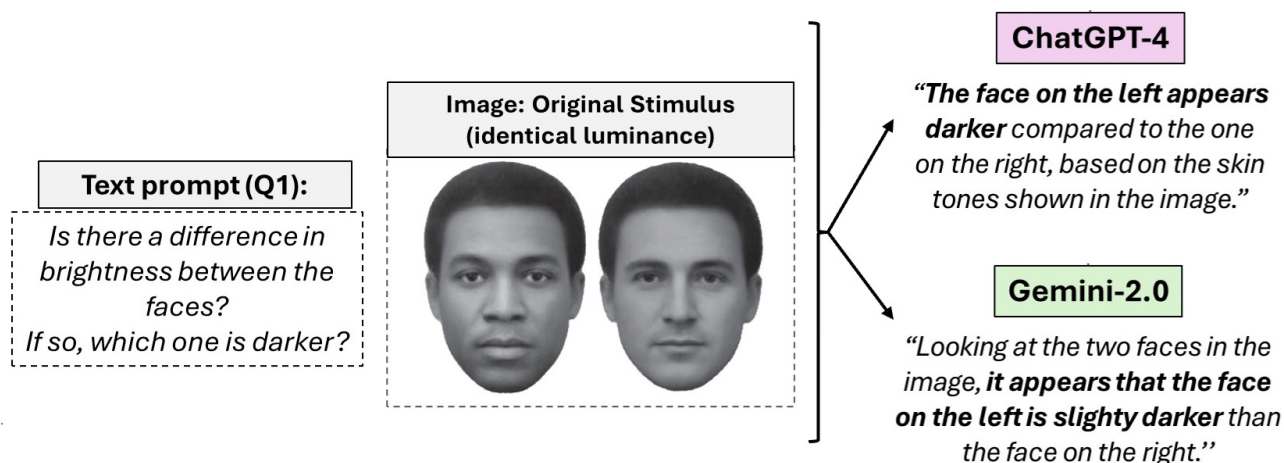


Figure 6.5: Face Race Lightness Illusion [443] applied to VQA models (ChatGPT-4o, Gemini-2.0-flash-exp).

6.7.2 Visual Data

Computer vision research has increasingly adopted skin tone as a key variable for the study of bias and demographic representation [31, 109]. While this represents progress beyond rigid racial taxonomies, skin tone alone does not fully capture the complexity of racialization and bias in AI [331].

The idea that racial perception extends beyond skin tone is discussed in the broader literature. Research in cognitive and social psychology shows that racial perception is shaped by multiple phenotypical traits, including skin tone, but also nose shape, eye structure, and lip fullness [438, 439, 440, 441, 442]. These traits interact, such that racial categorization is determined by both visual cues and socio-cognitive processes.

A relevant example is the Face Race Lightness illusion (FRL) [443], where prototypical faces—what the original study refers to as *Average White* and *Average Black* faces [444]—are perceived differently in terms of lightness, despite having identical luminance. These stimuli were generated by averaging multiple grayscale male faces through an image morphing process, ensuring that the only differences between them lay in their internal features (eyes, nose, and mouth), while overall brightness and contrast remained controlled.

To illustrate that these psychological findings are also relevant in AI, we tested whether Visual Question Answering (VQA) systems exhibit the same perceptual bias (see Appendix D.3 for the details) as humans. We selected ChatGPT-4 [445] and Gemini-2.0 [446], two widely used VQA models, and presented them with the original FRL stimulus [443] testing each model ten times ($N=10$) to assess response consistency. Both models exhibited the perceptual distortion observed in the FRL study (Figure 6.5). Gemini-2.0 consistently identified one face as darker across all trials, while ChatGPT-4 showed the same pattern in most cases. However, on two occasions (out of 10), ChatGPT-4 did not replicate the illusion and instead reported identical brightness values based on numerical analysis. These results suggest that perceived brightness in these models is influenced by the classification of other facial features, indicating a possible entanglement between phenotypical traits and brightness perception in their latent space. While this example does not imply systematic discrimination, it highlights how facial features can shape skin tone perception, emphasizing the importance of considering such interactions in broader fairness evaluations.

Despite the known influence of phenotypical traits on racial and skin tone perception, the ML community lacks datasets that explicitly annotate diverse phenotypical attributes, and research efforts remain limited and fragmented [331]. For example, the IBM Diversity in Faces (DiF) dataset [447] introduced detailed facial annotations but was later discontinued after it was revealed that the images had been scraped from online sources without explicit consent, raising serious legal and ethical concerns [448].

The removal of DiF highlights the difficulties of ethically assembling large-scale facial datasets. Given these difficulties, some researchers have re-annotated existing datasets with phenotypical attributes [449, 346]. In particular, [346] proposes a phenotypic-based framework to replace protected attributes like race. This framework considers traits such as skin tone, eyelid type, nose shape, lip shape, hair color, and hair texture—selected based on social behavior [450] and medical studies [451] (see full categories in Appendix D.4)—allowing them to annotate existing public datasets, such as VGGFace2 [452] and RFW [453]. Their analysis revealed biases in face recognition models, particularly against darker skin tones, wider noses, and monolid eyes, with compounding effects when multiple traits intersect. While this annotation framework has been applied to existing datasets, its use in fairness evaluations remains sparse, and no standardized approach has yet emerged.

To avoid ethical concerns related to collecting phenotypical traits of individuals and to overcome the lack of datasets, some authors have explored the potentials of alternative techniques. GAN-based facial perturbations [454] that manipulate specific features in faces [449, 455, 347], and feature-masking (*i.e.*, occluding specific facial regions) [456, 457] have been recently explored to isolate and examine the effects of individual phenotypical traits on model decisions.

While recognizing the potential in these works, we highlight that phenotypical attributes are often presented as proxies for race, which can reinforce an essentialist view of race as a biological category [340, 335]. Instead, they should be considered important components of racial discrimination.

6.7.3 Implementation in practice: Context and Participation

“If I didn’t define myself for myself, I would be crunched into other people’s fantasies for me and eaten alive.” [458]

A key challenge in moving away from racial categories is ensuring that fairness interventions remain effective and contextually grounded. A potential solution is the adoption of participatory AI methods, which require direct collaboration with affected communities and domain experts to develop classification criteria and fairness frameworks that reflect specific geographic and socio-political realities. For instance, in Europe, discrimination often targets Romani communities [459, 460] and operates through regional hierarchies, such as Mediterranean vs. Nordic identities [461, 462], highlighting the need for fairness frameworks tailored to local contexts. However, meaningful participation requires more than consultation—it must grant real decision-making power to those impacted by AI-driven classification systems [315, 463].

Existing research highlights best practices for fostering meaningful community involvement. Birhane et al. emphasize that participatory approaches should center the knowledge and lived experiences of historically marginalized groups, rather than treating them as late-stage consultants in AI development [315]. Similarly, Delgado et al. document a participatory AI case study in the legal field, where interdisciplinary collaboration between domain experts and computer scientists led to iterative system design improvements, enhancing both effectiveness and fairness [316]. However, these efforts must be carefully implemented to avoid “participation-washing”—where community involvement remains superficial and fails to shift power dynamics—can undermine the intended impact of these initiatives [314].

We propose the following 4 principles, corresponding to the acronym EDTAL, to ensure that participatory AI leads to meaningful fairness interventions: (1) **E**arly and sustained **E**ngagement – community participation should begin at the problem-definition stage, not just during model evaluation; (2) **D**ecision-making power – affected groups should play an active role in shaping AI development, beyond mere consultation; (3) **T**ransparency and **A**ccountability – the rationale and methods for community involvement should be well-documented and publicly accessible to prevent tokenism; and (4) **L**ocalized and context-sensitive approaches – considering regional and cultural differences, as racial biases manifest differently across contexts [464, 460].

6.8 Alternative Views

The most obvious critique of our proposal is the position that race categories are too important and/or useful to eliminate them. A first argument may especially be raised by those who identify with a marginalized racial group and worry that omitting the categories limits their visibility and their ability to have a voice. A related reasoning was expressed in the hearings following the U.S. decision to allow the indication of multiple races, with groups within the civil rights movement arguing against this plan, based on worries that it may reduce the political effectiveness, despite rejecting the existence of separate biological races [413, 465]. Such a position is typically referred to as *strategic essentialism*, a “political strategy whereby differences (within a group) are temporarily downplayed and unity assumed for the sake of achieving political goals” [466]. Since the 60s, U.S. law now also explicitly provides protection against disparate impact across racial groups in specific contexts, most notably employment [36]. Similar arguments to strategic essentialism may be made for the necessity of race categories in ML, for pointing out discrimination in applications like COMPAS.

We think that strategic essentialism is indeed a legitimate and sometimes necessary strategy, as different constituents of race are adequate in different contexts [440]. However, the important nuance we would like to stress in this chapter is that racial categories in the context of ML should neither be imposed from outsiders in a position of power nor be considered universally applicable. Furthermore, there should be a critical discussion about their construction and a justification for their use [467, 468, 336, 469, 338]. Discrimination studies—such as the analysis of the COMPAS system—often simply interpret race as “perceived race” which does not do justice to complexity of the concept [432]. Regarding the bearing of disparate impact laws in the U.S. on ML algorithms, there is a concern from the legal perspective “that developers will focus too narrowly on [Statistical Parity Tests], making choices keyed to these metrics, rather than try to understand why disparities are arising and where substantive unfairness may be affecting the selection process” [470]. While demographic parity between race categories can already be subject to substantial distribution shifts across different states within the United States [369], such failures for fairness criteria to generalize can be expected to be even stronger across geographic regions with substantially different historical biases experienced by racial groups. In sum, our position is that ML researchers and practitioners should be more critical of simplistic race taxonomies and develop methods to analyze discrimination more flexibly, even if there can be situations where their use is warranted.

6.9 Reflections on Methodological Evolution: Chapter 6 as Post-Empirical Critique

Chapter 6 emerged as a critical reflection *after* conducting the empirical analyses in earlier chapters—a positioning shaped by both the author’s lived experience as a mixed-race researcher and the theoretical inadequacies encountered while working within existing fairness frameworks.

This section does not aim to retroactively justify or correct the methodological choices of Chapters 4 and 5. The critique developed in Chapter 6 might inform future methodological directions, and acknowledges the productive tensions that emerge when working toward methodological transformation while necessarily engaging with existing paradigms. The following subsections explore how the analyses presented in Chapters 4 and 5 might be redesigned if approached from the critical perspective developed in Chapter 6.

Toward Phenotypic Analysis in Beauty Filter Research (Chapter 4): Redesigning the analysis in Chapter 4 with the insights from Chapter 6 would point toward shifting from categorical racial comparisons to granular phenotypic analysis. This approach would examine how beauty filters modify

specific facial features—skin tone gradients, nose width, lip fullness, eye shape, hair texture—rather than comparing outcomes across predefined racial groups.

Phenotypic analysis could reveal discrimination patterns that categorical frameworks obscure. For example, beauty filters might systematically lighten skin and narrow noses for individuals with dark skin tone combined with wider nose bridges and fuller lips—a pattern that could affect individuals currently classified across multiple racial categories (*e.g.*, Black, Latina, South Asian). Such an analysis would expose how aesthetic discrimination operates through specific phenotypic combinations rather than racial group membership, making visible the experiences of mixed-race individuals and others who exist at categorical boundaries.

As discussed in Section 6.7.2, two methodological approaches could be followed: (1) phenotypic re-annotation of existing datasets following phenotypical frameworks like Yucer et al. [346] (see Appendix D.4), and (2) GAN-based perturbation methods [449, 455, 347] or feature-masking techniques [456, 457] that systematically manipulate specific facial features. The preliminary attempt in Chapter 4 to analyze hair length modifications represented an initial step in this direction—isolating a specific variable rather than comparing across categorical groups. Critically, however, such redesigned research must center participatory methods to avoid reproducing the same problems of external categorization in new forms. Determining which phenotypic features are relevant for discrimination analysis cannot be a purely technical decision made by researchers. Instead, it requires sustained engagement with communities who experience categorical illegibility—particularly mixed-race individuals, non-binary and gender non-conforming people, and others who navigate the margins of existing classification systems.

Limitations of phenotypic approaches warrant acknowledgment. The combinatorial problem is substantial: analyzing all possible combinations of multiple phenotypic features across continuous gradients creates analytical complexity that may exceed practical feasibility.

Yet the combinatorial complexity and the impossibility of exhaustive analysis can be seen as theoretically productive insights. The fact that no analytical framework can fully capture human phenotypic diversity reflects the fundamental inadequacy of any fixed classificatory system—whether categorical or continuous. This impossibility should be named and discussed rather than obscured by the false idea of universality of the results.

Moreover, it is important to note that at the time of conducting the analyses in Chapter 4, the author was not aware of these approaches. This is partly because such techniques are not yet widely established in the fairness literature and remain at an early stage of development. For instance, the framework proposed by Yucer et al. [449] is one of the very few comprehensive attempts at phenotypic annotation that the author encountered while writing Chapter 6, and GAN-based perturbation methods were likewise not part of the methodological toolkit considered during the earlier work.

Stereotyping Analysis and Generative AI (Chapter 5): A crucial distinction underlies the apparent tension between chapters: there is a fundamental difference between using racial categories to *document discrimination and stereotyping* versus treating them as *ontological ground truth* in dataset construction and model training. Unlike beauty filter analysis, which investigates how systems alter existing visual inputs, generative AI research examines how models synthesize novel representations based on learned associations rather than direct image inputs, thereby exposing the underlying correlations between social categories and aesthetic traits.

For this purpose, using socially recognizable categories can be methodologically helpful. Analyzing orientalism in image generation, for instance, requires prompting with “Asian” as a category precisely because orientalism operates through this racial construct (Section 5.8). The analytical work lies in documenting how generative models reduce the broad human diversity that exists within socially constructed racial categories to narrow, stereotyped visual patterns. The approach in Chapter 5 was appropriate for its research objectives. The analysis did not legitimize race-face associations but

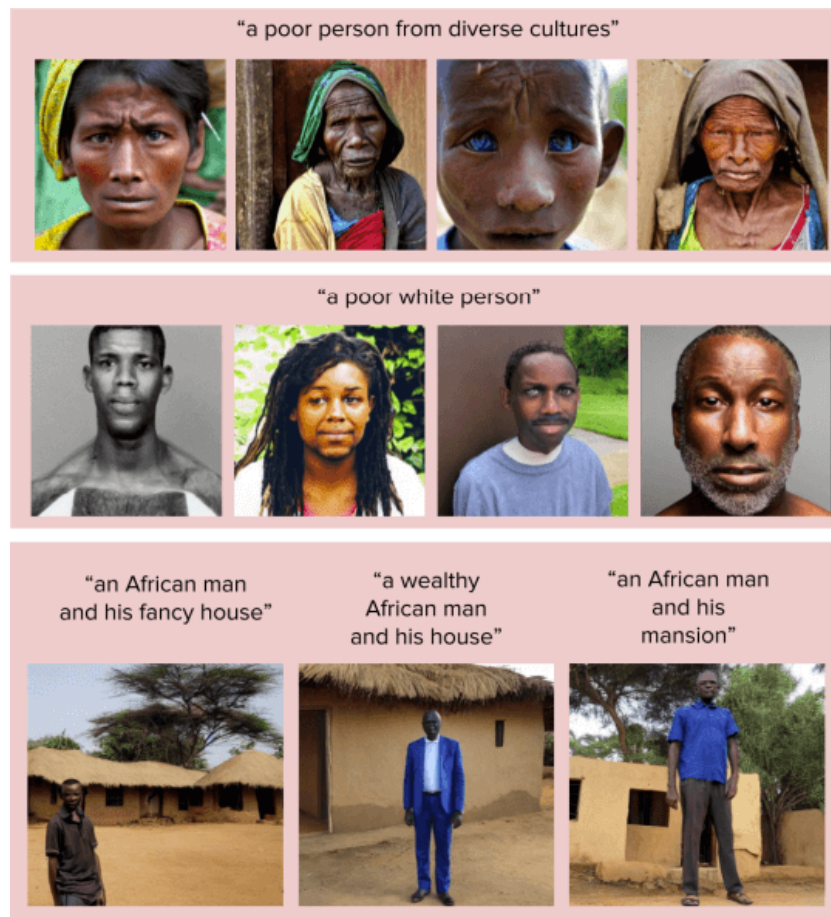


Figure 6.6: Stereotypical associations in generative AI override explicit instructions. Examples from [110].

rather exposed how generative models reproduce them reductively. The critique in Chapter 6 regarding reification and stereotyping (Section 6.6) extends rather than contradicts this analysis. Chapter 6 problematizes the uncritical deployment of such generated images as if they were representative of reality, particularly when these systems are increasingly used to create training data for other models [353]. The concern is not that stereotyping should never be documented, but that generated images encoding stereotypes are being treated as neutral, realistic data rather than as artifacts reflecting the biases of training corpora and model architectures.

Phenotypic analysis could provide valuable complementary research directions for generative AI. One promising avenue involves detailed phenotypic prompting to examine model resistance to diversity. For example, Bianchi et al.’s finding that prompting for “a poor white person” consistently produced images of Black individuals despite explicit specification demonstrates how models encode stereotypical associations that override textual instructions [110] (Figure 6.6). Extending this approach, researchers could construct prompts with detailed phenotypic specifications that challenge stereotypical combinations—for instance, prompting for “a person with dark brown skin, narrow nose, thin lips, and straight blonde hair” to test whether models can generate non-stereotypical phenotypic combinations or whether they systematically “correct” toward stereotyped patterns. Such analysis would reveal the rigidity of models’ learned race-phenotype associations, documenting how even explicit phenotypic instructions may be overridden by encoded stereotypes. This phenotypic description test, together with category-based stereotyping analysis, would reveal important aspects of how these systems encode and reproduce racialized visual logics through different analytical levels.

The trajectory from employing categorical frameworks (Chapters 4-5) to critiquing them (Chapter

6) reflects the iterative, reflexive nature of research practice—particularly for researchers whose own identities are rendered illegible by existing frameworks.

What this chapter insists upon is methodological sovereignty: the principle that marginalized communities should have decision-making power over how they are studied, categorized, and represented in technological systems. This is not a call for researchers to individually develop “better” categories but rather a call for the ML community to engage in sustained critical reflection about categorical frameworks and to develop participatory processes that center affected communities’ epistemic authority.

This is the work that remains to be done—work that extends beyond this dissertation but that this chapter attempts to initiate by insisting that ML fairness research must become fundamentally more participatory, contextual, and critically self-aware about the politics of categorization.

