A Subject Based Methodology for Measuring Interclass Bias in Facial Recognition Verification Systems’

By

Aramael Andrés Peña-Alcántara

B.S. Drama and Theatre Arts with a Concentration in Design,
Columbia College, Columbia University, 2017
B.S. Mechanical Engineering, The Fu Foundation School of
Engineering and Applied Science, Columbia University, 2017
S.M. Civil and Environmental Engineering, School of Engineering,
Massachusetts Institute of Technology, 2021

SUBMITTED TO THE CIVIL AND ENVIRONMENTAL ENGINEERING DEPARTMENT IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF SCIENCE IN CIVIL AND ENVIRONMENTAL ENGINEERING

AT THE

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2022

Signature of Author: ____
Civil and Environmental Engineering
May 13, 2022

Certified by: ____
Dr. John Williams, Thesis Supervisor
Professor, Civil and Environmental Engineering

Accepted by: ____
Dr. Colette L. Heald,
Professor of Civil and Environmental Engineering
Chair, Graduate Program Committee

This Page Left Intentionally Blank

A Subject Based Methodology for Measuring Interclass Bias in Facial Recognition Verification Systems'

By

Aramael Andrés Peña-Alcántara

Submitted to the Civil and Environmental Engineering Department on May 13, 2022, in partial fulfillment of the requirements for the degree of Doctor of Science in Civil and Environmental Engineering

ABSTRACT

Rapid progress in automated facial recognition has led to a proliferation of the use of algorithms to support decision-making in high-stakes applications, such as immigration and border control, hiring, and the criminal justice system. Recent research has uncovered serious concerns about equality and transparency in facial recognition algorithms, finding performance disparities between groups of people based on their phenotypes, such as gender presentation and skin tone. These challenges can result in loss of employment opportunities, extra scrutiny in transactions, and even loss of freedom, raising the need of deeper analysis of facial recognition's shortcomings.

This dissertation proposes a novel methodology and a general test statistic to measure facial recognition algorithm interclass bias. The test uses distance-based variance to capture shape-related differences in an algorithm's accuracy at multiple operating points.

The author assesses the performance of the test to evaluate the interclass bias for skin tone and gender, in commercial facial verification algorithms. Using a dermatologist-approved classification for skin tone system and a simple masculine and feminine classification for gender presentation, thirteen commercial-off-the-shelf facial verification algorithms are evaluated, utilizing a subset of the IARPA Janus Benchmark C dataset, and it's 1:1 verification protocol. The analyses show that darker-skinned people have the least accurate results, with interclass bias measures up to 7.2 times higher than lighter-skinned people. Additionally, the results show that one evaluated commercial facial verification algorithm statistically eliminates the interclass bias for skin tone. Yet, all thirteen commercial facial verification algorithms evaluated experienced worse performance for feminine presenting persons compared to masculine presenting persons. The author believes this new measure of interclass bias can be incorporated into an algorithm's design to remove this bias. The present biases in classifying darker-skinned and feminine presenting people require urgent attention, if commercial companies are to build genuinely equal, transparent, and accountable facial verification algorithms.

Thesis Supervisor: John Williams
Title: Professor of Civil and Environmental Engineering

Acknowledgements

Research reported in this dissertation was in part supported by funding from the Accenture Security Technology Group. Its contents are solely the responsibility of the authors and does not necessarily represent the official views of Accenture.

I am also indebted to all my colleagues, mentors, and friends who have helped develop these ideas over the years. There are too many to list, but I have tried to give them credit for the ideas they helped me through together. I apologize in advance if I have inadvertently not given credit where it is due.

To my parents, Dr. Feniosky Peña-Mora and Dr. Minosca Alcántara, thank you for giving me my drive to ask questions, you are the reason I am here today. To my sisters, Amnahir and Giramnah Peña-Alcántara, you are the greatest gift life has given me. Also, I know I can't wait to see you at your defenses.

To my thesis committee, Prof. John Williams, Prof. Admir Masic, Prof. Caitlin Mueller, thank you for setting high expectations, refining my thinking, sharing your knowledge, time, and insights. This dissertation is a reflection of your mentorship.

To Dr. Malek Ben Salem, thank you for your support, latitude, and guidance throughout this research.

To those who worked with me to pave the road to the opportunities now before me:

Caroline Bartles
Gabriel Calatrava
Santiago Calatrava
Dr. Darleny Cepin-Bailey
Prof. Maria Feng
Claudine Freard-Rose
Prof. Sandra Goldmark
Elwood Howard
Amy Jupiter
Dr. Tom Kelly
Prof. Mike Massimino
Richard Maxfield
Mike McCullough
Joe Muskin
Dr. Stephen Palfrey
Craig Russell
Dr. David Schiller
Joel Sherry
Prof. Thomas Sullivan
Dr. Jeffery Weitz
Prof. Nancy Workman
Jackie Ziff

To my colleagues, mentors, and professors who helped formulate my understanding of risk, safety, and ethics that underpin the ideas in this dissertation:

Mike Arevalo
Kristin Athans
Dr. Haluk Ay
Spencer Banks
Samantha Catanzaro
Michael Cawlina
Jordan Coleman
Kathy Drach
Tim East
Corinne Meade Gallagher
Stephen (Toaster) Gregory
Emily Holden
Holger Irmler
Przemyslaw Iwanowski
Jared Kleier
David Lester
Lauren Mertz
Sarah Nesbitt

Bob Peitzman
Dani Bellar Pilukas
Rodney Rayborne
Cody Shackelford
Dr. Emily Sparer-Fine
Robert Spielvogel
Jeremy Stolars
Raymond Warner III
April Warren
Chris Wessels
Janice Wiener

To all of my friends who have listened to me talk through this work, who directly and indirectly provided me the inspiration and valuable suggestions that resulted in this study:

Alex Randall
Chris Gabriel
Christina Kyrillios
Elizabeth Goodstein
Emerald Gearing
Emily Salvador
Emily Van Belleghem
José Ricardo Moreno
Julia Warren
Kameron Burke
Kyle Haley
Logan Light
Michael Tornato
Olivia El Sadr-Davis
Piyali Mukherjee
Sean Walsh
Simone Getty
Stefan Palios

To my colleagues at MIT who made the Institute a home, and helped me find my way through the infinite:

Dr. Angi Acocella
Andrew Foster
Elias Johnson
Dr. Bruce Jones
Dr. Sam Raymond
Emily Salvador
Stephan Stansfield
Dylan Wooton

Para Ramon Peña and Melchor Alcántara por inculcarme el valor de la educación e impulsarme a seguir adelante. A la próxima generación, ya tú tienes un mandato.

Table of Contents

Introduction

Recently, the proliferation of a particular application of artificial intelligence, facial recognition systems, into the public and private sphere has triggered an intense debate and critique for their differential treatment of various demographic groups, in particular along racial and gender lines (Furl et al. 2002; Fussell 2020; Garvie et al. 2016; Gentzel 2021; Grother et al. 2019; Harwell 2021; Lohr 2018; Norval and Prasopoulou 2017; Rhue 2018; Wang et al. 2018b). The author motivated by this active area of research, in algorithmic equality (fairness), focuses on one particular use of facial recognition systems for verification, where the system is asked to compare two images to verify if they are the same person. In other words, to compare faces as it may be done in an access control scenario (e.g., a security guard who is tasked with letting authorized persons into a facility).

Facial recognition systems have been the topic of intense research for over fifty years, but in the last decade, due to improvements to artificial intelligence in general, they have improved considerably (Guo and Zhang 2019; Li and Jain 2011). In order to understand facial recognition systems success in numerous applications for private companies and governments ranging from authentication services on mobile consumer devices to use by law enforcement or armed forces, one first needs to understand artificial intelligence (Garvie et al. 2016; Parge et al. 2021; Robertson et al. 2016; Smith et al. 2015).

The Rise of Artificial Intelligence Agents

AI, or Artificial Intelligence, needs no introduction. We have all come across it in some form; in science fiction stories, real life, or popular culture. Yet, AI remains one of the most elusive subjects in Computer Science. The term is frequently used without a clear definition. In both popular and research definitions, the term encapsulates machine’s ability to perform tasks that have been considered to require human intelligence. Perhaps it’s more useful to consider AI as a meta-construct to describe a variety of different technologies that enable machines to sense, comprehend, act, or learn in order to maximize its chance of achieving its goals.

Our contemporary understanding of AI includes technologies like machine learning, deep learning, natural language processing, and computer vision among others. Each one evolving along its own path and when applied in conjunction with data and automation can drive meaningful innovation and empower people.

When understood in this manner, it’s easy to note that AI is no longer part of some distant future in popular culture’s imagination, but rather an integrated technology already prevalent in many facets of our day. On the personal, individual level, AI is used to compose responses to our emails (Google 2017), to enable our cameras to take better pictures than the one we shoot (Apple 2019) or a camera that becomes its own photographer and seeks out the best moments to take pictures (Payne 2017), headphones that can instantly translate any foreign language to empower us to communicate smoothly (Timekettle 2020; Waverly Labs 2019), wrist watches that are able to call for assistance when we have a hard fall and cannot get up (Apple 2020; Samsung 2020), and thermostats that perfectly adjust our home's heating and cooling systems to our preferred temperatures (ecobee 2020; Nest Labs 2020). In each of these products, we have personally delegated our choices to the AI to make decisions, or actions either in digital or physical environments that we would have otherwise performed ourselves (Puntoni et al. 2021).These consumer AI technologies are a far cry from the "transhumanist" nightmares imagined in Marry Shelley's *Frankenstein; or, The Modern Prometheus*, Isaac Asimov's *I, Robot,* or James Cameron's *Terminator*, cautionary tales where technology is able to overcome fundamental human limitations and molds new ideals of technological perfection. Rather, and perhaps most importantly, each of these examples is considered unremarkably normal by consumers and users. These small delegation experiences, where we take advantage of AI's capability to act as a substitute for human labor, have primed us individually, and as a culture to embrace AI as an inevitable and indispensable tool.

As AI has become entrenched in consumer's lives, it has excelled is as an integrated technology deployed in a menagerie of sectors: finance, national security, health care, criminal justice, transportation, and smart cities (West and Allen 2018). At the corporate level, AI is used to optimize underwriting processes for re/insurance companies by drawing from large unstructured collections of data to understand risk-related insights for a business or customers (Andriotis 2019; Planck Re 2017; Upstart 2021; Waddell 2019); help medical insurers and pharmaceutical companies automate the prior authorization of medical procedures or filling costly prescription by deciding whether it's necessary for a patient to undergo a given procedure or take a particular medication (Lash Group 2018; Salian 2019); powering medical imaging platforms or testing regimens to detect high-risk patients earlier for healthcare providers (Freenome 2021; Putcha et al. 2020; Zebra Medical Vision 2021); writing marketing copy for online advertisements to better target customers (Ives 2019; Persado 2019; Phrasee 2021; Wilhelm 2021) or even orchestrating where these ads should be placed (The Trade Desk 2018); filter and recommend applicants for a job based on scanning resumes or understanding culture fits through additional metrics collected from third-party information (Leoforce 2019; pymetrics 2018). In each of these products, similar to personal uses of AI, corporations have partially automated roles by delegating to the AI decisions, or actions that would otherwise be performed by a human (for ease of reference we will describe these as AI agents), and in the process uncovered labor force productivity improvements. These widespread implementations of AI are transforming industries and societies writ large, with increased efficiencies and are projected to boost corporate profitability in 16 industries across 12 economies by an average of 38% by 2035 (Purdy and Daugherty 2017). This is expected to boost the global economic output by more than $15 trillion by 2030 (Rao and Verweij 2017).

Safe Guards Against Artificial Intelligence Agents

Yet while these AI implementations are able to learn in narrow ways, a pioneer in AI, Yoshua Bengio, complained in WIRED magazine "[AI needs] much more data to learn a task than human examples of intelligence, and they still make stupid mistakes" (Simonite 2019). This narrow learning can be conceptualized as requiring three distinct safeguards to ensure that the AI agent is working as intended:

1. AI agents require supervision to prevent mistakes throughout their development and production lifecycle.
2. AI agents require carefully constructed goals to solve intended problems
3. AI agents require carefully controlled datasets to learn how to solve intended problems.

AI agents require supervision throughout their lifecycle

The requirement that AI agents require supervision throughout their lifecycle is best exemplified by the following anecdote: an expert loader who packs pallets onto an aircraft can expertly align irregularly shaped pallets together or stack pallets multiple levels high; on the contrary, an AI system, deployed by DHL, trained to perform the same task makes mistakes, especially in the early stages of deployment, requiring human oversight, in the form of a human taking control of the AI controlled robotic arm for some time to improve the accuracy of the algorithm (Knight 2020). One can observe the same behavior even in more trivial examples, such as an AI learning how to play the two-dimensional table tennis video game Pong may let the ball fly past its paddle a few hundred times before learning that's not a good way of increasing its score (Reynolds 2017). Yet even after the AI would learn how to manipulate the paddle and improved its game play strategy it would periodically forget that lesson that letting the ball fly past the paddle is not an effective methodology and have to re-learn this fact (Lipton et al. 2018). This is all to say that AI agents require supervision to prevent mistakes throughout their lifecycle in both learning and execution environments.

Researchers have shown that when humans supervise AI agents, they frequently defer to the AI agent's decisions. Researchers have shown that after human supervisors have been told that face matching had been performed by AI agents, their supervisory accuracy deteriorates (Howard et al. 2020). That is to say, human supervisors' internal criteria when judging if two individuals are the same person is altered by knowing the AI agent's confidence of the match. This had regretful consequences in Detroit when the Police Department used grainy security footage of a suspect, who had stolen $3,800 worth of luxury timepieces, to find potential matches from the state's driver's license photos (Allyn 2020; Hill 2020a). The AI agent, developed by a company called DataWorks Plus, identified a match, and the police issued an arrest warrant based on it (Ryan-Mosley 2021). The man identified by the AI agent was taken to a detention center, photographed, fingerprinted, and held overnight. During questioning a detective showed the detained man, the photo of the suspect used by the AI agent, and asked "Is this you?" The blurry photo featured a heavyset man dressed in black standing in front of the watch display, but it was clearly not the detained man. “No, this is not me,” the detained man reportedly said. “You think all black men look alike?” The detective replied “The computer says it’s you.” The prosecutors dropped the case less than two weeks later, arguing that officers had relied on a bad match from the AI agent. Detroit Police Chief James Craig later apologized for what he called “shoddy” investigative work (Allyn 2020). Yet the harm was done, the wrongly accused man still sat in jail for hours, and the man’s daughters have been “traumatized” by the police officers arresting their father in front of them.

In another instance, a man in New Jersey was held in jail for ten days after he was falsely accused by an AI agent of shoplifting from a hotel gift shop in 2019 (Hill 2020b). This New Jersey man’s history with the criminal justice system compounded the effect, because it would have been his third felony, meaning he was at risk of a ten year prison sentence. The man almost took a guilty plea deal, because he was afraid to go to trial. The AI agents, though not culpable of the outcomes in these two cases, bear some responsibility. The police departments partially automated roles previously performed by police officers to AI agents, without fully supervising and controlling for mistakes that can be expected to happen throughout the AI agent’s execution. Understanding how to control for the complexities of the systems these AI agents are placed in, is explored in more detail near the end of this section.

AI agents require carefully constructed goals

AI systems are trained to exhibit superhuman performance. For instance, one video game may learn to intentionally kill itself in a way that *paradoxically* earns greater points in ways that a human gamer would never attempt (Bellemare et al. 2016). In an infamous event dramatized by the media (Beal and Jehring 2017; Field 2017; Kenna 2017; Sassoon 2017), Facebook AI Research ended a project where two AI robots equipped to simulate human-like English conversations via text (known colloquially as a “chatbot”) were challenged to negotiate with each other over a trade, attempting to swap hats, balls, and books. The chatbots quickly developed a kind of English shorthand that enabled them to execute successful negotiations; however, they failed to follow a traditional grammatical structure. The chatbots were skilled negotiators, and eventually learned to feign interest in an item just to sacrifice it later as a faux compromise. However, their exchange looked like gibberish to native English speakers (Lewis et al. 2017). An excerpt of the negotiation between the two chatbots is listed below:

Bob: i can i i everything else . . . .

Alice: balls have zero to me to me to me to me to me to me to me to me to

Bob: you i everything else . . . .

Alice: balls have a ball to me to me to me to me to me to me to me

Bob: i i can i i i everything else . . . .

Alice: balls have a ball to me to me to me to me to me to me to me

Bob: i . . . .

Alice: balls have zero to me to me to me to me to me to me to me to me to

Bob: you i i i i i everything else . . . .

Alice: balls have 0 to me to me to me to me to me to me to me to me to

Bob: you i i i everything else . . . .

Alice: balls have zero to me to me to me to me to me to me to me to me to

This may seem like an uneventful failure, but it underscores that when the AI agent was presented with the opportunity to learn language, it did; however, the chatbots decided it was more convenient to build their own dialect than to learn a vernacular English. Yet phenomenon of AI agents reworking the English language to better suit its purposes is not an isolated incident, the AI system behind Google Translate seems to have some sort of interlingua that encodes the semantic of the sentence that enable translations between languages (Johnson et al. 2017). Unlike the team at Facebook, the Google Translate team considered this a beneficial development.

In another AI agent, trained to distinguish between photos of Wolves and Huskies (i.e., Alaskan Malamutes, Siberian Huskies, Greenland Dogs, or other dogs who work as sled dogs in the polar regions). The AI agent performed successfully on its training data, and easily separated canines that are incredibly visually similar (Ribeiro et al. 2016). When the AI agent was presented with new images it was not trained on, it was discovered that the AI agent was making its decisions based on the backgrounds of the image not the dogs themselves. Images of wolves typically had a snowy background, and huskies were generally indoors or in less snowy climates. Instead of classifying the dogs in the images, the AI agent discovered an easier problem to solve: is there snow or a light colored background in the image? That is to say, the huskies classifying AI agent was unwittingly a snow detector. These examples illustrate why AI agents require carefully constructed goals in order to ensure that the AI actually solves the intended problem instead of a similar problem of its own definition.

AI agents require carefully controlled datasets

These mistakes (or potentially *innovations*, as that judgement is formed in the eyes of the beholder) caused by AI agents can also be attributed to the fact that AI agents require "more data to learn a task than human examples of intelligence" (Simonite 2019). That is to say that if there exist any undesirable issues inside the data that the AI agent uses to learn, then the AI agent will learn those issues. In 2016 Microsoft was preparing to release a new chatbot, Tay, designed to engage people in dialogue on Twitter (Microsoft 2016). Tay was designed to learn more about language over time enabling them to discover patterns of language and emulate Internet speech patterns. At first Tay engaged with a growing number of followers with banter and lame jokes, but within sixteen hours Tay tweeted statements like "Ricky Gervais learned totalitarianism from Adolf Hitler, the inventor of atheism" (Schwartz 2019). Microsoft immediately suspended the account. In the following weeks reports of how Tay became so vile emerged detailing how Tay who was trained on cleaned, filtered, and anonymized public data with editorial content provided by improvisational comedians became so vile (Microsoft 2016). A coordinated group of users exploited a "repeat after me" feature where Tay would repeat racist, misogynistic, and antisemitic language (Hunt 2016; Schwartz 2019). In most cases Tay was only repeating other users' reprehensible statements, but Tay's built in capacity for learning ensured that it learned from those interactions also and incorporated them into its future messages (Ohlheiser 2016). In this example, the AI agent was purposefully exploited to learn undesirable issues by incorporation of undesirable data; however, these concerns can manifest even without an adversary to exploit it. In 2015, Amazon retired an experimental algorithm designed to provide recommendations to recruiters to find talented candidates after discovering that the system penalized resumes that included "women's", as in "women's volleyball captain," penalized graduates of all-women's colleges, and favored candidates who described themselves using verbs that were more commonly found on male engineers' resumes, such as "executed" and "captured" (Dastin 2018; Hsu 2020).

In AI agents trained to identify people in photos, researchers have discovered that AI agents designed in Western countries recognized Caucasian faces more accurately than East Asian faces. Correspondingly, AI agents designed in East Asian countries recognized East Asian faces more accurately than Caucasian faces (Furl et al. 2002; Klare et al. 2012; Phillips et al. 2009). Researchers have theorized that these differences are due to the racial composition of the training datasets for Western and East Asian algorithms (Cavazos et al. 2020). That is to say, facial recognition systems that are trained within the narrow context of a specific dataset inevitably optimize to learn the specific attributes of that dataset. This narrow context creates systematic errors as the system skews towards learning those specific attributes; and the issue is believed to stem from under-representation or over-representation of groups in the dataset. These examples elucidate why AI agents require large and carefully controlled datasets to learn how to solve the intended problem in an acceptable fashion to avoid later discoveries of undesirable solutions.

Expanding Definitions of Algorithmic Errors to include Harm

As indicated earlier and as demonstrated through these examples, AI implementations are currently able to learn in only narrow ways, potentially creating accidents and creating new hazards. More importantly, as AI agents have been delegated partial or full decision-making power by individuals or corporations the consequences of these hazards can grow and can cause unintended consequences and undesirable losses. These losses may involve equipment, financial, and information losses but can also involve other major losses such as human death, injury, imprisonment, and reputational loss.

Taken all together, one can gather that AI agents could pose danger. Even the best-intentioned AI agents require careful design considerations and continued monitoring and management to ensure that they produce desirable solutions to the intended problem. Yet, we continue to empower AI agents with more decision-making power, and insert them into highly complex and highly coupled systems that are difficult for its designers to consider all the potential system states and handling all normal and abnormal situations safely and effectively. Though this is not a new situation, as humans have long adopted technologies before fully understanding their scientific underpinnings and engineering knowledge (Leveson 2016). Indeed, astronomer Carl Sagan, called "America's most effective salesman of science" by Time magazine, states this especially well:

> *Many of the dangers we face indeed arise from science and technology—but, more fundamentally, because we have become powerful without becoming commensurately wise. The world-altering powers that technology has delivered into our hands now require a degree of consideration and foresight that has never before been asked of us. (Sagan 1997)*

To gain this wisdom, developers of AI agents need to expand the definitions of errors, beyond simply errors that a system can make, to include the harms that the AI agent is capable of perpetuating. When AI agents are designed, they are often created with assumptions that they'll be used for benign purposes, or that checks and controls for their decisions will be monitored by a separate program, person, or organization[1]. However, we know from the examples provided earlier, that that these tools may be used in more hazardous ways than their developers envisioned. In these context's the AI agents can work as intended but still result in financial, reputational, and injurious losses. The problem in these instances is the overall system design. However, attempts to restructure criminal justice systems, loan servicing providers, corporate labor practices, or any of these societal issues exacerbated by AI are beyond the scope of any engineering dissertation. Efforts to re-conceptualize how end-users attempt to use and employ AI agents, when their use is so commonplace and unremarkable is similarly a Sisyphean task[2]. Yet, the hazards in these systemic issues, once acknowledged, are uniquely capable of being addressed by AI developers; because in these particular cases increasing the reliability of the AI

[1] Amazon Web Services ("AWS"), a commercial off the shelf provider of facial recognition AI tools, service terms denote that "Law Enforcement Agencies that use Amazon Rekognition [AWS' trade name for its facial recognition AI agent] to assist personnel in making decisions that could impact civil liberties or equivalent human rights must ensure such personnel receive appropriate training on responsible use of facial recognition systems…" (Amazon Web Services, Inc 2021). However, it's important to note that while AWS recommends trainings, it does not undertake any efforts to provide that training itself or content standards.

[2] Any attempts put the genie back in the bottle, are likely to fail in this authors opinion. AI is so widespread and has helped individuals and business realize tremendous efficiencies, that few would entertain the idea of giving up those productivity gains. This idea is further explored in the section entitled ***A Brief Aside on the Nature of Risk in the Adoption of Technology*** in the ***Appendix***.

agents or protecting against the unintended consequences would have prevented these hazards because the components created hazardous conditions though they did not fail.

It is therefore imperative to adjust developers of AI agents' understanding of what a failure consist of; in order to encapsulate the risks visible only at a higher level of hierarchy developers' may not be exposed to. Principal to this effort is the incorporation of the actual uses and reasonably expected uses of the system by both the system's end-users and owners to the enumeration of failures.

Investigative Focus of this Dissertation

This dissertation is motivated by these larger problems concerning the use of AI agents, and demonstrates how developers can better understand failures of their system. In particular, the author focuses on one particular use of AI agents: facial recognition systems for verification, where the AI agent is asked to compare two images to verify if they are the same person. In other words, to compare faces as it may be done in an access control scenario (e.g., a security guard who is tasked with letting authorized persons into a facility). These facial recognition systems have been an active problem in computer science for over half a century, but recent improvements to AI in general, have given rise to numerous applications of facial recognition systems for companies and governments. It has been useful in the private sector to secure a nuclear research facility (Scheeres 2002), to protect an artist from stalkers at her shows (Knopper 2018), in the public sector to confirm the identity of Osama bin Laden (Reuters Staff 2011), and convicting an armed thief in Chicago (Main 2014). Yet, this deployment has triggered an intense debate and critique for their differential treatment of various demographic groups, in particular along racial and gender lines (Furl et al. 2002; Fussell 2020; Garvie et al. 2016; Gentzel 2021; Grother et al. 2019; Harwell 2021; Lohr 2018; Norval and Prasopoulou 2017; Rhue 2018; Wang et al. 2018b).

While in general, one cannot expect precise numbers quantifying the risks due to AI agents, as the risks are complex, unpredictable and generally cannot be approximated by a long-term frequency or simple statistical analysis. Here the author builds a methodology to quantify the risks associated with a specific AI agent, namely, a facial recognition system performing a verification task, and the specific harm of failing to be recognized due to one's gender or skin tone. That is to say, the author proposes a methodology to measure facial recognition systems interclass bias within a classification schema.

The author believes this measure of interclass bias will engender comprehensive analyses of facial recognition systems verification algorithms biases that can be incorporated into an algorithm's design, implementation, or training processes and an end user's testing and commissioning processes. For the latter case, the author understands that many detection and recognition systems are not built in-house, but instead make use of commercial off the shelf cloud-based platforms offered by large technology corporations. The implementation details of those systems are not exposed to the end user and even if they were, quantifying their failure modes would be difficult. As such, the author undertakes a case study measuring the interclass bias of thirteen commercial off the shelf facial recognition systems algorithms from Amazon Web Services and Microsoft Azure Cognitive Services. This method uses a “black-box” approach and does not require any knowledge of the internal properties, configuration, or architecture of the underlying facial recognition system, only access to its outputs.

Overview of this Dissertation

The following chapter, titled ***Literature Review*** on page 18, presents the evolution of facial recognition systems and an overview of current the state of the art for evaluating their performance. This chapter also situates these topics within the evolving discussions concerning equality (and fairness) in artificial intelligence. The last section of this chapter, ***Prior Work in Evaluating Bias in Facial Recognition Systems*** on page 29, is focused on reviewing other related work in measuring bias in facial recognition systems, and shows how this work addresses the unsatisfactory gaps in their methodology.

The design of the proposed interclass metric is discussed in detail in the chapter titled ***Methodology*** on page 32. This includes the methodology for the case study of the commercial off the shelf facial recognition systems provided by Amazon Web Services and Microsoft Azure Cognitive Services. The details of these systems performance with regard to skin tone and gender presentation is detailed in the chapter titled ***Commercial*** Facial Verification Algorithms Audit Findings on page 49 and ***Discussion*** on page 136. This includes a discussion, in the section titled ***The Importance of Building in Public*** 142on page 142, of the ways in which facial recognition system developers committed to building equal (fair) systems can build trust without voiding their intellectual property claims.

A summary of the findings in this dissertation, and the contributions to the field are presented at the end of this dissertation in the chapters titled ***Conclusion*** and ***Thesis Contributions*** on page 152 and 154.

Literature Review

This dissertation examines the evaluation of interclass bias within a classification schema. Of primary concern are, (i) automated facial analysis algorithms, (ii) the means for evaluating algorithmic performance, and (iii) the datasets of faces used to train and benchmark algorithmic performance. The author reviews the evolution of facial recognition systems to highlight key breakthroughs and their implications for the task of facial verification, where the system is asked to compare two images to verify if they are the same person.

Researchers have studied bias in facial recognition systems for the past two decades (Furl et al. 2002; Phillips et al. 2003). Early work focused on single-demographic effects of either race or gender, whereas more recent work has focused on intersectional analyses between gender and skin tone (El Khiyari and Wechsler 2016; Garvie et al. 2016; Klare et al. 2012; Norval and Prasopoulou 2017; O'Toole et al. 2011) The latter works have been and continue to be hugely impactful both within academia and the industry, and have inspired works on remedying the ills of these socially impactful technology and unequal systems (Buolamwini and Gebru 2018; Gentzel 2021; Grother et al. 2019; Rhue 2018; Wang et al. 2018b).

Finally, the author presents current practices for evaluating facial recognition systems performance along with the current limitations of existing approaches.

Automated Facial Recognition Systems

Deep neural networks have achieved remarkable successes in computer vision tasks like image classification, object detection and instance segmentation. Today these systems often play a role in determining who to hire and fire, who to grant a loan to and for how much and how long to sentence someone to prison. They have been inserted into tasks traditionally performed by humans and as such require a keen eye to ensure they are operating correctly and fairly (O'Neil 2017). Recent years have seen particular success in the implementation of automated systems that use deep neural networks to make a positive identification of a face against a pre-existing database of faces (a "facial recognition system") deployed by both government agencies, and private companies.

Before exploring this further, it is important to lay out some distinctions and specialized terminology from the field of facial recognition that differ from the field of computer vision. For more general classes of objects such as cars or dogs, the term "recognition" often refers to the problem of recognizing a "member of the larger class, rather than a specific instance" (i.e., "recognizing" a cat in the context of computer vision research is used to denote that one has identified a particular object as a cat, rather than one has identified a particular cat) (Huang et al. 2007). In the literature covering facial recognition systems, the term "recognition" is used to refer to the identification of a particular individual, not just any human being. This dissertation is concerned with facial recognition, as such, the author adopts this latter terminology.