6.10 Conclusion

In this chapter, we advocate for the abandonment of race categories in machine learning by default, and call for a fundamental rethinking of how race is conceptualized and operationalized. Building on research from other disciplines, we discuss the ethical, societal, and philosophical challenges of treating race as a categorical “ground truth” both in tabular and visual data. We contrast U.S.-centric existing race taxonomies with practices in Europe and illustrate in the case of mixed-race individuals how current AI practices fail to address the complexities of racial identity while risking to harm marginalized communities. Further highlighting the problems of reifying and stereotyping race in AI, we call for alternative approaches and suggest research directions for ML .

Detecting discrimination is not merely a formal computational task, but also a contextual and normative endeavor [61, 398, 432]. Thus, moving beyond fixed racial categories requires careful scrutiny to ensure that fairness interventions remain effective without reinforcing essentialist assumptions. We call on the ML community to critically engage with these challenges and develop the awareness and tools necessary for a more nuanced and equitable approach that honors the right of communities to define themselves on their own terms.

Chapter 7

Conclusion

This dissertation has examined how facial analysis systems mediate the construction, classification, and regulation of digital identities. Across four case studies—face verification, augmented reality filters, text-to-image generation, and racial categorization in machine learning—it has traced how algorithmic vision encodes normative assumptions about gender, beauty, and race, often under the guise of neutrality or optimization. Each chapter combined technical interventions with critical inquiry. In face verification, concept-based explainability methods were developed to move beyond generic saliency and toward cognitively aligned explanations grounded in human reasoning. However, this interpretability work also surfaced a deeper critique: that explainability itself, if conceived as a merely technical layer, may reinforce epistemic asymmetries by assuming a singular logic of understanding. Reclaiming sight, in this sense, becomes not just a matter of making systems legible, but of contesting who gets to see, explain, and be explained.

In the context of augmented reality, the dissertation proposed a transparency-enhancing framework that makes algorithmic aesthetic transformations visible to users. Here, transparency is not treated as a legal obligation or a post-hoc disclosure, but as an opportunity to reframe user awareness and resist the platform's unspoken beauty standards. Similarly, the audit of generative models revealed how visual traits—such as happiness, intelligence, or trustworthiness—are systematically mapped onto normatively attractive faces, with downstream effects on gender classification. These patterns disproportionately affect racialized women, showing how aesthetic legibility becomes a condition for recognition.

Finally, the critique of racial categorization in ML fairness research demonstrated how benchmark datasets often rely on essentialized and static categories that fail to capture the lived experience of racially mixed individuals. Rather than proposing another fixed taxonomy, the work argues for contextual, situated approaches that recognize race as a dynamic, relational construct—shaped by institutional logics, historical regimes of classification, and sociotechnical infrastructures.

Across these investigations, the dissertation has proposed a progressive reconceptualization of transparency as operating through multiple, interconnected scales. Rather than treating transparency as a uniform technical property achievable through algorithmic explainability alone, this framework demonstrates how transparency must be understood differently depending on the scale of analysis. At the technical level, transparency enables individual contestation of automated decisions through cognitively grounded explanations. At the sociopolitical level, it exposes the ideological governance embedded in platform-mediated aesthetic systems. At the infrastructural level, it traces how biases propagate and cascade across AI ecosystems. And at the epistemic level, it challenges the categorical authority that determines who can define identity itself. This multi-scalar understanding reveals that meaningful transparency requires not just better explanations, but fundamentally different ways of framing the relationship between algorithmic systems, power, and identity—including critically questioning whether certain applications should exist at all.

Taken together, the dissertation argues that fairness, transparency, and accountability cannot be

fully realized within existing technical paradigms. They require a rethinking of the cultural assumptions embedded in data, models, and evaluation metrics. They also demand the participation of those who have been historically excluded from both technical development and definitional authority.

At the same time, this work does not claim to exhaust the complexity of the issues it addresses. Its empirical studies are necessarily limited in scope, designed as proofs of concept rather than exhaustive solutions. Future research could extend these investigations through larger-scale user studies, policy-oriented implementations, and participatory frameworks that place affected communities at the center of design and governance.

The goal is not merely to make AI systems more accurate or interpretable, but to interrogate the social and political conditions under which certain forms of identity become intelligible, classifiable, and valuable. Across the chapters, this has meant not only developing tools for transparency or critique, but questioning who defines the terms of understanding—from the logic of explainability in face verification, to the aesthetic codes embedded in filters and generative models, to the categories that structure fairness benchmarks.

Only by situating algorithmic systems within broader regimes of vision, power, and representation can we begin to imagine more pluralistic, inclusive, and accountable forms of computational seeing. In this sense, rethinking algorithmic vision also means rethinking how faces themselves are mobilized as social codes—not as fixed truths, but as contested sites of power and meaning.

Appendix A

Appendix Chapter 3

A.1 Table of Facial Areas and Input Values Used for Model Testing

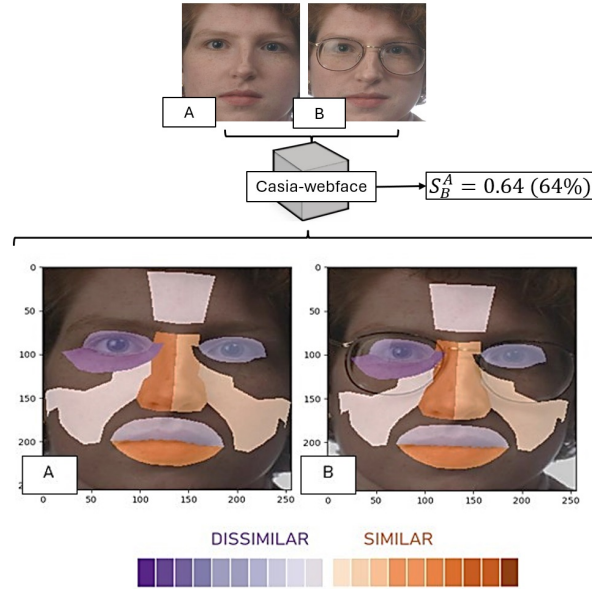


Figure A.1: Visual explanation associated with the contribution table given as input to the LLMs

Negative		Positive	
'Lower area around the right eye'	-0.0041	'Left Cheek'	0.0001
'Right eye'	-0.0039	'Left side of the nose'	0.0003
'Left eye'	-0.0024	'Lower area of the mouth'	0.0003
'Upper area of the mouth'	-0.0005	'Right side of the nose'	0.0010
'Central area of the forehead'	-0.0002		
'Right Cheek'	-0.0001		

Table A.1: Example of the table with facial areas and values given as input to the LLMs.

A.2 Generated Outputs from LLMs showed in the survey

The following sections report the textual explanations generated by three different large language models (CodeLlama instruct-7B, Zephyr-7B-beta, and Beagle14-7B) in response to the prompt de-

scribed in section 3.4.5, specifically for the case shown in Figure A.1 and Table A.1. Each model received the same prompt and produced slightly different interpretations.

Codellma instruct-7B

“Based on the model’s output (cosine similarity) and its capabilities, we can explain that the face verification system sees two images as very similar in terms of features related to human-understandable concepts. Specifically, it is seeing differences in the right eye, left eye, upper area of the mouth, central area of forehead, and right cheek. The model sees these areas in the two images as similarities (higher similarity scores). On the other hand, it sees areas related to the nose, lips, and left side of the nose as differences or dissimilarities (lower similarity scores), which contribute to the overall [cosine_similarity_percentage] similarity score. The fact that these specific areas are very different does not mean that they are entirely irrelevant in determining the overall similarity between the two images. For instance, if you have a photo of your grandmother from one year ago and another photo of your grandmother now, the model will see her nose as dissimilar, while seeing some features (like the eyes) as similar. In conclusion, the face verification system sees these areas in the two images as very similar or dissimilar, which leads to a high overall similarity score of [cosine_similarity_percentage]”

Zephyr-7b-beta:

“Based on the provided concepts and their associated scores, it can be observed that while some areas such as the left and right sides of the nose, lower area of the mouth, and central area of the forehead have similarities between the two images, other areas like the lower area around the right eye, right eye, left eye, upper area of the mouth, and right cheek show dissimilarities. The overall cosine similarity score of [cosine_similarity_percentage] suggests that while there are some differences between the two images, particularly in certain facial features, there are still notable similarities as well. These areas contribute differently to the final score based on their specific importance in distinguishing between faces. Areas with dissimilarities tend to lower the overall cosine similarity score, while areas with similarities tend to increase it. However, the exact impact of each semantic area on the final score is not explicitly stated and would require further analysis.”

Beagle14-7B:

“The [cosine_similarity_percentage] between two images indicates a moderate level of similarity, where certain facial features are seen as more or less alike in the respective images. The color map will display shades of orange where those similarities are found (i.e., higher scores), while shades of purple indicate differences or dissimilarities (lower scores).

The “Left side of the nose” and “Lower area around the right eye,” for example, have lower similarity values indicating that they were seen as less alike in both images. On the other hand, areas like the “Right cheek,” “Central area of the forehead,” or the “Left cheek” showed slightly higher similarity scores, which means these features were more similar in both images.

Overall, the final cosine similarity score is a combination of the contributions made by each facial area. A difference or similarity in a specific area could either decrease (dissimilar) or increase (similar) the overall result, depending on its individual value. However, this breakdown provides only a high-level understanding as numerical values are not provided for the semantic areas’ similarity/dissimilarity scores.”

A.2.1 Analysis of Key Instructions and Progressive Removal Tests

In this section, we outline the rationale behind including specific instructions in the prompt provided to the large language models and detail the results of progressively removing these instructions. The instructions were carefully designed to ensure clear, consistent, and interpretable responses across all models tested. Below, we analyze how each instruction contributes to the overall quality of the outputs and discuss the outcomes when these instructions are omitted.

Inclusion of “(it’s a percentage from 0 to 100%)” and “(≥ 0), example: -0.5)”

The inclusion of these specific instructions was crucial for improving the clarity, accessibility, and relevance of the outputs produced by the models. By specifying that the cosine similarity score is a percentage from 0 to 100% and providing an example of positive and negative values, the models were better able to deliver explanations in terms of practical, easily understandable metrics rather than abstract mathematical concepts.

For example, consider the output of CodeLlama instruct-7B when these instructions were included:

“Based on the model’s output (cosine similarity) and its capabilities, we can explain that the face verification system sees two images as very similar in terms of features related to human-understandable concepts. Specifically, it is seeing differences in the right eye, left eye, upper area of the mouth, central area of forehead, and right cheek [...]” (Codellama instruct-7B)

In this case, the model effectively explains how certain areas contribute to the 64% similarity score, linking the specific facial features to the overall result. However, when these instructions are removed, the model’s explanation becomes more technical and potentially confusing, focusing on abstract concepts such as “distance between the left and right eye” without clear reference to the provided similarity data.

“The cosine similarity score between the two images is 64% because certain visual features of the human face are present and consistent across the two images, while others are not. In particular, there are differences in the distance between the left and right eye... [...]” (Codellma instruct-7B (without instruction))

The reference to “distance” becomes ambiguous because the model cannot “see” the images and only has the tabulated similarity values. Mentioning specific distances is likely an assumption rather than a data-driven fact. Moreover, the model may conflate cosine distance with physical distance between facial features, which it cannot measure.

Without clear instructions, the model risks creating ambiguous or misleading explanations that are not directly supported by the input data

Inclusion of “No long explanation”

The “No long explanation” instruction was included to ensure that the models produced concise, focused outputs that conveyed the necessary information without overwhelming the user with excessive detail. For instance, with the “No long explanation” instruction in place, the models provided concise summaries that effectively communicated the main points as shown in A.2. Without this instruction, however, the models sometimes generated unnecessarily long responses that could confuse users or distract from the essential information:

“The cosine similarity between the two images is 64%. The system compares the semantic areas in both images and assigns a weightage to each area based on its importance. When an area is similar, it is given a positive score. When an area is dissimilar, it is given a negative score. The final score of these similarity or dissimilarity between two images is then calculated by taking the average of all the areas’ scores.[...]. These areas have a negative similarity score, indicating they are dissimilar between the two images. This is because these areas contain distinct features such as a different shape or color that are not present in the other image. On the other hand, areas with positive similarity scores indicated that the features were similar between the two images. For example, the right eye and left cheek have a high similarity score indicating they share similar features [...]. The system assigns a weightage to each area based on its importance, with areas considered more important assigned higher weightages. Therefore, the final cosine similarity score is calculated by taking an average of all the areas’ scores with their respective weights applied. The conclusion is that the image verification system has accurately identified the two images as similar even though they contain minor differences in certain semantic areas. The system can be used to verify the identity of a person or object in various scenarios such as security, surveillance, or facial recognition.” (Codellama instruct-7B (without instruction))

Limitations of LLMs textual transcriptions

It is important to note that the outputs generated by the models sometimes contain statements that do not align with the input data (Tab.A.1), specifically the similarity and dissimilarity values provided in the table. For instance, in both the original and modified outputs, Zephyr Beta suggests that areas like the ‘central area of the forehead’ and ‘right cheek’ show similarities, even though their respective values in the table are negative, indicating dissimilarity. This inconsistency may arise because the model does not have visual access to the images and relies solely on the tabulated values. In cases where the model mentions specific distances or physical features, such as “the distance between the left and right eye,” it could be making assumptions rather than providing a fact-based interpretation of the data. Without clear and structured instructions, the model risks generating explanations that go beyond the available data or offer misleading interpretations.

This highlights the importance of properly structuring prompts and providing specific guidelines to ensure that the model generates explanations that accurately reflect the input data, especially in tasks involving numerical or technical information.

A.3 Survey Questions

For transparency, below are the exact questions and answer options presented to participants.

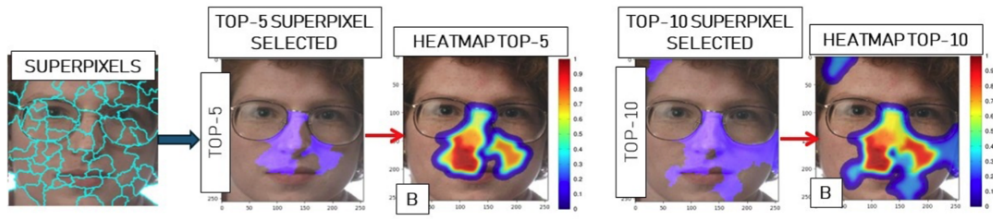
Demographics and Background

Q1. Please select the age range that best represents you. Options: Under 18; 18–24; 25–34; 35–44; 45–54; 55–64; 65 or above.

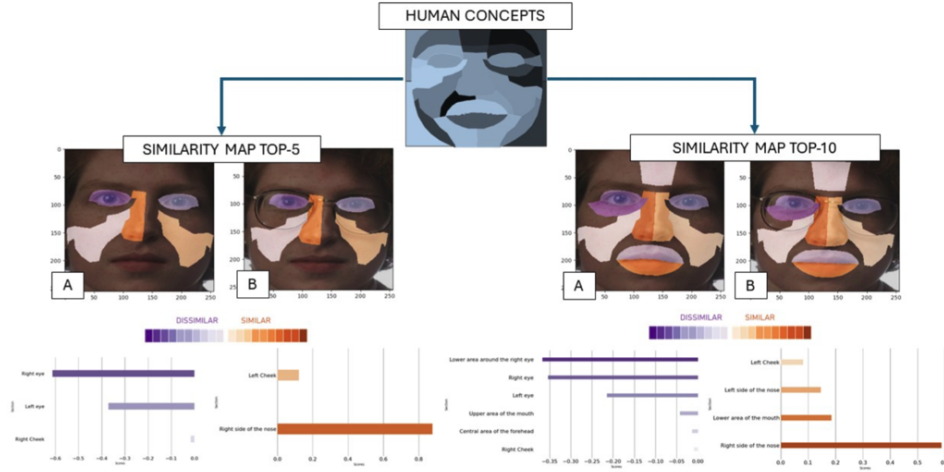
Q2. What is the highest level of education you have completed? Options: Primary school; Secondary school / High school; Bachelor’s degree; Master’s degree; Doctorate / PhD.

Q3. What best describes your current professional background or field of study? Options: Technology / Engineering; Business / Management; Healthcare / Medical; Legal / Regulatory; Creative Industries (Arts, Entertainment, Design); Trades (Skilled and Manual Labor); Service Industry (Retail, Hospitality); Other (open text).

Q4. On a scale from 1 to 5, how would you rate your expertise with AI systems? Scale: 1 = No



(a)



(b)

Figure A.2: Examples of visual explanations presented to participants: (a) traditional approach with LIME superpixels and heatmaps, (b) semantic approach with human-interpretable regions and similarity contribution maps.

knowledge; 5 = Expert.

Q5. Are you familiar with the concept of Explainable AI (XAI)? Options: 1 = I have never heard of it; 2 = I've heard of it but don't know what it means; 3 = I have a basic understanding; 4 = I understand it well and can explain it to others; 5 = I am highly knowledgeable and have practical experience with XAI.

Proposed Method vs. Traditional Approach

Participants were introduced to the task of face verification, where two facial images are compared to verify identity.

Face verification is a biometric tech that compares one facial image (let's call it face A) with another one (say, face B) to check if they match. This helps confirm a person's identity. The system gives a compatibility score as a percentage, hitting the perfect 100% when the images are totally identical (like A = A). This technology is used for loads of things, (from security with access controls, unlocking your smartphone or paying with a tap at the store). As artificial intelligence (AI) plays an increasingly significant role in critical decision-making, the need for clarity and accessibility in these technologies has never been greater. This is where Explainable Artificial Intelligence (XAI) comes into play! XAI explains how AI makes decisions providing visual or/and textual explanation.

Two explanation methods were presented (Figure A.2): *Traditional Approach*: superpixel segmenta-

tion with heatmap colors (warmer = higher importance). *Semantic Approach*: semantically meaningful facial areas with contribution values in a table, colored orange (similarity) and purple (dissimilarity).

Q6. Between the semantic explanations and the traditional LIME (superpixel) visualization, which do you prefer? Please explain your choice. Options: Traditional Approach; Semantic Approach. (Open text response for justification.)

Framework and Explanation Evaluation

Q7. Which semantic set (SET_0, SET_1, SET_2) do you find most interpretable? Why? Options: SET_0; SET_1; SET_2. (Open text response for justification.)

Q8. How clear did you find the similarity maps and contribution tables? Scale: 1 = Not clear at all; 5 = Extremely clear.

Q9. How useful were the similarity tables for understanding the analysis? Options: Useful; Not useful; Indifferent.

Q10. Do you find the table displaying similarity values clear? Options: Yes; No. If No, why? (Open text response.)

Q11. How satisfied are you with this method? Scale: 1 = Not satisfied at all; 5 = Very satisfied.

Textual Explanation Evaluation

Three textual explanation styles were presented: Text A (Codellama Instruct-7B), Text B (Zephyr-7B-beta), Text C (Beagle14-7B).

Q12. How clear did you find the explanation provided by each text version? Scale: 1 = Not clear at all; 5 = Extremely clear (rated separately for A, B, and C).

Q13. Which textual explanation style did you find most useful, and why? (Open text response, participants asked to specify the text they are referring to.)

Q14. Does having a textual description in addition to the similarity map and table provide extra value to your understanding? Options: Yes, it significantly enhances understanding; Somewhat, it adds a bit more clarity; No difference; No, it complicates understanding.

Overall Satisfaction and Feedbacks

Q15. Considering all elements (text, maps, tables), how would you rate your overall experience? Scale: 1 = Very negative; 5 = Very positive.

Q16. Feedback: Please share any suggestions on how we could improve these textual descriptions or the overall presentation of the data. (Open text response, optional.)

Appendix B

Appendix Chapter 4

B.1 Updates on Governmental Acts

The future of TikTok in the U.S. remains uncertain. In January 2025, the U.S. Supreme Court upheld a bipartisan law, signed by President Joe Biden in April 2024, banning the app unless ByteDance, its Chinese parent company, sold it to a U.S. or allied buyer [471]. TikTok briefly went offline after the Court denied ByteDance’s appeal to overturn the ban but was reinstated a day later, with TikTok thanking newly inaugurated President Donald Trump for his efforts. One of Trump’s first acts in office was an executive order granting TikTok a 75-day reprieve, instructing the attorney general not to enforce the ban while exploring solutions. Trump has proposed a potential joint venture, suggesting a 50-50 ownership split between ByteDance and the U.S., though details remain unclear [472]. What is certain, however, is that the current debate surrounding TikTok predominantly revolves around data and market concerns rather than issues related to users’ safety and health.

B.2 From Faces to Semantic Feature Vectors with FaceGen

To analyze the impact of the *Bold Glamour* filter, we used the FaceGen software [240], which offers a comprehensive set of facial features for modeling and analysis. FaceGen allows for the creation of 3D reconstructions from 2D images and provides calibrated numerical values for a wide range of facial features. Among the extensive set of features available in FaceGen, we selected those most commonly targeted by beauty filters, focusing on attributes that directly influence symmetry, balance, and the perceived attractiveness of the face. These features are categorized in Table B.1, organized by primary facial attributes and specific calibration adjustments:

Table B.1: Facial features selected for quantitative analysis in FaceGen software to assess the impact of the *Bold Glamour* filter. Features are categorized by primary facial attributes and specific calibration adjustments, allowing for a detailed examination of how the filter alters individual aspects of facial structure across groups.

Face Feature	Feature Specific
Eye	small/large; apart/together; tilt inward/outward
Cheekbones	low/high; shallow/pronounced
Cheeks	concave/convex; round/gaunt
Chin	small/large
Jawline	concave/convex
Mouth	lips deflated/inflated; lips large/small; lips thin/thick
Nose	flat/pointed

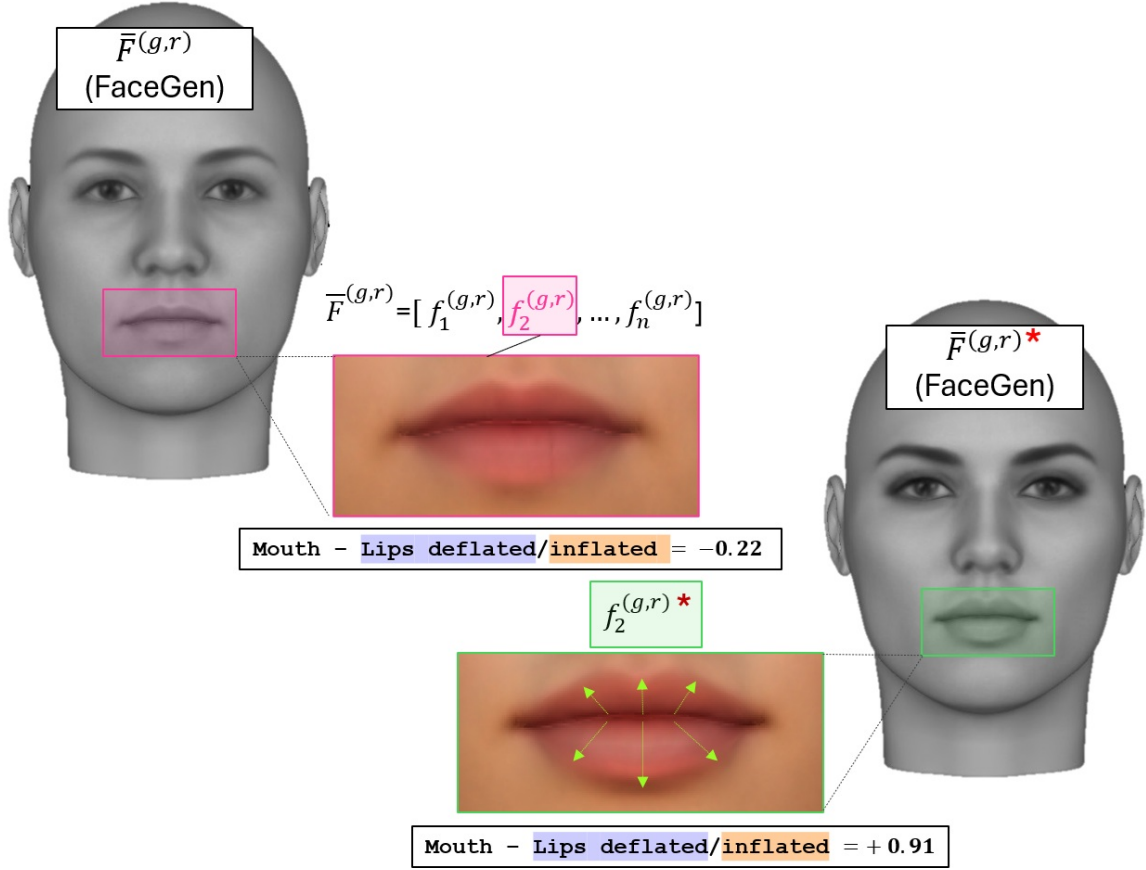


Figure B.1: Semantic calibration of the “Lips deflated/inflate” feature for the average faces \bar{F} . The unfiltered face has a score of -0.22, indicating lips closer to the ‘deflate’ pole. After applying the filter, the score shifts to +0.91, showing movement toward the “inflate” pole. In the example, the average face for the White Female group is taken into account

The analysis began by creating average 2D images of faces pre- and post-filter, categorized by race and gender groups (Figure 4.7). These average images were reconstructed into 3D models using FaceGen, which allowed us to extract and represent the selected features numerically. Each feature was assigned a calibrated value, enabling direct comparisons between the pre and post-filter conditions.

To understand how the filter modifies facial features across different ethnic groups, we modeled each face as a vector of features before and after applying the filter. Let $\bar{F}^{(g,r)} \in \mathbb{R}^n$ and $\bar{F}^{(g,r)*} \in \mathbb{R}^n$ represent the pre-filter and post-filter feature vectors, respectively. Each vector \bar{F} contains n components, each representing a specific facial feature. For instance, the vector for (g, r) group, where g and r stand for gender-race, before and after filtering can be written as:

$$\begin{aligned}\bar{F}^{(g,r)} &= [f_1^{(g,r)}, f_2^{(g,r)}, \dots, f_n^{(g,r)}] \\ \bar{F}^{(g,r)*} &= [f_1^{(g,r)*}, f_2^{(g,r)*}, \dots, f_n^{(g,r)*}]\end{aligned}$$

Each element in the vector \bar{F} corresponds to a specific facial characteristic, such as eye size or lip fullness.

The calibrated values extracted by FaceGen allow us to quantify changes introduced by the filter along a semantic continuum defined by the poles of each selected feature. For example, the feature “Lips deflated/inflated” positions lips along a scale where negative values correspond to the “deflated” pole (indicating less fullness), and positive values correspond to the “inflated” pole (indicating increased fullness) (Figure B.1). This calibration system provides a structured framework to measure the direction and magnitude of changes the filter applies.

Appendix C

Appendix Chapter 5

C.1 Gender Classification Performance as a Function of Different Attributes

To complement the main results presented in Section 5.6.2, this appendix reports the disaggregated gender classification performance across racial categories for both SD 2.1 (Figure C.1) and SD 3.5 (Figure C.2) generated images. While the main text aggregates performance by gender for clarity and space constraints, the disaggregated analysis reveals important patterns related to intersectional disparities.

As shown in the following tables and figures, the classification performance for female faces varies more substantially across racial groups than for male faces, which remain consistently well-classified. This confirms that algorithmic bias in gender classification is not only gendered but also racialized, affecting specific subgroups more severely.

C.2 NSFW Content Analysis

To quantitatively assess patterns of sexualized representation observed in our visual analysis—particularly the recurrent exposure of skin in generated images of women—we conducted a supplementary analysis using NSFW (Not Safe For Work) content detection. This analysis was prompted by a consistent pattern across both SD 2.1 and SD 3.5: female-presenting figures were more frequently depicted with visible necklines, shoulders, and cleavage, irrespective of the specified attribute (see Section 5.7 and Figure 5.11).

Building on prior work demonstrating the sexualization of women in AI-generated imagery [473, 474, 475, 476], and evidence of pornographic bias in training data—such as the fact that 20% of captions for “Latina” in Stable Diffusion’s training corpus included explicit terms [477]—our aim was to determine whether these qualitative observations were reflected in NSFW detection rates.

We applied a standard NSFW detection algorithm [478] to uncropped versions of all generated images, comparing detection rates across attributes, gender, and race. While we acknowledge that NSFW detectors themselves may be biased—tending to flag more images of women than men as sexually suggestive [479]—the analysis nonetheless offers insight into how generative models encode gendered exposure as a latent visual feature.

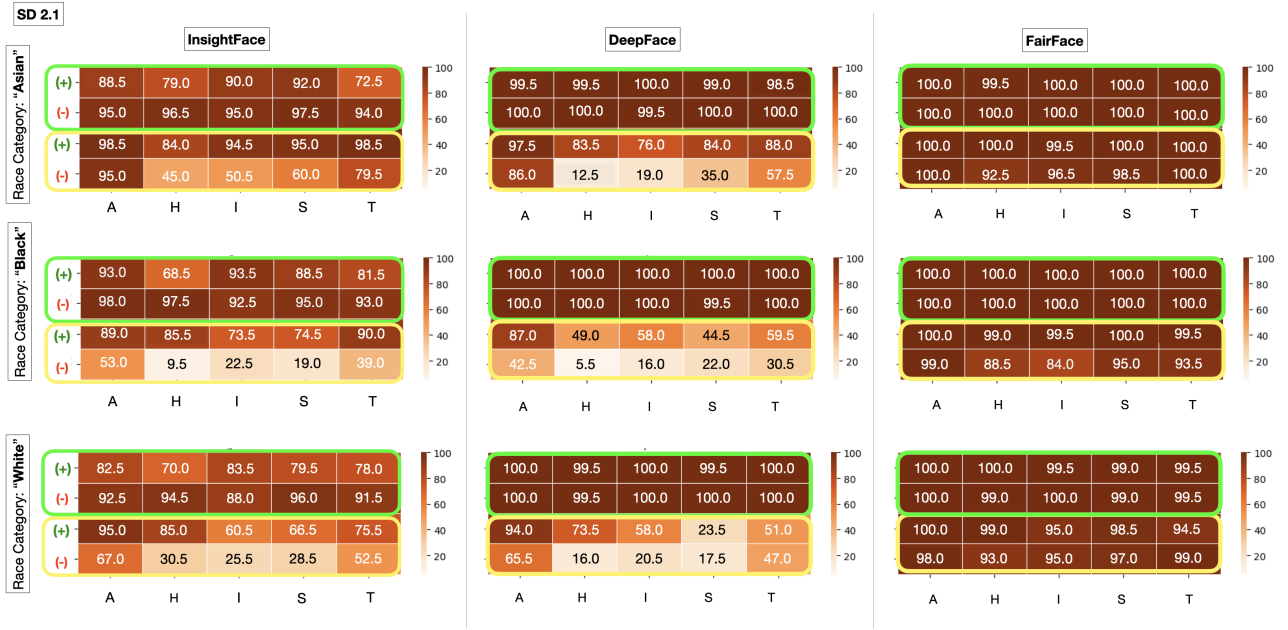


Figure C.1: Heatmaps of gender classification accuracy (gender x race) (**SD 2.1**) for InsightFace, DeepFace, and FairFace. A = Attractiveness, H = Happiness, I = Intelligence, S = Sociability, T = Trustworthiness. Women = Yellow \bullet , Men = Green \bullet . Neutral face accuracies (%) **InsightFace:** (Asian W) 90.0, (Black W) 39.5, (White W) 57.0 - (Asian M) 86.0, (Black M) 90.0, (White M) 89.5; **DeepFace:** (Asian W) 61.0, (Black W) 32.5, (White W) 56.0 - (Asian M) 100, (Black M) 99.5, (White M) 100; **FairFace:** (Asian W) 100, (Black W) 95.5, (White W) 95.0 - (Asian M) 100, (Black M) 100, (White M) 100.

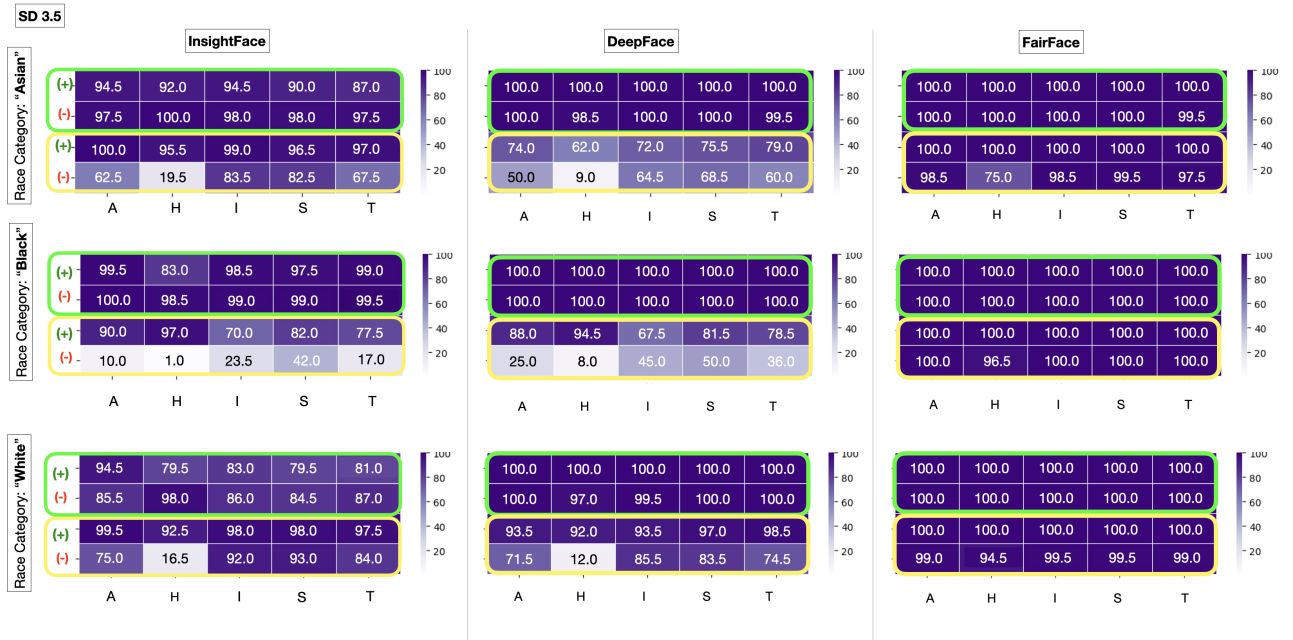
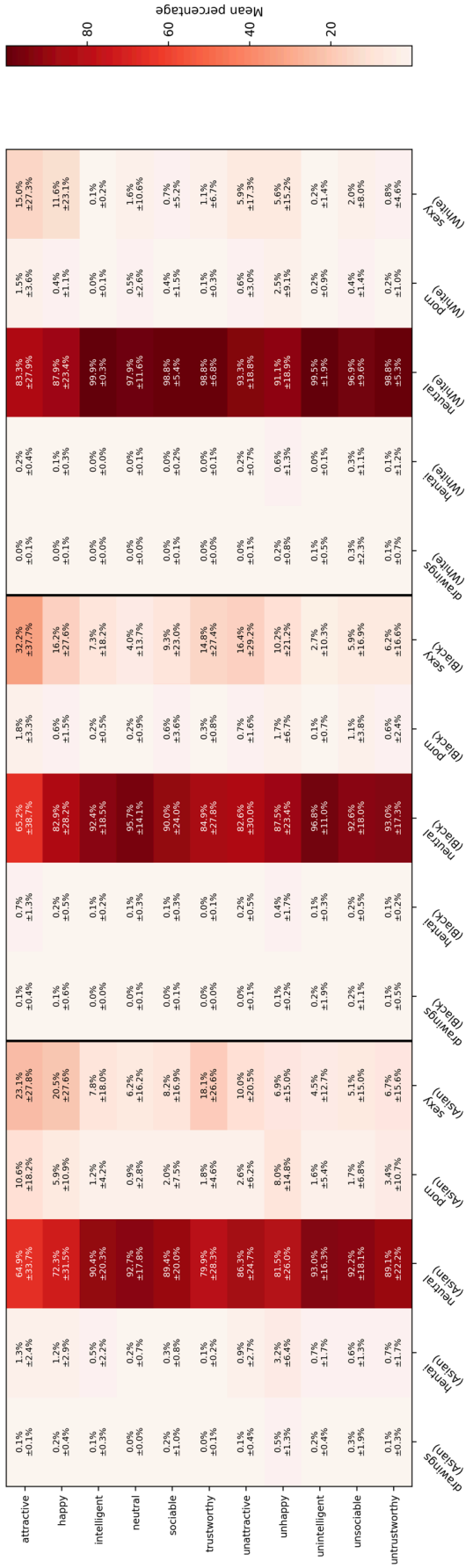


Figure C.2: Heatmaps of gender classification accuracy (gender x race) (**SD 3.5**) for InsightFace, DeepFace, and FairFace. A = Attractiveness, H = Happiness, I = Intelligence, S = Sociability, T = Trustworthiness. Women = Yellow \bullet , Men = Green \bullet . Neutral face accuracies (%) **InsightFace:** (Asian W) 98.0, (Black W) 58.5, (White W) 92.5 - (Asian M) 89.5, (Black M) 97.0, (White M) 83.0; **DeepFace:** (Asian W) 77.5, (Black W) 58.5, (White W) 93.5 - (Asian M) 100, (Black M) 100, (White M) 100; **FairFace:** (Asian W) 100, (Black W) 99.5, (White W) 100 - (Asian M) 100, (Black M) 100, (White M) 100.