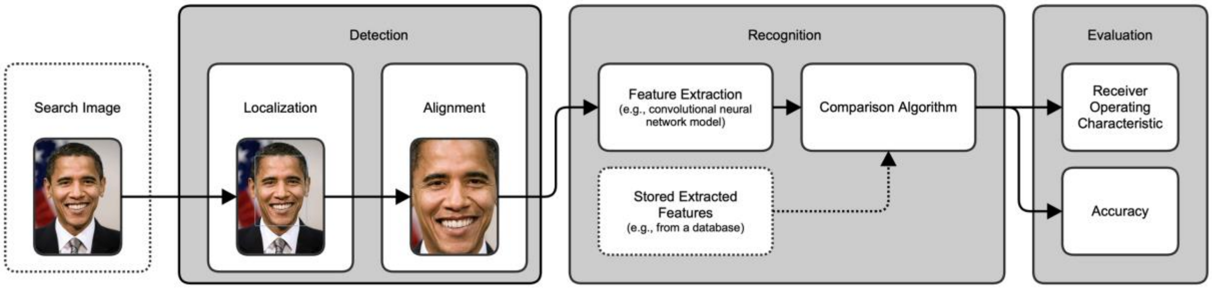


***FIGURE 1** The pipeline of a typical facial recognition system. First a detection algorithm is used to locate a face in a provided search image, and then the face is aligned to some normalized canonical coordinate system. Subsequently, the aligned facial image is passed to an algorithm that extracts learned discriminative features. These features are compared to some stored features to generate a similarity score representing the facial recognition system's confidence that the search image contains the same person as represented in the stored features. Lastly, the facial recognition system's performance on a series of these tasks can be evaluated by metrics of success (i.e., the receiver operating characteristic, or accuracy). The gray boxes represent the scope of this dissertation's evaluation.*

Facial recognition systems are considered a classical problem and an active area of computer vision research. Generally, a facial recognition system is an automated system that takes an input image with facial imagery. This system is divided into two main tasks: distinguishing human faces from other items of an image or "detection" and "recognition" of those detected faces through computationally analyzing facial features in terms of the spatial relationships between common landmarks (for example, the center of the pupil, the bridge of a nose, the ends of an eyebrow) (Chen et al. 2018; Wang and Deng 2020). Finally, the system compares these extracted features against a gallery of previously known individuals in a database to determine the individual in an input image. Facial recognition systems are subdivided into two different protocols at the recognition stage: "verification" where the facial recognition system is asked to compare two input images belong to the same person, or "identification" where the facial recognition system tries to recognize the person from a gallery of face images of different people. An overview of a typical facial recognition system performing a verification protocol is shown in **FIGURE 1**.

Early work in automated facial recognition began with manually curated subjective features (e.g., interpupillary distance and lip size) to achieve partially computerized facial identification (Goldstein et al. 1971). Advances in pattern recognition, led to improvements that avoided both manual definitions of relevant facial features and manual coding of specific features, in favor of an approach that automated dimensionality reduction using the linear algebra technique of principal component analysis (Sirovich and Kirby 1987). This innovation propelled appearance-based approaches, or approaches that treat facial imagery globally instead of focusing on specific facial regions like eyes or mouth, for facial recognition. The facial imagery is projected into orthogonal (i.e., uncorrelated) lower dimensional sub-spaces known as “eigenfaces”[3], and can be precisely represented as a weighted sum of these eigenfaces. The success of appearance-based techniques led to traditional facial recognition algorithms like Eigenfaces (Turk and Pentland 1991), Fisherfaces (Belhumeur et al. 1997), Bayesian face (Moghaddam et al. 2000), Metaface (Yang et al. 2010), algorithms using support vector machines (Guo et al. 2000) and boosting (Guo and Zhang 2001), among others. These facial recognition algorithms were somewhat constrained by environmental factors (e.g., image illumination, and facial expression), and required near-frontal facial imagery.

These facial recognition algorithms were supplanted by deep neural networks, which are more capable of learning, better at generalizing, and more robust to variation in input facial imagery (Guo and Zhang 2019). Deep neural networks have exhibited impressive results and have been shown to learn essential feature representation of data by constructing high-level features from the low-level pixel information encoded in an image. Furthermore, gains in facial recognition systems performance stem from well-capitalized artificial intelligence research in both industry and academia, which has led to the development of convolutional neural networks, and open-source implementations thereof (e.g. Caffe, TensorFlow, PyTorch) (Abadi et al. n.d.; Jia et al. 2014; Paszke et al. 2019). Additionally facial recognition systems have benefited from the availability of a large number of identity labeled images from the internet and curated datasets, and the availability of powerful computation hardware (i.e., many CPU cores and / or GPUs) to support those convolutional neural networks (Grother et al. 2019). An extensive overview of recent research on modern facial recognition systems using deep neural networks can be found in Wang and Deng (2020) and Guo and Zhang (2019).

The parallel advances in recent years in facial recognition systems alongside improvements in digital video camera technology, have led to these technologies becoming ever-present. Coffee shops are using facial recognition systems to identify repeat customers and their orders (BBC 2019; Bolger 2018). Shopping malls and retailers are using facial recognition systems to identify audience demographics to display tailored advertisements (Gillespie 2019). Airports are screening travelers by matching facial scans to online images, watch lists, criminal databases, and social media (Burt 2018; Delta Airlines 2021). Facial recognition systems is also being used by law enforcement to identify persons of interest or suspects in ongoing investigations (Garvie et al. 2016; Valentino-DeVries 2020). In their totality facial recognition systems are a powerful, pervasive, and ubiquitous technology that promises a myriad of benefits and conveniences to consumers, businesses, and governments.

[3] Sirovich and Kirby (1987) denote the coordinate system of these sub-spaces as “eigenpictures,” it is not until 1991 when Turk and Pentland that the now common term “eigenfaces” was introduced.

Evaluation of Facial Recognition Systems

These newer convolutional neural networks require datasets of sufficient size that include facial imagery and so called “distracting imagery”, images that do not contain a face, in order to build their high-level features that form the system’s internal representations of a human face. These datasets are crucially important as they empower the advancement of the field (Forczmański and Furman 2012). Generating these datasets is a resource and time-intensive job, as such there exist a large number of datasets available publicly to researchers in facial recognition systems. Many of these datasets are tailored to the specific needs of the algorithm under development, so many researchers supplement the training of their facial recognition system with independently collected, private datasets (Gross 2005).

The addition of private datasets to a facial recognition systems training complicates efforts to compare different facial recognition systems. This spurred the development of a new type of public datasets called benchmarks. These common benchmarks, a public dataset used for testing the performance of a facial recognition system, form a standard basis for researchers to be able to directly compare the results of different facial recognition systems. These benchmarks are overseen by government agencies, conference organizations, and research institutes which set up corresponding public challenges for researchers to compete for the best performance on these benchmarks to decide the current state-of-the-art for specific tasks. Benchmarks galvanize research and development activity and stimulate researchers in both the private sector and academia to achieve certain milestones in facial recognition systems.

Influential Facial Recognition Benchmarks

Benchmarks from the National Institute of Standards and Technology

One of the first of these benchmarks was the Face Recognition Technology (FERET) dataset published by the United States Department of Defense and the National Institute of Standards and Technology (NIST), an US government agency tasked with promoting innovation and advancing national competitiveness through advancing standards, in 1996. This dataset was comprised of 14,126 still portraits of 1,199 individuals collected over fifteen controlled photo shoots. At the moment of its release, FERET was the largest and most comprehensive effort to create a benchmark to accurately compare and evaluate existing facial recognition systems (Phillips et al. 2000; Zafeiriou et al. 2015). Additionally, FERET provided participants, a protocol for performing a standard set of experiments, the code for scoring their facial recognition systems performance, and the answers—known as the ground truth. By providing the answers NIST allowed participants in the competition the ability to improve their facial recognition systems or develop new facial recognition systems. The dataset provided researchers the data they required to make progress in the field, and successfully stimulated not only research interests in facial recognition but commercial applications of facial recognition systems (Raji and Fried 2021).

After FERET, NIST released a series of Facial Recognition Vendor Tests (FVRT) in 2000, 2002, 2006, 2010, and 2013 aimed at evaluating emerging commercial facial recognition systems "applied to a wide range of civil, law enforcement and homeland security applications including verification of visa images, de- duplication of passports, recognition across photojournalism images, and identification of child exploitation victims" (Blackburn et al. 2001; Phillips et al. 2003, 2007; Grother et al. 2010; Grother and Ngan 2014). Each iteration of the benchmark yielded improved test procedures and facial recognition system performance, but the early commercialization attempts that participated in these tests were somewhat constrained by environmental factors, as mentioned earlier. These facial recognition systems were crippled by small changes in image illumination, a person's facial expression, partial occlusions (e.g., a scarf, mask, or a pair of glasses) (Yang et al. 2002).

Other Influential Benchmarks

The introduction of the Labeled Faces in the Wild was a notable departure from previously collected datasets and benchmarks for facial recognition systems (Raji and Fried 2021). Labeled Faces in the Wild addressed researchers' desires to have access to naturally situated and varied input images that mirrored the large variation seen in everyday life (hence the "in the wild" moniker). The dataset leveraged the Internet to curate a dataset of 13,233 "previously existing images" that is to say, images not taken for the special purpose of facial recognition (Huang et al. 2007).

Labeled Faces in the Wild sparked a flurry of datasets for facial recognition that collected images from the Internet—often sourcing images without consent from the companies that hosted the images online, the owners (photographers) of the intellectual property in the images, or the people featured in the images—like Google Image search (Bainbridge et al. 2013; Cao et al. 2018a; Han et al. 2017), Flickr catalogues (Kemelmacher-Shlizerman et al. 2016; Merler et al. 2019), photojournalist published images by Yahoo News (Huang et al. 2007; Jain and Learned-Miller 2010), and still images from uploaded YouTube videos (Chen et al. 2017; Dantcheva et al. 2012; Wolf et al. 2011). The success of the deep neural networks that relied on unstructured and relatively unconstrained "in the wild" datasets, led to a proliferation of benchmarks. Researchers in search of even larger datasets to train their deep neural networks, compiled even larger datasets such as Oxford's VGG-Face dataset with 2.6 million images of 2,622 people (Cao et al. 2018b), Microsoft's 1M MS Celeb dataset with 8,456,240 images of 99,892 people (Guo et al. 2016), the CASIA WebFace dataset with 494,414 images of 10,575 people, and the MegaFace dataset with 1,027,060 photos of 690,572 people (Kemelmacher-Shlizerman et al. 2016).

Datasets that more closely resembled real world conditions, led to commercial facial recognition systems that performed better in the real world, and to new benchmarks to evaluate these commercial products. The Facial Recognition Vendor Test evolved extensively to move the state-of-the-art for facial recognition systems, growing from 13,872 images of 1,462 people in 2000 to 30.2 million images of 14.4 million people in 2013 (Blackburn et al. 2001; Grother and Ngan 2014). Overviews of publicly available datasets and the differences between these datasets are covered by Raji and Fried (2021), Forczmański and Furman (2012), and Gross (2005).

Measuring and Reporting Accuracy of Facial Recognition Systems

These benchmarks successfully galvanized researchers in both the private sector and academia to achieve certain milestones in facial recognition systems, and form a standard basis for researchers to directly compare the results of different facial recognition systems. While the benchmarks serve as the basis of these comparisons, the receiver operating characteristic serves as the standard practice in reporting performance metrics for facial recognition systems (Kemelmacher-Shlizerman et al. 2016; Maze et al. 2018; Phillips and O'Toole 2014).

The receiver operating characteristic graphically represents the transition between the true accept rate (also known as the sensitivity or recall), calculated from the fraction of genuine comparisons that correctly exceed a given threshold (the independent variable), and the false accept rate which is similarly calculated from the fraction of imposter comparisons that incorrectly exceed the threshold. When multiple facial recognition systems are analyzed on the same dataset, the receiver operating characteristic helps a system owner easily identify which facial recognition system has the highest true accept rate at a specified false accept rate.

Additionally, to render simpler comparisons, many benchmark evaluations report the true accept rate at some specified false accept rate. This is obtained by determining the lowest threshold value that yields a specified false accept rate, and using that threshold to calculate the true accept rate. Typical system owners of facial recognition systems target false accept rate spanning several decades from 10^{-6} to as 10^{-2}, but NIST has encouraged facial recognition systems using its benchmarks to report the true accept rate at a false accept rate of 1% and 0.1% (Whitelam et al. 2017). Results for the Labeled Faces in the Wild dataset are typically reported at the threshold where the false accept rate and the false rejection rate, calculated from the fraction of genuine comparisons that incorrectly falls below a given threshold, are equal. However, for the best facial recognition systems this would imply a false accept rate of 1% to 5%, which may be too high for many system owners (Kemelmacher-Shlizerman et al. 2016).

However, Krishnapriya et al. (2019) points out that while receiver operating characteristics compare true accept rates at the same false accept rates; different populations, with different priors, can achieve the same false accept rate at different thresholds. The populations that perform better on the receiver operating characteristics is simply determined by which population has the better true accept rate at the threshold that realizes the specified false accept rate for that population. That is to say, receiver operating characteristics can obscure important information as it may show different cohorts as having better accuracy when in fact there is a consistent difference in the underlying true accept rates and false accept rates (Krishnapriya et al. 2019). In keeping with this understanding, this dissertation eschews the use of the receiver operating characteristic in favor of the true accept rate and the false accept rate as a function of the threshold.

Bias in Facial Recognition Systems

Nevertheless, there are growing concerns about facial recognition systems with regards not only to the technology's shortcomings but also to how its use compromises civil rights, and raise issues of diminished accountability. Chief among these concerns is the potential and consequences of misrecognition. It has been shown that facial recognition systems can learn human-like biases unless actively controlled for during training dataset selection (Caliskan et al. 2017; Steed and Caliskan 2021) or in architecture selection for the deep neural network (Costa-jussà et al. 2020). There have been concerns of facial recognition systems failing to recognize Black and dark-skinned faces due unbalanced datasets that comprise the training data the facial recognition systems learn from (Noble 2018) and large-scale bias in the form of systematic misrecognition by skin color or ethnic background, and gender classification (Buolamwini and Gebru 2018; Crawford and Paglen 2019; Klare et al. 2012; Ngan and Grother 2015). That is to say, facial recognition systems that are trained within only the narrow context of a specific dataset inevitably optimize to learn the specific attributes of that dataset. This narrow context creates systematic errors as the system skews towards learning those specific attributes; and appears as under-representation or over-representation of groups in the dataset.

Additionally the benchmarks for facial recognition systems contain significant demographic bias, for example the Labeled Faces in the Wild benchmark containing celebrity faces has been estimated to be 77.4% male and 83.5% White (Han and Jain 2014). **TABLE 1**, reproduced from Merler et al. (2019), shows the distribution of gender and skin color for eight popular facial recognition datasets. It is important to note that different methods were used for characterizing skin color, as such the definitions of darker and lighter skin colors is inconsistent across the datasets (Merler et al. 2019).

Even less is known about the large private datasets listed in **TABLE 3**, which are built with unknown epistemological and metaphysical assumptions about the images, labels, categorization, representation, or demographics. Furthermore, few facial recognition systems report their accuracy by race or gender, which taken together with the skewed benchmark datasets, mean that there is little to no documentation about whether the reported high accuracy applies to people who are not well represented in the benchmark.

In recent years there have been endeavors to create diverse sets of collected faces, such as the Pinellas County Sheriff's Office (PCSO) and Michigan State Police (MSP) datasets containing mugshots of recidivists who are labelled with a binary black and white racial classification (Deb et al. 2017). In 2015, NIST together with the Intelligence Advanced Research Projects Activity (IARPA), an organization within the Office of the Director of National Intelligence, released the IARPA Janus Benchmark-A (IJB-A) facial recognition benchmark.

TABLE 1 *An overview of known public face datasets and the distribution of gender and skin color. This table shows seven prominent face datasets available at the time of writing. This list is not meant to be exhaustive, nor to describe the datasets in detail, but merely to provide a sampling of the datasets that are available. Where possible a peer-reviewed paper or technical report was cited, and otherwise a citation referring to the webpage for the database is given when available. Reproduced from Merler et al. (2019). (Buolamwini and Gebru 2018; Huang et al. 2007; Kumar et al. 2011; Liu et al. 2015; Maze et al. 2018; Moschoglou et al. 2017; Wang et al. 2018a; Zhang et al. 2017)*

	Gender Presentation		***Skin Tone***	
Dataset	*Masculine*	*Feminine*	*Lighter*	*Darker*
Labeled Faces in the Wild (LFW)	77.4%	22.5%	81.2%	18.8%
IARPA Janus Benchmark C (IJB-C)	62.7%	37.4%	79.6%	20.4%
Pubfig	49.2%	50.8%	82.0%	18.0%
CelebFaces Attributes Dataset (CelebA)	42.0%	58.1%	85.8%	14.2%
University of Tennessee Knoxville Face (UTKface)	52.2%	47.8%	64.4%	35.6%
AgeDB	59.5%	40.6%	94.6%	5.4%
Pilot Parliaments Benchmark (PPB)	55.4%	44.6%	53.6%	46.4%
IMDb-Face	55.0%	45.0%	88.0%	12.0%

The benchmark, and it's updates the IJB-B, IJB-C benchmarks, ran from 2015 to 2017. The IJB-C, the latest update to the benchmark, is a dataset of 138,000 images. The 3,531 people included in the dataset were selected to not overlap with other popular facial recognition benchmarks (Oxford's VGG-Face and the CASIA WebFace dataset) in order to prevent overfitting. However the IJB-A, which classified skin color on a six point scale (Klare et al. 2015; Maze et al. 2018; Whitelam et al. 2017),still has a super-majority of light skinned people: 79.6% of the IJB-A dataset images are light skinned (Findley 2020).

Understanding What is Meant by Bias

These concerns, and the remedies proposed, can be primarily understood to address questions of equality. Equality (and its closely related concept of fairness) occupy a prominent place in moral philosophy, social choice theory, economics and law.

For philosophers, the central problem is to define a just distribution of resources, rights, and duties in society, that is to say, define a just social order. This distribution problem requires that an individual's success or welfare in life be independent of irrelevant characteristics, that the individual could not be responsible for. This grand problem has animated philosophers since antiquity, from Plato's and Aristotle's conceptions of the ideal state, to the social contract theories of Hobbes, Locke, and Rousseau, to the modern theories of Rawls, Nozick, and Walzer

(Young 1994). The difference among philosophers pondering this question is mainly about which characteristics should be considered irrelevant. This debate is often summarized as asking the question of "equality of what?" (Roemer and Trannoy 2013).

The wide variety of interpretations of equality correspond to a wide spectrum of beliefs regarding what constitutes an irrelevant attribute. For example, welfarists would argue that all characteristics are irrelevant (i.e., all people should enjoy the same success); whereas, Libertarians would argue that the only non-productive characteristics (e.g., race) are the only irrelevant characteristics (i.e., one's success in life should be independent of one's race) (Calsamiglia 2005). The period since 1970 has been one in which, in political philosophy, non-welfarist theories have flourished, on both the right and left ends of the political spectrum. This equality literature changed the focus by pointing out that only some kinds of inequality are ethically objectionable, and to the extent that researchers ignore this distinction "they may be measuring something that is not ethically salient" (Roemer and Trannoy 2013).

This dissertation is purposely silent on the specific irrelevant attributes that should be measured or which kinds of inequality are objectionable. The author believes that these concerns are best left to local policymakers who are better equipped with the knowledge of how locals hold views of justice necessary to make these determinations; and practitioners who are most acutely aware of the harms that can be perpetrated by their facial recognition systems.

The Classification Systems of Irrelevant Attributes

Furthermore, and perhaps more importantly, specific irrelevant attributes require a taxonomy and a method of categorization or classification (for ease of reference, in this section, the author refers from now on only to categorization). This dissertation understands that categorization is a social construction. Michel Foucault, noted French philosopher and historian, argues that categories are a matter of invention in his book The Order of Things (1973). Foucault claims that categories are the result of a priori historical systems of classification, invented (or constructed) by societies. Categories are the result of the application of a system, which is like a set of criteria that when applied to experience, makes us think of the world in certain ways (Gracia 2001). Categories "[m]ake it possible for us to name, speak, and think" (Foucault 1973). These categorizations can affect and condition how societies view themselves. Furthermore, attempts to classify have reflected the social, cultural, religious, and political order of the time.

Foucault introduces an encyclopedia in which it is written that "animals are divided into: (a) belonging to the Emperor, (b) embalmed, (c) tame, (d) sucking pigs, (e) sirens, (f) fabulous, (g) stray dogs, (h) included in the present classification, (i) frenzied, (j) innumerable, (k) drawn with very fine camelhair brush, (1) et cetera, (m) having just broken the water pitcher, (n) that from a long way off look like flies" (Foucault 1973). Using this taxonomy, Foucault expresses wonder at the limitation of our own system of thought that results in the stark impossibility of employing this categorization. We experience this system as exotic and charming (Foucault's words) because it upends the basis of our own categories. The historical systems that allow us to make sense of the world, also cloud our ability to see it in any other way. Artificial intelligence, and facial recognition systems, similarly construct (invent) and inherit from the historical system, its own methods of categorization that represent the social order of the time. Labelled Faces in the Wild, generated its dataset based on photographs on Yahoo News from 2002 to 2204, which results in a dataset that is 77.4% masculine presenting and 71.2% lighter skin tones as mentioned in **TABLE 1** on page 25 above. Furthermore, the most represented face in this dataset is George W. Bush, who appears 530 times in more than thirteen thousand pictures. Labelled Faces in the Wild is an invention of the culture and social order at the time of its construction.

The Bifurcated Theories of Inequality

As Young (1994) points out in the first chapter of his book Equity: "[social justice] theories in the large have little to say about … how to solve concrete, everyday distributive problems such as how to adjudicate a property dispute, who should get into medical school, or how much to charge for a subway ride." Yet, anti-discrimination laws provide a rich jurisprudence to understand how some local policymakers attempt to distinguish which attributes are irrelevant. These laws can be widely understood to bifurcate between two theories of discrimination: disparate treatment (i.e., intentional discrimination) and / or disparate impact (i.e., the discriminatory consequences of a neutral policy). In theory, disparate impact is a straightforward concept: discrimination can occur when a facially neutral policy is implemented, without a legitimate justification, that disproportionately affects a group that shares a specific irrelevant attribute protected by statute (often termed a "protected group"). Importantly, disparate impact does not require discriminatory motive or intent (*Faulkner v. Super Valu Stores, Inc.* 1993).

A number of recent studies which investigated quantifying and guaranteeing equality in machine learning (Dwork et al. 2011; Feldman et al. 2015; Hardt et al. 2016; Kamishima et al. 2012; Kleinberg et al. 2016; Luong et al. 2011; Zemel et al. 2013), have borrowed extensively from these legal concepts when designing or problematizing automated machine learning algorithms (Heidari and Krause 2018; Zafar et al. 2017). This dissertation also adopts this disparate impact formulation of the problem; said simply, bias is defined as disparate outcome based only or partially on a specific irrelevant (i.e., protected) attribute. In order to be concordant with specialized terminology from the field of fairness in machine learning, this dissertation will use "protected attribute" in the same way that a specific irrelevant attribute has been defined.

Highlighting the Finer Points of Some Terminology

Some researchers when defining bias introduce the concept of privilege (and unprivileged) values of protected attributes. These privileged values of a protected attribute are specifically formulated to indicate groups that have been historically at a systemic advantage (e.g., Whites for race in the United States, Men for gender in the United Kingdom, Brahmin for caste in India, Catholicism for religion in France) (Bellamy et al. 2018). Bias is then formulated as a systematic error that places groups that shares a specific privileged attribute at systematic advantage over unprivileged groups. Introducing historicity of privilege into the definitions unnecessarily winnows the classification schemas used and protected attributes of bias that are worth measuring and improving. Therefore, in keeping with the author's aforementioned purposeful silence on which protected attributes should be measured or which kinds of inequality are objectionable, the author rejects this distinction provided by Bellamy et al. (2018).

Additionally, though this work incorporates understandings of the accuracy of a decision taken by a facial recognition system into its formulation of the disparate impact problem in the measurement of interclass bias within a classification schema, and might be reasonably considered to be what Zafar et al. (2017) term "disparate mistreatment," the author continues to use the term disparate impact. In particular, the author takes issue with the naïve formulation that ignores the legal precedent of disparate impact which provides exceptions if the policy is justified or otherwise consistent with a business necessity; and subsequently uses this ignored exception to conceptualize the term "disparate mistreatment."

Technical Implications of Bias

Definitions of bias have technical implications. Narayanan (2018) described at least twenty-one mathematical definitions of equality from the literature. These differing definitions are not merely theoretical differences in how to measure equality, but can yield entirely different outcomes depending on the measure used (Narayanan 2018).

Facial recognition systems generally commit two kinds of errors, the Type I Error where an individual is incorrectly associated with another, and Type II Error where an individual is incorrectly not associated with themselves. The nature of these errors is explored in more detail in the section entitled ***Quantitative Analytical Methods*** on page 42. When a deep neural network is trained, a cost function is defined that can optimize for the absence of Type I Errors or Type II Errors. Ideally, the deep neural network would want to minimize all errors, yet Kleinberg et al. (2016) show that balancing these errors while maintaining a high predictive accuracy is impossible. Therefore, practitioners must make subjective decisions based on their engineering judgement about the perceived impact of Type I Errors or Type II Errors. For example, in a security checkpoint the probability of Type II Errors from a facial recognition system represents a measure of the system security, while the probability of Type I Errors represents the user inconvenience level. In some instances, like screening for *persona non grata* at a border entry point the risks associated with a Type I Error, failing to recognize a *persona non grata*, are often much higher than the harms resulting from a Type II Error, requiring additional visitors to submit to a secondary screening. Whereas in supermarket that uses a facial recognition system to attach a customer's loyalty number to a transaction, the risks associated with a Type I Error, attaching a transaction to another's loyalty number might be considered negligible to the harms resulting from a Type II Error, the added time to a transaction incurred by a customer typing in their loyalty number.

In order to facilitate the evaluation of these tradeoffs and the cautionary note from Krishnapriya et al. (2019) regarding the receiver operating characteristic, covered in more detail in the section entitled ***Measuring and Reporting Accuracy of Facial Recognition Systems*** on page 23, this dissertation takes the position that any measure of bias must independently measures the bias of the true accept rate and the false accept rate as a function of the threshold.

Furthermore, this dissertation postulates that in order to achieve a full picture on the nature of interclass bias within a classification schema it is important to evaluate the true accept rate and the false accept rate across the entire range of the threshold. NIST restricts its evaluation of the true accept rate and the false accept rate to what it believes are common targets, namely targeting false accept rates spanning several decades from 10^{-6} to as 10^{-2}. However, this approach shifts responsibility for threshold management to the system owner rather than the developer of the facial recognition system. That may sound appropriate, but it imposes a responsibility on the system owner to determine what the thresholds should be based on their targeted false accept rate via some appropriate testing.

In practice, the only system owner controlled independent variable is the threshold, and the meaning of the threshold is easily confused. Facial recognition system developers report that their system is set to report matches it is “99% confident in,” as such, operators might assume that each match has a 99% chance of being a genuine match, when in actuality the majority of the alerts the facial recognition system generates are likely to be false (Crumpler and Lewis 2021). The operators and system owners are likely to assume that the commercial facial recognition system works directly “off the shelf,” and not make any changes to the default configuration. In fact, in one installation of a facial recognition system at the Washington County Sheriff’s Office in Oregon, the system owner clarified that “we do not set nor do we utilize a confidence threshold” (Menegus 2019). Furthermore, the facial recognition system vendor only supplied documentation (characterized as “very lacking or wrong” by an analyst employed to evaluate the facial recognition system by the system owner), and did not provide any direct training to investigators using the system to clarify the meaning of the thresholds. This means that any reported matches by the facial recognition system, regardless of the confidence score expressed by the system, may be interpreted a match by the system owners or operators.

Using this evaluation, the author sees benefits in having a facial recognition system for which true accept rates and false accept rates are homogenous across the entire range of the threshold (i.e., do not vary over any protected attribute). To the authors knowledge, this is the first investigation that takes such a stance.

Prior Work in Evaluating Bias in Facial Recognition Systems

The broad effects exemplified in this report concerning gender presentation have been known as far back as 2003 (Phillips et al. 2003). These findings were re-affirmed by NIST in 2017 for Type I Error bias, where an individual is incorrectly associated with another, across gender presentation and skin tone (Grother et al. 2017), and again in 2019 (Grother et al. 2019). Most of the research being performed has focused on evaluations of facial recognition systems performing a verification task with standardized facial images (i.e., images collected for visas, border crossings, or law enforcement booking photographs) (Cook et al. 2019; Grother et al. 2017, 2019; Howard et al. 2019; Krishnapriya et al. 2019). These images are collected with cooperating subjects with controlled cameras or dedicated capture equipment and lighting, and are generally in reasonable conformance with the ISO/IEC 19794-5 Biometric data interchange formats. This standardization work included consideration of cameras, lights, and geometry, and with explicit consideration of the need to capture light and dark skinned individuals. This report is the first to focus on measuring bias within the relatively unconstrained "in the wild" facial imagery, like in Labeled Faces in the Wild. See the section entitled ***Assumptions About the Use of Unconstrained Images*** on page 145 for a more detailed understanding of how this report addresses "in the wild" imagery.

Furthermore, while this dissertation evaluates false accept rates across the entire range of the threshold, some researchers don't report the incidence of Type II Error, where an individual is incorrectly not associated with themselves, at all (Cavazos et al. 2020); others report it only at fixed false accept rates burying threshold related bias (El Khiyari and Wechsler 2016); and many report it at a specific threshold (Cook et al. 2019; Grother et al. 2019).

Lastly, other researchers have focused on racial, ethnic, ancestry, or country of origin classifications for measuring bias. These classification schemas are difficult to compare due to the large intraclass variation, and the inconsistencies of the schemas utilized: East Asian vs Caucasian faces (Cavazos et al. 2020), Black or African Americans vs. Caucasian faces (Cook et al. 2019; El Khiyari and Wechsler 2016; Krishnapriya et al. 2019), and Asian, Black, Indian, vs. White (Grother et al. 2019). Additionally, the methodologies of how each racial or ethnic label is applied vary widely, some use subject self-identification (Cook et al. 2019; Howard et al. 2019), others use the country of origin (Grother et al. 2019), and others do not explain how the labels were assigned (El Khiyari and Wechsler 2016; Krishnapriya et al. 2019). This dissertation adopts the six skin tone classification used by Buolamwini and Gebru (2018); Klare et al. (2015); Maze et al. (2018); Whitelam et al. (2017) to account for skin tone diversity with a more reproducible classification system.

Summary

Taken together, these studies clearly indicate the importance of creating techniques to identify biases and to measure these biases before these algorithms are integrated into both government agencies, contractors, and private companies. The researchers believe this measure of interclass bias will engender comprehensive analyses of facial verification algorithms biases that can be incorporated into an algorithm's design, implementation, or training processes and an end user's testing and commissioning processes.

For this dissertation, the author argues that to adequately assess bias, one needs disaggregated evaluation metrics to evaluate the success of an artificial intelligence model. Furthermore, these metrics that that incorporate a realistic bias model.

In order to trouble the existing evaluation metrics, this dissertation adopts the disparate impact formulation of the problem and defines bias as a disparate outcome based only or partially on a protected attribute. Furthermore, the author argues that an unbiased facial recognition system is one for which true accept rates and false accept rates are homogenous across the entire range of the threshold (i.e., do not vary over any protected attribute). To the authors knowledge, this is the first investigation that takes such a stance.

This dissertation is purposely silent on the protected attributes that should be measured or which kinds of inequality are objectionable, and adopts a notion of bias that can be used to measure not only to intrinsic physical characteristics (e.g., skin color, hair color, wearing vision eyewear) but also to extrinsic characteristics, limited by a physical presentation recordable in an image.