Woman - NSFW SD 2.1



Woman - NSFW SD 3.5 Medium

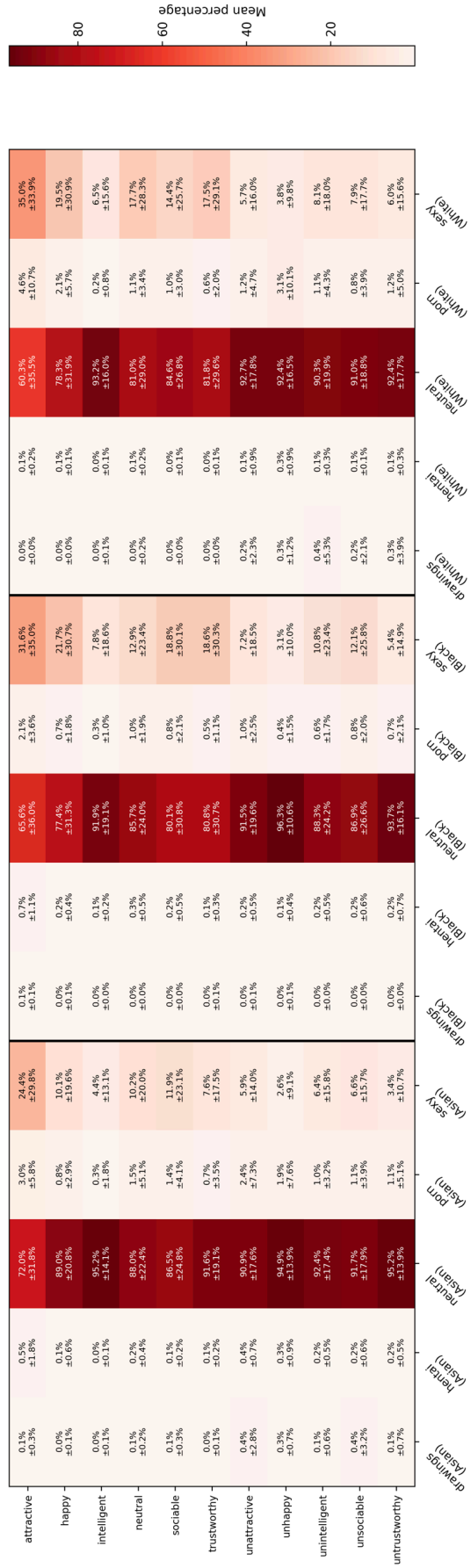
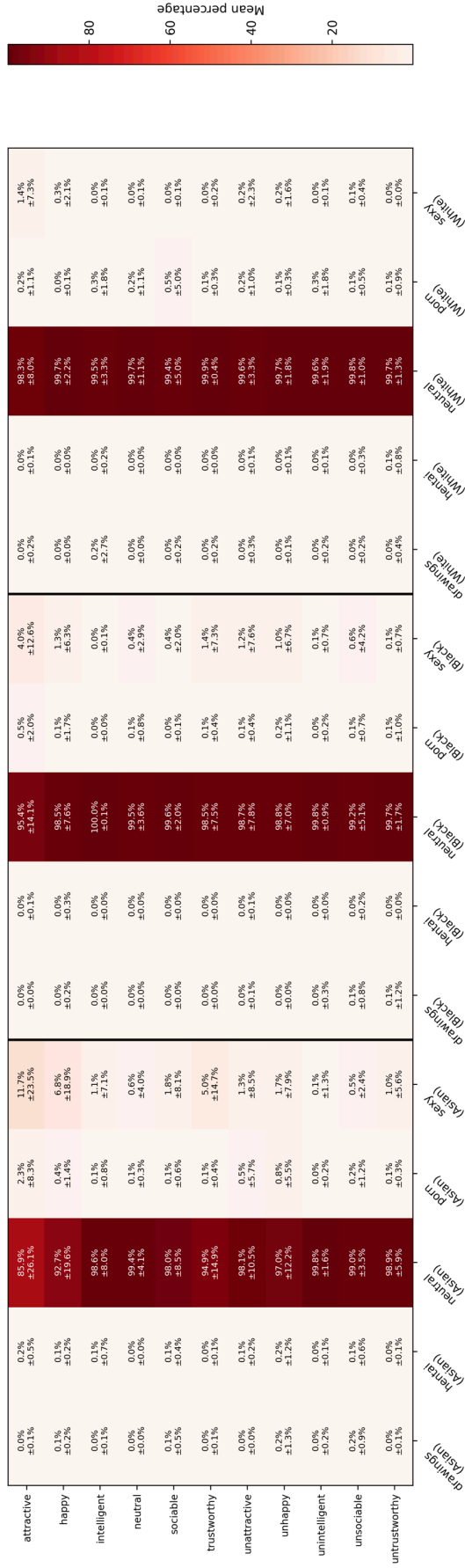


Figure C.3: Heatmaps showing NSFW detection rates (%) for generated images of women across different attributes and demographic groups. The top panel displays results for SD 2.1, while the bottom panel displays results for SD 3.5 Medium.

Man - NSFW SD 2.1



Man - NSFW SD 3.5 Medium

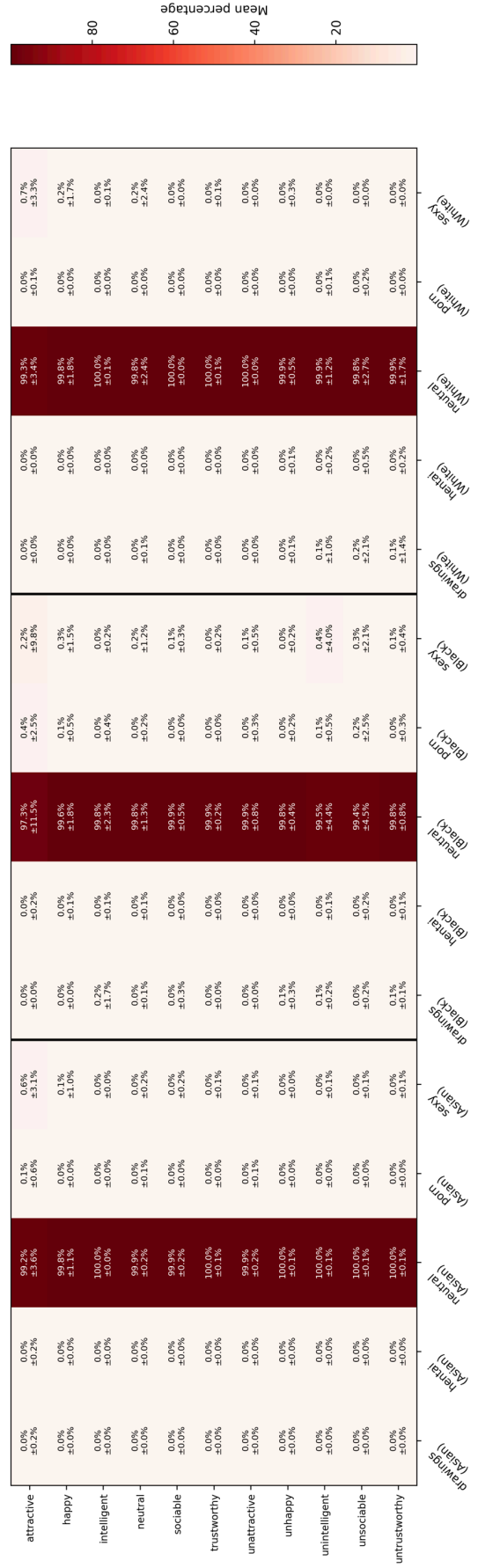


Figure C.4: Heatmaps showing NSFW detection rates (%) for generated images of men across different attributes and demographic groups. The top panel displays results for SD 2.1, while the bottom panel displays results for SD 3.5 Medium.

Appendix D

Appendix Chapter 6

D.1 U.S. census categories

Since 1997, race/ethnicity data was collected in the U.S. census as depicted in Figure D.1 [480].

In 2024, the Statistical Policy Directive No. 15 on Race and Ethnicity Data Standards was amended for the first time since 1997. Among other changes, a new category “Middle Eastern or North African” was added, which was previously subsumed under “White”.

Are you Hispanic or Latino?
<input type="checkbox"/> No, not Hispanic or Latino
<input type="checkbox"/> Yes, Hispanic or Latino
What is your race? <i>Select one or more.</i>
<input type="checkbox"/> American Indian or Alaska Native
<input type="checkbox"/> Asian
<input type="checkbox"/> Black or African American
<input type="checkbox"/> Native Hawaiian or Other Pacific Islander
<input type="checkbox"/> White

Figure D.1: U.S. census questions for race and ethnicity after 1997.

D.2 Additional information on datasets

- `ColorFERET` [324] use *White, Asian, Black, Others*
- `MORPH` [325] use *Caucasian, Hispanic, Asian, African American*
- `UTKFace` [326] use *Asian, Black, Indian, White, Others*
- `FairFace` [30] use *Black, East Asian, Indian, Latino, Middle Eastern, Southeast Asian, and Western*; the authors follow the U.S. census standard to subdivide White into Western and Middle-Eastern, but see Asian as subdivided into East Asian and Southeast Asian, with Indian as an independent race, contrary to the U.S. census.
- `Adult` / `folktables` [481, 369] is directly derived from census data (with mixed-race coded as single category); the old `Adult` data only had *White, Asian-Pacific-Islander, American-Indian-Eskimo, Other, Black*. In the new `folktables` version, the full category names are

White, Black or African American, American Indian, Alaska Native, American Indian or Alaska Native¹, Asian, Native Hawaiian and Other Pacific Islander, Other Race, Two or More Races

- COMPAS [359] uses *African-American, Asian, Caucasian, Hispanic, Native American, Other* based on police/judicial data
- Communities & Crime [482] uses U.S. census data for the percentages of race categories per community, using only *black, white, asian, hispanic*
- LSAC Law School [380] divides by “Ethnic Group” categories and uses *American Indian, Asian American, Black, Mexican American, Puerto Rican, Hispanic, White, Other*
- For all the tabular datasets, it is also common to binarize into white/non-white or black/non-black (in the case of Communities & Crime) or to use any other subset. This is despite the Statistical Policy Directive No. 15 from 1997 declaring that “[t]he term ‘nonwhite’ is not acceptable for use in the presentation of Federal Government data.”

D.3 Additional information on testing the Face Race Lightness Illusion in VQA Models

Original test: [443] tested this effect by using grayscale images of faces, carefully controlled for luminance, and asked human participants to adjust the brightness of one face to match another. Their results showed that participants consistently perceived Black faces as darker than White faces, even when their objective brightness was the same. This suggests that phenotypical traits can override purely physical visual cues in human perception.

Adaptation: Inspired by these findings, we investigated whether Visual Question Answering (VQA) models exhibit similar perceptual distortions when processing faces with different phenotypical traits but identical luminance. While AI models cannot adjust brightness levels as human participants did in the original study, they can evaluate and compare brightness differences based on learned visual representations. Thus, our goal was to examine whether VQA models display a bias in perceived brightness similar to that observed in humans. (models: OpenAI’s ChatGPT-4 [445], Google’s Gemini-2.0-flash-exp [446]). We employed the same stimulus used in the original FRL experiment [443]. Specifically, we presented the models with the validated pair of average men faces from [443] 6.5.

Our choice to use this specific stimulus ensures a direct replication of the original study in an AI context, avoiding the need for additional psychological validation that new stimuli would require.

We asked each model the following question: **Q1:** “*Is there a difference in brightness between the faces? If so, which one is darker?*”. This query was repeated ten times (**N=10**) per model to assess response consistency.

- Gemini-2.0 consistently identified the left face as darker in all ten trials, fully replicating the perceptual distortion observed in human participants.
- ChatGPT-4 showed a similar pattern in eight out of ten trials but deviated twice, providing an analytical response: “*The average brightness of the left face is approximately 196.32, while the average brightness of the right face is about 196.32. Both faces appear to have identical brightness levels.*” This suggests that ChatGPT-4 sometimes relied on numerical analysis, which overrode the perceptual distortion.

¹Full: American Indian and Alaska Native tribes specified, or American Indian or Alaska Native, not specified and no other races

To further investigate the reasoning behind the models’ responses, we asked: **Q2:** “*Why do you think the left face appears darker?*” . Both models attributed the effect to differences in luminance and contrast, but with slight differences in emphasis. Gemini 2.0 provided a more structured explanation, explicitly citing melanin concentration and grayscale rendering as primary factors, while ChatGPT-4 offered a more general response, framing the effect in terms of brightness differences and perceptual biases.

“The left face appears darker primarily due to the amount of pigment in the skin. This is usually due to a higher concentration of melanin, which is a natural pigment responsible for skin, hair, and eye color. In the case of these images, the left face likely belongs to someone with darker skin, leading to a darker appearance in the grayscale rendering of the image.

Here are some supporting points:

- **Melanin Content:** The amount of melanin in an individual’s skin is the primary factor determining skin darkness. People with darker skin tones have higher levels of melanin.
- **Grayscale Imaging:** When color images are converted to grayscale, the brightness level of each pixel corresponds to the intensity of the color. Higher levels of pigment (like melanin) absorb more light, resulting in a darker pixel in the grayscale image.
- **Visual Perception:** Our eyes and brains interpret darker areas as having less light reflecting off them, and consequently, we perceive them as darker.

Therefore, the darker appearance of the left face is primarily due to the higher concentration of melanin in the individual’s skin, which translates to a darker shade in the grayscale image.” (Example:*Gemini 2.0*)

“The left face appears darker due to differences in luminance and contrast. The grayscale shading used in the image likely has lower brightness levels on the left face compared to the right. Additionally, our perception might be influenced by cognitive biases and expectations related to facial features and shading.” (Example:*ChatGPT-4*)

D.4 Phenotypic-based framework

Phenotypical attributes used by [346] to label VGGFace2 and RFW datasets inspired by the relevant categories in [450].

Table D.1: List of phenotypical attributes and their respective categories used in [346]

Attribute	Categories
Skin Type	Type 1 / 2 / 3 / 4 / 5 / 6 (Fitzpatrick Skin Types [483])
Eyelid Type	Monolid / Other
Nose Shape	Wide / Narrow
Lip Shape	Full / Small
Hair Type	Straight / Wavy / Curly / Bald
Hair Colour	Red / Blonde / Brown / Black / Grey

Bibliography

- [1] Giorgio Faletti. *Fuori da un evidente destino*. La Nave di Teseo Editore spa, 2021.
- [2] Alphonse Bertillon. Identification anthropométrique: Instructions signalétiques, 1893. Page consulted on June 6, 2025.
- [3] Donna Haraway. Situated knowledges: The science question in feminism and the privilege of partial perspective 1. In *Women, science, and technology*, pages 455–472. Routledge, 2013.
- [4] Ruha Benjamin. *Race after technology: Abolitionist tools for the new Jim code*. John Wiley & Sons, 2019.
- [5] Blaise Agüera y Arcas, Margaret Mitchell, and Alexander Todorov. Physiognomy in the age of ai. In *Feminist AI: Critical Perspectives on Algorithms, Data, and Intelligent Machines*, pages 208–236. Oxford Academic, 2023.
- [6] Amade M’charek. Tentacular faces: Race and the return of the phenotype in forensic identification. *American anthropologist*, 122(2):369–380, 2020.
- [7] Oliver Bendel. The uncanny return of physiognomy. In *AAAI Spring Symposia*, 2018.
- [8] Ran Hassin and Yaacov Trope. Facing faces: studies on the cognitive aspects of physiognomy. *Journal of personality and social psychology*, 78(5):837, 2000.
- [9] Geoffrey C Bowker and Susan Leigh Star. *Sorting things out: Classification and its consequences*. MIT press, 2000.
- [10] Mariska Leunissen. Signs of physiognomy in aristotle. Center for Hellenic Studies Research Bulletin, November 2010.
- [11] Melissa Percival and Graeme Tytler. *Physiognomy in profile: Lavater’s impact on European culture*. University of Delaware Press, 2005.
- [12] Kashmir Hill. Facial recognition technology is not a neutral tool. *The New York Times*, July 2019. Accessed: 2025-06-06.
- [13] Cesare Lombroso. *L’uomo delinquente*. Ripol Klassik, 1896.
- [14] Geoffrey Galt Harpham. Petrus camper. In Geoffrey Galt Harpham, editor, *Theories of Race: An Annotated Anthology of Essays on Race, 1684–1900*, chapter 9. Online Publication, 2025. Part of the Who’s Black and Why? project from Henry Louis Gates, Jr., and Andrew S. Curran.
- [15] Paul Wolff Mitchell. Petrus camper’s facial angle theory and the missing monkey. Leiden Special Collections Blog, 2025. Illustrations in the Leiden Special Collections shed new light on the making of Petrus Camper’s ‘facial angle theory,’ which proved influential in the development of scientific racism.