Methodology for Measuring the Interclass Bias

This dissertation proposes a novel methodology to measure facial recognition systems interclass bias within a classification schema. Throughout this dissertationthe term bias is used to demarcate any propensity or prepossession towards a particular class within a classification schema. The researchers believe this measure of interclass bias will engender comprehensive analyses of facial recognition systems' biases that can be incorporated into an algorithm's design, implementation, or training processes and an end user's testing and commissioning processes. Due to practical constraints, this study focuses on a subset of facial recognition systems, facial verification algorithms, where the algorithm is asked to compare a facial imagery from two images to verify if they are the same person. In other words, to compare faces as it may be done in an access control type scenario (e.g., a security guard who is tasked with letting authorized persons into a facility).

The author sets out to design this measure of interclass bias not as a function of specific images submitted for analysis, as many current benchmarks do, but rather, as a measure of a facial verification algorithms' performance for everyone who could be classified under a given classification schema. That is to say, the author believes that this measure shall generalize well to an entire class of people. This author plans to achieve this affect by adopting two principles in the creation of the interclass bias measure. First, the author postulates that multiple comparisons of a diverse set of facial imagery of the same individual will elucidate the facial verification algorithms' performance for a specific individual. These multiple comparisons are powerful because they can be used to measure the facial verification algorithms' performance independent of the images provided. Second, this process after repeated with multiple individuals from the same class could be used to generalize the facial verification algorithms' performance on individuals from that class. The author believes that a facial verification algorithms' performance is a distribution that must be modeled to account for the complexity of the similarity threshold. The similarity threshold is the independent variable selected by end-users to simplify the facial verification algorithms' performance to a binary output (e.g., the aforementioned security guard has reasonable doubt the person attempting to enter is authorized to enter the building and prohibits him, her, or them from entering) and is configurable by the end user based on the needs of the intended application (e.g., a bank might set a higher threshold for authorizing a withdrawal than a grocery store pulling up your loyalty account). The scope for this dissertation is illustrated in **FIGURE 2** below.

This dissertation begins by defining the calculation of this novel measure of interclass bias and then goes on to provide a case study of the efficacy of this proposed interclass bias measure, and the evaluation of two commercial off the shelf facial verification algorithms (for ease of reference, the author will use the term "commercial facial verification algorithms") to test our findings.

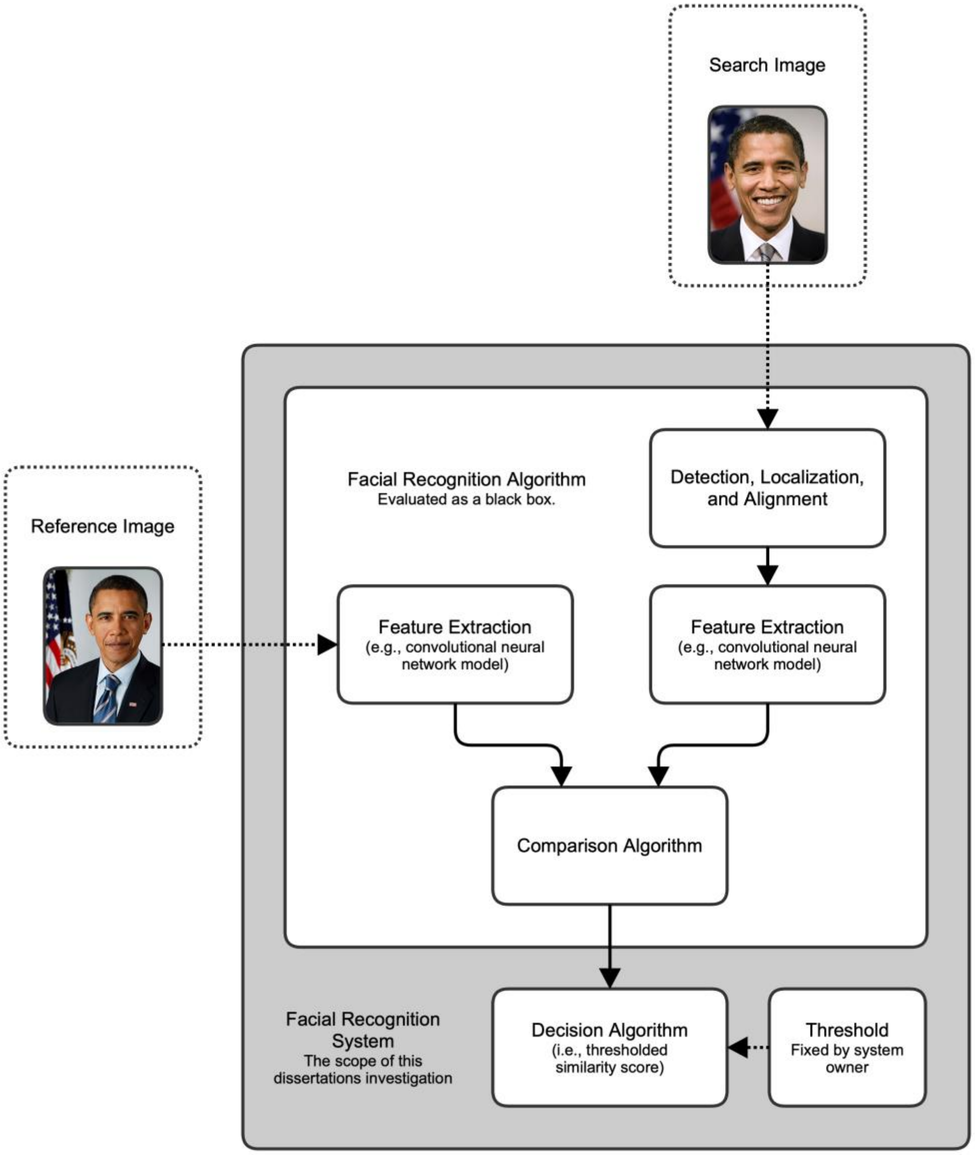

Search Image
Reference Image
Facial Recognition Algorithm
Evaluated as a black box.
Detection, Localization, and Alignment
Feature Extraction
(e.g., convolutional neural network model)
Feature Extraction
(e.g., convolutional neural network model)
Comparison Algorithm
Facial Recognition System
The scope of this dissertations investigation
Decision Algorithm
(i.e., thresholded similarity score)
Threshold
Fixed by system owner

FIGURE 2 *A facial recognition system performing a verification protocol. The gray box represent the scope of this dissertation's evaluation.*

Data Collection and Analysis for the Proposed Metric

The section below moves on to describe in greater detail how to generate and report the novel measure of interclass bias.

Principal to this approach is a determination, generation, or selection of the classification schema used to evaluate facial verification algorithms for interclass bias. It is important to note that these classifications can refer not only to intrinsic physical characteristics (e.g., skin color, hair color, wearing vision eyewear) but also to extrinsic characteristics limited by a physical presentation recordable in an image, as facial recognition systems ultimately depend on what can be captured by a camera. This can be illustrated briefly by focusing on gender classification schemas. Commonly reported sex classifications with “male” and “female” classes are unfeasible as these labels cannot be discerned by facial or unconstrained imagery. However, a gender presentation classification schema is in fact possible. Labels such as “high confidence of presenting as masculine” and “high confidence of presenting as feminine” can be discerned from facial imagery based on cultural understandings of masculine and feminine presenting traits.

Next, a dataset of facial images of classified subjects according to the selected classification schema must be collected, generated, and/or labelled. This dataset should contain a reasonable number of subjects and each subject in the dataset should contain numerous samples of facial images for that subject. The multitude of facial imagery in effect minimizes the variations in an algorithms performance that might be attributed to an individual photo (i.e., pose, lighting, image size, orientation, occlusions, etc.) and empowers further analyses that can achieve a subject specific understanding of the facial verification algorithms’ performance. As such, there is an obvious preference for the largest number of facial images for each subject desired, but the author recognizes the tradeoffs that must be made to generate such a dataset. Furthermore, the author understands that data collection can be difficult if not unfeasible and as such, this process can be reversed with the dataset constraining the selection of a classification schema.

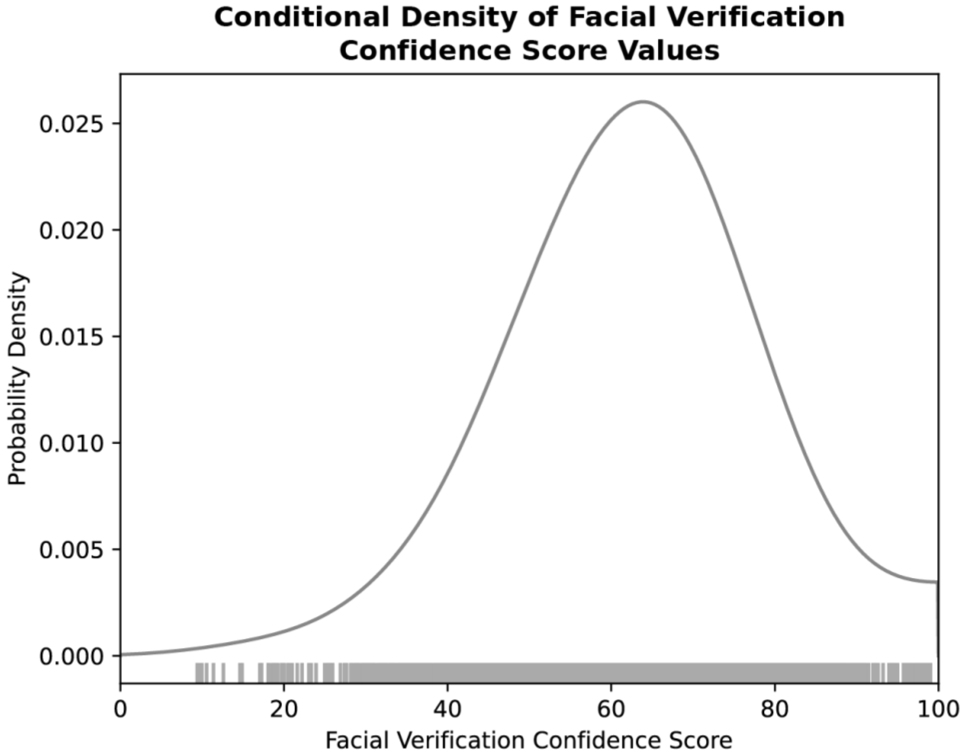
After the dataset is compiled, the list of comparisons can be generated. Each comparison is a set of two facial images provided to the facial verification algorithms to evaluate if they represent the same person by a score. These evaluation result in scores that are collected and subsequently used to calculate the novel measure of interclass bias within a classification schema for a given facial verification algorithm. In order to achieve the aforementioned subject specific understanding, each facial image for given subject must be compared to every other facial image for the subject, in the style of double round-robin tournament. In a more precise description, each test subject ($i, 1 \leq i \leq I, i \in \mathbb{Z}$), has n_i total “templates”: an input facial image converted into a proprietary template (i.e., a feature vector) representing the face in the image, which is subsequently stored in an internal database. The goal of facial verification is for the algorithm to determine if a sample template generated from a query image (commonly referred to as the “probe”) is the same person as the template stored in a database (commonly referred to as the “gallery”). For each subject (i), one image is sampled as the gallery template ($n_i^g, 1 \leq n_i^g < n_i$)

and the remaining templates are allocated to the probe set ($n_i^p + n_i^g = n_i$). Each of the templates in the probe set is queried against its corresponding gallery template, and then another image is sampled without replacement as the gallery template, until all templates in the set have been used as the gallery template. This yields a total of $\sum_{i=1}^{I} n_i(n_i - 1)$ comparisons for analysis. It is important to note that this process assumes that the comparisons submitted to the facial verification algorithm are not commutative, that is to say comparing probe template A to gallery template B is different from comparing probe template B to gallery template A. These reversed comparisons are made to protect from algorithms implementations that may return different results based on the order.

These similarity scores for the list of comparisons are collected and then evaluated for interclass bias within a classification schema for the given facial verification algorithm. A similarity score is a statistical measure of how likely the gallery and probe templates are the same person, when analyzed by the facial verification algorithm. These scores are typically reported as ranges from 0 to 1 (or an equivalent scale) with larger numbers indicating higher similarity and importantly are relevant and comparable to other scores exclusively to the algorithm that generated them. These similarity scores are typically thresholded, such that any similarity score lower than a given threshold is rejected as a match and any similarity score greater than the threshold is accepted as a match. The standard practice in reporting performance metrics for comparison protocols is to report the receiver operating characteristic, which at a given threshold (the independent variable) measures the true accept rate ("TAR"), calculated from the fraction of genuine comparisons that correctly exceed the threshold, and the false accept rate which is similarly calculated from the fraction of imposter comparisons that incorrectly exceed the threshold. However, due to the larger impact that false negatives can have in environments that utilize facial verification algorithms, and practical constraints, this dissertation focuses on measuring the interclass bias for Type II Errors , encapsulated by the TAR.

A Formal Definition of Interclass Bias

We can formalize this understanding by introducing some basic notation to define the problem. Each comparison of two facial images submitted to the facial verification algorithm has a known state $\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_K$, where $K = \sum_{i=1}^{I} n_i(n_i - 1)$, the total number of comparisons submitted for analysis. Where $\mathcal{X}_k$, for comparison k, represents one of two states: "confirming that the two submitted facial imagery represents the same person" is denoted as 1, and "rejecting that the two submitted facial imagery represents the same person" is 0.



<b>FIGURE 3</b> Illustrative example of similarity score densities for a sample facial verification algorithm.

Additionally let, $x_1, x_2, \dots, x_K$, represent the set of similarity scores, returned by a facial verification algorithm. The similarity scores are reported on a continuous scale that the facial verification algorithm has assigned to the set of two comparison templates. The author restricts their considerations to the case of continuous scores as this facilitates notation and reasoning, and is in keeping with the scores reported by many facial verification algorithms. Similarity scores with discrete values can be treated in a similar way. Furthermore, we can define the conditional distributions given the two values of $\mathcal{X}_k$ can take by specifying the conditional densities f_0 and f_1 respectively. The probability that a similarity score x_k is greater than some threshold t given the predicted value $\hat{X}_k$ is x, say $x = 1$, can be expressed as an integral of the density f_1.

$$F_x(t) = P\left[x_k \leq t | \hat{X}_k = x\right] = \int_{-\infty}^{t} f_x(u)\ du\,, x \in \{0,1\}$$

FIGURE 3 above illustrates a plot of what the conditional density f_1 might look like for a facial verification algorithm. From an end-user's perspective, the whole proposition looks like a decision problem. Assume that we consider two templates with facial imagery and have been provided the facial verification algorithms similarity score. How can the end-user infer whether or not to confirm or reject that the two submitted facial imagery represents the same person? This represents a case of binary classification with a one-dimensional co-variate. Formally, a realization $(x_k, \mathcal{X}_k)$ has been randomly sampled, x_k is observed, $\mathcal{X}_k$ is not yet known: Is $\mathcal{X}_k = 0$ or $\mathcal{X}_k = 1$? The typical way that the end-user would come to a decision would be to select some set $\mathcal{A}$ of similarity score value such that they infer the match is confirmed, if the score is in $\mathcal{A}$. If the similarity score is not in $\mathcal{A}$, then the end-user would conclude that the match is rejected.

$$x_k \in \mathcal{A} \;\Rightarrow \mathcal{X}_k = 1 \quad x_k \notin \mathcal{A} \Rightarrow \mathcal{X}_k = 0$$

The end-user is then just left to determine an appropriate selection set $\mathcal{A}$. This can be stated as having to discriminate between conditional score distributions on the $P[x_k \in \cdot \,|\mathcal{X}_k = 1]$ confirmed and $P[x_k \in \cdot \,|\mathcal{X}_k = 0]$ rejected match sub-populations, respectively. The end-user desires a high certainty in the case of a decision to reject the presumption that the two facial images submitted are different people. Formally this can be stated as the null hypothesis is that the two facial images submitted are the not the same person, or the state $\mathcal{X}_k = 0$. The alternative hypothesis is that the two facial images submitted are the same person, or the state variable is $\mathcal{X}_k = 1$.

$$H_0: P[x_k \in \cdot \,|\mathcal{X}_k = 0] \text{ versus } H_A: P[x_k \in \cdot \,|\mathcal{X}_k = 1]$$

The end-user can subsequently perform a statistical test on the null hypothesis against the alternative. Under this test, the decision the end-user chooses could be wrong in one of two ways. The end-user could reject the presumption that the two facial images submitted are different people, or reject $\mathcal{X}_k = 0$ when $\mathcal{X}_k$ is actually 0 (a Type I Error). Alternatively, the end-user would accept the presumption that the two facial images submitted are different people when the two facial images submitted are of the same person, or accept $\mathcal{X}_k = 0$ when $\mathcal{X}_k$ is actually 1 (a Type II Error). The end-user might want to create an optimal selection set $\mathcal{A}$ to manage the probabilities of the two possible erroneous decisions. The probability of a Type I Error is the probability under the rejected match score distribution that the comparison's similarity score will be an element of the acceptance set $\mathcal{A}$. Similarly, the probability of a Type II Error is the probability under the confirmed match score distribution that a comparison's similarity score will not be an element of the acceptance set $\mathcal{A}$.

$$P[\text{Type I Error }] = P[x_k \in \mathcal{A}|\mathcal{X}_k = 0] \quad P[\text{Type II Error }] = P[x_k \notin \mathcal{A}|\mathcal{X}_k = 1]$$

Typically, the Type I Error is bounded by a small constant (commonly illustrated as one or five percent) determined by the cost associated with a Type I Error incurred by the end-user. We shall denote this constant by α. Once this value is bonded, the objective for an end-user is to minimize the Type II Error probability. Additionally, because the similarity scores are continuous, this can be described in terms of some threshold t. The end-user can set a threshold based on the ratio of the rejected and confirmed comparisons conditional score densities f_0 and f_1, respectively, as a function to the similarity score variable itself.

$$1 - \alpha = P\left[\frac{f_1(x_k)}{f_0(x_k)} \leq t|\mathcal{X}_k = 0\right]$$

This can be solved for the threshold t . Furthermore, this yields a decision rule for the end-user: “Reject the presumption that the two facial images submitted are different people if the similarity score is greater than the threshold” that is optimal amongst all the decision rules that guarantee a probability of Type I Error not greater than α. Formally, if the similarity score is greater than the t then the end-user infers the match is confirmed. Otherwise, the match is rejected.

$$x_k \geq \ t \ \Rightarrow \mathcal{X}_k = 1 \quad x_k < t \Rightarrow \mathcal{X}_k = 0$$

That means, that for a given threshold t, $\hat{X}_k$, the predicted value of $\mathcal{X}_k$, is given by the following:

$$\hat{X}_k = \begin{cases} 1 & \text{if} \quad x_k \geq t \\ 0 & \text{if} \quad x_k < t \end{cases}$$

An important note to the reader, the above transformation from the set $\mathcal{A}$ to threshold t, relies on the assumption that the likelihood ratio $f_1(x_k)/f_0(x_k)$ is monotonous. In theoretical examples where the rejected and confirmed comparisons conditional score densities f_0 and f_1, respectively, are normal with equal standard deviation, it is clear that the likelihood ratio is monotonous. However, in practice the monotonicity of the likelihood ratio or the conditional densities is difficult to ensure, but the clarity and economic value of using a threshold as a cut-off means that the author of this dissertation adopts an assumption of monotonicity.

Now, we introduce basic notation to define the terminology used in the TAR analysis. For every possible threshold value selected to discriminate between two cases, some comparisons will be correctly classified as confirming that the two submitted facial imagery represent the same person (we denote this outcome as a positive case, and because it was rightly detected as positive, we denote it as a true positive or TP). Other comparisons will be incorrectly classified as rejecting that the two submitted facial imagery represent the same person (we denote this outcome as a negative case, and because it was incorrectly detected as negative, we denote it as a false negative or FN). Conversely some data will be correctly classified as negative (a true negative or TN), but some will be incorrectly classified as positive (a false positive or FP). The various outcomes for each comparison are represented in **TABLE 2** below.

	Predicted Outcome	
Actual Outcome	*Confirm Match*	*Reject Match*
Confirm Match	True positive (TP)	False negative (FN)
Reject Match	False positive (FP)	True negative (TN)

TABLE 2 *The confusion matrix for a binary classifier modified to reflect the use case in the current dissertation.*

From these definitions, we can estimate $\vartheta(t)$, the TAR (also called the sensitivity or the true positive rate), which is the probability of accurately predicting a positive outcome ($\hat{X}_k = 1$), conditional that the observation is truly positive ($\mathcal{X}_k = 1$):

$$\vartheta(t) = \ P\big(\hat{X}_k = 1 \big| \mathcal{X}_k = 1\big) = \frac{1}{K} \sum_{i=1}^{K} I\big(\hat{X}_k = 1 \big| \mathcal{X}_k = 1\big)$$

A facial verification algorithm with a higher TAR has a lower Type II Error rate. Empirically, this is the proportion of positive instances classified correctly:

$$\hat{\vartheta}(t) = \hat{P}\left(\hat{X}_k = 1 \middle| \mathcal{X}_k = 1\right) \approx \frac{TP}{TP + FN}$$

where TP is the number of positive instances correctly classified, and FN is the number of positive instances misclassified. Further, this TAR point estimate can be evaluated across the set of all possible threshold values (τ) to yield a TAR curve. **FIGURE 4** below shows examples of how a TAR curve may look like. Note that though this notation describes the TAR ($\vartheta(t)$) as a function of the threshold (t), in practice as explained earlier an end-user of a facial verification algorithm would select the acceptable TAR and adjust the threshold to meet the selected TAR; as such, all plots of the TAR are presented in this way, with the threshold as a function of the TAR.

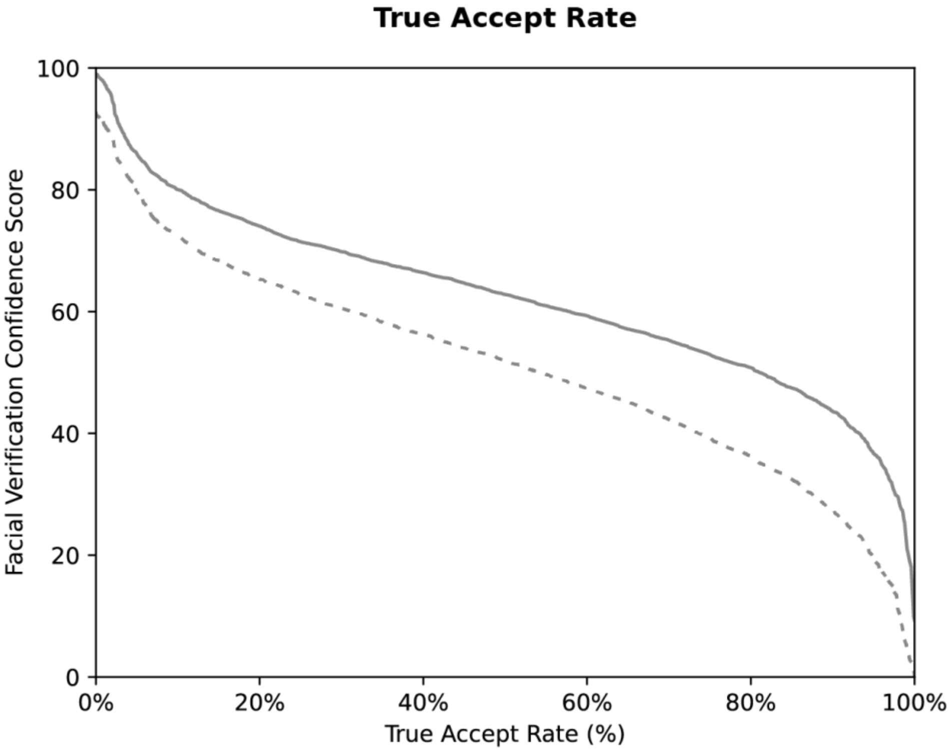


FIGURE 4 *Illustrative example of the True Accept Rate (TAR) curves of life-like (solid line), a worse performing curve (dashed line), and perfect score (dotted line) variables as explained in the main text. For the life-like similarity scores variable associated with the conditional densities shown in* ***FIGURE 3****.*

A perfect facial verification algorithm that can always correctly confirm that the two submitted facial imagery represents the same person and never rejects that the two submitted facial imagery represent the same person, is represented by the dotted straight line at $t = 100$, shown in **FIGURE 4** above. The solid curve belongs to the score variable whose conditional density is shown in **FIGURE 3** on page 36. The solid curve represents a facial verification algorithm that performs better than the dashed curve. Results towards the upper region of the plot are better. For example, if one wanted to achieve an 80% TAR—a default lower limit of consideration (IBM Corporation 2022)—on the corresponding threshold value on the default (solid curve) facial verification algorithm is significantly higher than on the worse (dashed curve) facial verification algorithm. The author wants to introduce a caveat when comparing two different facial verification algorithm TAR curves, the conclusions about the more accurate or "better" algorithm are limited, as the similarity score values are specific to each algorithm and cannot be compared without additional information on the algorithms Type I Error.

However, let's consider that these same two curves represented TAR curves for populations of two classes of interest in a classification schema with similarity scores generated under the same facial verification algorithm. Let's denote a new classification schema with two classes Class $\mathbb{A}$ and Class $\mathbb{B}$, and posit that both classes are drawn from the population expected by the facial verification algorithm.

In this scenario, again assuming an 80% TAR, the corresponding threshold value on the solid curve, representing the TAR curve for Class $\mathbb{A}$, is significantly lower than on the dashed curve, representing the TAR curve for Class $\mathbb{B}$. If the end-user wishes to achieve comparable TAR rates for both classes, a reasonable assumption as both classes are drawn from the same population, is that the end-user could set a different threshold for each class. However, it might be infeasible if not impossible to know which class a given comparison belongs to when making the decision to apply the threshold. This proposed solution adds complexity as the end-user would need to classify comparisons to one of the two classes, this could be done by an automated system or by a human, and store the information alongside the comparisons. Additionally, using class specific thresholds, also introduces an additional source of error, as mis-labelled comparisons might be subject to a looser standard than they were otherwise intended to. Additionally, the people whose facial imagery is retained in a comparison typically can be classified by multiple classes, so this problem of multiple thresholds would only increase as our taxonomy for classifying people increases. This complexity and potential for error means that many end-users select just one threshold that applies for all. Some commercial facial verification algorithms promote a single threshold in their marketing materials and developer resources. Under a single threshold, in the illustrated TAR curves, it's obvious that there will be a tradeoff no matter what threshold is selected. If the threshold is selected such that the end-user achieves an 80% TAR for Class $\mathbb{B}$, a small fraction of comparisons from Class $\mathbb{A}$ would be accepted. Similarly, if the threshold is selected such that the end-user achieves an 80% TAR for Class $\mathbb{A}$, almost all comparisons from Class B would be accepted. A threshold somewhere in-between these two extremes would mitigate these differences but not eliminate them. The author terms this phenomenon, the differences between these two TAR curves, "interclass bias". In the subsequent sections, the author describes qualitative and quantitative measures of the interclass bias.

Once all of the genuine comparisons submitted for analysis by the facial verification algorithm are returned with a similarity score, the novel measure of interclass bias within a classification schema can be calculated. Under the limitation that this dissertation only covers Type II Errors, all comparisons must therefore be genuine; however, the methodology described below would equally apply to any indicator functions such as the correlation coefficient, quantile, conditional value-at-risk, to prediction error measurement, etc. A confidence interval for the TAR for each class in the classification schema is evaluated using the bootstrap method. Generally, the bootstrap method enables empirical estimates of accuracy (bias, variance, and confidence intervals) to sample estimates through random sampling with replacement (“resamples”) (i.e., mimicking the experimental sampling process) (Davison and Hinkley 1997; Efron 1979; Efron and Tibshirani 1994; Hall and Martin 1988). This is particularly important as it allows estimation of the sampling distribution for the interclass bias and provides a foundation for analysis of the statistical power the measure of interclass bias. Furthermore, this method is advantageously data driven instead of requiring detailed model knowledge, which can be difficult to ascertain with a presented facial verification algorithm. Under the bootstrap method, the similarity scores are partitioned into sets based on the respective class in the classification schema. Then from each of the partitioned sets, a large number of resamples are constructed from the respective similarity, from there the threshold is varied over a sufficiently large range tabulating the resulting TAR (or another indicator of interest).

Precisely, this can be stated as given the similarity scores, returned by a facial verification algorithm, of size $k = \sum_{i=1}^{I} n_i(n_i - 1)$, say $X_1, X_2, \dots, X_k$. We further assume these similarity scores are independent and identically distributed (i.i.d). Additionally, we aim to construct a confidence interval for the TAR $\vartheta := \vartheta(P)$, where P is the similarity scores distribution and $\vartheta: \mathcal{P} \to \mathbb{R}$ is a function with $\mathcal{P}$ as the set of all distributions on the data domain. We can construct a point estimate of $\vartheta(P)$, $\hat{\vartheta}_k(P) := \vartheta(\hat{P}_k)$ where $\hat{P}_k(t) := {}^{1}\!/_{n} \sum_{i=1}^{k} I(X_i \in t)$, s.t $t \in \tau$ is the empirical distribution constructed from the data, $I(\cdot)$ denotes the indicator function, t is the threshold, and τ is the set of all possible threshold values.

Next, we construct a confidence interval for ϑ as follows. For each replication $b = 1, 2, \dots, B$, we independently and uniformly sample with replacement from $\{X_1, X_2, \dots, X_k\}$ k times, to obtain $\{X_1^{*b}, X_2^{*b}, \dots, X_k^{*b}\}$ (i.e., resample the dataset), and evaluate the resample estimate $\vartheta_k^{*b} := \vartheta(P_k^{*b})$, where $P_k^{*b}(t) := {}^{1}\!/_{n} \sum_{i=1}^{k} I(X_i^{*b} \in t)$ is the resample empirical distribution.

Therefore, the two-sided distribution free conservative $100(1-\alpha)\%$ confidence interval can be stated as:

$$\mathcal{J} = \left[\hat{\vartheta}_k - t_{B,1-\alpha/2}S \quad \hat{\vartheta}_k + t_{B,1-\alpha/2}S\right]$$

Where the critical value $t_{B,1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of t_B, the student t-distribution with degree of freedom B, and S^2 is the sample variance of the resample estimates, centered at the original point estimate instead of the resample mean:

$$S^2 = \frac{1}{B-1}\sum_{b=1}^{B}\left(\vartheta_k^{*b} - \hat{\vartheta}_k\right)^2$$

This process can be used to computed the bootstrap confidence intervals for each class in the classification schema. Taken all together, the curves represent the facial verification algorithms TAR for individuals across all classes in the classification schema. This information may be represented graphically to qualitatively understand the severity or lack thereof of interclass bias between classes in the same classification schema. While the author believes that there is significant power to a graphical viewpoint of this interclass bias, as it provides a visual notion of distance between the various curves and their confidence intervals, this allows for more qualitative analysis of the features in the TAR plots.

Quantitative Analytical Methods

The author posits that creating a measure of interclass bias can be incorporated into an algorithm's design, implementation, or training processes and an end user's testing and commissioning processes; as such, they recognize that it's important to reduce these distributions to a score. This score should be numerical and continuous to support the use of the score during a facial verification algorithm's training process.

The author considers the problem of detecting differences between TAR curves representative of differing classes in the classification schema. The underlying assumption therefore is that, under the null hypothesis of equality between curves, the area between them is zero commonly stated as:

$$H_0\colon \vartheta\left(P^{(i)}\right) = \vartheta\left(P^{(j)}\right) \text{ versus } H_A\colon \vartheta\left(P^{(i)}\right) \neq \vartheta\left(P^{(j)}\right)$$

For all $t \in \tau$, where P is the similarity scores distribution as defined above; and $P^{(i)}, P^{(j)}$ are the similarity scores distributions for classes in the defined classification schema (i) and (j) respectively. Under this hypothesis the area between the TAR curves must be zero under the null:

$$d_{l_2}\left(\vartheta\left(P^{(i)}\right), \vartheta\left(P^{(j)}\right)\right) \equiv \left(\int_{\tau} \left(\vartheta\left(P^{(i)};\ t\right) - \ \vartheta\left(P^{(j)};\ t\right)\right)^2 dt\right)^{\frac{1}{2}}$$

This d_{l_2}can be calculated for all classes in the classification schema; however, the author posits it is beneficial to consider the interclass bias not between each and every class in a classification schema, but instead between a class and the overall performance.

As such, the author describes one method for calculating an overall performance TAR curve. The author postulates that in an ideal facial verification algorithm all classes in a classification schema should perform equally; however, the author acknowledges that depending on the classification schema selected it may be appropriate to weigh the performance of certain classes over others. Therefore, a simple overall performance TAR curve can be constructed as the average of the TAR for all classes. Once the overall performance is established—any deviations from the facial verification algorithm for a specific class from the overall performance are an anomaly to measure—which combined with the d_{l_2} measure established above can yield a continuous score.

In order to create this score, the author modifies the aforementioned methodology to create the TAR plots and corresponding confidence intervals described above. Similarly, confidence intervals for the TAR for each class in the classification schema are evaluated using the bootstrap method. However, in this iteration, for each of the replication $b = 1, 2, \ldots, B$, bootstrap samples generated, not only are the similarity scores resampled for each class in the classification schema but also the overall performance curve is calculated for that resample, following the methodology outlined above. Furthermore, for this replication, the difference between a given class and the overall performance is calculated for all classes. Now bootstrap confidence intervals, in a similar fashion outlined above, can be calculated for the difference between each class in the classification schema and the overall performance. Lastly, the author calculates the d_{l_2} measure for the absolute value of each of these difference confidence intervals. So far, we have described the interclass measure of bias for a specific class in a classification schema, but to truly characterize the performance across all classes this information must be reduced once more. Here the author makes a point that in order to evaluate a facial verification algorithm's interclass bias, it is prudent to evaluate it by its worst performing class, that is to say the outlier is ***the*** problem to address. Furthermore, as articulated earlier, due to the impact that false negatives can have in environments that utilize facial verification algorithms, the author believes that it is best to characterize an algorithm by its worst performance, as that is where its effects will be most greatly felt. Therefore, the author states that the interclass measure of bias for a given classification schema is the maximum the difference between a class and the overall performance for any class in a classification schema.

Data Collection for the Commercial Facial Verification Algorithms Case Study

To show how this measure can be implemented, the author undertakes a case study measuring the interclass bias of two commercial facial verification algorithms. This section focuses on the generation of a set of image comparisons to be submitted to the commercial facial verification algorithm providers for evaluation that proceeds as follows: a) dataset selection, and b) classification schema selection, followed by c) dataset subject and image sanitization, d) subject template selection, and concludes with e) generating the comparison protocol.

Selection of Commercial Facial Verification Algorithm providers

The author limited their discussion of this case study to commercial facial verification algorithm providers sold in programing interface ("API") bundles. Microsoft Azure Cognitive Services Face API ("*Microsoft Face API*") and Amazon Web Services (AWS) Rekognition were selected as both companies have made large investments in artificial intelligence and facial recognition systems and provide public access and demonstrations of their commercial facial verification algorithms. Following the methodology outlined above, to conduct the audit of the selected commercial facial verification algorithm providers, *Microsoft Face API* and *AWS Rekognition*, a classification schema must first be selected.

Classification Schema and Dataset Selection

It is now well established from a variety of studies that facial recognition systems are failing to recognize Black and dark-skinned faces due to unbalanced datasets that comprise the training data the facial recognition systems learn from (Noble 2018), large-scale bias in the form of systematic misrecognition by skin color or ethnic background, and gender classification (Buolamwini and Gebru 2018; Crawford and Paglen 2019; Klare et al. 2012; Ngan and Grother 2015). Based on claims made by these author as to the amount of bias exhibited by these providers, and on the distribution of data provided in earlier toy analyses conducted by the author, the author understands the value of utilizing a large corpus of images to generate these comparisons. Based on the asymptotic relative efficiency adjustment to the Mann-Whitney method for determining sample size, the author estimated that 2,790 to 8,044 subjects might be necessary to measure the gender or skin color bias evidenced by these commercial facial verification algorithm providers, respectively (Conover 1971; Mann and Whitney 1947). However, there are few publicly available datasets of images large enough to meet the bar of statistical significance created by the above measure. As illustrated by **TABLE 3** the largest datasets for facial recognition systems are primarily restricted to private corporations such as Facebook, Google and Megavii. Furthermore, only one is labeled sufficiently large as to generate conclusions about the gender or skin color bias of the commercial facial verification algorithm providers: the IARPA Janus Benchmark C ("IJB-C") dataset.

TABLE 3 *An overview of known public and private face datasets. This table shows some of the face datasets available at the time of writing. This list is not meant to be exhaustive, nor to describe the datasets in detail, but merely to provide a sampling of the types of datasets that are available. Where possible, a peer-reviewed paper or technical report was cited, and otherwise a citation referring to the webpage for the database is given when available. (Chen et al. 2014; Huang et al. 2007; Maze et al. 2018; Schroff et al. 2015; Sun et al. 2013; Taigman et al. 2014a; Wang et al. 2018a; Yi et al. 2014; Zhou et al. 2015)*

Dataset	*Identities*	*Images*	*Availability*
Google Face Dataset	8M	260M+	Private
Megavii Face Classification	20,000	5.0M	Private
Social Face Classification	4,030	4.4M	Private
VGG Face Dataset	2,622	2.6M	Public
IMDb-Face	59,000	1.7M	Public
CASIA-WebFace	10,575	494k	Public
CelebFaces	10,177	202k	Private
Cross-Age Celebrity Dataset (CACD)	2,000	163k	Public
IARPA Janus Benchmark C (IJB-C)	3,531	21k	Public
Labeled Faces in the Wild (LFW)	5,749	13k	Public

Dataset Selection

The IJB-C was selected for this study. This is a US government benchmark released by the National Institute of Standards and Technology (NIST) in 2018. The dataset consists of unconstrained still images, frames, and videos of celebrities and internet personalities “in the wild” collected from the web. In this dataset each subject ($\mathrm{n} = 3{,}531$) has a variety of individually labeled enrollment templates. The subject pool is the largest, most geographically diverse of any public recognition dataset.

Classification Schema Selection

The constraint imposed by the publicly accessible dataset IJB-C drives some of the choices of the case study used to evaluate the success of this measure. One such constraint is the definition of the classification schema used to measure interclass bias. Each of the images in the IJB-C is individually labeled with one of six skin tone classes (Light Pink, Light Yellow, Medium Pink / Brown, Medium Yellow / Brown, Medium-Dark Brown, and Dark Brown). Importantly, annotations of the imagery were added utilizing the crowdsourcing service, Amazon Mechanical Turk, which caused labels to be defined in laymen’s terms instead of strict scientific definitions.

Dataset Subject and Image Sanitization

Some subjects had conflicting skin tone labels from one labeled image to another, explained either due to the variability in labeling due to the layman’s definition or the crowd-sourced nature of the dataset labels, introducing some complexity. While there are a variety of conditions that may affect a person’s skin tone (e.g., depigmentation due to injuries to skin, vitiligo, tinea versicolor, albinism, or skin whitening), or affect the way a subject’s skin tone is represented in an image (e.g., poor illumination, poorly exposed images, default camera calibrations optimized for lighter skinned individuals), the author believes that any changes significant enough to yield a different label, would unnecessarily add complexity and error to the analysis undertaken in this dissertation. Given these circumstances, these subjects with conflicting labeled templates were removed from consideration reducing our pool to 3,514 from its original number of 3,531 ($n = 3{,}514$).

Prior work has established that commercial facial recognition technology is built upon “subject-specific modeling,” wherein a single template is generated for a subject based upon most, if not all, available pieces of facial images and media of that subject. The authors of the IJB-C set out to mimic that behavior in their testing suite by using multi-image templates. Additionally, they believed that “the inherent difficulty of the dataset is obfuscated”, as algorithms have the ability to pool information from multiple pieces of media through subject-specific modeling. However, it is important to note that while *AWS Rekognition* and *Microsoft Face API*, allow for multi-image templates, neither requires it, and both provide samples for end-users looking to use single image templates. As such this case study focuses of the “inherently difficult” challenge of one image gallery templates, these templates are used in the generated comparisons for measurement of the interclass bias. Similar to the authors of the IJB-C, the author of this dissertation strongly believes that it is important to evaluate these commercial facial verification algorithms in this case study in circumstances similar to the ones allowed for end-users by these algorithms. However, we believe this option requires testing single image templates, as this is still an accepted use-case of these algorithms and provide a more difficult testing criteria to evaluate them against, commensurate with the potential harm the subjects compared by these commercial facial verification algorithm algorithms might endure due to the relaxed comparisons allowed by the providers.

The authors of the IJB-C set a minimum face size of 36 pixels $\times 36$ pixels for all of the templates in the dataset, as they believe that there is a dearth of identifiable information at lower resolutions. This is in alignment with the minimum face size of commercial facial verification algorithm provider from *Microsoft Face API*. However, the other commercial facial verification algorithm provider, *AWS Rekognition*, requires a minimum template of 50 pixels $\times 50$ pixels. In order to enable stronger comparisons between multiple providers for this case study, the author strongly believe that both commercial facial verification algorithm providers should be assessed with the same facial images, as such any templates that failed to meet the more stringent (i.e., 50 pixels $\times 50$ pixels) of these requirements was eliminated from the dataset used for subsequent consideration. This yielded a total of $11{,}075$ remaining templates.

Subject Template Selection

After undergoing the procedure for sanitizing and ensuring consistency and agreement between subject templates above, the dataset provides a minimum of 4 image templates per subject, but imposes no upper bound on the number of templates for each subject. A plain reading of the documentation for the IJB-C does not outline a rationale or provide a justification for their choice of having a minimum of four image templates per subject. As explained earlier and as general rule, the greater the number of templates provided for a subject the better one can model how well a specific facial verification algorithm performs an individual subject. However, due to the fixed limitations of this dataset, each additional template required per subject reduces the number of subjects available for analysis. As such the author proceeds in a brief exploration of statistical power efficiency for the IJB-C, to evaluate if the same minimum template size is appropriate for this comparison protocol. Considering, each subject (s_i) in the dataset has (n_i) templates, then it is trivial to compute the number of subjects who have at least j templates ($\{s_i | n_i > j, \forall i\}$). If this is conducted for each possible value of j, a curve can be drawn showing how many subjects can be analyzed under these constraints. This analysis can be repeated for each of the subject presentation covariates under investigation to yield a more detailed view of the dataset, as evidenced in **FIGURE 5**, below.

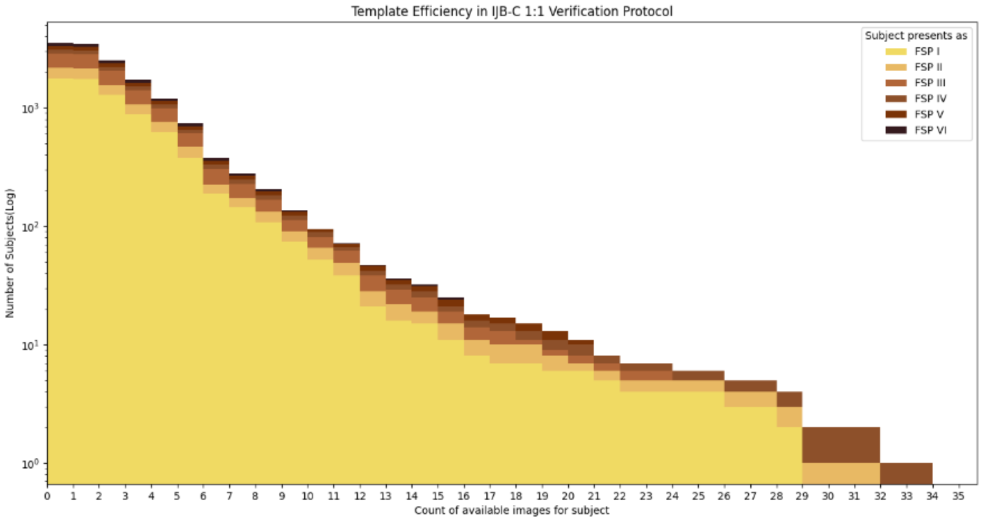


FIGURE 5 A log-scale histogram of the number of templates each subject has in the dataset after undergoing the collection methodology, coded by subject skin tone presentation.

Due to the nature of the dataset, there is no inflection point where the dataset loses a number of subjects. Therefore, the author ensures to select a minimum required number of templates (j) per subject to maintain a minimum number of subjects in each of the six labelled skin tone classifications to ensure robust statistical power in the calculation for interclass bias. The author decided to require that at least 30 subjects, as some of the assumptions which underpin the desired statistical analyses, based on the Central Limit theorem fall apart with fewer subjects (Hogg et al. 2015). It's noteworthy to point out that the IJB-C dataset has significantly more light skinned subjects (i.e., the proportion of light skinned subjects in the IJB-C is 79.6%) than dark skinned subjects, and the subjects are not evenly distributed amongst the six classifications. Furthermore, after removing templates that did not meet the minimum face size of $50 \text{ pixels } \times 50 \text{ pixels}$, as established earlier, the problem is exacerbated. With this requirement it was determined that for the interclass bias measure for the six (6) class skin tone classification schema used by the IJB-C no more than 5 templates could be required per subject. Given these constraints, there seemed little benefit to introduce stricter standards than those initially set by the IJB-C, that is to say the author believes the reduction in the number of subjects did not seem commensurate with the marginal benefit of one additional template per remaining subject. As such, any subjects who had fewer than 4 image templates were eliminated from the sample ($n = 2{,}311$).

Generating the Comparison Protocol

The protocol used seeks to compare facial imagery from two images to verify whether or not they are the same person. In other words, to compare faces as may be done in an access control type scenario. This section specifies the approach used for generating the test protocol. The protocol consists of creating a series of genuine comparisons between two facial images of the same person, for submission to a commercial facial verification algorithm provider.

Following these criteria, a set of four images was created for each subject by taking a simple random sample without replacement from all the still imagery available for that subject. Therefore, the selected set may include some of the highest or lowest quality pieces of media for that subject. The author believes that this random selection without replacement to generate a fixed set size for a subject, mirrors the random decisions exhibited by the original media acquisition process for the IJB-C that yielded some subjects that only had four still images that met the inclusion criteria established earlier. Finally, the collection process yields a dataset with 1,203 subjects and a total of 4,812 templates, with each subject guaranteed to have 4 image templates ("the dataset").

The comparison protocol and its list of comparisons and subject templates were submitted to commercial facial verification algorithms: *Microsoft Face API* and *AWS Rekognition* for evaluation. A total of $14,436$ comparisons were submitted to each algorithm. Importantly, *Microsoft Face API* provides end-users with a choice of three detection models (i.e., detection_01, detection_02, detection_03), used to detect faces in a submitted image, and four recognition models (i.e., recognition_01, recognition_02, recognition_03, recognition_04), used to extract face features to facilitate comparisons. These models are continually supported by Microsoft to ensure backwards compatibility, and if an end-user does not specify otherwise, they default to detection_01 and recognition_01. For each comparison, *Microsoft Face API* reported a confidence score between 0 to 1 representing the algorithm's "confidence of whether two faces belong to the same person"; additionally, each comparison also reported a binary state that represented if the two faces belong to the same person, which was set to report true if the confidence score was greater that 0.5 (Microsoft n.d.). Similarly, *AWS Rekognition* reported a similarity score between 0 and 100 representing a "statistical measure of how likely two faces in an image are the same person, when analyzed by the algorithm" (Amazon Web Services, Inc n.d.; Amazon Web Services, Inc. n.d.). AWS recommends a 99% threshold for the similarity score in end-user use cases where highly accurate face similarity matches are important.

Failure to Extract Features

During the comparison protocol, some facial verification algorithms may fail to convert facial imagery to a template. The author adopts NIST's treatment of these failed templates, where any comparison that involves an image for which a failure to extract occurred as producing a zero similarity score (Grother et al. 2019).

Commercial Facial Verification Algorithms Audit Findings

As an initial pass, *Microsoft Face API*, under its default configuration of detection_01 and recognition_01, accepted 75% of the comparisons at its recommended confidence score threshold of 0.5; and *AWS Rekognition* accepted 50% of the comparisons at its recommended similarity score threshold of 99%. For simplicity in making comparisons between the two commercial facial verification algorithm providers, the author has scaled the *AWS Rekognition* similarity score to a value between 0 and 1. For the remainder of this dissertation, the author uses "confidence score" to refer to the *Microsoft Face API* reported confidence scores and *AWS Rekognition* similarity score.

IJB-C Six Skin Tone Audit of Commercial Facial Verification Algorithms

A total of 14,436 comparisons were submitted to each algorithm for scoring, yielding 14,436 confidence scores for each of the commercial facial verification algorithms: (a) *Microsoft Face API* under its default configuration of detection_01 and recognition_01 released in 2017, (b) *Microsoft Face API* with a configuration of detection_01 and recognition_02, (b) *Microsoft Face API* with a configuration of detection_01 and recognition_03, (c) *Microsoft Face API* with a configuration of detection_01 and recognition_04, (d) *Microsoft Face API* with a configuration of detection_02 and recognition_01, (e) *Microsoft Face API* with a configuration of detection_02 and recognition_02 released in 2019, (f) *Microsoft Face API* with a configuration of detection_02 and recognition_03 released in 2020, (h) *Microsoft Face API* with a configuration of detection_02 and recognition_04, (i) *Microsoft Face API* with a configuration of detection_03 and recognition_01, (j) *Microsoft Face API* with a configuration of detection_03 and recognition_02, (k) *Microsoft Face API* with a configuration of detection_03 and recognition_03, (l) *Microsoft Face API* with its latest released a configuration of detection_03 and recognition_04 released in 2021, and (m) *AWS Rekognition*. For each of these commercial facial verification algorithms, the confidence scores reported were partitioned into six skin tone classifications as defined by the IJB-C skin tone classification schema.

True Accept Rates

Subsequently, confidence bounds for the true accept rate ("TAR"), calculated from the fraction of genuine comparisons that correctly exceed the threshold, for each of six skin tone classifications as defined by the IJB-C skin tone classification schema, using the bootstrap method. The author set $\alpha = 10\%$ for the two-sided confidence interval, α was selected based on the expected statistical power from the aforementioned asymptotic relative efficiency adjustment to the Mann-Whitney method for determining sample size. It has been shown that for this significance level that a minimum of 599 bootstrap resamples must be conducted (Davidson and MacKinnon 2000; Wilcox 2010), as computational power and resources were freely available, the author conducted 9999 resamples in keeping with the default choice for the statistical software package used to generate the resamples (Pedregosa et al. 2011). Using these resampled confidence scores, the threshold is varied across the entire domain of the confidence scores to plot the TAR and the confidence bounds for the facial verification algorithms performance for each of the six skin tone classifications. **Figure 6**, **Figure 9**, **Figure 12**, **Figure 15**, **Figure 18**, **Figure 21**, **Figure 24**, **Figure 27**, **Figure 30**, **Figure 33**, **Figure 36**, **Figure 39**, and **Figure 42** plot the TAR and the corresponding confidence score threshold to achieve it under the thirteen commercial facial verification algorithms. **Figure 7**, **Figure 10**, **Figure 13**, **Figure 16**, **Figure 19**, **Figure 22**, **Figure 25**, **Figure 28**, **Figure 31**, **Figure 34**, **Figure 37**, **Figure 40**, and **Figure 43** isolate the TAR for light skinned persons, collectively **Skin Tone I** (Light Pink), **Skin Tone II** (Light Yellow), and **Skin Tone III** (Medium Pink / Brown), and the corresponding confidence score threshold to achieve it under the thirteen commercial facial verification algorithms. **Figure 8**, **Figure 11**, **Figure 14**, **Figure 17**, **Figure 20**, **Figure 23**, **Figure 26**, **Figure 29**, **Figure 32**, **Figure 35, Figure 38**, **Figure 41**, and **Figure 44** isolate the TAR for dark skinned persons, collectively **Skin Tone IV** (Medium Yellow / Brown), **Skin Tone V** (Medium-Dark Brown), and **Skin Tone VI** (Dark Brown), and the corresponding confidence score threshold to achieve it under the thirteen commercial facial verification algorithms.

Microsoft Face API, Released 2017

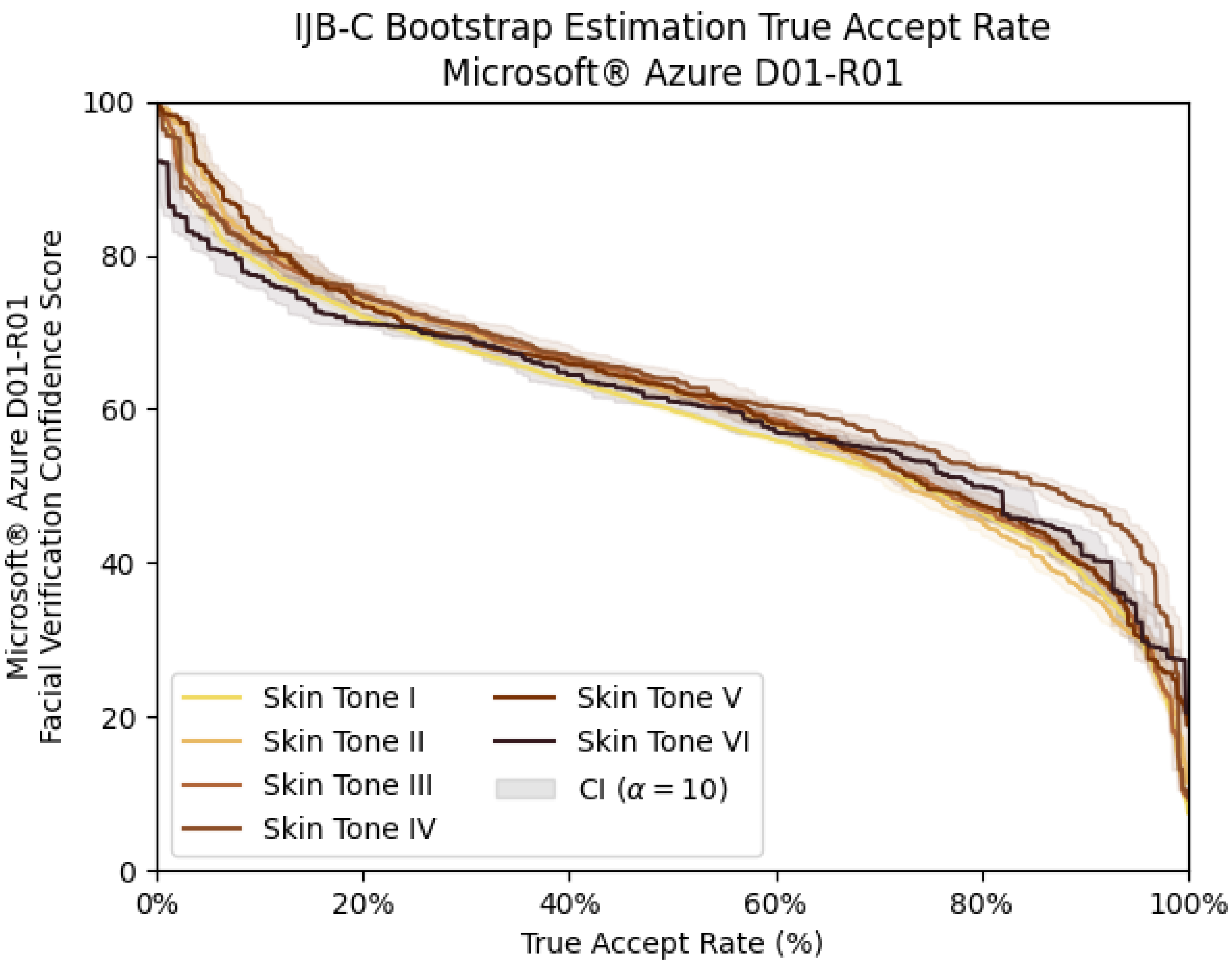


FIGURE 6 The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under Microsoft Face API under its default configuration of detection_01 and recognition_01, released in 2017, for the six skin tone classifications as defined by the IJB-C skin tone classification schema.

Microsoft Face API, Released 2017

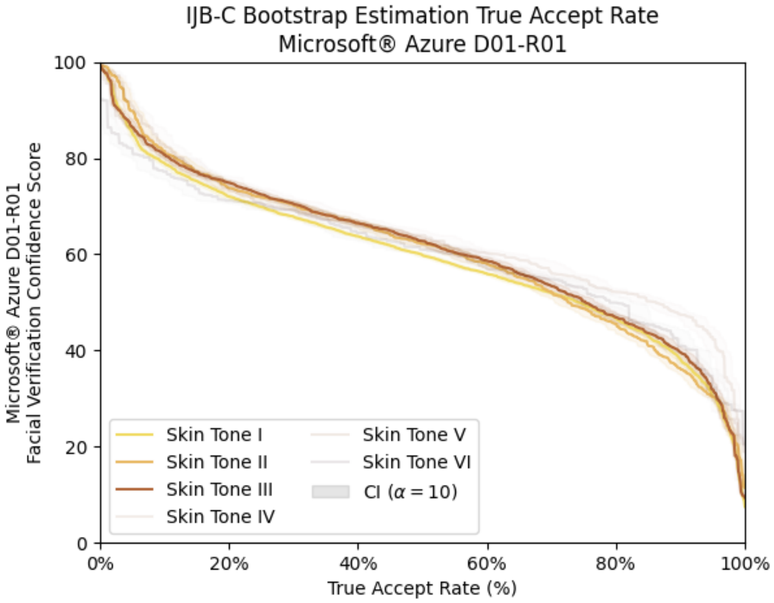


***FIGURE 7** The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under* Microsoft Face API *under its default configuration of* detection_01 *and* recognition_01, *released in 2017, for the three light skinned skin tone classifications, collectively **Skin Tone I** (Light Pink), **Skin Tone II** (Light Yellow), and **Skin Tone III** (Medium Pink / Brown), using the IJB-C skin tone classification schema.*

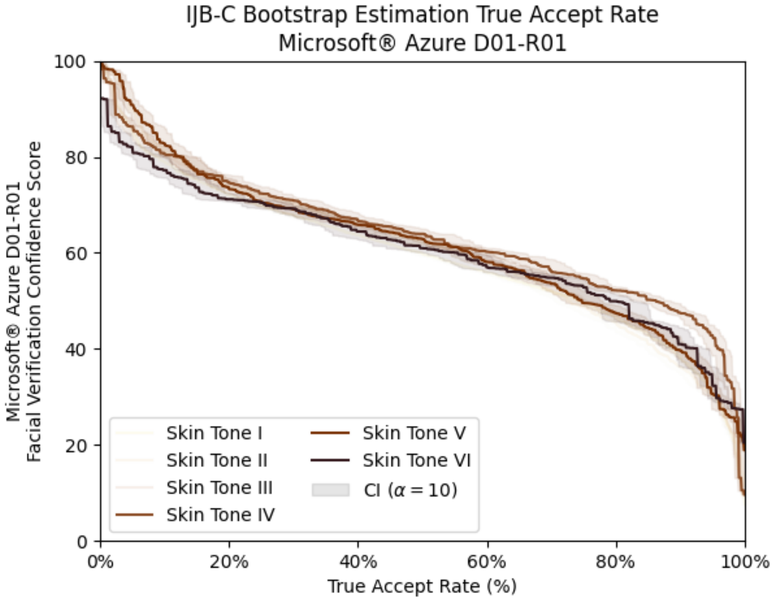


***FIGURE 8** The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under* Microsoft Face API *under its default configuration of* detection_01 *and* recognition_01, *released in 2017, for the three dark skinned skin tone classifications, collectively **Skin Tone IV** (Medium Yellow / Brown), **Skin Tone V** (Medium-Dark Brown), and **Skin Tone VI** (Dark Brown), using the IJB-C skin tone classification schema.*