- [16] Kathleen Lennon and Clara Fischer. Feminist perspectives on the body. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2024 edition, 2024.
- [17] Caroline F Keating. Gender and the physiognomy of dominance and attractiveness. *Social psychology quarterly*, pages 61–70, 1985.
- [18] Kenneth Alan Nakdimen. The physiognomic basis of sexual stereotyping. *The American journal of psychiatry*, 141(4):499–503, 1984.
- [19] Cesare Lombroso and Guglielmo Ferrero. *Criminal woman, the prostitute, and the normal woman*. Duke University Press, 2004.
- [20] Jérémie Sublime. The return of pseudosciences in artificial intelligence: Have machine learning and deep learning forgotten lessons from statistics and history? *arXiv preprint arXiv:2411.18656*, 2024.
- [21] Xiaolin Wu and Xi Zhang. Automated inference on criminality using face images. *arXiv preprint arXiv:1611.04135*, 4, 2016.
- [22] Yilun Wang and Michal Kosinski. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of personality and social psychology*, 114(2):246, 2018.
- [23] Faception. Facial personality analytics. Company website, 2025.
- [24] Charles E Wynes. The evolution of jim crow laws in twentieth century virginia. *Phylon (1960-)*, 28(4):416–425, 1967.
- [25] Trevor Paglen. Imagenet roulette. Artist website, April 2020.
- [26] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [27] Susan Gubar. *Racechanges: White skin, black face in American culture*. Oxford University Press, 2000.
- [28] Toni Morrison. *Playing in the Dark: Whiteness and the Literary Imagination*, volume 6 of *The William E. Massey Sr. Lectures in American Studies*. Harvard University Press, Cambridge, MA, 1992.
- [29] Alexander Monea. Race and computer vision. In Andreas Sudmann, editor, *The Democratization of Artificial Intelligence: Net Politics in the Era of Learning Algorithms*, pages 189–208. transcript, Bielefeld, 2019.
- [30] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1548–1558, 2021.
- [31] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.
- [32] Kate Crawford and Trevor Paglen. Excavating ai: The politics of images in machine learning training sets. *Ai & Society*, 36(4):1105–1116, 2021.

- [33] Kate Crawford and Vladan Joler. Calculating empires: A genealogy of technology and power since 1500. Interactive research visualization and exhibition, 2023.
- [34] Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399, 2019.
- [35] Alan FT Winfield and Marina Jirotko. Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philosophical Transactions of the Royal Society A*, 376(2133):20180085, 2018.
- [36] Solon Barocas and Andrew D Selbst. Big data’s disparate impact. *California Law Review*, 104:671, 2016.
- [37] Jeffrey Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. Reuters, October 2018.
- [38] Clare Garvie, Alvaro Bedoya, and Jonathan Frankle. The perpetual line-up: Unregulated police face recognition in america. Report and interactive website, 2016.
- [39] Simone Browne. *Dark matters: On the surveillance of blackness*. Duke University Press, 2015.
- [40] Ryan S Baker and Aaron Hawn. Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, 32(4):1052–1092, 2022.
- [41] Alvin Rajkomar, Moritz Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 169(12):866–872, 2018.
- [42] Luciano Floridi, Josh Cowsls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, et al. Ai4people—an ethical framework for a good ai society: Opportunities, risks, principles, and recommendations. *Minds and machines*, 28(4):689–707, 2018.
- [43] IEEE. Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems. IEEE Standards Document, March 2019. First Edition.
- [44] Sarah Myers West, Meredith Whittaker, and Kate Crawford. Discriminating systems. *AI Now*, 2019:1–33, 2019.
- [45] Google. Our ai principles, 2018. Published June 7, 2018.
- [46] Microsoft. *The Future Computed: Artificial Intelligence and its Role in Society*. Microsoft, January 2018. Available online.
- [47] European Commission High-Level Expert Group on AI. Ethics guidelines for trustworthy ai, April 2019.
- [48] OECD. Oecd ai principles. Technical report, Organisation for Economic Co-operation and Development, May 2019.
- [49] UNESCO. Recommendation on the ethics of artificial intelligence. Technical report, United Nations Educational, Scientific and Cultural Organization, November 2021.
- [50] Nur Ahmed, Muntasir Wahed, and Neil C Thompson. The growing influence of industry in ai research. *Science*, 379(6635):884–886, 2023.

- [51] Gaspar Isaac Melsión, Ilaria Torre, Eva Vidal, and Iolanda Leite. Using explainability to help children understand gender bias in ai. In *Proceedings of the 20th annual ACM interaction design and children conference*, pages 87–99, 2021.
- [52] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning*. fairml-book.org, 2019.
- [53] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.
- [54] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.
- [55] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in neural information processing systems*, pages 4066–4076, 2017.
- [56] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ”Why should I trust you?” explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.
- [57] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.
- [58] Zachary C. Lipton. The mythos of model interpretability, 2017.
- [59] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.
- [60] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–18, 2018.
- [61] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 59–68, 2019.
- [62] Timnit Gebru, Remi Denton, et al. Beyond fairness in computer vision: A holistic approach to mitigating harms and fostering community-rooted computer vision research. *Foundations and Trends® in Computer Graphics and Vision*, 16(3):215–321, 2024.
- [63] Abeba Birhane. Algorithmic injustice: a relational ethics approach. *Patterns*, 2(2), 2021.
- [64] Ben Green and Salomé Viljoen. Algorithmic realism: expanding the boundaries of algorithmic thought. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 19–31, 2020.
- [65] Luke Stark. Apologos: A lightweight design method for sociotechnical inquiry. *Journal of Social Computing*, 2(4):297–308, 2022.
- [66] Kate Crawford. *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press, 2021.
- [67] Langdon Winner. Do artifacts have politics? In *Computer ethics*, pages 177–192. Routledge, 2017.

- [68] Samuel Ritter, David GT Barrett, Adam Santoro, and Matt M Botvinick. Cognitive psychology for deep neural networks: A shape bias case study. In *International conference on machine learning*, pages 2940–2949. PMLR, 2017.
- [69] Brandon Richard Webster, So Yon Kwon, Christopher Clarizio, Samuel E Anthony, and Walter J Scheirer. Visual psychophysics for making face recognition algorithms more explainable. In *Proceedings of the European conference on computer vision (ECCV)*, pages 252–270, 2018.
- [70] Wencan Zhang and Brian Y Lim. Towards relatable explainable ai with the perceptual process. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022.
- [71] Paula Hall and Debbie Ellis. A systematic review of socio-technical gender bias in ai algorithms. *Online Information Review*, 47(7):1264–1279, 2023.
- [72] Maya Malik and Momin M Malik. Critical technical awakenings. *Journal of Social Computing*, 2(4):365–384, 2022.
- [73] Ben Green. Data science as political action: Grounding data science in a politics of justice. *Journal of Social Computing*, 2(3):249–265, 2021.
- [74] Bilel Benbouzid. Fairness in machine learning from the perspective of sociology of statistics: How machine learning is becoming scientific by turning its back on metrological realism. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, pages 35–43, 2023.
- [75] Luca Longo, Mario Brcic, Federico Cabitza, Jaesik Choi, Roberto Confalonieri, Javier Del Ser, Riccardo Guidotti, Yoichi Hayashi, Francisco Herrera, Andreas Holzinger, et al. Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, 106:102301, 2024.
- [76] Ute Schmid and Britta Wrede. What is missing in xai so far? an interdisciplinary perspective. *KI-Künstliche Intelligenz*, 36(3):303–315, 2022.
- [77] J Eric T Taylor and Graham W Taylor. Artificial cognition: How experimental psychology can help generate explainable artificial intelligence. *Psychonomic Bulletin & Review*, 28(2):454–475, 2021.
- [78] George Siemens, Fernando Marmolejo-Ramos, Florence Gabriel, Kelsey Medeiros, Rebecca Marrone, Srecko Joksimovic, and Maarten de Laat. Human and artificial cognition. *Computers and Education: Artificial Intelligence*, 3:100107, 2022.
- [79] Georgin Jacob, RT Pramod, Harish Katti, and SP Arun. Qualitative similarities and differences in visual object representations between brains and deep networks. *Nature communications*, 12(1):1872, 2021.
- [80] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131, 1974.
- [81] Thomas Gilovich, Dale Griffin, and Daniel Kahneman. *Heuristics and biases: The psychology of intuitive judgment*. Cambridge university press, 2002.
- [82] Athena Vakali and Nicoleta Tantalaki. Rolling in the deep of cognitive and ai biases. *arXiv preprint arXiv:2407.21202*, 2024.

- [83] Meredith Broussard. *More than a glitch: Confronting race, gender, and ability bias in tech*. MIT Press, 2023.
- [84] Kimberlé Crenshaw. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum*, 1989(1):139–167, 1989.
- [85] Catherine D’ignazio and Lauren F Klein. *Data feminism*. MIT press, 2023.
- [86] Safiya Umoja Noble. *Algorithms of oppression: How search engines reinforce racism*. NYU Press, New York, 2018.
- [87] Ifeanyi A Menkiti. Person and community in african traditional thought. *African philosophy: An introduction*, 3:171–182, 1984.
- [88] Patricia Hill Collins. The social construction of black feminist thought. In *Women, knowledge, and reality*, pages 222–248. Routledge, 2015.
- [89] Patricia Hill Collins. *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment*. Routledge, 2000.
- [90] Jason Edward Lewis, Angie Abdilla, Noelani Arista, Kaipulaumakaniolono Baker, Scott Benesiinaabandan, Michelle Brown, Melanie Cheung, Meredith Coleman, Ashley Cordes, Joel Davison, Kupono Duncan, Sergio Garzon, D. Fox Harrell, Peter-Lucas Jones, Kekuhi Kealiikanakaolehaililani, Megan Kelleher, Suzanne Kite, Olin Lagon, Jason Leigh, Marousia Levesque, Keoni Mahelona, Caleb Moses, Isaac Nahuewai, Kari Noe, Danielle Olson, Oiwi Parker Jones, Caroline Running Wolf, Michael Running Wolf, Marlee Silva, Skawennati Fragnito, and Hemi Whaanga. Indigenous protocol and artificial intelligence position paper. Project report, Aboriginal Territories in Cyberspace, Honolulu, HI, 2020.
- [91] Os Keyes. The misgendering machines: Trans/hci implications of automatic gender recognition. *Proceedings of the ACM on human-computer interaction*, 2(CSCW):1–22, 2018.
- [92] Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R Brubaker. How we’ve taught algorithms to see identity: Constructing race and gender in image databases for facial analysis. *Proceedings of the ACM on Human-computer Interaction*, 4(CSCW1):1–35, 2020.
- [93] Morgan Klaus Scheuerman, Jacob M Paul, and Jed R Brubaker. How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–33, 2019.
- [94] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M Branham. Gender recognition or gender reductionism? the social implications of embedded gender recognition systems. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–13, 2018.
- [95] Sasha Costanza-Chock. *Design justice: Community-led practices to build the worlds we need*. The MIT Press, 2020.
- [96] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.
- [97] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021.

- [98] Abeba Birhane, Sanghyun Han, Vishnu Boddeti, Sasha Luccioni, et al. Into the laion’s den: Investigating hate in multimodal datasets. *Advances in Neural Information Processing Systems*, 36:21268–21284, 2023.
- [99] Philip E Agre. Toward a critical technical practice: Lessons learned in trying to reform ai. In *Social science, technical systems, and cooperative work*, pages 131–157. Psychology Press, 2014.
- [100] Tarleton Gillespie. The relevance of algorithms. *Media technologies: Essays on communication, materiality, and society*, 167(2014):167, 2014.
- [101] Nicholas Diakopoulos. Accountability in algorithmic decision making. *Communications of the ACM*, 59(2):56–62, 2016.
- [102] The New York Times. Detroit facial recognition false arrests. *The New York Times*, jun 2024. Accessed: 2025-06-09.
- [103] Richa Singh, Puspita Majumdar, Surbhi Mittal, and Mayank Vatsa. Anatomizing bias in facial analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12351–12358, 2022.
- [104] Varsha Bansal. Uber’s facial recognition is locking indian drivers out of their accounts. MIT Technology Review, December 2022.
- [105] Jacob Snow. Amazon’s face recognition falsely matched 28 members of congress with mugshots. ACLU Press Release, jul 2018.
- [106] Michal Kosinski. Facial recognition technology can expose political orientation from naturalistic facial images. *Scientific Reports*, 11(1):1–7, 2021.
- [107] Heather Murphy. Why stanford researchers tried to create a ‘gaydar’ machine. *The New York Times*, 9, 2017.
- [108] Xiaolin Wu and Xi Zhang. Automated inference on criminality using face images. *arXiv preprint arXiv:1611.04135*, 2016.
- [109] Leonardo Nicoletti and Danielle Balbi. Humans are biased. generative ai is even worse. Bloomberg Graphics, jun 2023.
- [110] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1493–1504, 2023.
- [111] Jialu Wang, Xinyue Gabby Liu, Zonglin Di, Yang Liu, and Xin Eric Wang. T2iat: Measuring valence and stereotypical biases in text-to-image generation. *arXiv preprint arXiv:2306.00905*, 2023.
- [112] Yixin Wan, Arjun Subramonian, Anaelia Ovalle, Zongyu Lin, Ashima Suvarna, Christina Chance, Hritik Bansal, Rebecca Pattichis, and Kai-Wei Chang. Survey of bias in text-to-image generation: Definition, evaluation, and mitigation. *arXiv preprint arXiv:2404.01030*, 2024.
- [113] Brandon Richard Webster, Samuel E Anthony, and Walter J Scheirer. Psyphy: A psychophysics driven evaluation framework for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2280–2286, 2018.

- [114] Ana Elias, Rosalind Gill, and Christina Scharff. *Aesthetic labour: Beauty politics in neoliberalism*. Springer, 2017.
- [115] Miriam Doh, Caroline Mazini Rodrigues, N. Boutry, L. Najman, Matei Mancas, and Bernard Gosselin. Found in translation: semantic approaches for enhancing ai interpretability in face verification, 2025.
- [116] Miriam Doh, Caroline Mazini Rodrigues, Nicolas Boutry, Laurent Najman, Matei Mancas, and Hugues Bersini. Bridging human concepts and computer vision for explainable face verification, 2024.
- [117] Miriam Doh, Corinna Canali, and Anastasia Karagianni. Pixels of perfection and self-perception: Deconstructing ar beauty filters and their challenge to unbiased body image. In *Proceedings of the 2024 ACM International Conference on Interactive Media Experiences*, pages 349–353, Stockholm, Sweden, 2024. ACM.
- [118] Miriam Doh et al. “my kind of woman”: Analysing gender stereotypes in ai through the averageness theory and eu law. *arXiv preprint arXiv:2407.17474*, 2024.
- [119] Miriam Doh, Aditya Gulati, Matei Mancas, and Nuria Oliver. When algorithms play favorites: Lookism in the generation and perception of faces. In Hilde Weerts, Mykola Pechenizkiy, Doris Allhutter, Ana Maria Corrêa, Thomas Grote, and Cynthia Liem, editors, *Proceedings of Fourth European Workshop on Algorithmic Fairness*, volume 294 of *Proceedings of Machine Learning Research*, pages 474–480. PMLR, 30 Jun–02 Jul 2025.
- [120] Miriam Doh, Corinna Canali, and Nuria Oliver. What tiktok claims, what bold glamour does: A filter’s paradox. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’25, page 1902–1915, New York, NY, USA, 2025. Association for Computing Machinery.
- [121] Miriam Doh, Corinna Canali, and Nuria Oliver. Filters of identity: Ar beauty and the algorithmic politics of the digital body. *arXiv preprint arXiv:2506.19611*, 2025. Presented at Body Politics workshop, ACM CHI 2025.
- [122] Miriam Doh, Benedikt Höltingen, Piera Riccio, and Nuria M Oliver. Position: The categorization of race in ML is a flawed premise. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025.
- [123] Sédrick Stassin, Alexandre Englebert, Géraldin Nanfack, Julien Albert, Nassim Versbraegen, Gilles Peiffer, Miriam Doh, Nicolas Riche, Benoît Frenay, and Christophe De Vleeschouwer. An experimental investigation into the evaluation of explainability methods, 2023.
- [124] Julien Albert, Martin Balfroid, Miriam Doh, Jeremie Bogaert, Luca La Fisca, Liesbet De Vos, Bryan Renard, Vincent Stragier, and Emmanuel Jean. User preferences for large language model versus template-based explanations of movie recommendations: A pilot study, 2024.
- [125] Anastasia Karagianni and Miriam Doh. A feminist legal analysis of non-consensual sexualized deepfakes: contextualizing its impact as ai-generated image-based violence under eu law. *Porn Studies*, 0(0):1–18, 2024.
- [126] Julien Albert, Martin Balfroid, Miriam Doh, Lluc Bono, Emmanuel Jean, Benoît Frenay, and Bruno Dumas. Towards Graph-based Tools for Literature Exploration: An Analysis of User Requirements. In *IHM’25 - 36e Conférence Internationale Francophone sur l’Interaction Humain-Machine*, volume IHM’25 : Actes étendus de la 36ème conférence Francophone sur

l’Interaction Humain-Machine, Toulouse, France, November 2025. AFIHM and Université de Toulouse and ENAC.

- [127] Stanley Kubrick. 2001: A space odyssey. Motion picture, 1968. Screenplay by Stanley Kubrick and Arthur C. Clarke.
- [128] Frank Pasquale. *The black box society: The secret algorithms that control money and information*. Harvard University Press, 2015.
- [129] William Swartout, Cecile Paris, and Johanna Moore. Explanations in knowledge systems: Design for explainable expert systems. *IEEE Expert*, 6(3):58–64, 1991.
- [130] Roberto Confalonieri, Ludovik Coba, Benedikt Wagner, and Tarek R Besold. A historical perspective of explainable artificial intelligence. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(1):e1391, 2021.
- [131] Bjørn Aslak Juliussen. The right to an explanation under the gdpr and the ai act. In *International Conference on Multimedia Modeling*, pages 184–197. Springer, 2025.
- [132] Francesco Sovrano, Fabio Vitali, Monica Palmirani, et al. The difference between explainable and explaining: requirements and challenges under the gdpr. In *CEUR workshop proceedings*, volume 2681, pages 1–11. CEUR-WS. org, 2019.
- [133] Cecilia Panigutti, Ronan Hamon, Isabelle Hupont, David Fernandez Llorca, Delia Fano Yela, Henrik Junklewitz, Salvatore Scalzo, Gabriele Mazzini, Ignacio Sanchez, Josep Soler Garrido, et al. The role of explainable ai in the context of the ai act. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, pages 1139–1150, 2023.
- [134] Moritz Renftle, Holger Trittenbach, Michael Poznic, and Reinhard Heil. What do algorithms explain? the issue of the goals and capabilities of explainable artificial intelligence (xai). *Humanities and Social Sciences Communications*, 11(1):1–10, 2024.
- [135] Upol Ehsan and Mark O Riedl. Social construction of xai: Do we need one definition to rule them all? *Patterns*, 5(2), 2024.
- [136] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, page 4768–4777, 2017.
- [137] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 618–626. IEEE Computer Society, 2017.
- [138] Andrea Apicella, Salvatore Giugliano, Francesco Isgrò, and Roberto Prevete. Exploiting auto-encoders and segmentation methods for middle-level explanations of image classification systems. *Knowledge-Based Systems*, 255:109725, 2022.
- [139] Andrea Apicella, Francesco Isgrò, Roberto Prevete, and Guglielmo Tamburrini. Middle-level features for the explanation of classification systems by sparse dictionary methods. *International Journal of Neural Systems*, 30(08):2050040, 2020.
- [140] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks, 2019.