Microsoft Face API with a configuration of detection_01 and recognition_02

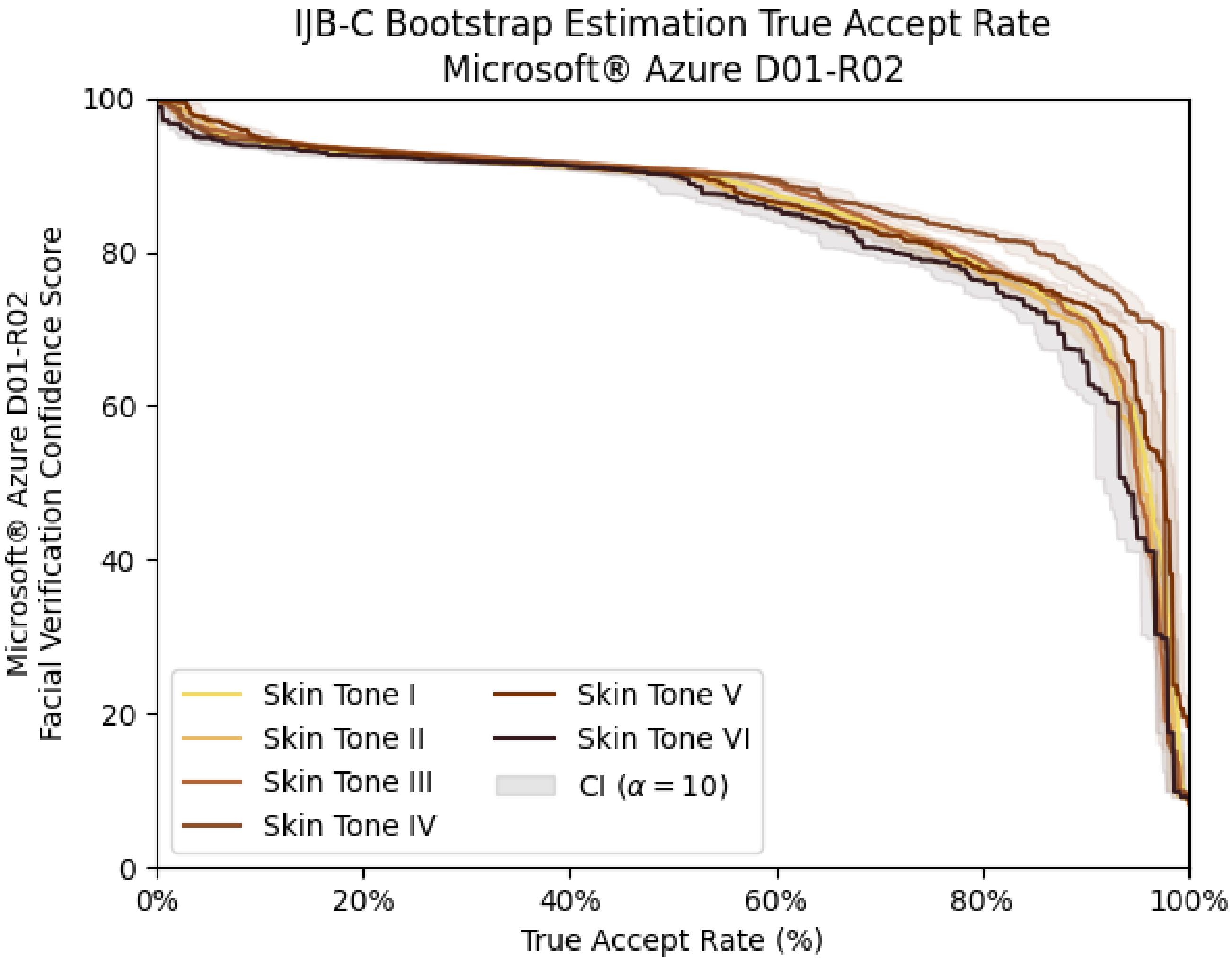


FIGURE 9 The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under Microsoft Face API *with a configuration of* detection_01 *and* recognition_02*, for the six skin tone classifications as defined by the IJB-C skin tone classification schema.*

Microsoft Face API with a configuration of detection_01 and recognition_02

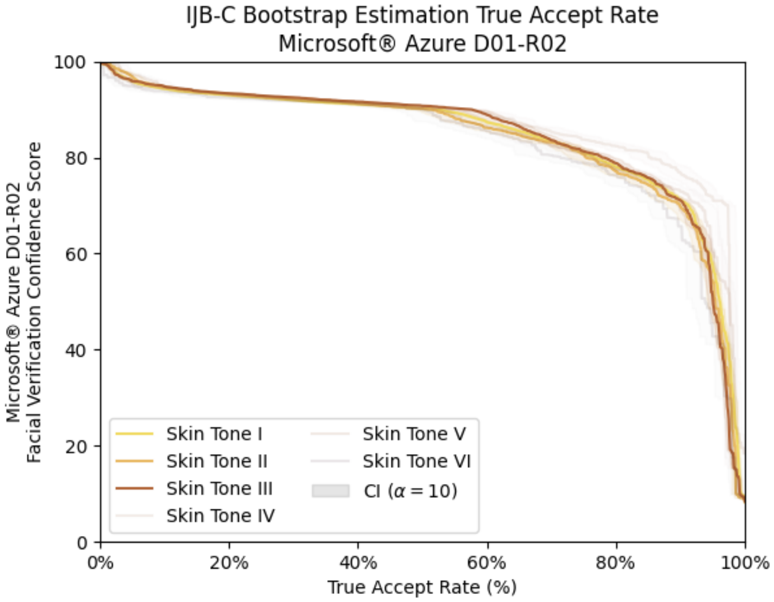


FIGURE 10 *The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under* Microsoft Face API *with a configuration of* detection_01 *and* recognition_02, *for the three light skinned skin tone classifications, collectively* ***Skin Tone I*** *(Light Pink),* ***Skin Tone II*** *(Light Yellow), and* ***Skin Tone III*** *(Medium Pink / Brown), using the IJB-C skin tone classification schema.*

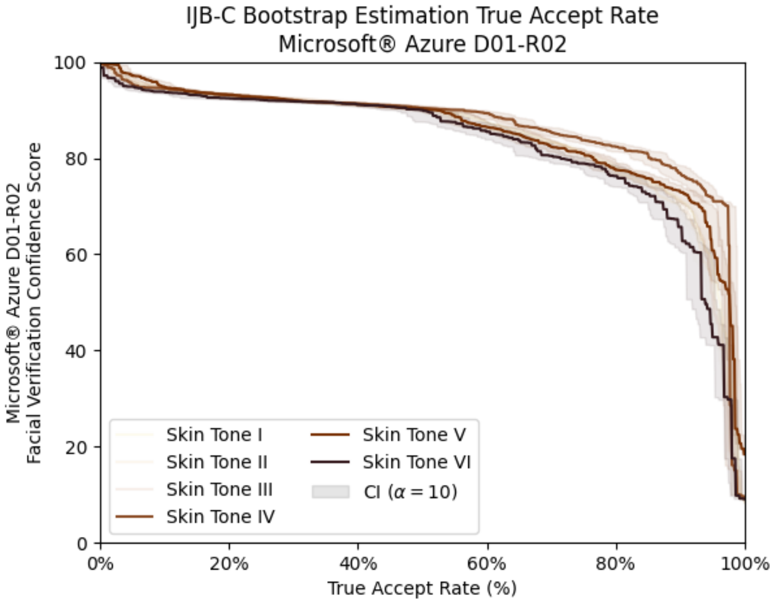


FIGURE 11 *The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under* Microsoft Face API *with a configuration of* detection_01 *and* recognition_02, *for the three dark skinned skin tone classifications, collectively* ***Skin Tone IV*** *(Medium Yellow / Brown),* ***Skin Tone V*** *(Medium-Dark Brown), and* ***Skin Tone VI*** *(Dark Brown), using the IJB-C skin tone classification schema.*

Microsoft Face API with a configuration of detection_01 and recognition_03

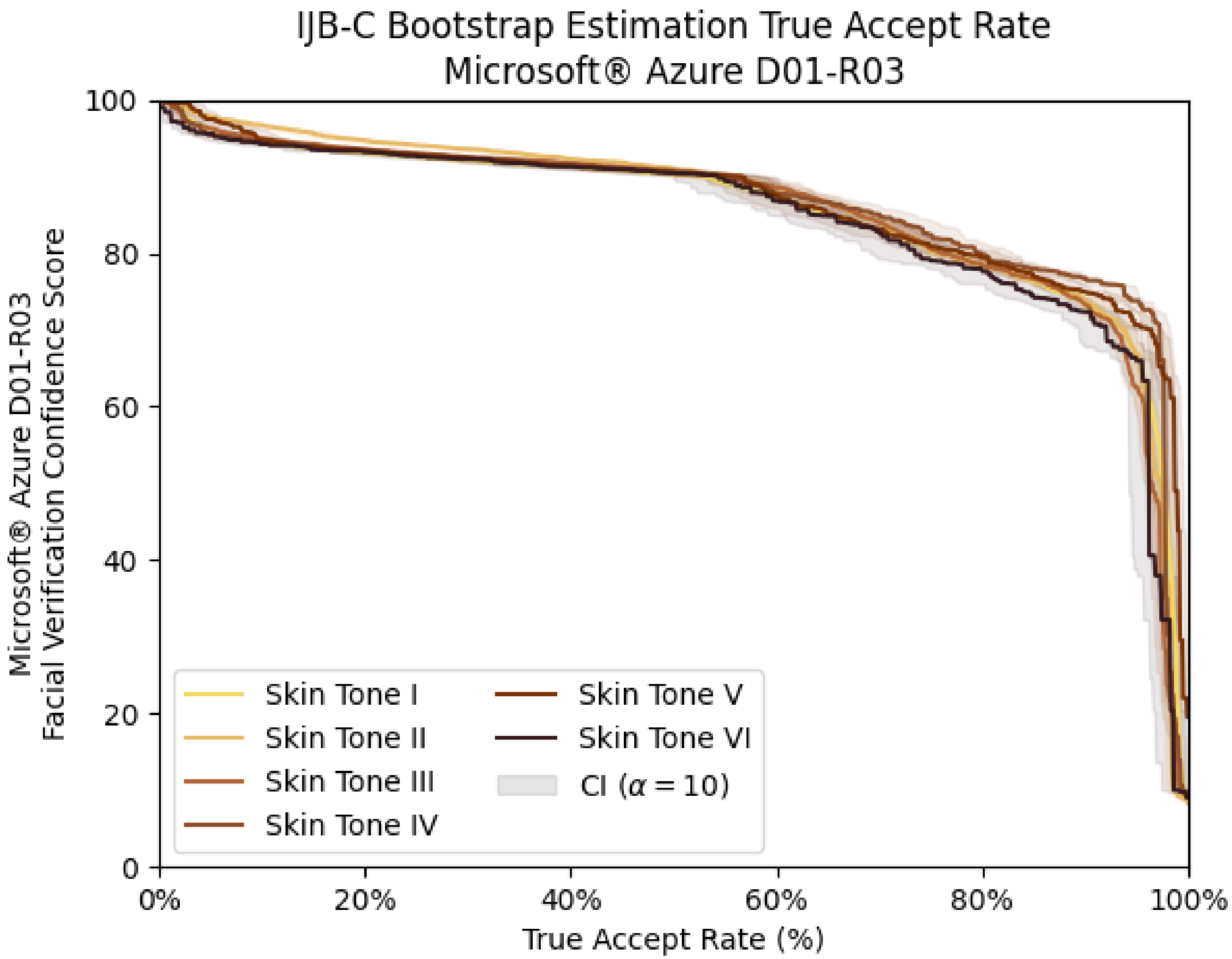


FIGURE 12 The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under Microsoft Face API *with a configuration of* detection_01 *and* recognition_03*, for the six skin tone classifications as defined by the IJB-C skin tone classification schema.*

Microsoft Face API with a configuration of detection_01 and recognition_03

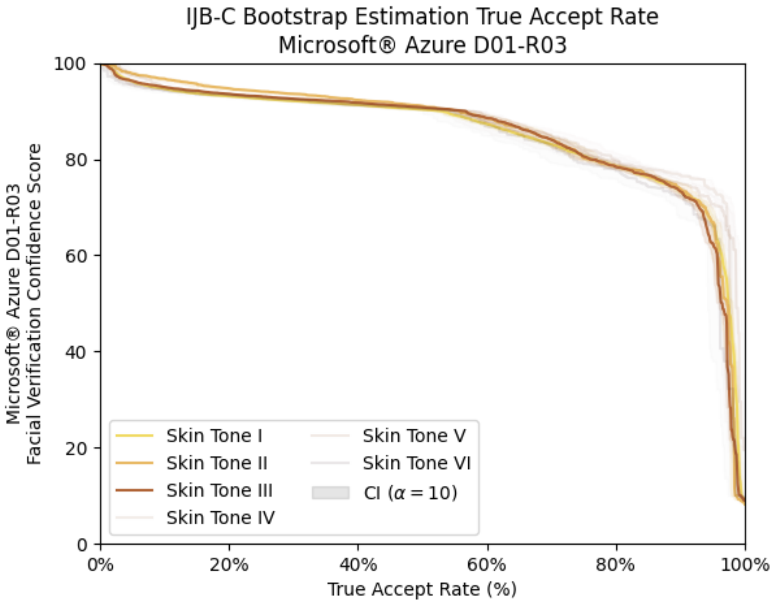


***FIGURE 13** The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under* Microsoft Face API *with a configuration of* detection_01 *and* recognition_03*, for the three light skinned skin tone classifications, collectively **Skin Tone I** (Light Pink), **Skin Tone II** (Light Yellow), and **Skin Tone III** (Medium Pink / Brown), using the IJB-C skin tone classification schema.*

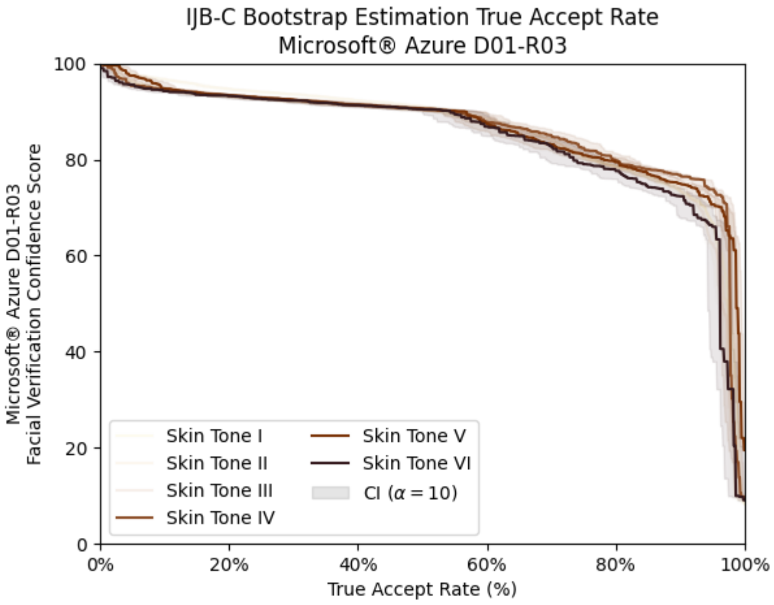


***FIGURE 14** The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under* Microsoft Face API *with a configuration of* detection_01 *and* recognition_03*, for the three dark skinned skin tone classifications, collectively **Skin Tone IV** (Medium Yellow / Brown), **Skin Tone V** (Medium-Dark Brown), and **Skin Tone VI** (Dark Brown), using the IJB-C skin tone classification schema.*

Microsoft Face API with a configuration of detection_01 and recognition_04

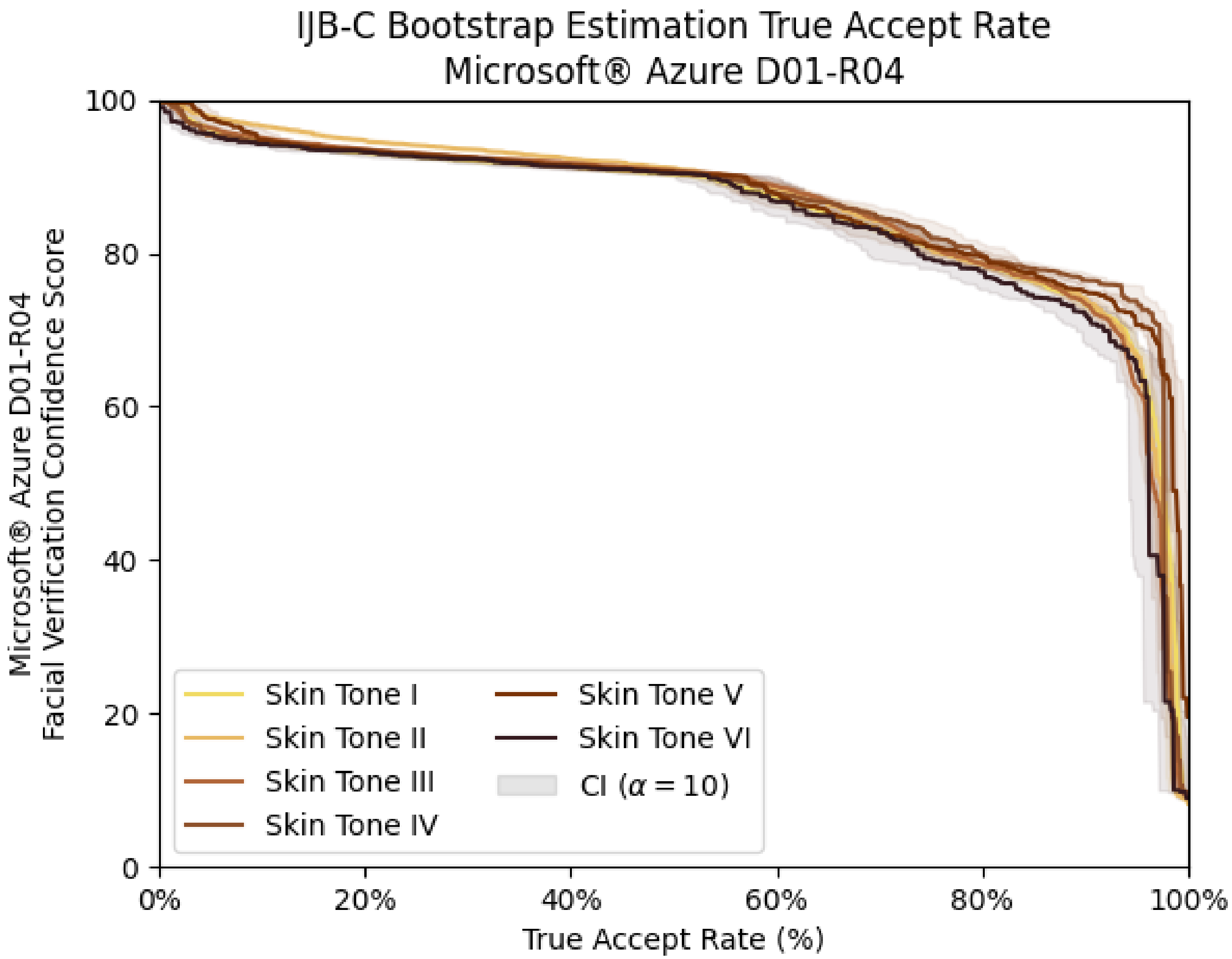


Figure 15 *The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under* Microsoft Face API *with a configuration of* detection_01 *and* recognition_04, *for the six skin tone classifications as defined by the IJB-C skin tone classification schema.*

Microsoft Face API with a configuration of detection_01 and recognition_04

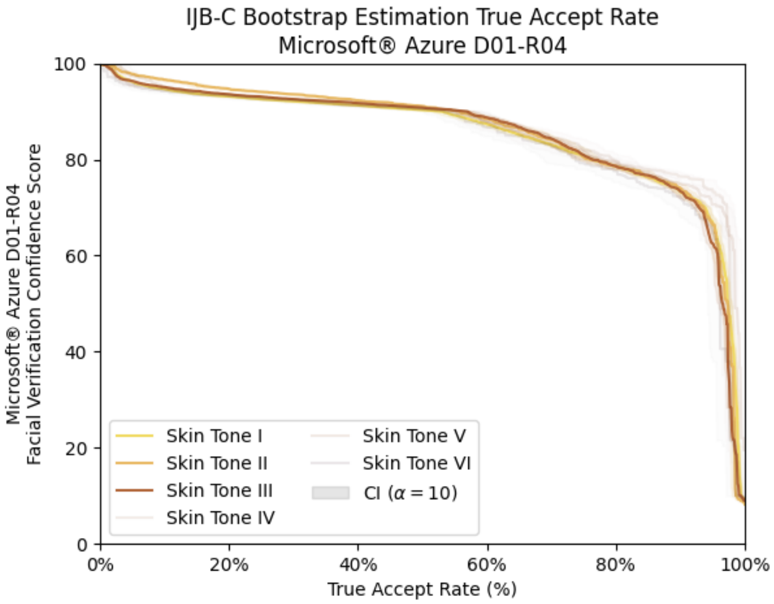


***FIGURE 16** The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under* Microsoft Face API *with a configuration of* detection_01 *and* recognition_04*, for the three light skinned skin tone classifications, collectively **Skin Tone I** (Light Pink), **Skin Tone II** (Light Yellow), and **Skin Tone III** (Medium Pink / Brown), using the IJB-C skin tone classification schema.*

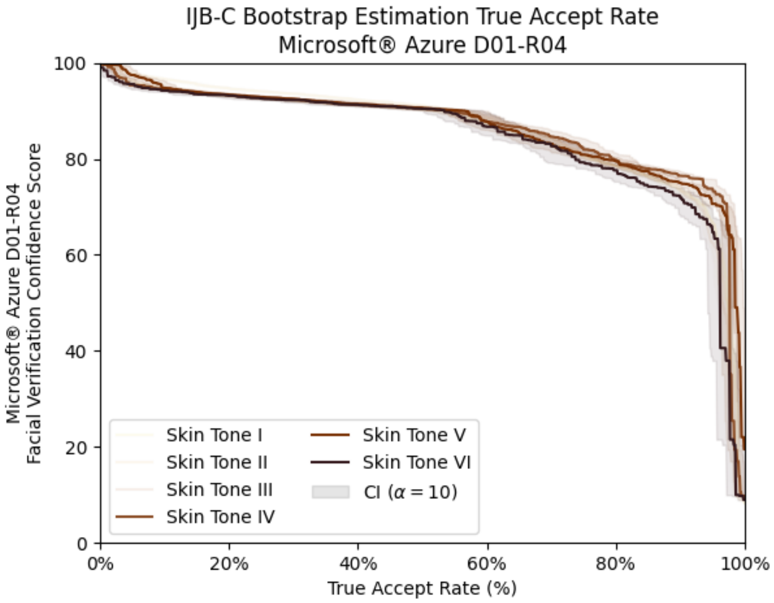


***FIGURE 17** The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under* Microsoft Face API *with a configuration of* detection_01 *and* recognition_04*, for the three dark skinned skin tone classifications, collectively **Skin Tone IV** (Medium Yellow / Brown), **Skin Tone V** (Medium-Dark Brown), and **Skin Tone VI** (Dark Brown), using the IJB-C skin tone classification schema.*

Microsoft Face API with a configuration of detection_02 and recognition_01

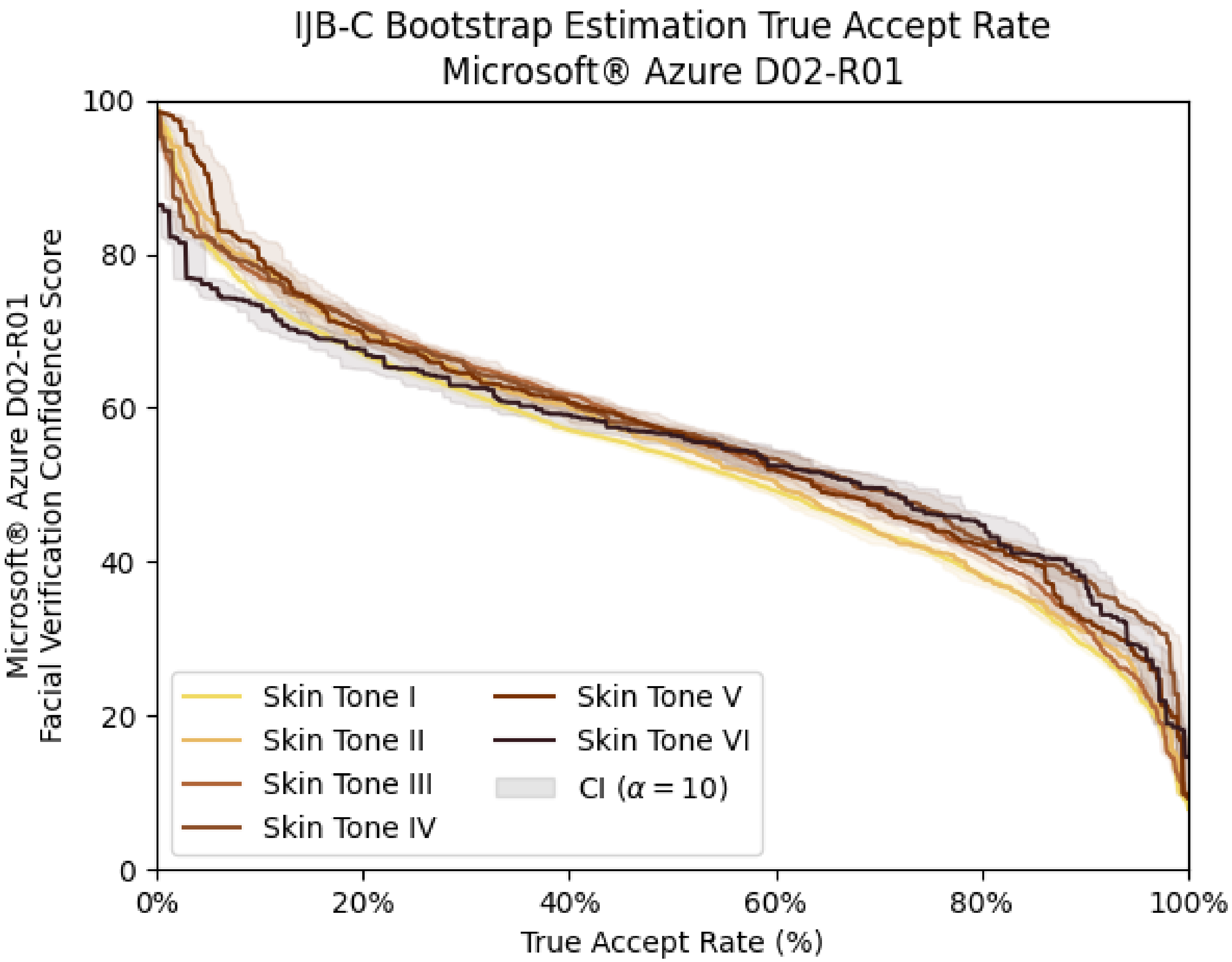


FIGURE 18 The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under Microsoft Face API *with a configuration of* detection_02 *and* recognition_01, *for the six skin tone classifications as defined by the IJB-C skin tone classification schema.*

Microsoft Face API with a configuration of detection_02 and recognition_01

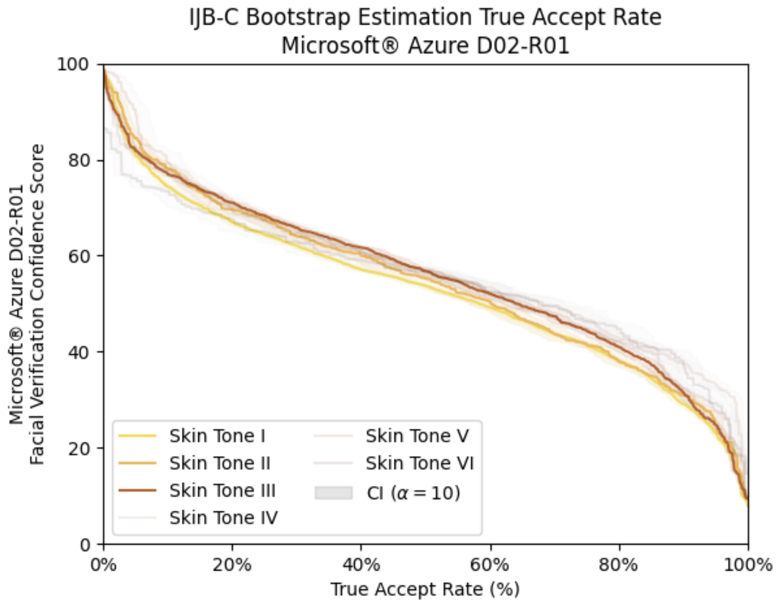


***FIGURE 19** The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under* Microsoft Face API *with a configuration of* detection_02 *and* recognition_01*, for the three light skinned skin tone classifications, collectively **Skin Tone I** (Light Pink), **Skin Tone II** (Light Yellow), and **Skin Tone III** (Medium Pink / Brown), using the IJB-C skin tone classification schema.*

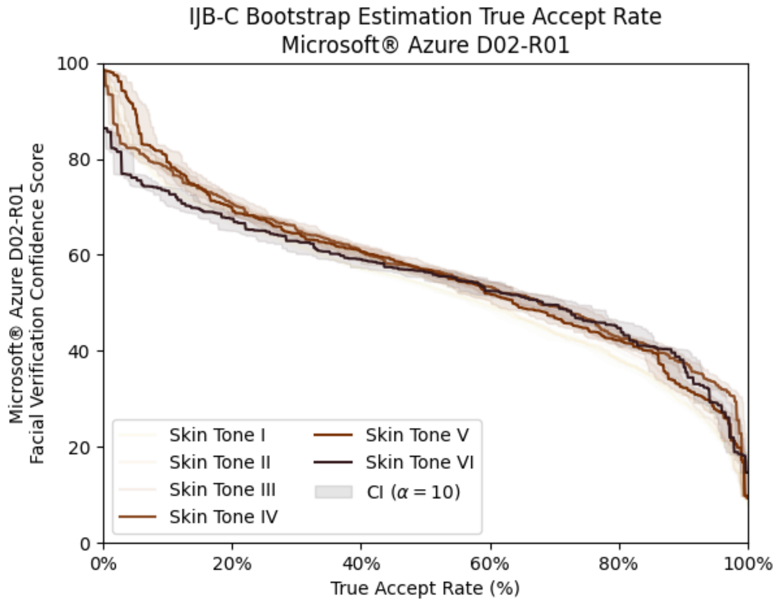


***FIGURE 20** The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under* Microsoft Face API *with a configuration of* detection_02 *and* recognition_01*, for the three dark skinned skin tone classifications, collectively **Skin Tone IV** (Medium Yellow / Brown), **Skin Tone V** (Medium-Dark Brown), and **Skin Tone VI** (Dark Brown), using the IJB-C skin tone classification schema.*