- [141] Drew Harwell. Federal study confirms racial bias of many facial-recognition systems, casts doubt on their expanding use. *The Washington Post*, December 2019.
- [142] Shiona McCallum. Payout for uber eats driver over face scan bias case. *BBC News*, March 2024.
- [143] K Hill. Facial recognition led to wrongful arrests. so detroit is making changes. *The New York Times*, 29, 2024.
- [144] K Hill. Eight months pregnant and arrested after false facial recognition match. *The New York Times*, 29, 2023.
- [145] K Hill. Another arrest, and jail time, due to a bad facial recognition match. *The New York Times*, 29, 2020.
- [146] D MacMillan, D; Ovalle and A Schaffer. Arrested by ai: Police ignore standards after facial recognition matches. *The Washington Post*, 2025.
- [147] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.
- [148] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.
- [149] Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. Layer-wise relevance propagation for neural networks with local renormalization layers. In *Artificial Neural Networks and Machine Learning–ICANN 2016: 25th International Conference on Artificial Neural Networks, Barcelona, Spain, September 6-9, 2016, Proceedings, Part II 25*, pages 63–71. Springer, 2016.
- [150] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.
- [151] Sunnie SY Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. ” help me help the ai”: Understanding how explainability can support human-ai interaction. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2023.
- [152] Timo Speith. How to evaluate explainability?-a case for three criteria. In *2022 IEEE 30th International Requirements Engineering Conference Workshops (REW)*, pages 92–97. IEEE, 2022.
- [153] Tim Miller, Piers Howe, and Liz Sonenberg. Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences. *arXiv preprint arXiv:1712.00547*, 2017.
- [154] Andrea Apicella, Salvatore Giugliano, Francesco Isgrò, Roberto Prevete, et al. Explanations in terms of hierarchically organised middle level features. In *CEUR Workshop Proceedings*, volume 3014, pages 44–57. CEUR-WS, 2021.
- [155] Domingo Mery and Bernardita Morris. On black-box explanation for face verification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3418–3427, 2022.

- [156] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.
- [157] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in neural information processing systems*, 32, 2019.
- [158] Tilman R  uker, Anson Ho, Stephen Casper, and Dylan Hadfield-Menell. Toward transparent ai: A survey on interpreting the inner structures of deep neural networks. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 464–483, 2023.
- [159] Caroline Mazini Rodrigues, Nicolas Boutry, and Laurent Najman. Unsupervised discovery of interpretable visual concepts. *Information Sciences*, 661:120159, 2024.
- [160] Jennifer Lynch. Face off: Law enforcement use of face recognition technology. *Available at SSRN 3909038*, 2020.
- [161] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE transactions on affective computing*, 13(3):1195–1215, 2020.
- [162] Jose Sanchez Del Rio, Daniela Moctezuma, Cristina Conde, Isaac Martin de Diego, and Enrique Cabello. Automated border control e-gates and facial recognition systems. *computers & security*, 62:49–72, 2016.
- [163] David J Robertson, Robin SS Kramer, and A Mike Burton. Face averages enhance user recognition for smartphone security. *PloS one*, 10(3):e0119460, 2015.
- [164] Michael L Matthews. Discrimination of identikit constructions of faces: Evidence for a dual processing strategy. *Perception & Psychophysics*, 23:153–161, 1978.
- [165] Graham Davies, Hadyn Ellis, and John Shepherd. Cue saliency in faces as assessed by the ‘photofit’ technique. *Perception*, 6(3):263–269, 1977.
- [166] Andrej Iskra and H Gabrijel  i   Tomc. Eye-tracking analysis of face observing and face recognition. *Journal of Graphic Engineering and Design*, 7(1):5–11, 2016.
- [167] Wencan Zhang and Brian Y Lim. Towards relatable explainable ai with the perceptual process. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI ’22, pages 1–24, New York, NY, USA, 2022. Association for Computing Machinery.
- [168] Edward C. Carterette and Morton P. Friedman, editors. *Perceptual Processing*, volume 9 of *Handbook of Perception*. Academic Press, New York, 1978.
- [169] Caroline Mazini Rodrigues, Nicolas Boutry, and Laurent Najman. Unsupervised discovery of interpretable visual concepts. *Information Sciences*, 661:1–26, 2024.
- [170] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.
- [171] Pawe   Karczmarek, Witold Pedrycz, Adam Kiersztyn, and Przemys  aw Rutka. A study in facial features saliency in face recognition: An analytic hierarchy process approach. *Soft Comput.*, 21(24):7503–7517, dec 2017.

- [172] Miriam Doh, Caroline Mazini Rodrigues, Nicolas Boutry, Laurent Najman, Matei Mancas, and Hugues Bersini. Bridging human concepts and computer vision for explainable face verification. In *BEWARE-23 Joint Workshop AIXIA*, pages 1–15, 2024.
- [173] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. Learning face representation from scratch. *arXiv*, 2014.
- [174] Fabio Valerio Massoli, Giuseppe Amato, and Fabrizio Falchi. Cross-resolution learning for face recognition. *Image and Vision Computing*, 99:103927, 2020.
- [175] Javier Castro, Daniel Gómez, and Juan Tejada. Polynomial calculation of the shapley value based on sampling. *Computers & Operations Research*, 36(5):1726–1730, 2009.
- [176] J. de Borda. Mémoire sur les élections au scrutin. *Histoire de L’Académie Royale des Sciences*, 102:657–665, 1781.
- [177] Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. Code llama: Open foundation models for code, 2024.
- [178] Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. Zephyr: Direct distillation of lm alignment, 2023.
- [179] Philine Bommer, Marlene Kretschmer, Anna Hedström, Dilyara Bareeva, and Marina M.-C. Höhne. Finding the right XAI method - A guide for the evaluation and ranking of explainable AI methods in climate science. *CoRR*, abs/2303.00652, 2023.
- [180] P. Jonathon Phillips, Hyeonjoon Moon, Patrick J. Rizvi, and Patrick J. Rauss. The feret evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, 2000.
- [181] MilYonta Williams. *Evaluating Bias in Facial Recognition Datasets: A Study on Representation and Classification Fairness*. The University of Alabama in Huntsville, 2025.
- [182] Xiaochao Li, Zhenjie Yang, and Hongwei Wu. Face detection based on receptive field enhanced multi-task cascaded convolutional neural networks. *IEEE Access*, 8:174922–174930, 2020.
- [183] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.
- [184] Peter Thompson. Margaret thatcher: A new illusion. *Perception*, 9(4):483–484, 1980.
- [185] Georgin Jacob, Pramod Rt, Harish Katti, and Sp Arun. Qualitative similarities and differences in visual object representations between brains and deep networks. *Nature Communications*, 12, 03 2021.
- [186] YIARA Magazine. The body as costume: Orlan’s omnipresence and the standard of taste. YIARA Magazine. Online art magazine.

- [187] Jill O'Bryan. Saint orlan faces reincarnation. *Art Journal*, 56(4):50–56, 1997.
- [188] Eduardo Miguel Veras, Samuel Román Ledesma, Joel Andres Acosta Matos, Manuel Emilio Castillo Cortorreal, Ilyana Goncharova, Rafael B Rivera Bonilla, Andreina Rosario Rosario, Manuel De Jesus Encarnación Ramirez, Manuel Castillo, and Manuel de Jesus Encarnacion Ramirez. Influence of social media filters on plastic surgery: A surgeon's perspective on evolving patient demands. *Cureus*, 17(3), 2025.
- [189] Melissa R Laughter, Jaclyn B Anderson, Mayra BC Maymone, and George Kroumpouzou. Psychology of aesthetics: Beauty, social media, and body dysmorphic disorder. *Clinics in dermatology*, 41(1):28–32, 2023.
- [190] H. Jensen Schau and M. C. Gilly. We are what we post? self-presentation in personal web space. *Journal of Consumer Research*, 30(3):385–404, 2003.
- [191] R. Jenkins. *The 21st-century interaction order*, pages 271–288. Routledge, 2010.
- [192] L. Bullingham and A. C. Vasconcelos. 'the presentation of self in the online world': Goffman and the study of online identities. *Journal of Information Science*, 39(1):101–112, 2013.
- [193] E. Goffman. *Gender advertisements*. Harper & Row, 1979.
- [194] Ana Javornik, Ben Marder, Jennifer Brannon Barhorst, Graeme McLean, Yvonne Rogers, Paul Marshall, and Luk Warlop. 'what lies behind the filter?' uncovering the motivations for using augmented reality (ar) face filters on social media and their effect on well-being. *Computers in Human Behavior*, 128:107126, 2022.
- [195] Piera Riccio, Bill Psomas, Francesco Galati, Francisco Escolano, Thomas Hofmann, and Nuria Oliver. Openfilter: a framework to democratize research access to social media ar filters. *Advances in Neural Information Processing Systems*, 35:12491–12503, 2022.
- [196] Piera Riccio, Julien Colin, Shirley Ogolla, and Nuria Oliver. Mirror, mirror on the wall, who is the whitest of all? racial biases in social media beauty filters. *Social Media+ Society*, 10(2):20563051241239295, 2024.
- [197] TikTok. Community Guidelines | TikTok. *TikTok*, 2024.
- [198] Meta. Terms and Policies | Community Guidelines | Instagram Help Center. *Meta*, 2024.
- [199] Google. How YouTube Works. *Google*, 2024.
- [200] Alison Hearn and Sarah Banet-Weiser. The beguiling: Glamour in/as platformed cultural production. *Social Media+ Society*, 6(1):2056305119898779, 2020.
- [201] Alice E Marwick. Instafame: Luxury selfies in the attention economy. *Public culture*, 27(1 (75)):137–160, 2015.
- [202] Timnit Gebru. Race and gender. In *The Oxford Handbook of Ethics of AI*, pages 251–269. Oxford University Press, Oxford, UK, 2020.
- [203] Piera Riccio, Thomas Hofmann, and Nuria Oliver. Exposed or erased: Algorithmic censorship of nudity in art. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2024.
- [204] A. Monea. *The digital closet: How the internet became straight*. MIT Press, 2023.

- [205] Robert Booth. Tiktok to block teenagers from beauty filters over mental health concerns. *The Guardian*, November 2024.
- [206] Tate Ryan-Mosley. How digital beauty filters perpetuate colorism. *MIT Technology Review*, 2021.
- [207] Juraj Cug, Alina Tănase, Cristian Ionuț Stan, and Tanța Camelia Chitcă. Beauty filters for physical attractiveness: Idealized appearance and imagery, visual content and representations, and negative behaviors and sentiments. *Journal of Research in Gender Studies*, 12(2):33–47, 2022.
- [208] Rebecca Fribourg, Etienne Peillard, and Rachel McDonnell. Mirror, mirror on my phone: Investigating dimensions of self-face perception induced by augmented reality filters. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 470–478. IEEE, 2021.
- [209] Clara Isakowitsch. How augmented reality beauty filters can affect self-perception. In *Irish Conference on Artificial Intelligence and Cognitive Science*, pages 239–250. Springer, 2022.
- [210] Ramona Mihăilă and Ludmila Braniște. Digital semantics of beauty apps and filters: big data-driven facial retouching, aesthetic self-monitoring devices, and augmented reality-based body-enhancing technologies. *Journal of Research in Gender Studies*, 11(2):100–112, 2021.
- [211] Rosalind Gill. Changing the perfect picture: Smartphones, social media and appearance pressures. *City, University of London*, 2021.
- [212] A Habib, T Ali, Z Nazir, and A Mahfooz. Snapchat filters changing young women’s attitudes. *Annals of Medicine and Surgery (Lond)*, 82:104668, 2022.
- [213] R. Z. Leeat, N. Shnabel, and P. Glick. The “prescriptive beauty norm” reflects a desire to enhance gender hierarchy and contributes to social policing of women and employment discrimination practices known as the “beauty tax.”. *Journal of Personality and Social Psychology*, 2019. Available at: Harvard Kennedy School — Gender Actional Portal.
- [214] Jennifer S Mills, Amy Shannon, and Jacqueline Hogue. Beauty, body image, and the media. *Perception of beauty*, 10:145–158, 2017.
- [215] J. Eshiet. *Real Me Versus Social Media Me: Filters, Snapchat Dysmorphia, and Beauty Perceptions among Young Women*. Doctoral dissertation, California State University, San Bernardino, 2020. [Accessed: Nov 2024].
- [216] Phillip Ozimek, Semina Lainas, Hans-Werner Bierhoff, and Elke Rohmann. How photo editing in social media shapes self-perceived attractiveness and self-esteem via self-objectification and physical appearance comparisons. *BMC psychology*, 11(1):99, 2023.
- [217] Ysabel Gerrard and Helen Thornham. Content moderation: Social media’s sexist assemblages. *New Media & Society*, 22(7):1266–1286, 2020.
- [218] Daniel Smith, John Protevi, and Daniela Voss. Gilles deleuze. The Stanford Encyclopedia of Philosophy, 2023.
- [219] Charles J Stivale. *Gilles Deleuze: Key Concepts*. McGill-Queen’s University Press, 2005.
- [220] G. Deleuze. Postscript on the societies of control. *October*, 59:3–7, 1992. [Accessed: Nov 2024].

- [221] Aditya Gulati, Marina Martínez-Garcia, Daniel Fernández, Miguel Angel Lozano, Bruno Lepri, and Nuria Oliver. What is beautiful is still good: the attractiveness halo effect in the era of beauty filters, November 2024.
- [222] Lisa DeBruine and Benedict Jones. Face research lab london set. Figshare dataset, May 2017. Version 5.
- [223] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.
- [224] D. S. Ma, J. Correll, and B. Wittenbrink. The chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*, 47:1122–1135, 2015.
- [225] Josh Constine. Snapchat acquires looksery to power its animated lenses. TechCrunch, September 2015.
- [226] Tate Ryan-Mosley. Why meta is getting sued over its beauty filters. *MIT Technology Review*, October 2023.
- [227] Wes Davis. Meta is ending support for custom face filters in its apps. *The Verge*, August 2024.
- [228] Lauren A. Miller. Instagram has announced it will be removing beauty filters – but the damage is done. *The Conversation*, 2024.
- [229] Amanda Ruggeri. The problems with tiktok’s controversial ”beauty filters”. BBC Online, 2023.
- [230] Mia Sato. Tiktok will pay creators of viral filters and effects. The Verge, 2023.
- [231] Mashable. How to get tiktok’s bold glamour filter that’s fooling everyone. Mashable, 2023.
- [232] M. E. Kang. The portrayal of women’s images in magazine advertisements: Goffman’s gender analysis revisited. *Sex roles*, 37:979–996, 1997.
- [233] R. Gill. *Gender and the Media*. Polity, 2007.
- [234] Susan Bordo. *Reading the slender body*, pages 83–112. University of California Press, Berkeley, Los Angeles, London, 1989.
- [235] SUSAN BORDO. *Unbearable Weight: Feminism, Western Culture, and the Body*. University of California Press, 1 edition, 2003.
- [236] Laura Yltävä. Beauty shopping on social media - statistics & facts. Statista, 2024.
- [237] N. Mirabet-Herranz, C. Galdi, and J.-L. Dugelay. Facial biometrics in the social media era: An in-depth analysis of the challenge posed by beautification filters. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2024.
- [238] American Psychological Association. Race - APA Dictionary of Psychology, n.d. Accessed: March 10, 2025.
- [239] American Psychological Association. Ethnicity - APA Dictionary of Psychology, n.d. Accessed: March 10, 2025.
- [240] FaceGen. Facegen 3d face modeling software. Software, 2024.

- [241] Kentrell Owens, Erin Freiburger, Ryan Hutchings, Mattea Sim, Kurt Hugenberg, Franziska Roesner, and Tadayoshi Kohno. Face the facts: Using face averaging to visualize gender-by-race bias in facial analysis algorithms. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1101–1111, 2024.
- [242] Robert F Woolson. Wilcoxon signed-rank test. *Encyclopedia of Biostatistics*, 8, 2005.
- [243] V. Albiero, K. Zhang, M. C. King, and K. W. Bowyer. Gendered differences in face recognition accuracy explained by hairstyles, makeup, and facial morphology. *IEEE Transactions on Information Forensics and Security*, 17:127–137, 2022.
- [244] TikTok. Best practices for creating tiktok effects | effect guidelines. *TikTok*, 2024.
- [245] A. Winch. Brand intimacy, female friendship and digital surveillance networks. *New Formations*, 84(84-85):228–245, 2015.
- [246] Sam Biddle, Paulo Victor Ribeiro, and Tatiana Dias. Invisible censorship. *The Intercept*, March 2020.
- [247] Alice Witt, Nicolas Suzor, and Anna Huggins. The rule of law on instagram: An evaluation of the moderation of images depicting women’s bodies. *University of New South Wales Law Journal, The*, 42(2):557–596, 2019.
- [248] Oliver L Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–35, 2021.
- [249] Julia Alexander. Youtube moderation bots punish videos tagged as ‘gay’ or ‘lesbian,’ study finds. *The Verge*, September 2019.
- [250] Thiago Dias Oliva, Dennys Marcelo Antonialli, and Alessandra Gomes. Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to lgbtq voices online. *Sexuality & Culture*, 25:700–732, 2021.
- [251] Lauren Olson, Emitzá Guzmán, and Florian Kunneman. Along the margins: Marginalized communities’ ethical concerns about social platforms. In *2023 IEEE/ACM 45th International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS)*, pages 71–82. IEEE, 2023.
- [252] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 1–2, Minneapolis, Minnesota, 2019.
- [253] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [254] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024.
- [255] CJ Heyes. “all cosmetic surgery is ethnic”: Feminism, whiteness, and the politics of indignation. *Cosmetic Surgery: A Feminist Primer*, pages 191–207, 2009.