Microsoft Face API, Released 2019

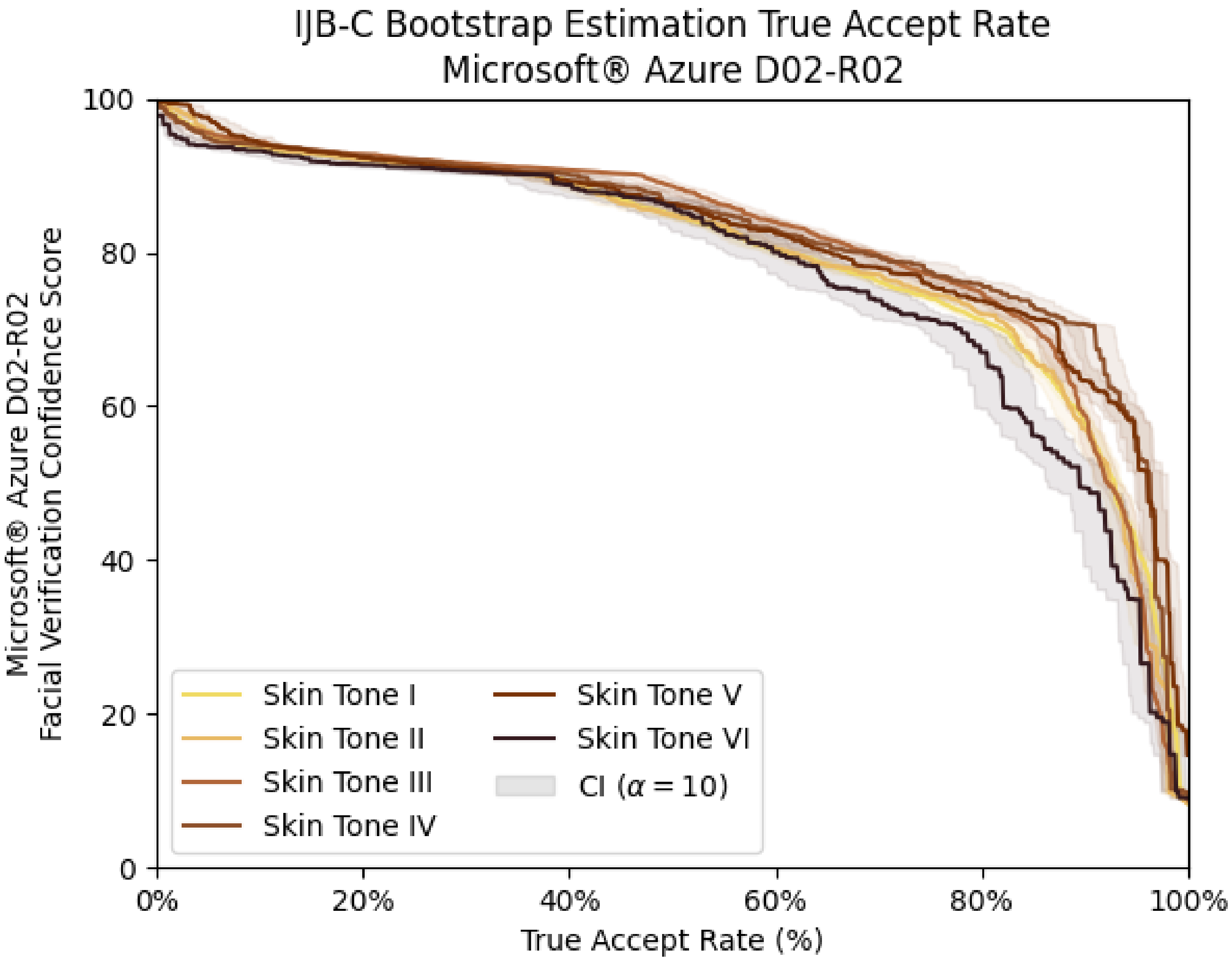


Figure 21 *The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under* Microsoft Face API *under its default configuration of* detection_02 *and* recognition_02, *released in 2019, for the six skin tone classifications as defined by the IJB-C skin tone classification schema.*

Microsoft Face API, Released 2019

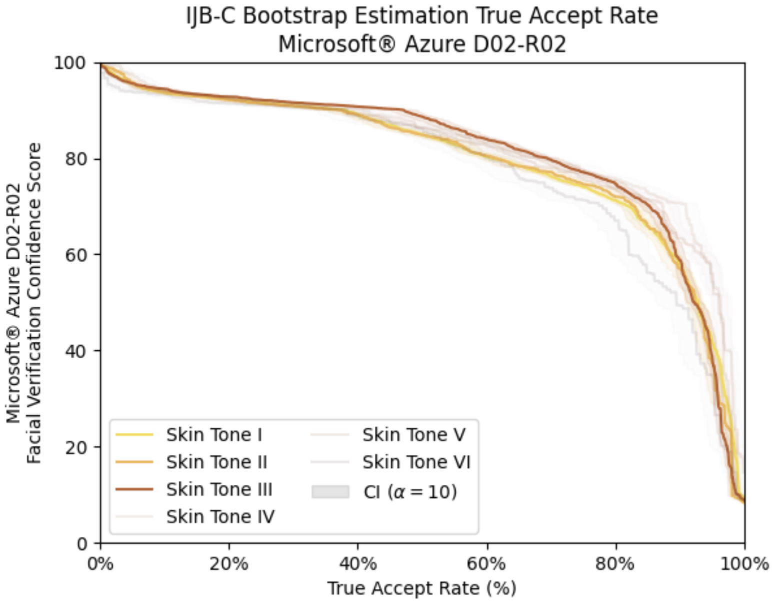


***FIGURE 22** The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under* Microsoft Face API *under its default configuration of* detection_02 *and* recognition_02*, released in 2019, for the three light skinned skin tone classifications, collectively **Skin Tone I** (Light Pink), **Skin Tone II** (Light Yellow), and **Skin Tone III** (Medium Pink / Brown), using the IJB-C skin tone classification schema.*

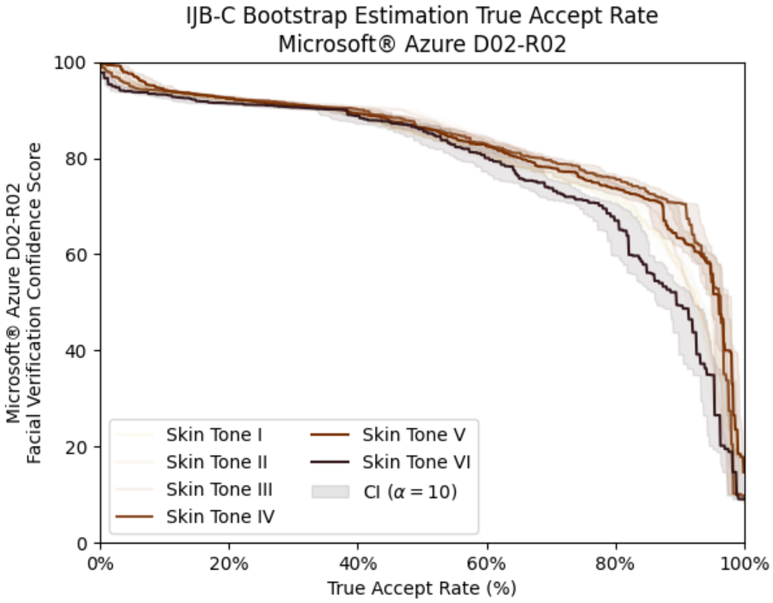


***FIGURE 23** The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under* Microsoft Face API *under its default configuration of* detection_02 *and* recognition_02*, released in 2019, for the three dark skinned skin tone classifications, collectively **Skin Tone IV** (Medium Yellow / Brown), **Skin Tone V** (Medium-Dark Brown), and **Skin Tone VI** (Dark Brown), using the IJB-C skin tone classification schema.*

Microsoft Face API, Released 2020

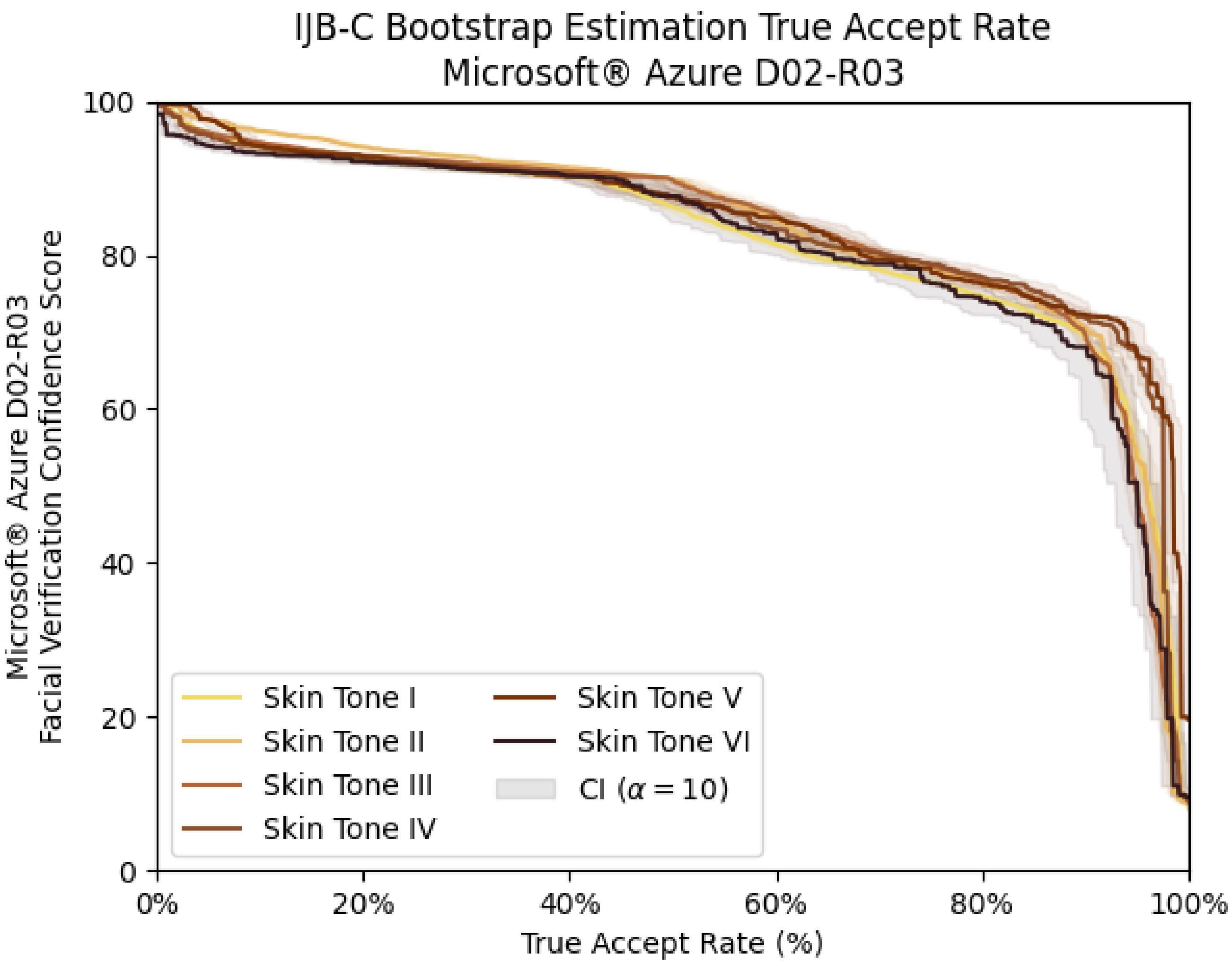


FIGURE 24 The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under Microsoft Face API *under its default configuration of* detection_02 *and* recognition_03, *released in 2020, for the six skin tone classifications as defined by the IJB-C skin tone classification schema.*

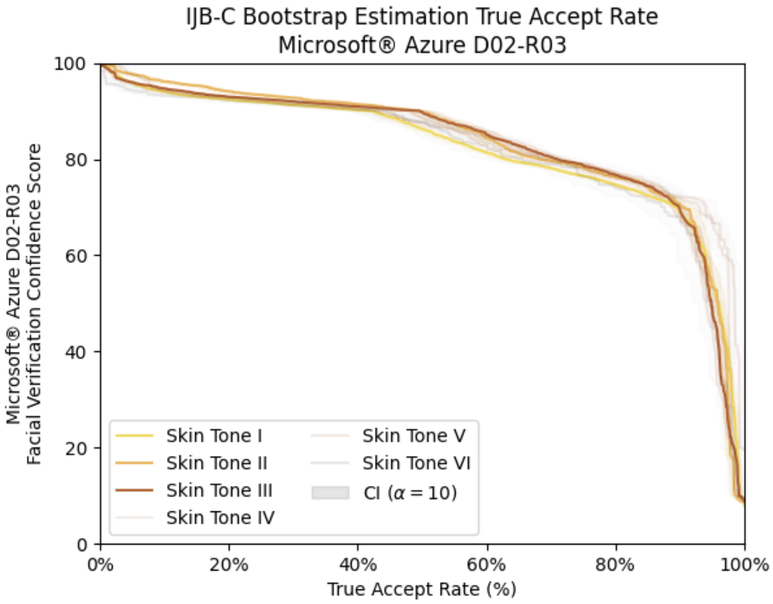


FIGURE 25 *The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under* Microsoft Face API *under its default configuration of* detection_02 *and* recognition_03, *released in 2020, for the three light skinned skin tone classifications, collectively* ***Skin Tone I*** *(Light Pink),* ***Skin Tone II*** *(Light Yellow), and* ***Skin Tone III*** *(Medium Pink / Brown), using the IJB-C skin tone classification schema.*

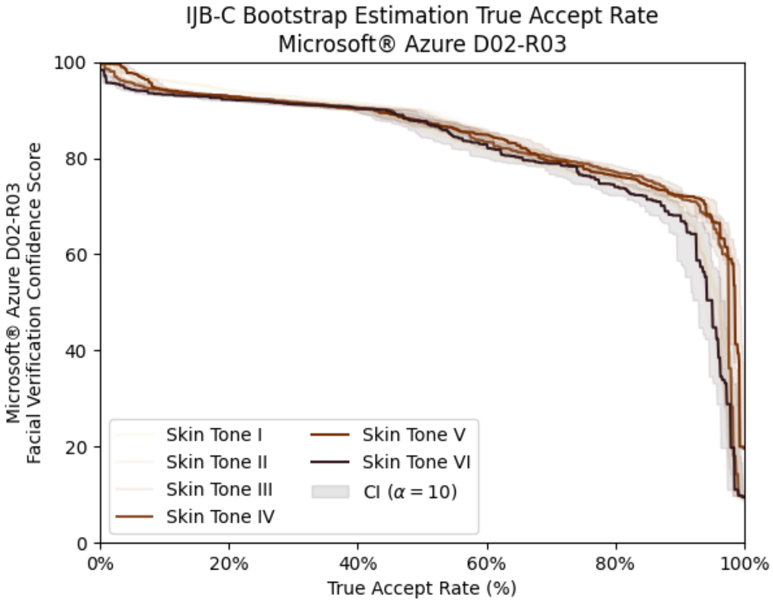


FIGURE 26 *The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under* Microsoft Face API *under its default configuration of* detection_02 *and* recognition_03, *released in 2020, for the three dark skinned skin tone classifications, collectively* ***Skin Tone IV*** *(Medium Yellow / Brown),* ***Skin Tone V*** *(Medium-Dark Brown), and* ***Skin Tone VI*** *(Dark Brown), using the IJB-C skin tone classification schema.*

Microsoft Face API with a configuration of detection_02 and recognition_04

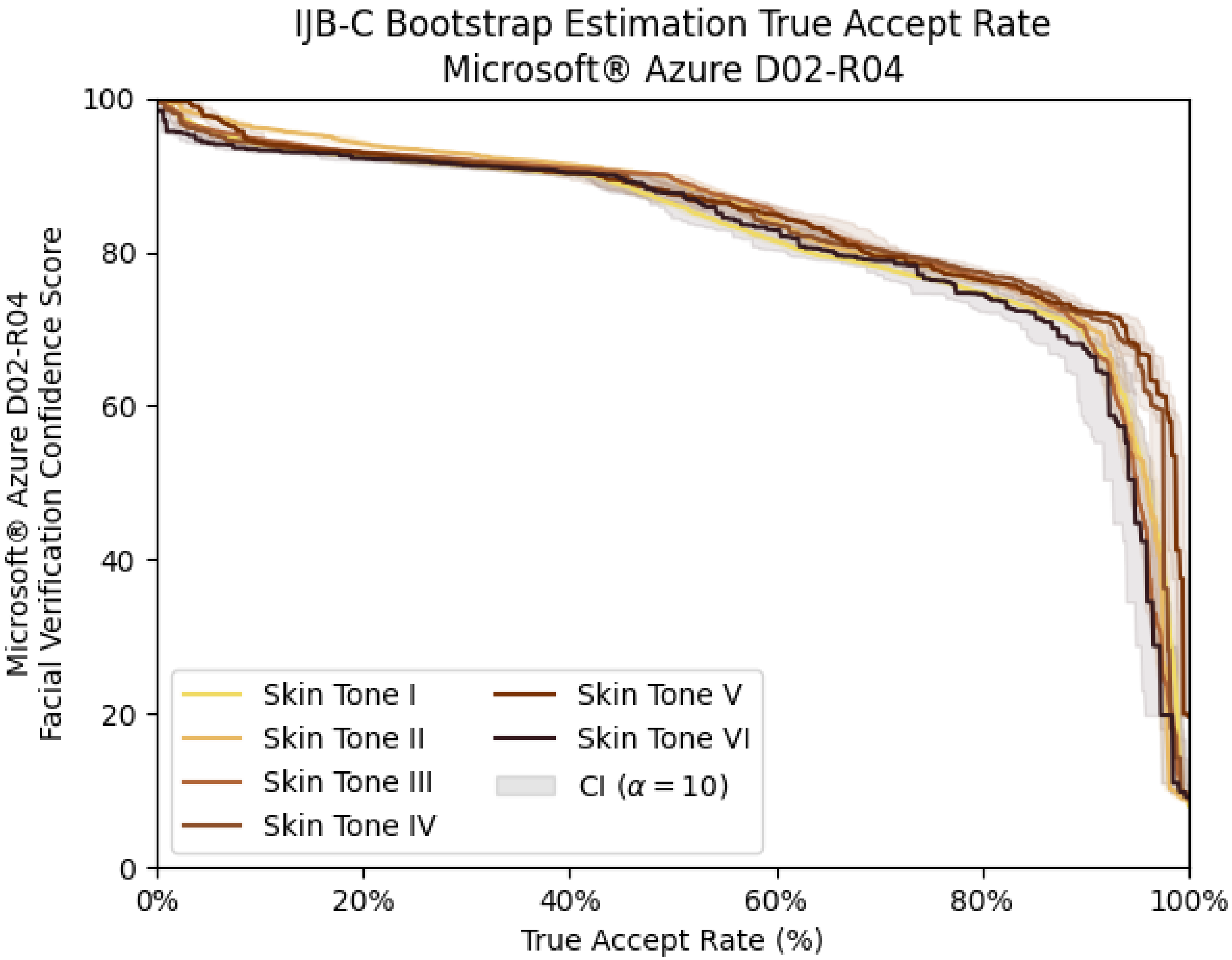


FIGURE 27 The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under Microsoft Face API *with a configuration of* detection_02 *and* recognition_04, *for the six skin tone classifications as defined by the IJB-C skin tone classification schema.*

Microsoft Face API with a configuration of detection_02 and recognition_04

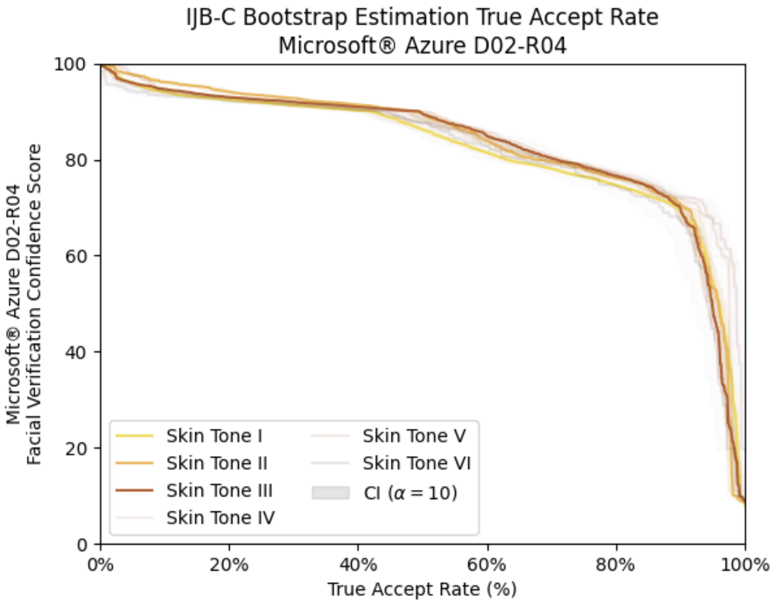


***FIGURE 28** The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under* Microsoft Face API *with a configuration of* detection_02 *and* recognition_04*, for the three light skinned skin tone classifications, collectively **Skin Tone I** (Light Pink), **Skin Tone II** (Light Yellow), and **Skin Tone III** (Medium Pink / Brown), using the IJB-C skin tone classification schema.*

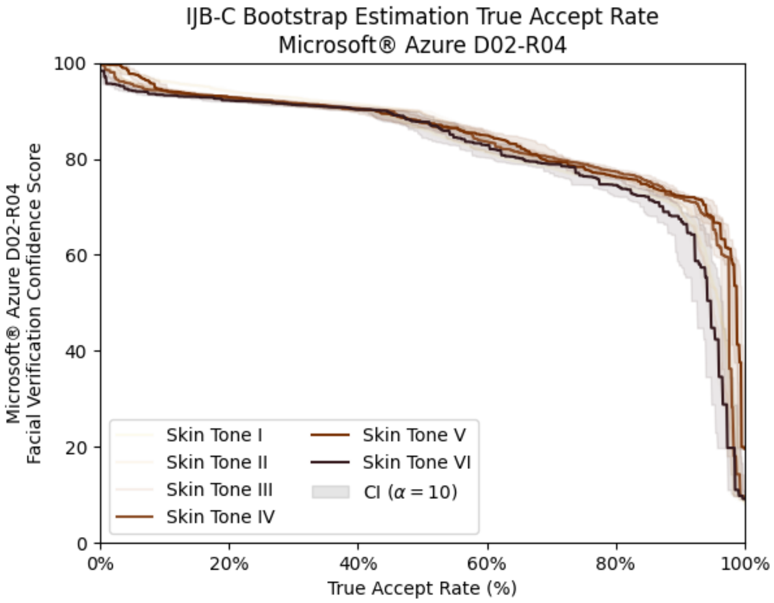


***FIGURE 29** The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under* Microsoft Face API *with a configuration of* detection_02 *and* recognition_04*, for the three dark skinned skin tone classifications, collectively **Skin Tone IV** (Medium Yellow / Brown), **Skin Tone V** (Medium-Dark Brown), and **Skin Tone VI** (Dark Brown), using the IJB-C skin tone classification schema.*

Microsoft Face API with a configuration of detection_03 and recognition_01

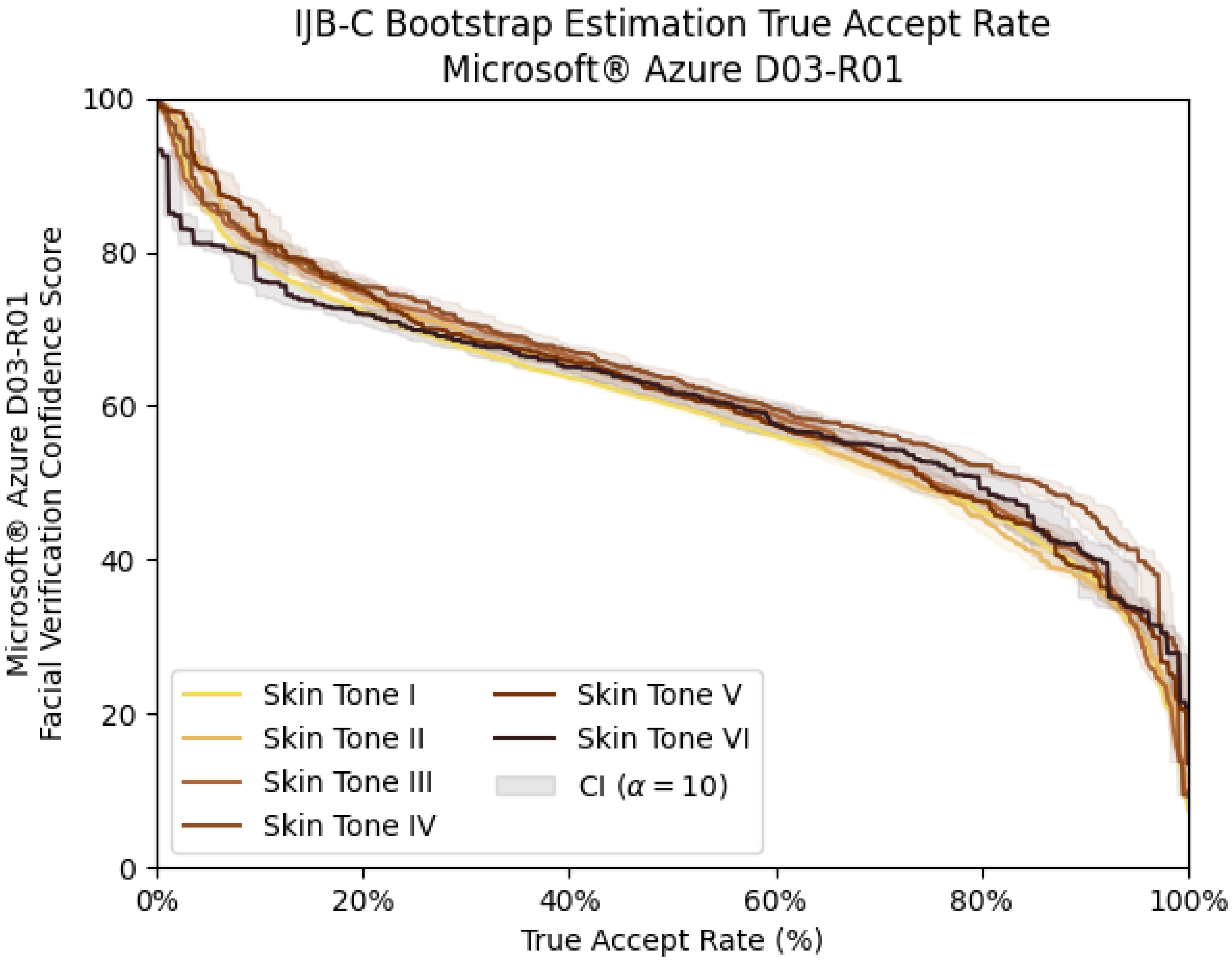


FIGURE 30 The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under Microsoft Face API *with a configuration of* detection_03 *and* recognition_01, *for the six skin tone classifications as defined by the IJB-C skin tone classification schema.*

Microsoft Face API with a configuration of detection_03 and recognition_01

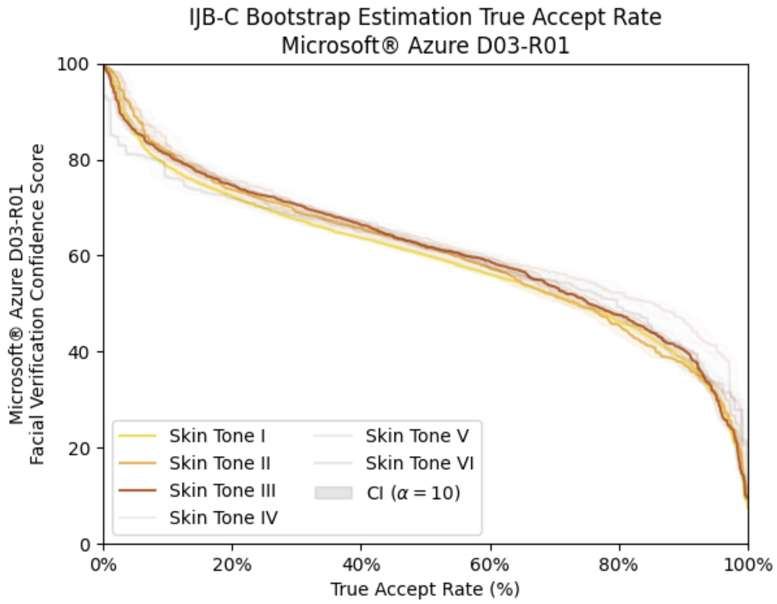


***Figure 31** The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under* Microsoft Face API *with a configuration of* detection_03 *and* recognition_01*, for the three light skinned skin tone classifications, collectively **Skin Tone I** (Light Pink), **Skin Tone II** (Light Yellow), and **Skin Tone III** (Medium Pink / Brown), using the IJB-C skin tone classification schema.*

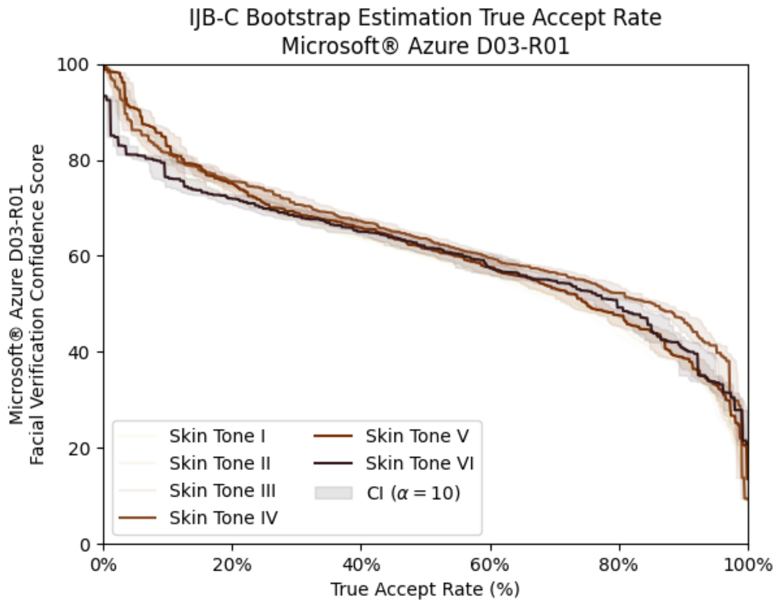


***Figure 32** The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under* Microsoft Face API *with a configuration of* detection_03 *and* recognition_01*, for the three dark skinned skin tone classifications, collectively **Skin Tone IV** (Medium Yellow / Brown), **Skin Tone V** (Medium-Dark Brown), and **Skin Tone VI** (Dark Brown), using the IJB-C skin tone classification schema.*

Microsoft Face API with a configuration of detection_03 and recognition_02

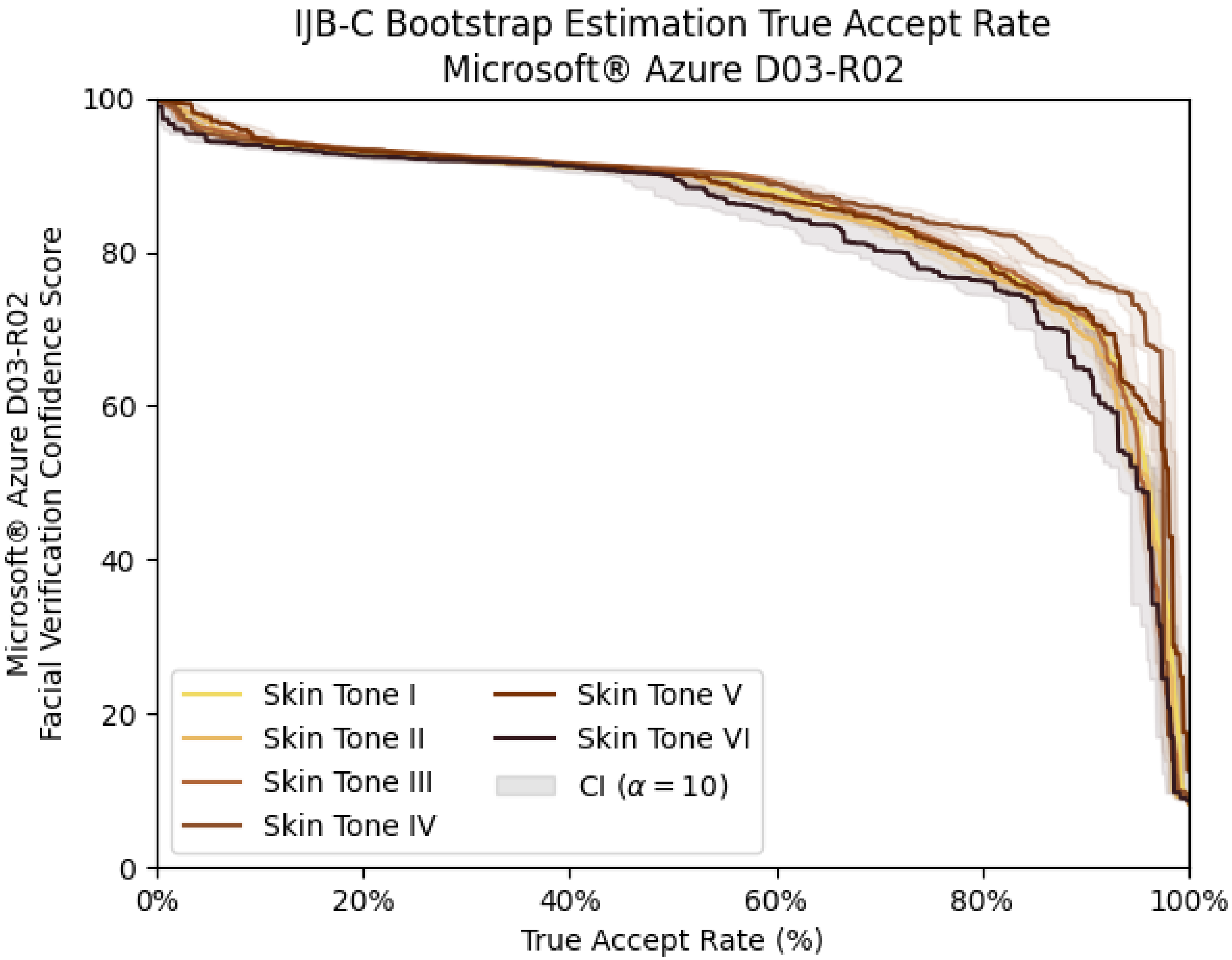


FIGURE 33 The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under Microsoft Face API *with a configuration of* detection_03 *and* recognition_02*, for the six skin tone classifications as defined by the IJB-C skin tone classification schema.*

Microsoft Face API with a configuration of detection_03 and recognition_02

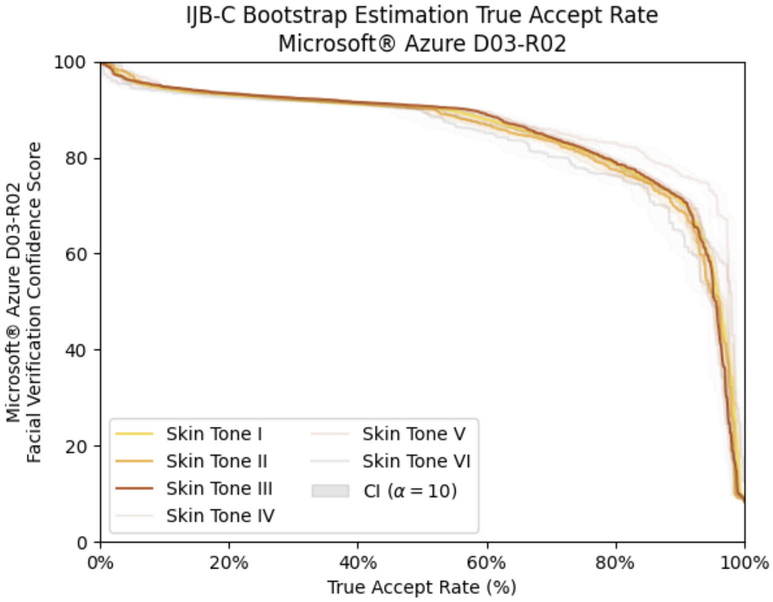


***FIGURE 34** The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under* Microsoft Face API *with a configuration of* detection_03 *and* recognition_02*, for the three light skinned skin tone classifications, collectively **Skin Tone I** (Light Pink), **Skin Tone II** (Light Yellow), and **Skin Tone III** (Medium Pink / Brown), using the IJB-C skin tone classification schema.*

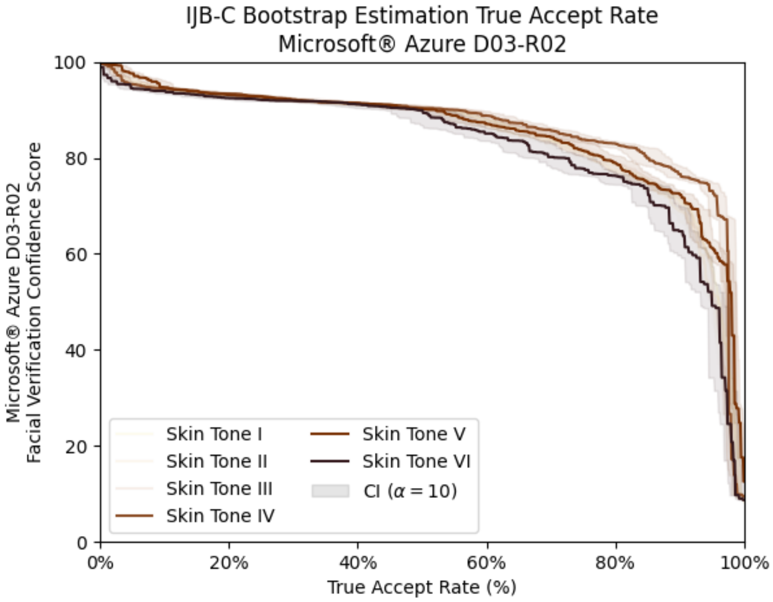


***FIGURE 35** The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under* Microsoft Face API *with a configuration of* detection_03 *and* recognition_02*, for the three dark skinned skin tone classifications, collectively **Skin Tone IV** (Medium Yellow / Brown), **Skin Tone V** (Medium-Dark Brown), and **Skin Tone VI** (Dark Brown), using the IJB-C skin tone classification schema.*

Microsoft Face API with a configuration of detection_03 and recognition_03

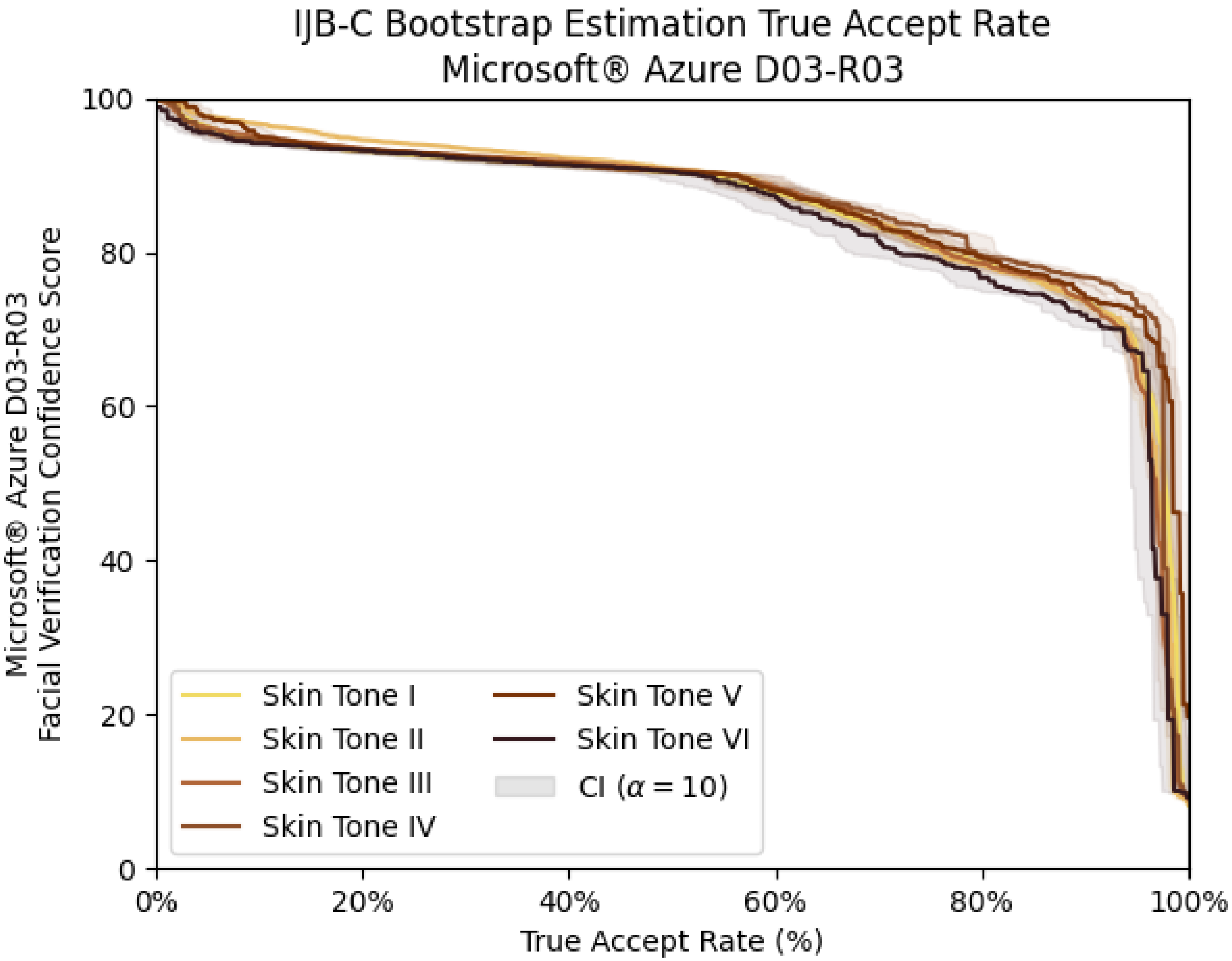
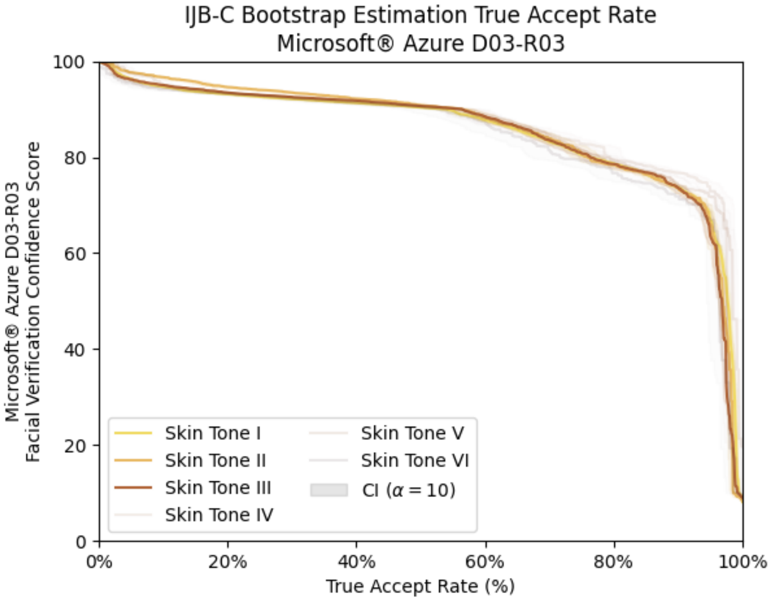
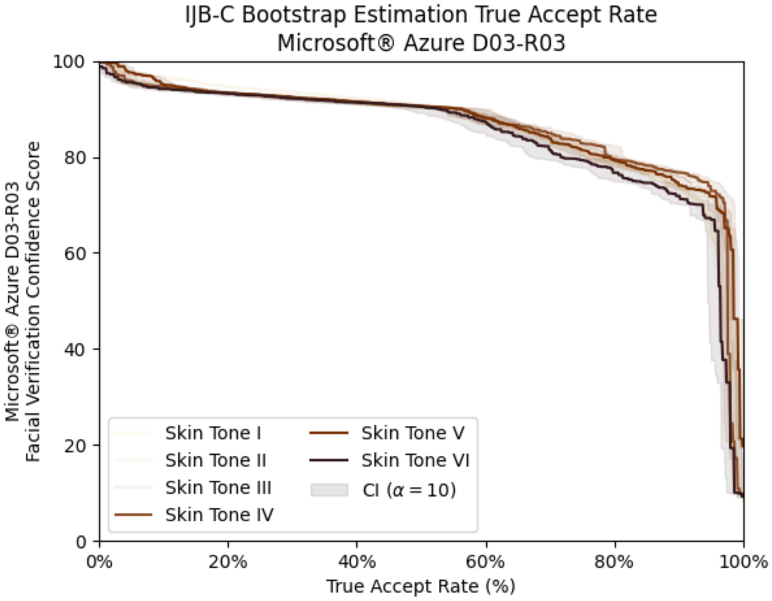


FIGURE 36 The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under Microsoft Face API *with a configuration of* detection_03 *and* recognition_03, *for the six skin tone classifications as defined by the IJB-C skin tone classification schema.*

Microsoft Face API with a configuration of detection_03 and recognition_03



F*IGURE *37 *The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under* Microsoft Face API *with a configuration of* detection_03 *and* recognition_03*, for the three light skinned skin tone classifications, collectively **Skin Tone I** (Light Pink), **Skin Tone II** (Light Yellow), and **Skin Tone III** (Medium Pink / Brown), using the IJB-C skin tone classification schema.*



F*IGURE *38 *The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under* Microsoft Face API *with a configuration of* detection_03 *and* recognition_03*, for the three dark skinned skin tone classifications, collectively **Skin Tone IV** (Medium Yellow / Brown), **Skin Tone V** (Medium-Dark Brown), and **Skin Tone VI** (Dark Brown), using the IJB-C skin tone classification schema.*

Microsoft Face API, Released 2021

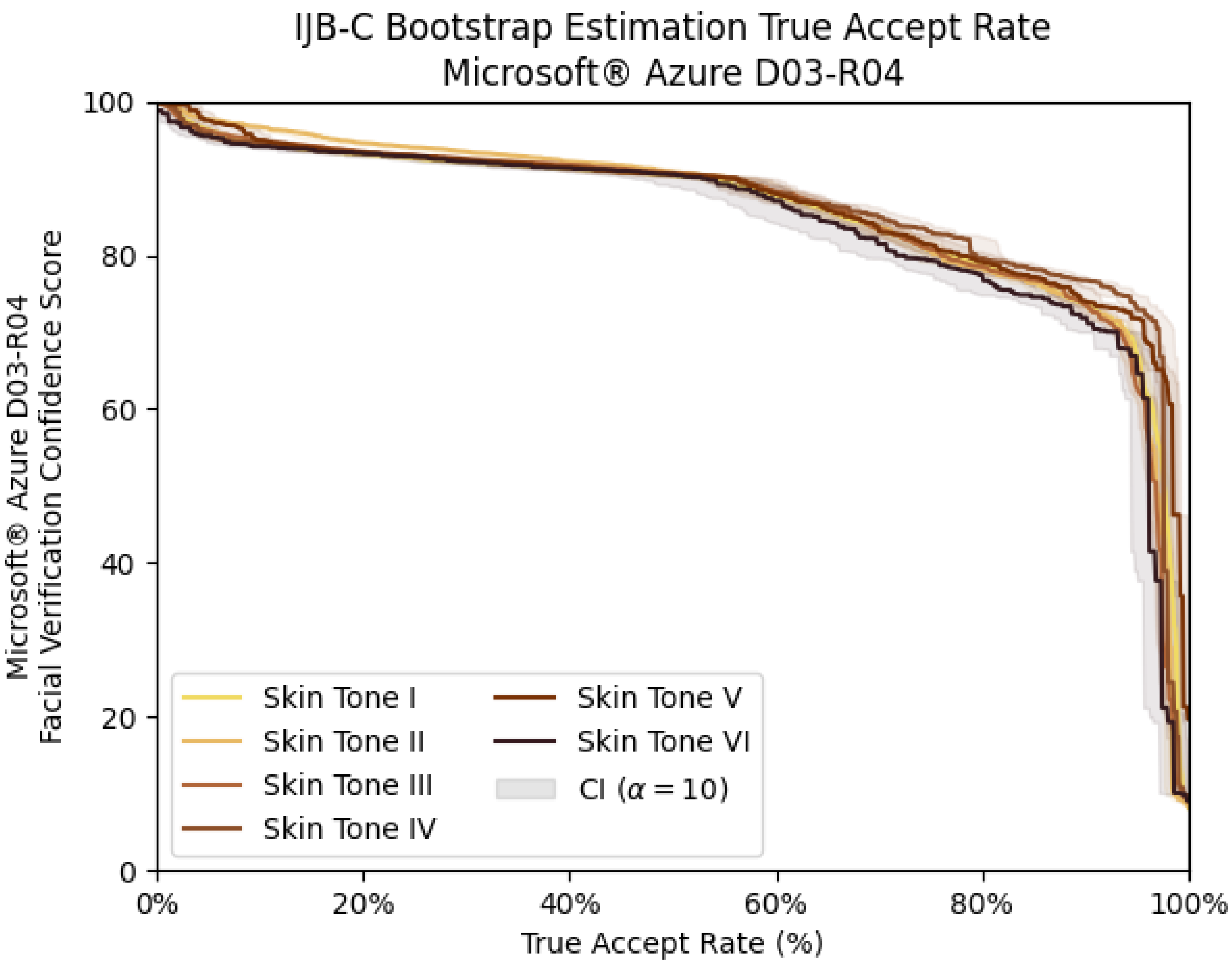


FIGURE 39 *The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under* Microsoft Face API *under its default configuration of* detection_03 *and* recognition_04*, released in 2021, for the six skin tone classifications as defined by the IJB-C skin tone classification schema.*

Microsoft Face API, Released 2021

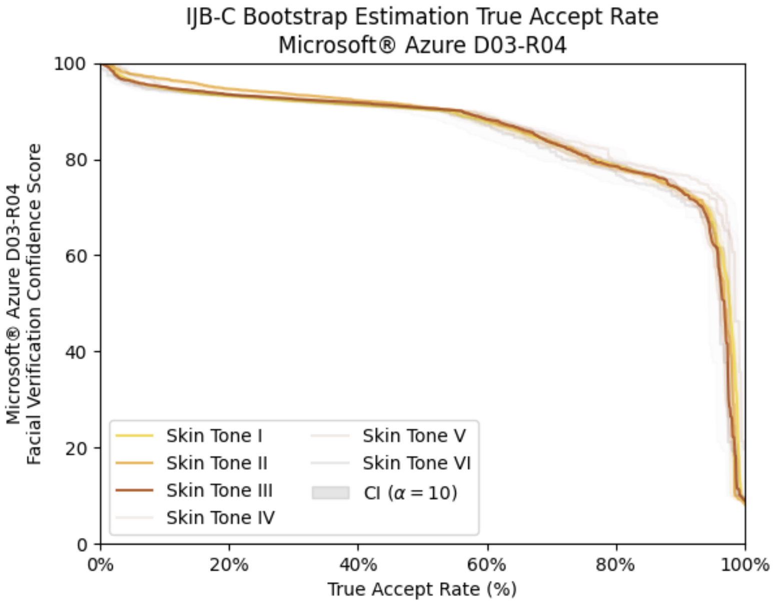


***FIGURE 40** The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under* Microsoft Face API *under its default configuration of* detection_03 *and* recognition_04, *released in 2021, for the three light skinned skin tone classifications, collectively **Skin Tone I** (Light Pink), **Skin Tone II** (Light Yellow), and **Skin Tone III** (Medium Pink / Brown), using the IJB-C skin tone classification schema.*

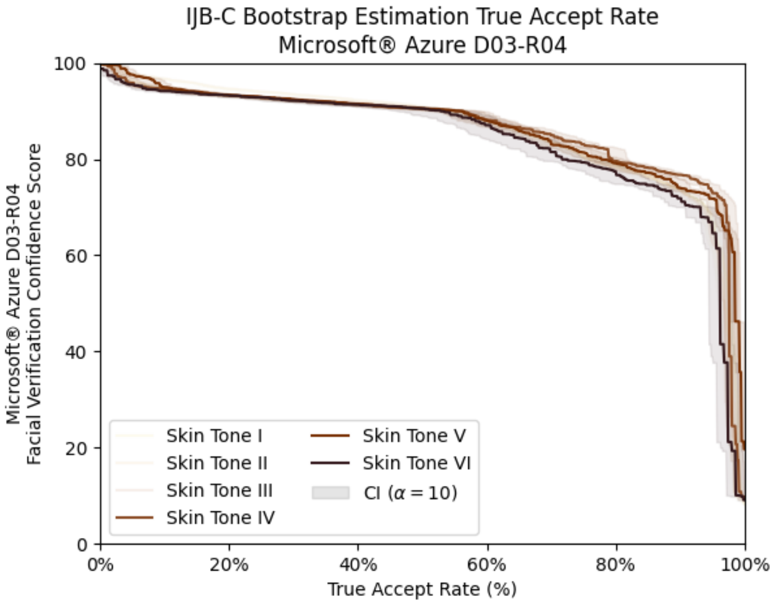


***FIGURE 41** The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under* Microsoft Face API *under its default configuration of* detection_03 *and* recognition_04, *released in 2021, for the three dark skinned skin tone classifications, collectively **Skin Tone IV** (Medium Yellow / Brown), **Skin Tone V** (Medium-Dark Brown), and **Skin Tone VI** (Dark Brown), using the IJB-C skin tone classification schema.*

AWS Rekognition, As of Fall 2021

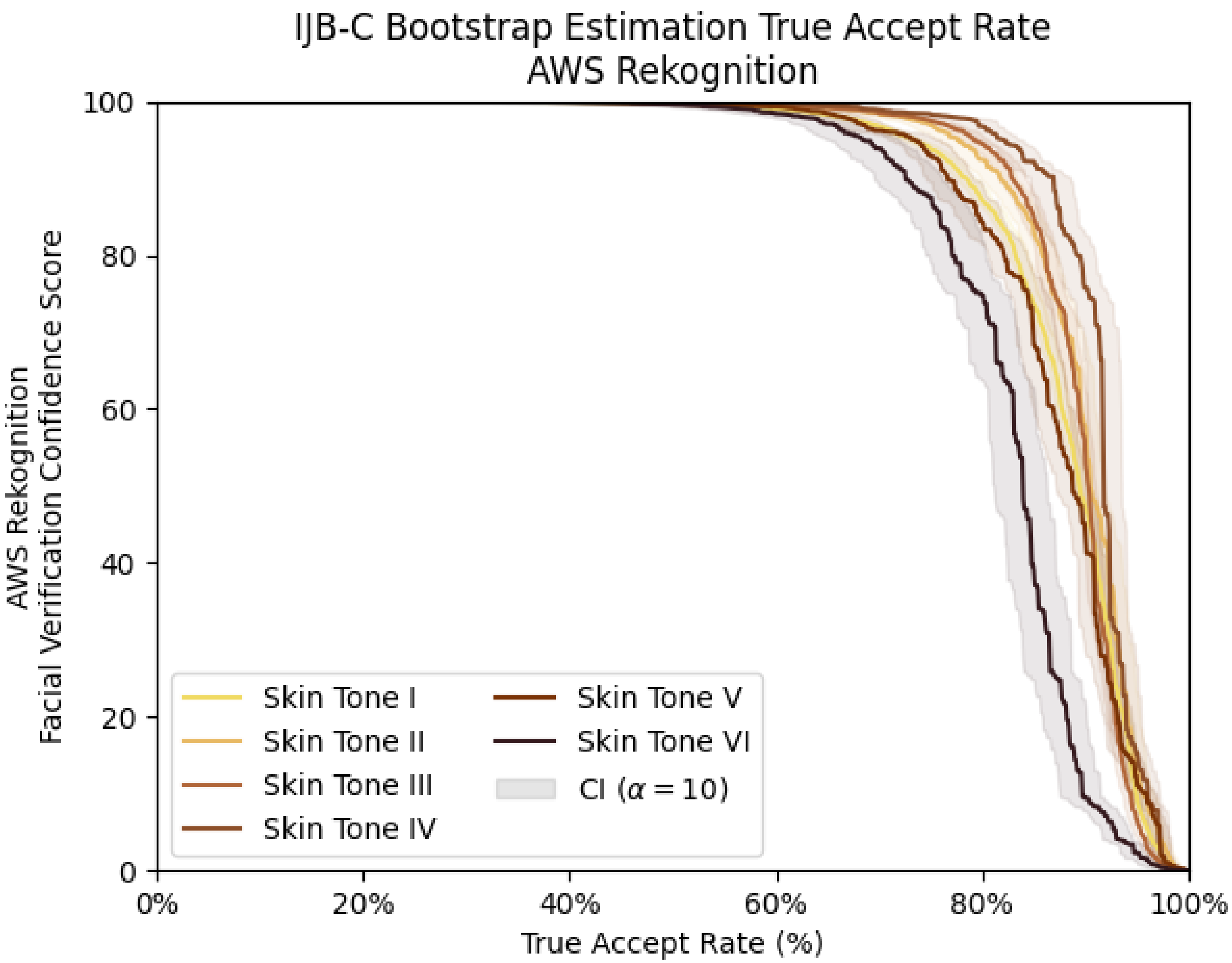


FIGURE 42 *The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under* AWS Rekognition *for the six skin tone classifications as defined by the IJB-C skin tone classification schema.*

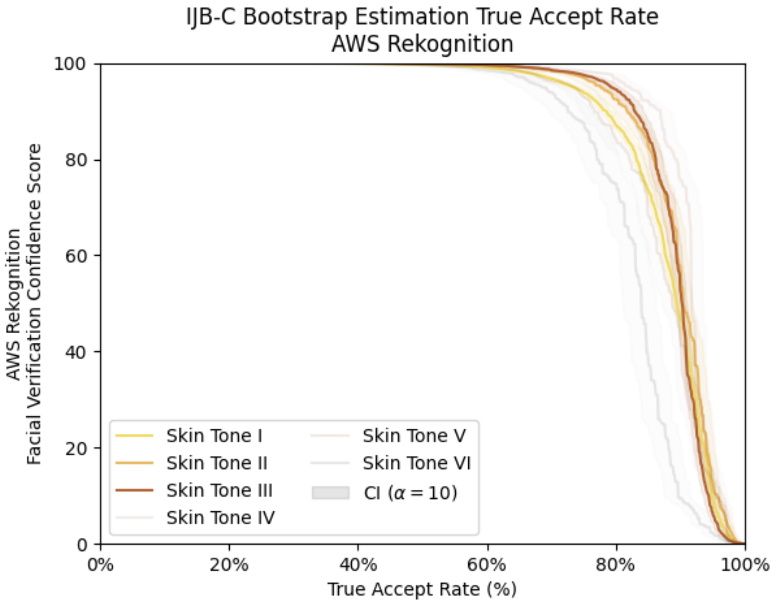


***FIGURE 43** The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under* AWS Rekognition *for the three light skinned skin tone classifications, collectively **Skin Tone I** (Light Pink), **Skin Tone II** (Light Yellow), and **Skin Tone III** (Medium Pink / Brown), using the IJB-C skin tone classification schema.*

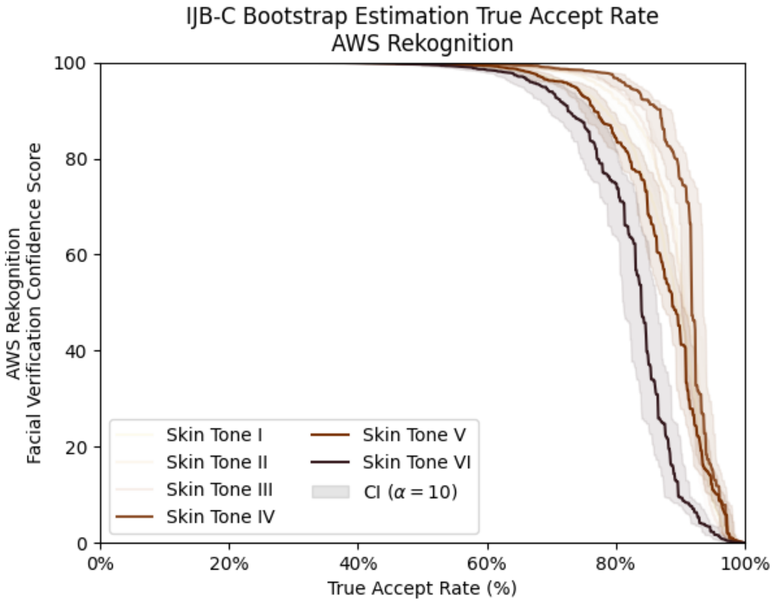


***FIGURE 44** The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under* AWS Rekognition, *for the three dark skinned skin tone classifications, collectively **Skin Tone IV** (Medium Yellow / Brown), **Skin Tone V** (Medium-Dark Brown), and **Skin Tone VI** (Dark Brown), using the IJB-C skin tone classification schema.*

Difference in Medians from Median True Accept Rate

These insights can be furthered, through calculation of a score detailing the interclass bias between one of the six skin tone classifications and the overall performance. Following the procedure outline earlier, the author adopt an overall performance TAR curve that assumes that all classes should perform equally. Therefore, for each of the bootstrap samples generated, the overall performance curve was calculated for that resample as the mean of the six classes thresholds at the TAR; and the difference between each skin tone class and the overall performance curve was measured. **FIGURE 45** ,**FIGURE 48**, **FIGURE 51**, **FIGURE 54**, **FIGURE 57**, **FIGURE 60**, **FIGURE 63**, **FIGURE 66**, **FIGURE 69**, **FIGURE 72**, **FIGURE 75**, **FIGURE 78**, and **FIGURE 81** plot the difference between the mean overall performance TAR and the corresponding confidence score threshold to achieve it under the thirteen commercial facial verification algorithms, to provide a clearer understanding of the interclass bias for each of the six skin tone classifications. **FIGURE 46**, **FIGURE 49**, **FIGURE 52**, **FIGURE 55**, **FIGURE 58**, **FIGURE 61**, **FIGURE 64**, **FIGURE 67**, **FIGURE 70**, **FIGURE 73**, **FIGURE 76**, **FIGURE 79**, and **FIGURE 82** isolate the difference between the mean overall performance TAR for light skinned persons, collectively **Skin Tone I** (Light Pink), **Skin Tone II** (Light Yellow), and **Skin Tone III** (Medium Pink / Brown), and the corresponding confidence score threshold to achieve it under the thirteen commercial facial verification algorithms. **FIGURE 47**, **FIGURE 50**, **FIGURE 53**, **FIGURE 56**, **FIGURE 59**, **FIGURE 62**, **FIGURE 65**, **FIGURE 68**, **FIGURE 71**, **FIGURE 74**, **FIGURE 77**, **FIGURE 80**, and **FIGURE 83** and isolate the difference between the mean overall performance TAR for dark skinned persons, collectively **Skin Tone IV** (Medium Yellow / Brown), **Skin Tone V** (Medium-Dark Brown), and **Skin Tone VI** (Dark Brown), and the corresponding confidence score threshold to achieve it under the thirteen commercial facial verification algorithms.