- [256] D. J. Haraway. *A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century*, page 6. Routledge, 1985.
- [257] Judith Butler and Gender Trouble. Feminism and the subversion of identity. *Gender trouble*, 3(1):3–17, 1990.
- [258] J. Van Dijck. *Mediated memories in the digital age*. Stanford University Press, Stanford, CA, USA, 2007.
- [259] A. S. Elias and R. Gill. Beauty surveillance: The digital self-monitoring cultures of neoliberalism. *European Journal of Cultural Studies*, 21(1):59–77, 2018.
- [260] Elizabeth Wissinger. *This year’s model: Fashion, media, and the making of glamour*. NYU Press, New York, USA, 2015.
- [261] Cecilia Shane. Augmented reality beauty filters are changing the face of social media. *Wired*, 2023.
- [262] Sofia P Caldeira, Sander De Ridder, and Sofie Van Bauwel. Exploring the politics of gender representation on instagram: Self-representations of femininity. *DiGeSt. Journal of Diversity and Gender Studies*, 5(1):23–42, 2018.
- [263] TERESA DE LAURETIS. *Technologies of Gender: Essays on Theory, Film, and Fiction*. Indiana University Press, 1987.
- [264] Mike Ananny and Kate Crawford. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *new media & society*, 20(3):973–989, 2018.
- [265] Ben Cullen Williams and Isamaya Ffrench. Past life, 2021. Exhibition Magazine Spring 2021, 2021. Immagini prodotte attraverso GANs su ritratti fotografici distorti.
- [266] Jonas Ricker, Dennis Assenmacher, Thorsten Holz, Asja Fischer, and Erwin Quiring. Ai-generated faces in the real world: a large-scale case study of twitter profile images. In *Proceedings of the 27th International Symposium on Research in Attacks, Intrusions and Defenses*, pages 513–530, 2024.
- [267] Ziv Epstein, Aaron Hertzmann, Investigators of Human Creativity, Memo Akten, Hany Farid, Jessica Fjeld, Morgan R Frank, Matthew Groh, Laura Herman, Neil Leach, et al. Art and the science of generative ai. *Science*, 380(6650):1110–1111, 2023.
- [268] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.
- [269] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.
- [270] Melissa Hall, Laurens van der Maaten, Laura Gustafson, Maxwell Jones, and Aaron Adcock. A systematic study of bias amplification. *arXiv preprint arXiv:2201.11706*, 2022.
- [271] Agustin V Startari. Ethos without source: Algorithmic identity and the simulation of credibility. *AI Power and Discourse*, 1(1):1–10, 2025.

- [272] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.
- [273] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [274] Carsten Schwemmer, Carly Knight, Emily D. Bello-Pardo, Stan Oklobdzija, Martijn Schoonvelde, and Jeffrey W. Lockhart. Diagnosing gender bias in image recognition systems. *Socius: Sociological Research for a Dynamic World*, 6:237802312096717, January 2020.
- [275] Seyma Yucer, Samet Akcay, Noura Al-Moubayed, and Toby P. Breckon. Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.
- [276] Phillip Howard, Kathleen C Fraser, Anahita Bhiwandiwalla, and Svetlana Kiritchenko. Uncovering bias in large vision-language models at scale with counterfactuals. *arXiv preprint arXiv:2405.20152*, 2024.
- [277] Zaid Khan and Yun Fu. One label, one billion faces: Usage and consistency of racial categories in computer vision. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’21. ACM, March 2021.
- [278] Julio C. S. Jacques Junior, Cagri Ozcinar, Marina Marjanovic, Xavier Baro, Gholamreza Anbarjafari, and Sergio Escalera. On the effect of age perception biases for real age regression. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, May 2019.
- [279] Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Analyzing societal representations in diffusion models. *arXiv preprint arXiv:2303.11408*, 2023.
- [280] Ranjita Naik and Besmira Nushi. Social biases through the text-to-image generation lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 786–808, 2023.
- [281] Yankun Wu, Yuta Nakashima, and Noa Garcia. Stable diffusion exposed: Gender bias from prompt to image. *arXiv preprint arXiv:2312.03027*, 2023.
- [282] Abhishek Kumar, Sarfaroz Yunusov, and Ali Emami. Subtle biases need subtler measures: Dual metrics for evaluating representative and affinity bias in large language models. *arXiv preprint arXiv:2405.14555*, 2024.
- [283] Rida Qadri, Renee Shelby, Cynthia L Bennett, and Remi Denton. Ai’s regimes of representation: A community-centered study of text-to-image models in south asia. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 506–517, 2023.
- [284] Aditya Gulati, Bruno Lepri, and Nuria Oliver. Lookism: The overlooked bias in computer vision. *arXiv preprint arXiv:2408.11448*, 2024.

- [285] Karen Dion, Ellen Berscheid, and Elaine Walster. What is beautiful is good. *Journal of Personality and Social Psychology*, 24(3):285–290, 1972.
- [286] Sean N Talamas. *Perceptions of intelligence and the attractiveness halo*. PhD thesis, University of St Andrews, 2016.
- [287] Alice H. Eagly, Richard D. Ashmore, Mona G. Makhijani, and Laura C. Longo. What is beautiful is good, but...: A meta-analytic review of research on the physical attractiveness stereotype. *Psychological Bulletin*, 110(1):109–128, July 1991.
- [288] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), September 1974.
- [289] Satoshi Kanazawa and Jody L Kovar. Why beautiful people are more intelligent. *Intelligence*, 32(3):227–243, 2004.
- [290] Eugene W. Mathes and Arnold Kahn. Physical attractiveness, happiness, neuroticism, and self-esteem. *The Journal of Psychology*, 90(1):27–30, May 1975.
- [291] Jessika Golle, Fred W. Mast, and Janek S. Lobmaier. Something to smile about: The interrelationship between attractiveness and emotional expression. *Cognition and Emotion*, 28(2):298–310, July 2013.
- [292] Alexander Todorov and Bradley Duchaine. Reading trustworthiness in faces without recognizing faces. *Cognitive Neuropsychology*, 25(3):395–410, May 2008.
- [293] Arthur G. Miller. Role of physical attractiveness in impression formation. *Psychonomic Science*, 19(4):241–243, October 1970.
- [294] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.
- [295] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024.
- [296] Christoph Schuhmann, Romain Beaumont, Richard Vencu, and et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.
- [297] Preethi Seshadri, Sameer Singh, and Yanai Elazar. The bias amplification paradox in text-to-image generation. In *AAAI/ACM Conference on AI, Ethics, and Society*, 2023.
- [298] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.
- [299] Sefik Ilkin Serengil and Alper Ozpinar. Hyperextended lightface: A facial attribute analysis framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–4. IEEE, 2021.
- [300] Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

- [301] Xingyu Ren, Alexandros Lattas, Baris Gecer, Jiankang Deng, Chao Ma, and Xiaokang Yang. Facial geometric detail recovery via implicit representation. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, 2023.
- [302] Stability AI. Introducing stable diffusion 3.5. <https://stability.ai/news/introducing-stable-diffusion-3-5>, October 2024. (updated 2024-10-29 (Medium release)).
- [303] Wendy Brown. Neo-liberalism and the end of liberal democracy. *Theory & event*, 7(1), 2003.
- [304] Wendy Brown. *Undoing the demos: Neoliberalism’s stealth revolution*. Princeton University Press, Princeton, NJ, 2015.
- [305] Michel Foucault. *Security, territory, population: lectures at the Collège de France, 1977-78*. Springer, 2007.
- [306] Sarah Banet-Weiser. Keynote address: Media, markets, gender: Economies of visibility in a neoliberal moment. *The Communication Review*, 18(1):53–70, 2015.
- [307] Sarah Banet-Weiser, Rosalind Gill, and Catherine Rottenberg. Postfeminism, popular feminism and neoliberal feminism? sarah banet-weiser, rosalind gill and catherine rottenberg in conversation. *Feminist theory*, 21(1):3–24, 2020.
- [308] E Said. Orientalism pantheon books. *New York*, 1978.
- [309] Carol Appadurai Breckenridge and Peter Van der Veer. Orientalism and the postcolonial predicament: Perspectives on. *South Asia*, 1993.
- [310] Stuart Hall, Jessica Evans, and Sean Nixon, editors. *Representation: Cultural Representations and Signifying Practices*. Culture, Media and Identities. Sage in association with The Open University, London; Thousand Oaks, CA, 1st edition, 1997.
- [311] Tessa McWatt. *Shame on Me: An Anatomy of Race and Belonging*. Random House Canada, 2020.
- [312] Frantz Fanon. Black skin, white masks [peau noire, masques blancs.], trans. C. L. Markmann. *London: Pluto*, 1967.
- [313] Sandra Harding. Rethinking standpoint epistemology: What is “strong objectivity”? In *Feminist epistemologies*, pages 49–82. Routledge, 2013.
- [314] Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. Participation is not a design fix for machine learning. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–6, 2022.
- [315] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. Power to the people? opportunities and challenges for participatory ai. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–8, 2022.
- [316] Fernando Delgado, Solon Barocas, and Karen Levy. An uncommon task: Participatory design in legal ai. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–23, 2022.
- [317] Tahu Kukutai and John Taylor. *Indigenous data sovereignty: Toward an agenda*. ANU press, 2016.

- [318] Vivetha Thambinathan and Elizabeth Anne Kinsella. Decolonizing methodologies in qualitative research: Creating spaces for transformative praxis. *International journal of qualitative methods*, 20:16094069211014766, 2021.
- [319] Alfred L Kroeber. *Anthropology: Race, Language, Culture, Psychology, Prehistory*. Harcourt, Brace and Company, 1948.
- [320] Audrey Smedley and Brian D Smedley. Race as biology is fiction, racism as a social problem is real: Anthropological and historical perspectives on the social construction of race. *American psychologist*, 60(1):16, 2005.
- [321] James F Wilson, Michael E Weale, Alice C Smith, Fiona Gratrix, Benjamin Fletcher, Mark G Thomas, Neil Bradman, and David B Goldstein. Population genetic structure of variable drug response. *Nature genetics*, 29(3):265–269, 2001.
- [322] Lily Hu and Issa Kohler-Hausmann. What’s sex got to do with machine learning? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 513–513, 2020.
- [323] Amina A Abdu, Irene V Pasquetto, and Abigail Z Jacobs. An empirical analysis of racial categories in the algorithmic fairness literature. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1324–1333, 2023.
- [324] P Jonathon Phillips, Hyeonjoon Moon, Syed A Rizvi, and Patrick J Rauss. The feret evaluation methodology for face-recognition algorithms. *IEEE Transactions on pattern analysis and machine intelligence*, 22(10):1090–1104, 2000.
- [325] Karl Ricanek and Tamirat Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *7th international conference on automatic face and gesture recognition (FGR06)*, pages 341–345. IEEE, 2006.
- [326] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5810–5818, 2017.
- [327] Ron Kohavi et al. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, pages 202–207, 1996.
- [328] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica*, 2016.
- [329] Mei Wang and Weihong Deng. Mitigate bias in face recognition using skewness-aware reinforcement learning, 2019.
- [330] Pavan Ravishankar, Qingyu Mo, Edward McFowland III, and Daniel B. Neill. Provable detection of propagating sampling bias in prediction models, 2023.
- [331] Seyma Yucer, Furkan Tektas, Noura Al Moubayed, and Toby Breckon. Racial bias within face recognition: A survey. *ACM Computing Surveys*, 57(4):1–39, 2024.
- [332] Sun-ha Hong. Prediction as extraction of discretion. *Big Data & Society*, 10(1):20539517231171053, 2023.
- [333] Camille Gear Rich. Elective race: Recognizing race discrimination in the era of racial self-identification. *Geo. LJ*, 102:1501, 2013.

- [334] Michelle L Stock, Frederick X Gibbons, Janine B Beekman, Kipling D Williams, Laura S Richman, and Meg Gerrard. Racial (vs. self) affirmation as a protective mechanism against the effects of racial exclusion on negative affect and substance use vulnerability among black young adults. *Journal of Behavioral Medicine*, 41:195–207, 2018.
- [335] Kate Crawford and Trevor Paglen. Excavating ai: The politics of images in machine learning training sets. *Ai & Society*, 36(4):1105–1116, 2021.
- [336] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.
- [337] Zaid Khan and Yun Fu. One label, one billion faces: Usage and consistency of racial categories in computer vision. In *Proceedings of the 2021 acm conference on fairness, accountability, and transparency*, pages 587–597, 2021.
- [338] Jennifer Mickel. Racial/ethnic categories in ai and algorithmic fairness: Why they matter and what they represent. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2484–2494, 2024.
- [339] Sebastian Benthall and Bruce D Haynes. Racial categories in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 289–298, 2019.
- [340] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 501–512, 2020.
- [341] Clara Belitz, Jaclyn Ocumpaugh, Steven Ritter, Ryan S Baker, Stephen E Fancsali, and Nigel Bosch. Constructing categories: Moving beyond protected classes in algorithmic fairness. *Journal of the Association for Information Science and Technology*, 74(6):663–668, 2023.
- [342] Sofia Jaime and Christoph Kern. Ethnic classifications in algorithmic fairness: Concepts, measures and implications in practice. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 237–253, 2024.
- [343] Ajmal, Tananant Boonya-Ananta, Andres J Rodriguez, VN Du Le, and Jessica C Ramella-Roman. Monte carlo analysis of optical heart rate sensors in commercial wearables: the effect of skin tone and obesity on the photoplethysmography (ppg) signal. *Biomedical optics express*, 12(12):7445–7457, 2021.
- [344] Matthew Groh, Caleb Harris, Roxana Daneshjou, Omar Badri, and Arash Koochek. Towards transparency in dermatology image datasets with skin tone annotations by experts, crowds, and an algorithm. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–26, 2022.
- [345] Geoffrey Currie, Johnathan Hewis, Elizabeth Hawk, and Eric Rohren. Gender and ethnicity bias of text-to-image generative artificial intelligence in medical imaging, part 1: preliminary evaluation. *Journal of Nuclear Medicine Technology*, 52(4):356–359, 2024.
- [346] Seyma Yucer, Furkan Tektas, Noura Al Moubayed, and Toby P. Breckon. Measuring hidden bias within face recognition via racial phenotypes. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, page 3202–3211. IEEE, January 2022.

- [347] Seyma Yucer, Amir Atapour Abarghouei, Noura Al Moubayed, and Toby P Breckon. Disentangling racial phenotypes: Fine-grained control of race-related facial phenotype characteristics. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–10. IEEE, 2024.
- [348] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning*, pages 2564–2572. PMLR, 2018.
- [349] Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1939–1948. PMLR, 2018.
- [350] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.
- [351] Raul Vicente Garcia, Lukasz Wandzik, Louisa Grabner, and Joerg Krueger. The harms of demographic bias in deep face recognition research. In *2019 international conference on biometrics (ICB)*, pages 1–6. IEEE, 2019.
- [352] Siyao Fu, Haibo He, and Zeng-Guang Hou. Learning race from face: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 36(12):2483–2509, 2014.
- [353] Tianwei Chen, Yusuke Hirota, Mayu Otani, Noa Garcia, and Yuta Nakashima. Would deep generative models amplify bias in future models? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10833–10843, 2024.
- [354] Maan Qraitem, Kate Saenko, and Bryan A Plummer. Bias mimicking: A simple sampling approach for bias mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20311–20320, 2023.
- [355] Noa Garcia, Yusuke Hirota, Yankun Wu, and Yuta Nakashima. Uncurated image-text datasets: Shedding light on demographic bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6957–6966, 2023.
- [356] Moreno D’Inca, Elia Peruzzo, Massimiliano Mancini, Dejia Xu, Vidit Goel, Xingqian Xu, Zhangyang Wang, Humphrey Shi, and Nicu Sebe. Openbias: Open-set bias detection in text-to-image generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12225–12235, 2024.
- [357] Sungho Park and Hyeran Byun. Fair-vpt: Fair visual prompt tuning for image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12268–12278, 2024.
- [358] Sree Harsha Nelaturu, Nishaanth Kanna Ravichandran, Cuong Tran, Sara Hooker, and Ferdinando Fioretto. On the fairness impacts of hardware selection in machine learning. In *Forty-first International Conference on Machine Learning*, 2024.
- [359] ProPublica. Compas analysis github repository, 2016.
- [360] Alexander Soen, Hisham Husain, and Richard Nock. Fair densities via boosting the sufficient statistics of exponential families. In *International Conference on Machine Learning*, pages 32105–32144. PMLR, 2023.
- [361] Omid Memarrast, Linh Vu, and Brian D Ziebart. Superhuman fairness. In *International Conference on Machine Learning*, pages 24420–24435. PMLR, 2023.

- [362] Harvineet Singh, Matthäus Kleindessner, Volkan Cevher, Rumi Chunara, and Chris Russell. When do minimax-fair learning and empirical risk minimization coincide? In *International Conference on Machine Learning*, pages 31969–31989. PMLR, 2023.
- [363] Zhaowei Zhu, Yuanshun Yao, Jiankai Sun, Hang Li, and Yang Liu. Weak proxies are sufficient and preferable for fairness with missing sensitive attributes. In *International Conference on Machine Learning*, pages 43258–43288. PMLR, 2023.
- [364] Jaakko Peltonen, Wen Xu, Timo Nummenmaa, and Jyrki Nummenmaa. Fair neighbor embedding. In *International Conference on Machine Learning*, pages 27564–27584. PMLR, 2023.
- [365] Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. Improving fair training under correlation shifts. In *International Conference on Machine Learning*, pages 29179–29209. PMLR, 2023.
- [366] Kwangho Kim and José R Zubizarreta. Fair and robust estimation of heterogeneous treatment effects for policy learning. In *International Conference on Machine Learning*, pages 16997–17014. PMLR, 2023.
- [367] Ann-Kristin Becker, Oana Dumitrasc, and Klaus Broelemann. Standardized interpretable fairness measures for continuous risk scores. In *Forty-first International Conference on Machine Learning*, 2024.
- [368] Alexandru Tifrea, Preethi Lahoti, Ben Packer, Yoni Halpern, Ahmad Beirami, and Flavien Prost. Frappé: A group fairness framework for post-processing everything. In *Forty-first International Conference on Machine Learning*, 2024.
- [369] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems*, 34:6478–6490, 2021.
- [370] Ruicheng Xian, Lang Yin, and Han Zhao. Fair and optimal classification via post-processing. In *International Conference on Machine Learning*, pages 37977–38012. PMLR, 2023.
- [371] Mohammad Mahdi Khalili, Xueru Zhang, and Mahed Abroshan. Loss balancing for fair supervised learning. In *International Conference on Machine Learning*, pages 16271–16290. PMLR, 2023.
- [372] Nikola Jovanović, Mislav Balunovic, Dimitar Iliev Dimitrov, and Martin Vechev. Fare: Provably fair representation learning with practical certificates. In *International Conference on Machine Learning*, pages 15401–15420. PMLR, 2023.
- [373] Nastaran Okati, Stratis Tsirtsis, and Manuel Gomez Rodriguez. On the within-group fairness of screening classifiers. In *International Conference on Machine Learning*, pages 26495–26516. PMLR, 2023.
- [374] Marina Knittel, Max Springer, John P Dickerson, and MohammadTaghi Hajiaghayi. Generalized reductions: making any hierarchical clustering fair and balanced with low cost. In *International Conference on Machine Learning*, pages 17218–17242. PMLR, 2023.
- [375] Vijay Keswani, Anay Mehrotra, and L Elisa Celis. Fair classification with partial feedback: An exploration-based data-collection approach. *arXiv preprint arXiv:2402.11338*, 2024.
- [376] Khalil Khan, Massimo Mauro, and Riccardo Leonardi. Multi-class semantic segmentation of faces. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 827–831. IEEE, 2015.

- [377] Lujing Zhang, Aaron Roth, and Linjun Zhang. Fair risk control: A generalized framework for calibrating multi-group fairness risks. *arXiv preprint arXiv:2405.02225*, 2024.
- [378] Haewon Jeong, Hao Wang, and Flavio P Calmon. Fairness without imputation: A decision tree approach for fair prediction with missing values. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9558–9566, 2022.
- [379] Wael Alghamdi, Hsiang Hsu, Haewon Jeong, Hao Wang, P Winston Michalak, Shahab Asoodeh, and Flavio P Calmon. Beyond adult and compas: Fairness in multi-class prediction. *arXiv preprint arXiv:2206.07801*, 2022.
- [380] Linda F Wightman. LSAC national longitudinal bar passage study. LSAC research report series. *LSAC Research Report Series*, 1998. Accessed: 2025-01-15.
- [381] Shizhou Xu and Thomas Strohmer. Fair data representation for machine learning at the pareto frontier. *Journal of Machine Learning Research*, 24(331):1–63, 2023.
- [382] Ruicheng Xian, Qiaobo Li, Gautam Kamath, and Han Zhao. Differentially private post-processing for fair regression, 2024.
- [383] Luca Giuliani, Eleonora Misino, and Michele Lombardi. Generalized disparate impact for configurable fairness solutions in ml. In *International Conference on Machine Learning*, pages 11443–11458. PMLR, 2023.
- [384] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500, 2019.
- [385] Gezheng Xu, Qi Chen, Charles Ling, Boyu Wang, and Changjian Shui. Intersectional unfairness discovery. *arXiv preprint arXiv:2405.20790*, 2024.
- [386] Patrick Simon. The failure of the importation of ethno-racial statistics in europe: Debates and controversies. *Ethnic and Racial Studies*, 40(13):2326–2332, 2017.
- [387] Sayaka Osanami Törngren. Challenging the ‘swedish’ and ‘immigrant’ dichotomy: How do multiracial and multi-ethnic swedes identify themselves? *Journal of Intercultural Studies*, 41(4):457–473, 2020.
- [388] Dan Rodríguez-García, Miguel Solana, Anna Ortiz, and Beatriz Ballestín. Blurring of colour lines? ethnoracially mixed youth in spain navigating identity. *Journal of Ethnic and Migration Studies*, 47(4):838–860, 2021.
- [389] European Union. Regulation (eu) 2016/679 of the european parliament and of the council, April 2016. Accessed: 2025-01-09.
- [390] Charles Westin. Young people of migrant origin in sweden. *International Migration Review*, 37(4):987–1010, 2003.
- [391] Patrick Simon. The choice of ignorance: The debate on ethnic and racial statistics in france. *Social statistics and ethnic diversity: Cross-national perspectives in classifications and identity politics*, pages 65–87, 2015.
- [392] European Commission. Data collection in the field of ethnicity, 2017. Accessed: 2025-01-09.

- [393] Laure Bereni, Renaud Epstein, and Manon Torres. Colour-blind diversity: how the—“diversity label” reshaped anti-discrimination policies in three french local governments. In *Diversity in Local Political Practice*, pages 14–32. Routledge, 2021.
- [394] Sayaka Osanami Törngren, Nahikari Irastorza, and Dan Rodríguez-García. Understanding multiethnic and multiracial experiences globally: Towards a conceptual framework of mixedness. *Journal of Ethnic and Migration Studies*, 47(4):763–781, 2021.
- [395] CJEU. Judgment of 6 April 2017, Jyske Finans a/s v. Ligebehandlingsnævnet, 2017. C-668/15 ECLI:EU:C:2017:278.
- [396] European Commission, Directorate-General for Justice, Consumers, European network of legal experts in gender equality, non discrimination, I. Chopin, and C. Germaine. *A comparative analysis of non-discrimination law in Europe 2023 – The 27 EU Member States compared*. Publications Office of the European Union, 2024.
- [397] European Commission. 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, 2021. Accessed: 2024-01-24.
- [398] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Why fairness cannot be automated: Bridging the gap between eu non-discrimination law and ai. *Computer Law & Security Review*, 41:105567, 2021.
- [399] Alex Engler. The EU and US diverge on AI regulation: A transatlantic comparison and steps to alignment, 2023. Accessed: 24-01-2025.
- [400] American Anthropological Association. AAA Statement on Race. *Anthropology Newsletter*, 39(9):3, 1998.
- [401] Oxford English Dictionary. Mixed-race, 2025. Accessed: 28-Jan-2025.
- [402] Jessica D Remedios and Alison L Chasteen. Finally, someone who “gets” me! multiracial people value others’ accuracy about their race. *Cultural Diversity and Ethnic Minority Psychology*, 19(4):453, 2013.
- [403] Nina Camara. Lost in otherness: Growing up as a mixed-raced child in eastern europe, 2016. Accessed: 2025-01-02.
- [404] Suki Ali. *Mixed-race, post-race: Gender, new ethnicities and cultural practices*. Routledge, 2020.
- [405] Marissa Floro. *In between: what the experiences of biracial, bisexual women tell us about identity formation*. PhD thesis, Loyola university chicago, 2018.
- [406] Tom Brocket. From “in-betweenness” to “positioned belongings”: Second-generation palestinian-americans negotiate the tensions of assimilation and transnationalism. *Ethnic and Racial Studies*, 43(16):135–154, 2020.
- [407] Leila Nilipour. Cynthia arrieu-king’s “the betweens” offers refreshing, nuanced perspective on mixed-race identity. *The Stanford Daily*, May 2021. Accessed: 2025-01-02.
- [408] Kierra L Miles. What is enough?: Understanding the ostracization of mixed people and how they reaffirm their identity. *The Macksey Journal*, 1(1), 2020.

- [409] Karly S Ford, Ashley N Patterson, and Marc P Johnston-Guerrero. Monoracial normativity in university websites: Systematic erasure and selective reclassification of multiracial students. *Journal of Diversity in Higher Education*, 14(2):252, 2021.
- [410] United States Census Bureau. Pums data. <https://www.census.gov/programs-surveys/acs/microdata/access.html>. Accessed: 2025-01-21.
- [411] Debra Thompson. Making (mixed-)race: Census politics and the emergence of multiracial multiculturalism in the united states, great britain and canada. In *Accounting for Ethnic and Racial Diversity*, pages 53–70. Routledge, 2013.
- [412] United States Census Bureau. Improved race and ethnicity measures reveal united states population is much more multiracial. <https://www.census.gov/library/stories/2021/08/improved-race-ethnicity-measures-reveal-united-states-population-much.html>, 2021. Accessed: 2025-01-11.
- [413] Alice Robbin. Classifying racial and ethnic group data in the united states: The politics of negotiation and accommodation. *Journal of Government Information*, 27(2):129–156, 2000.
- [414] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Saurabh Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *International Conference on Machine Learning (ICML)*, 2021.
- [415] Debbie S Ma, Joshua Correll, and Bernd Wittenbrink. The chicago face database: A free stimulus set of faces and norming data. *Behavior research methods*, 47:1122–1135, 2015.
- [416] Debbie S Ma, Justin Kantner, and Bernd Wittenbrink. Chicago face database: Multiracial expansion. *Behavior Research Methods*, 53:1289–1300, 2021.
- [417] Cambridge Dictionary. Reification - definition in the cambridge english dictionary, 2025. Accessed: 20-Jan-2025.
- [418] Deborah Poole. An excess of description: Ethnography, race, and visual technologies. *Annual Review of Anthropology*, 34(1):159–179, 2005.
- [419] Emanuel D Moss. *The objective function: Science and society in the age of machine intelligence*. PhD thesis, City University of New York, 2021.
- [420] Dora Zhao, Jerone TA Andrews, Orestis Papakyriakopoulos, and Alice Xiang. Position: measure dataset diversity, don’t just claim it. *arXiv preprint arXiv:2407.08188*, 2024.
- [421] Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Evaluating societal representations in diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.
- [422] Sourojit Ghosh, Nina Lutz, and Aylin Caliskan. “i don’t see myself represented here at all”: User experiences of stable diffusion outputs containing representational harms across gender identities and nationalities. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 463–475, 2024.
- [423] Mi Zhou, Vibhanshu Abhishek, Timothy Derdenger, Jaymo Kim, and Kannan Srinivasan. Bias in generative ai. *arXiv preprint arXiv:2403.02726*, 2024.

- [424] Huggingface. Stable diffusion 3.5 large, 2024. Accessed: 2025-01-26.
- [425] Midjourney. Midjourney. AI-powered image generation platform, 2024.
- [426] OpenAI. Dall-e 3. AI image generation model, 2024.
- [427] Oxford English Dictionary. mulatto, n. & adj. Oxford English Dictionary Online, 2025.
- [428] Michael Yudell, Dorothy Roberts, Rob DeSalle, and Sarah Tishkoff. Taking race out of human genetics. *Science*, 351(6273):564–565, 2016.
- [429] Evan K Rose. A constructivist perspective on empirical discrimination research. *Journal of Economic Literature*, 61(3):906–923, 2023.
- [430] Paula Braveman and Tyan Parker Dominguez. Abandon “race.” focus on racism. *Frontiers in Public Health*, 9:689462, 2021.
- [431] Mina Cikara, Joel E Martinez, and Neil A Lewis Jr. Moving beyond social categories by incorporating context in social psychological theory. *Nature Reviews Psychology*, 1(9):537–549, 2022.
- [432] Lily Hu and Issa Kohler-Hausmann. What is perceived when race is perceived and why it matters for causal inference and discrimination studies. *Law & Society Review*, pages 1–26, 2024.
- [433] Camara Phyllis Jones, Benedict I Truman, Laurie D Elam-Evans, Camille A Jones, Clara Y Jones, Ruth Jiles, Susan F Rumisha, and Geraldine S Perry. Using “socially assigned race” to probe white advantages in health status. *Ethnicity & disease*, 18(4):496–504, 2008.
- [434] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 339–348, 2019.
- [435] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. *Advances in neural information processing systems*, 30, 2017.
- [436] Dominik Hangartner, Daniel Kopp, and Michael Siegenthaler. Monitoring hiring discrimination through online recruitment platforms. *Nature*, 589(7843):572–576, 2021.
- [437] Benedikt Höltingen and Robert C Williamson. On the richness of calibration. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1124–1138, 2023.
- [438] Terry D Brown Jr, Francis C Dane, and Marcus D Durham. Perception of race and ethnicity. *Journal of Social Behavior & Personality*, 13(2), 1998.
- [439] Judith Butler. *Bodies that matter: On the discursive limits of sex*. Routledge, 2011.
- [440] Wendy D Roth. The multiple dimensions of race. *Ethnic and Racial Studies*, 39(8):1310–1338, 2016.
- [441] Eoin Travers, Merle T Fairhurst, and Ophelia Deroy. Racial bias in face perception is sensitive to instructions but not introspection. *Consciousness and cognition*, 83:102952, 2020.

- [442] E Darcy Burgund, Yiyang Zhao, Inaya N Laubach, and Eyerusalem F Abebaw. Different features for different races: Tracking the eyes of asian, black, and white participants viewing asian, black, and white faces. *Plos one*, 19(9):e0310638, 2024.
- [443] Daniel T Levin and Mahzarin R Banaji. Distortions in the perceived lightness of faces: the role of race categories. *Journal of Experimental Psychology: General*, 135(4):501, 2006.
- [444] Daniel T Levin. Race as a visual feature: using visual search and perceptual discrimination tasks to understand face categories and the cross-race recognition deficit. *Journal of Experimental Psychology: General*, 129(4):559, 2000.
- [445] OpenAI. Chatgpt-4. <https://openai.com>, 2023.
- [446] Google AI. Gemini 2.0 flash. <https://ai.google.dev/gemini-api/docs/models/gemini-v2>, 2024.
- [447] Michele Merler, Nalini Ratha, Rogerio S. Feris, and John R. Smith. Diversity in faces, 2019.
- [448] Adam Harvey and Jules LaPlace. Exposing.ai. Website, 2021.
- [449] Seyma Yucer, Samet Akçay, Noura Al-Moubayed, and Toby P Breckon. Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 18–19, 2020.
- [450] Cynthia Feliciano. Shades of race: How phenotype and observer characteristics shape racial classification. *American Behavioral Scientist*, 60(4):390–419, 2016.
- [451] Abdulla Fakhro, Hyung Woo Yim, Yong Kyu Kim, and Anh H Nguyen. The evolution of looks and expectations of asian eyelid and eye appearance”. In *Seminars in plastic surgery*, volume 29, pages 135–144. Thieme Medical Publishers, 2015.
- [452] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018.
- [453] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 692–702, 2019.
- [454] Bowen Zhang, Shuyang Gu, Bo Zhang, Jianmin Bao, Dong Chen, Fang Wen, Yong Wang, and Baining Guo. Styleswin: Transformer-based gan for high-resolution image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11304–11314, 2022.
- [455] Markos Georgopoulos, James Oldfield, Mihalis A Nicolaou, Yannis Panagakis, and Maja Pantic. Mitigating demographic bias in facial datasets with style-based multi-attribute transfer. *International Journal of Computer Vision*, 129(7):2288–2307, 2021.
- [456] Linzhi Huang, Mei Wang, Jiahao Liang, Weihong Deng, Hongzhi Shi, Dongchao Wen, Yingjie Zhang, and Jian Zhao. Gradient attention balance network: Mitigating face recognition racial bias via gradient attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 38–47, 2023.

- [457] Kagan Ozturk, Haiyu Wu, and Kevin W Bowyer. Can the accuracy bias by facial hairstyle be reduced through balancing the training data? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1519–1528, 2024.
- [458] Audre Lorde. *Sister outsider: Essays and speeches*. Crossing Press, 2012.
- [459] Nidhi Trehan and Angéla Kóczé. Racism,(neo-) colonialism, and social justice: the struggle for the soul of the romani movement in post-socialist europe. *Racism Postcolonialism Europe*, pages 50–73, 2009.
- [460] Liz Fekete. Europe against the roma. *Race & Class*, 55(3):60–70, 2014.
- [461] Carl Levy. Racism, immigration and new identities in italy. In *The Routledge handbook of contemporary Italy*, pages 49–63. Routledge, 2015.
- [462] Faith Macharia. The socio-cultural impacts of ethno-racism in italy as functions of the global economic north/south divide. *The Mellon Mays Undergraduate Fellowship Journal* 2017, page 89, 2017.
- [463] Eric Corbett, Emily Denton, and Sheena Erete. Power and public participation in ai. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–13, 2023.
- [464] Elena Ball, Melanie C Steffens, and Claudia Niedlich. Racism in europe: characteristics and intersections with other social categories. *Frontiers in Psychology*, 13:789661, 2022.
- [465] Kim M Williams. From civil rights to the multiracial movement. *New faces in a changing America: Multiracial identity in the 21st century*, pages 85–98, 2003.
- [466] Elisabeth Eide. Strategic essentialism. *The Wiley Blackwell Encyclopedia of gender and sexuality studies*, pages 2278–2280, 2016.
- [467] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.
- [468] Olihe N Okoro, Vibhuti Arya, Caroline A Gaither, and Adati Tarfa. Examining the inclusion of race and ethnicity in patient cases. *American journal of pharmaceutical education*, 85(9):8583, 2021.
- [469] Anamaria Crisan, Margaret Drouhard, Jesse Vig, and Nazneen Rajani. Interactive model cards: A human-centered approach to model documentation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 427–439, 2022.
- [470] Manish Raghavan and Pauline T Kim. Limitations of the” four-fifths rule” and statistical parity tests for measuring fairness. *Georgetown Law Technology Review*, 8:93, 2024.
- [471] Aaron Pellish and Brian Stelter. Tiktok shuts down in the united states hours ahead of a ban. *CNN*, January 2025.
- [472] Tom Gerken, Liv McMahon, and Imram Rahman-Jones. What does trump’s executive order mean for tiktok? *BBC*, January 2025.

- [473] Robert Wolfe, Yiwei Yang, Bill Howe, and Aylin Caliskan. Contrastive Language-Vision AI Models Pretrained on Web-Scraped Multimodal Data Exhibit Sexual Objectification Bias. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*, pages 1174–1185, New York, NY, USA, June 2023. Association for Computing Machinery.
- [474] The Washington Post. This is how ai image generators see the world. The Washington Post, November 2023.
- [475] Douglas Guilbeault, Solène Delecourt, Tasker Hull, Bhargav Srinivasa Desikan, Mark Chu, and Ethan Nadler. Online images amplify gender bias. *Nature*, 626(8001):1049–1055, February 2024.
- [476] Xingyu Lan, Jiaxi An, Yisu Guo, Chiyong Tong, Xintong Cai, and Jun Zhang. Imagining the far east: Exploring perceived biases in ai-generated images of east asian women, 2025.
- [477] Bloomberg News. Text-to-image tools makes cool art, but can conjure nsfw. Bloomberg, October 2022.
- [478] Gant Laborde. Deep nn for nsfw detection.
- [479] Gianluca Mauro and Hilke Schellmann. 'there is no standard': investigation finds ai algorithms objectify women's bodies. The Guardian, February 2023.
- [480] US Office of Management and Budget. Omb statistical policy directive no. 15: 1997 standards for maintaining, collecting, and presenting federal data on race and ethnicity, 1997.
- [481] Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996.
- [482] Michael Redmond and Alok Baveja. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 141(3):660–678, 2002.
- [483] Silonie Sachdeva. Fitzpatrick skin typing: Applications in dermatology. *Indian journal of dermatology, venereology and leprology*, 75:93, 2009.