Microsoft Face API, Released 2017

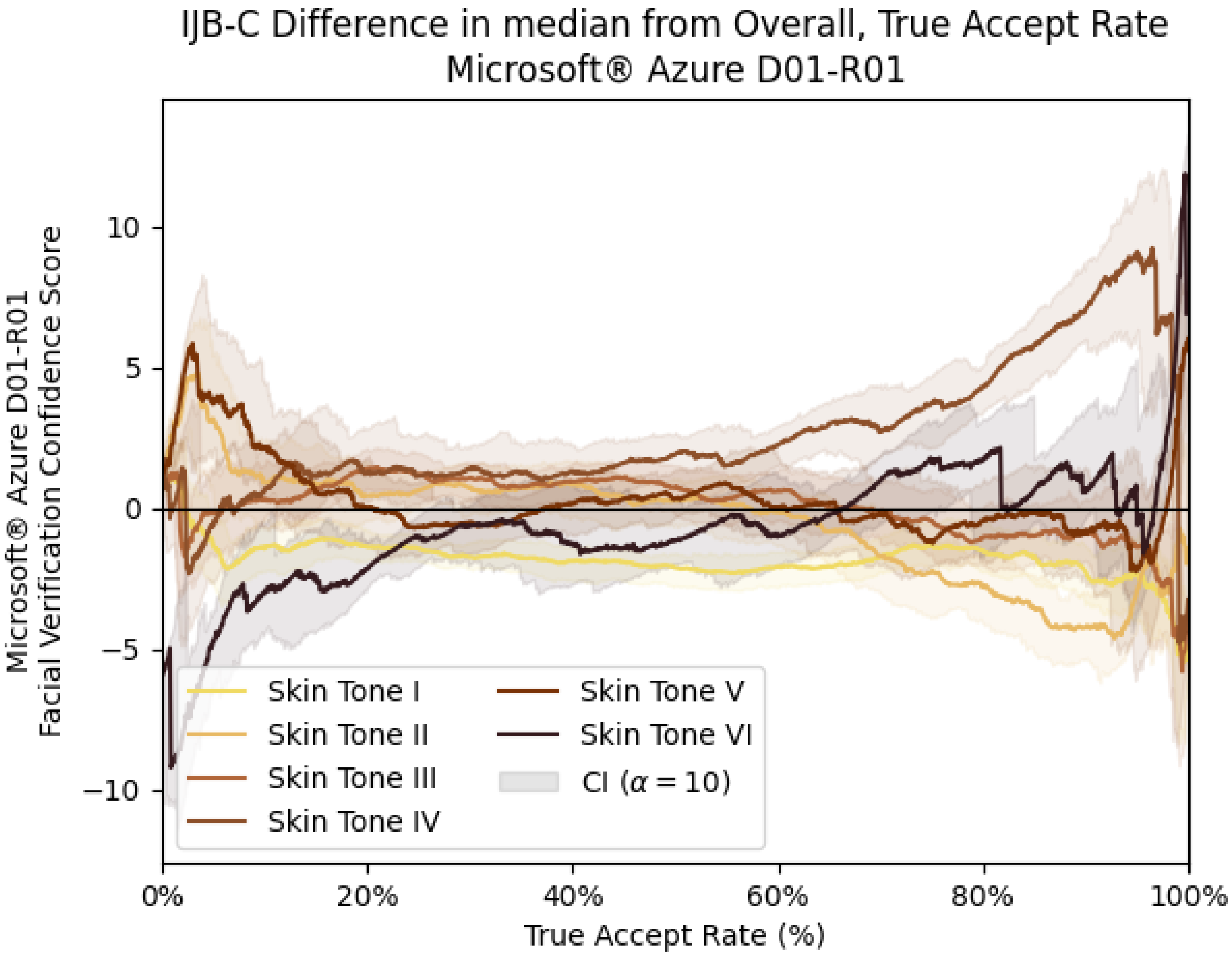


FIGURE 45 The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under Microsoft Face API *under its default configuration of* detection_01 *and* recognition_01, *released in 2017, for the six skin tone classifications as defined by the IJB-C skin tone classification schema.*

Microsoft Face API, Released 2017

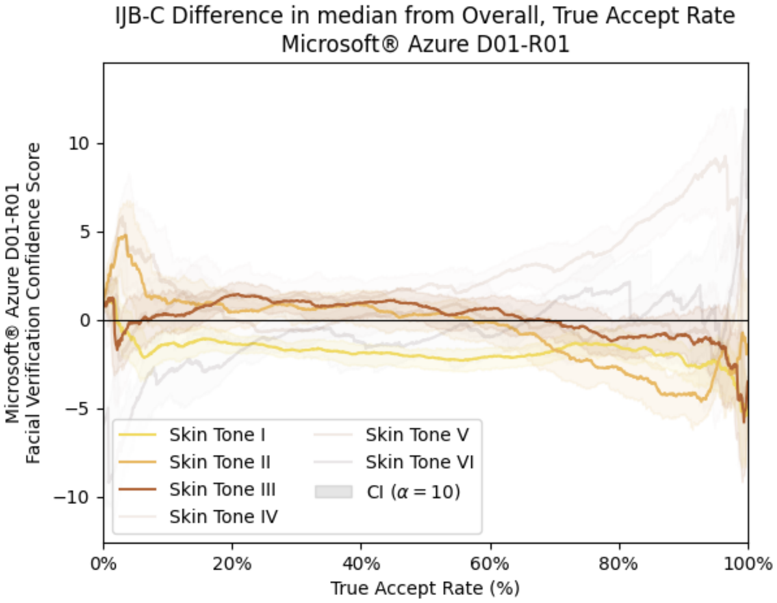


FIGURE 46 The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under Microsoft Face API *under its default configuration of* detection_01 *and* recognition_01, *released in 2017, for the three light skinned skin tone classifications, collectively* ***Skin Tone I*** *(Light Pink),* ***Skin Tone II*** *(Light Yellow), and* ***Skin Tone III*** *(Medium Pink / Brown), using the IJB-C skin tone classification schema.*

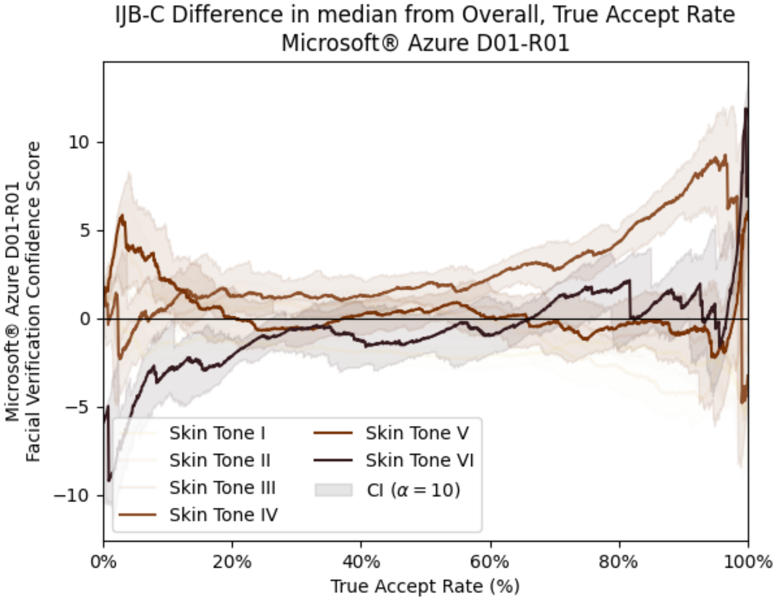


FIGURE 47 The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under Microsoft Face API *under its default configuration of* detection_01 *and* recognition_01, *released in 2017, for the three dark skinned skin tone classifications, collectively* ***Skin Tone IV*** *(Medium Yellow / Brown),* ***Skin Tone V*** *(Medium-Dark Brown), and* ***Skin Tone VI*** *(Dark Brown), using the IJB-C skin tone classification schema.*

Microsoft Face API with a configuration of detection_01 and recognition_02

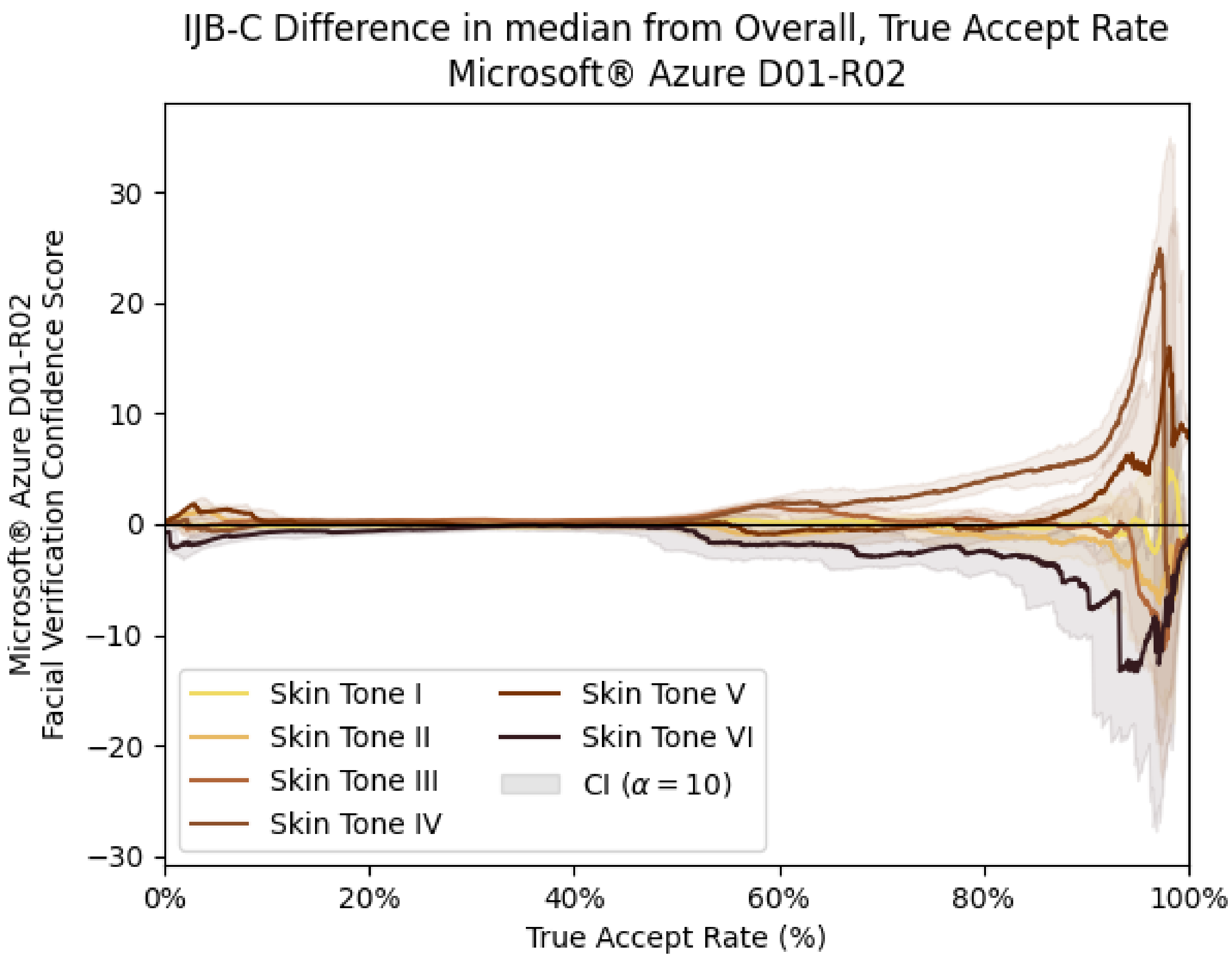


FIGURE 48 The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under Microsoft Face API *with a configuration of* detection_01 *and* recognition_02*, for the six skin tone classifications as defined by the IJB-C skin tone classification schema.*

Microsoft Face API with a configuration of detection_01 and recognition_02

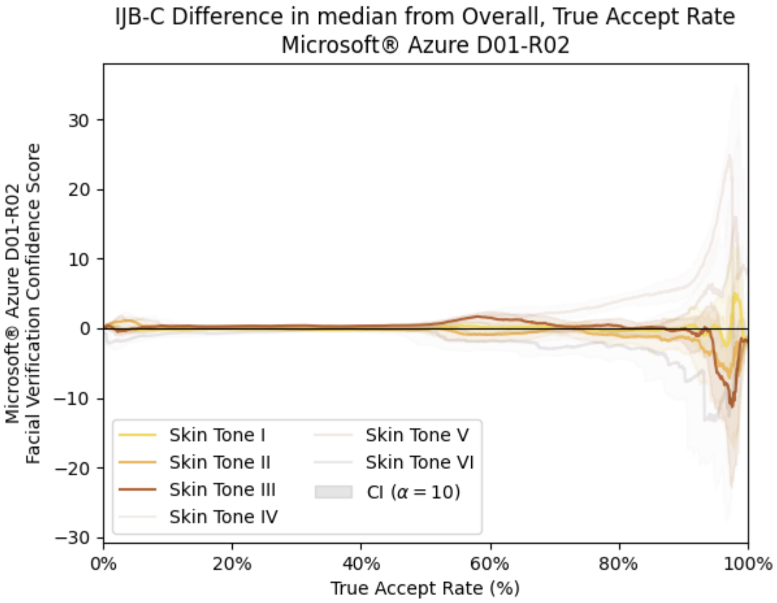


FIGURE 49 The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under Microsoft Face API *with a configuration of* detection_01 *and* recognition_02, *for the three light skinned skin tone classifications, collectively* ***Skin Tone I*** *(Light Pink),* ***Skin Tone II*** *(Light Yellow), and* ***Skin Tone III*** *(Medium Pink / Brown), using the IJB-C skin tone classification schema.*

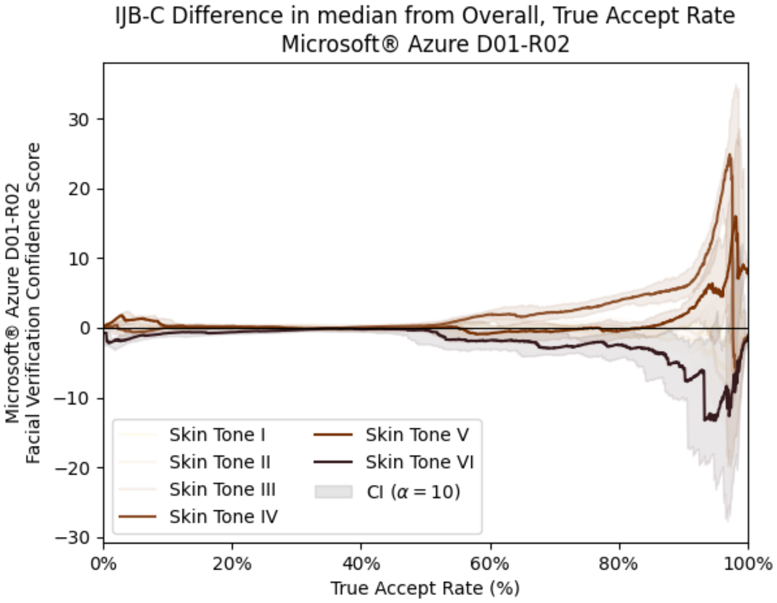


FIGURE 50 The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under Microsoft Face API *with a configuration of* detection_01 *and* recognition_02, *for the three dark skinned skin tone classifications, collectively* ***Skin Tone IV*** *(Medium Yellow / Brown),* ***Skin Tone V*** *(Medium-Dark Brown), and* ***Skin Tone VI*** *(Dark Brown), using the IJB-C skin tone classification schema.*

Microsoft Face API with a configuration of detection_01 and recognition_03

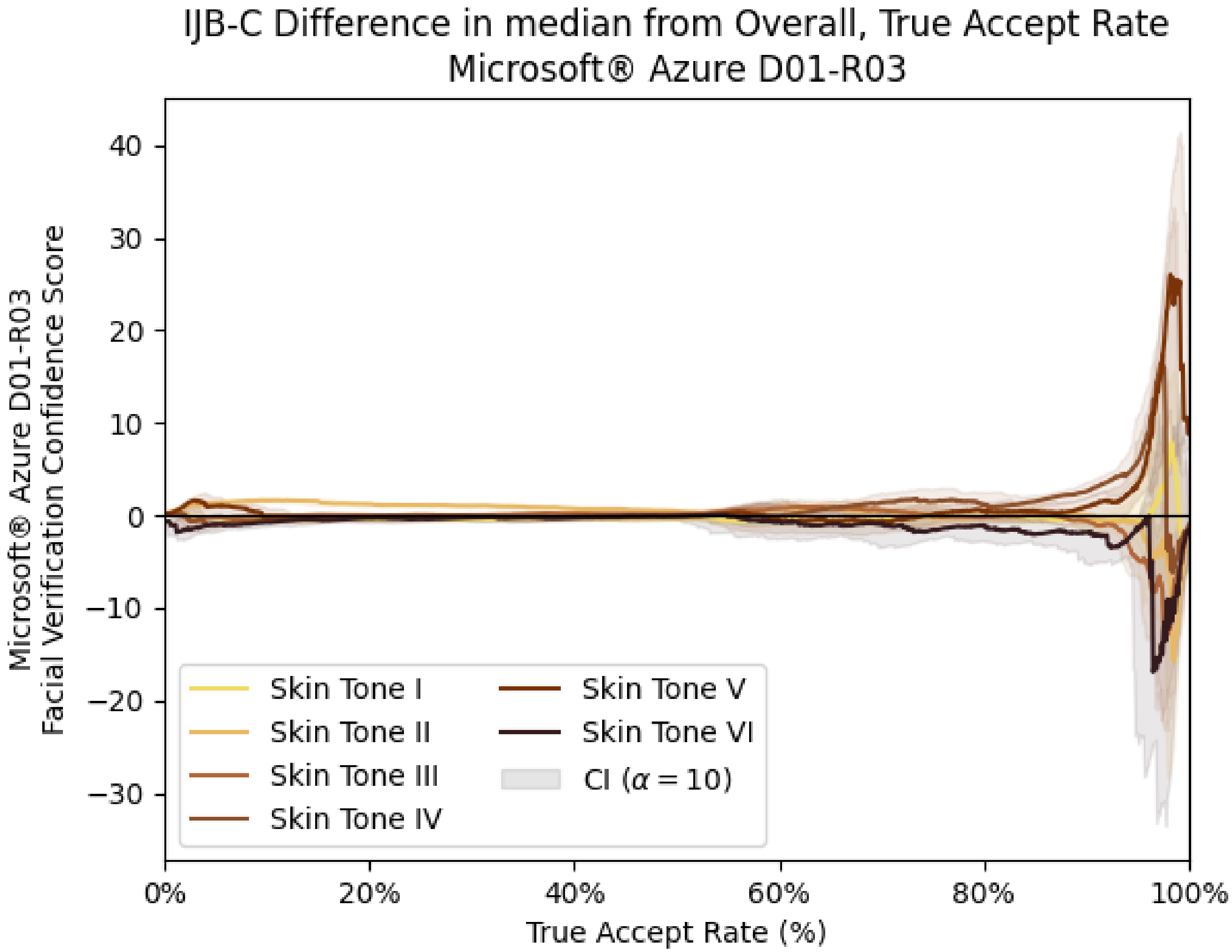


FIGURE 51 The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under Microsoft Face API *with a configuration of* detection_01 *and* recognition_03*, for the six skin tone classifications as defined by the IJB-C skin tone classification schema.*

Microsoft Face API with a configuration of detection_01 and recognition_03

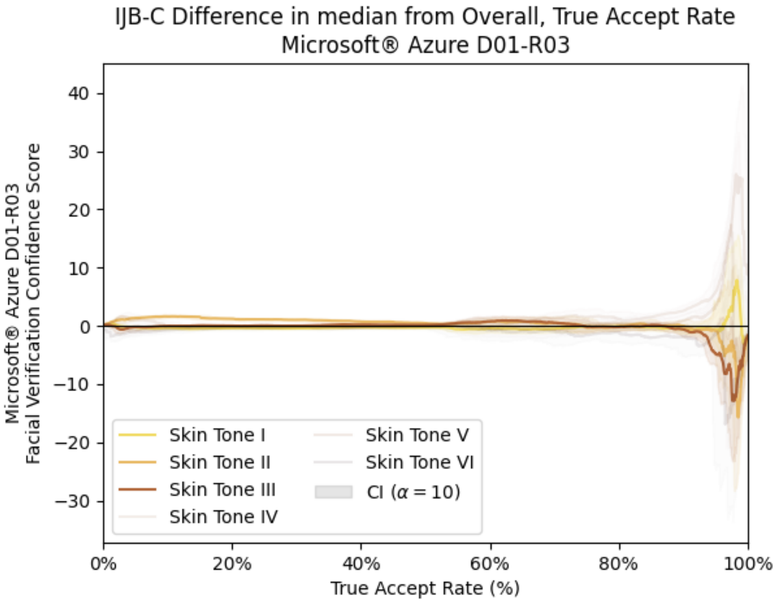


FIGURE 52 The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under Microsoft Face API *with a configuration of* detection_01 *and* recognition_03, *for the three light skinned skin tone classifications, collectively **Skin Tone I** (Light Pink), **Skin Tone II** (Light Yellow), and **Skin Tone III** (Medium Pink / Brown), using the IJB-C skin tone classification schema.*

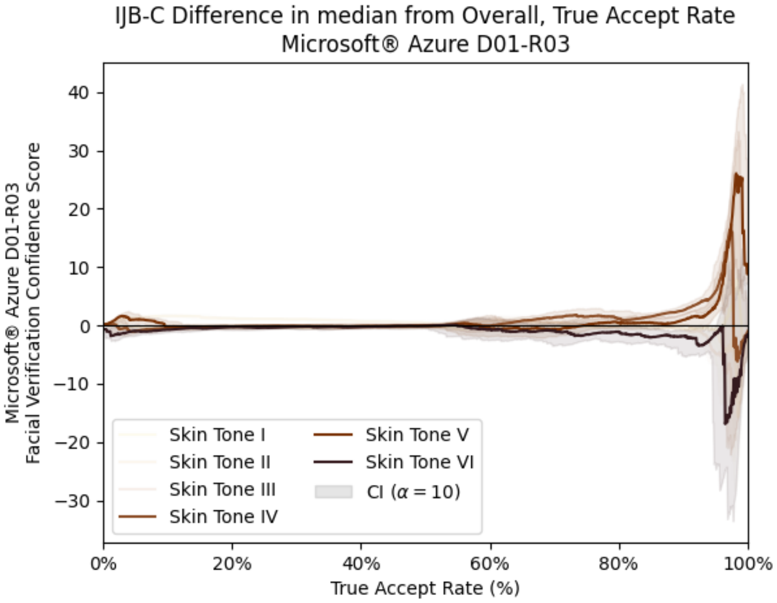


FIGURE 53 The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under Microsoft Face API *with a configuration of* detection_01 *and* recognition_03, *for the three dark skinned skin tone classifications, collectively **Skin Tone IV** (Medium Yellow / Brown), **Skin Tone V** (Medium-Dark Brown), and **Skin Tone VI** (Dark Brown), using the IJB-C skin tone classification schema.*

Microsoft Face API with a configuration of detection_01 and recognition_04

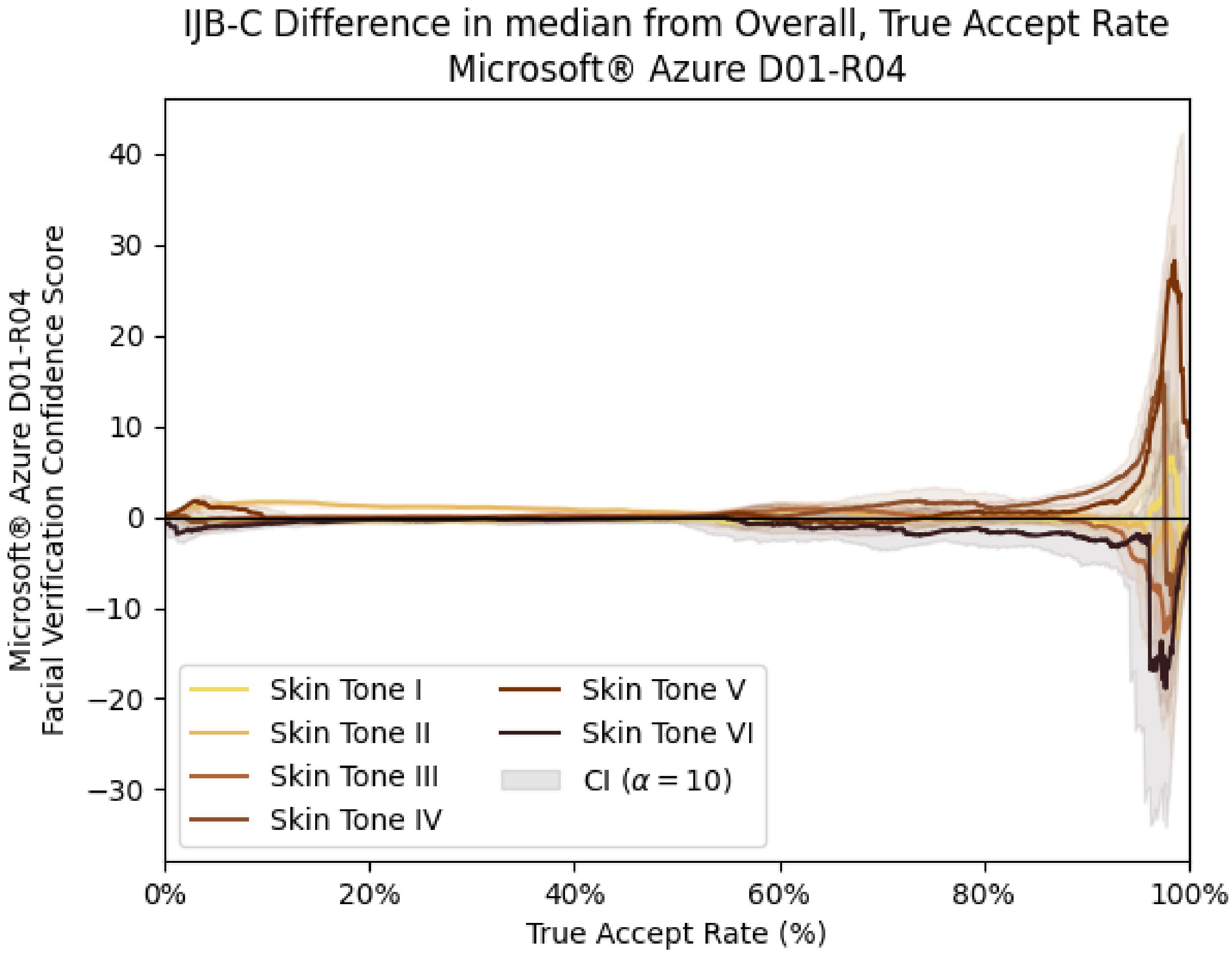


FIGURE 54 The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under Microsoft Face API *with a configuration of* detection_01 *and* recognition_04*, for the six skin tone classifications as defined by the IJB-C skin tone classification schema.*

Microsoft Face API with a configuration of detection_01 and recognition_04

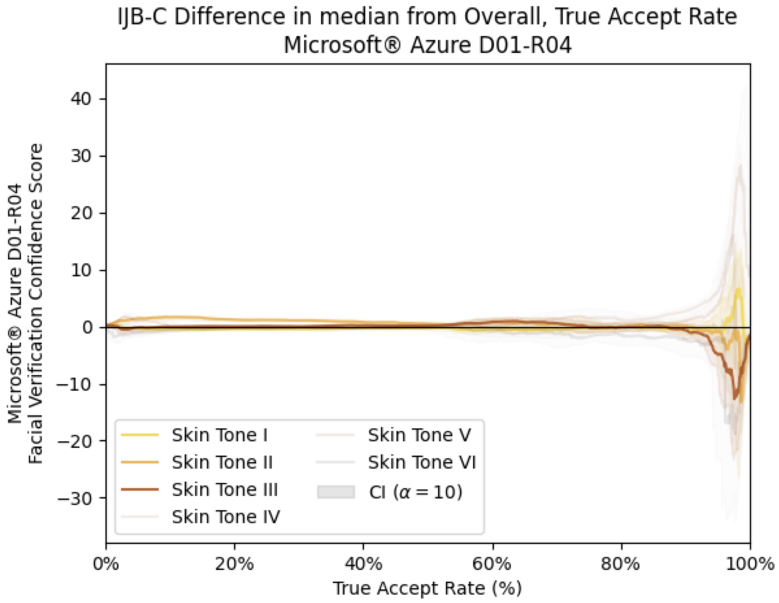


FIGURE 55 The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under Microsoft Face API *with a configuration of* detection_01 *and* recognition_04*, for the three light skinned skin tone classifications, collectively **Skin Tone I** (Light Pink), **Skin Tone II** (Light Yellow), and **Skin Tone III** (Medium Pink / Brown), using the IJB-C skin tone classification schema.*

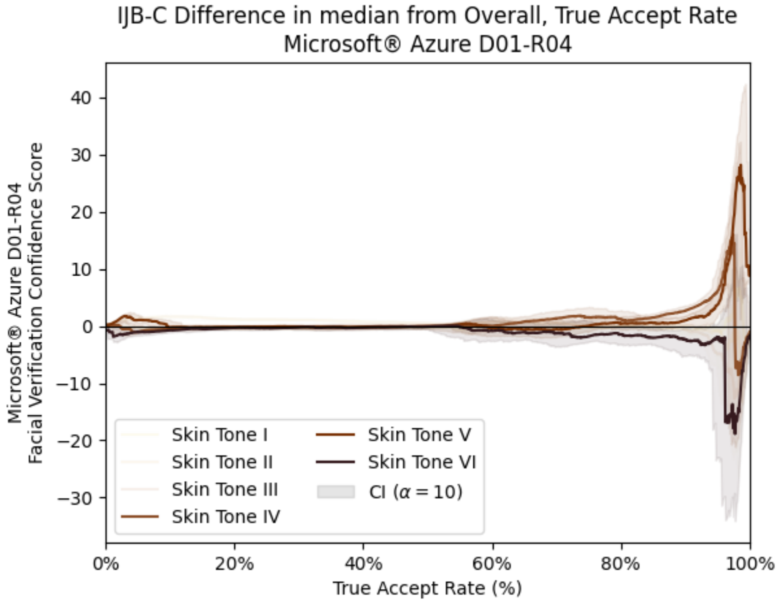


FIGURE 56 The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under Microsoft Face API *with a configuration of* detection_01 *and* recognition_04*, for the three dark skinned skin tone classifications, collectively **Skin Tone IV** (Medium Yellow / Brown), **Skin Tone V** (Medium-Dark Brown), and **Skin Tone VI** (Dark Brown), using the IJB-C skin tone classification schema.*

Microsoft Face API with a configuration of detection_02 and recognition_01

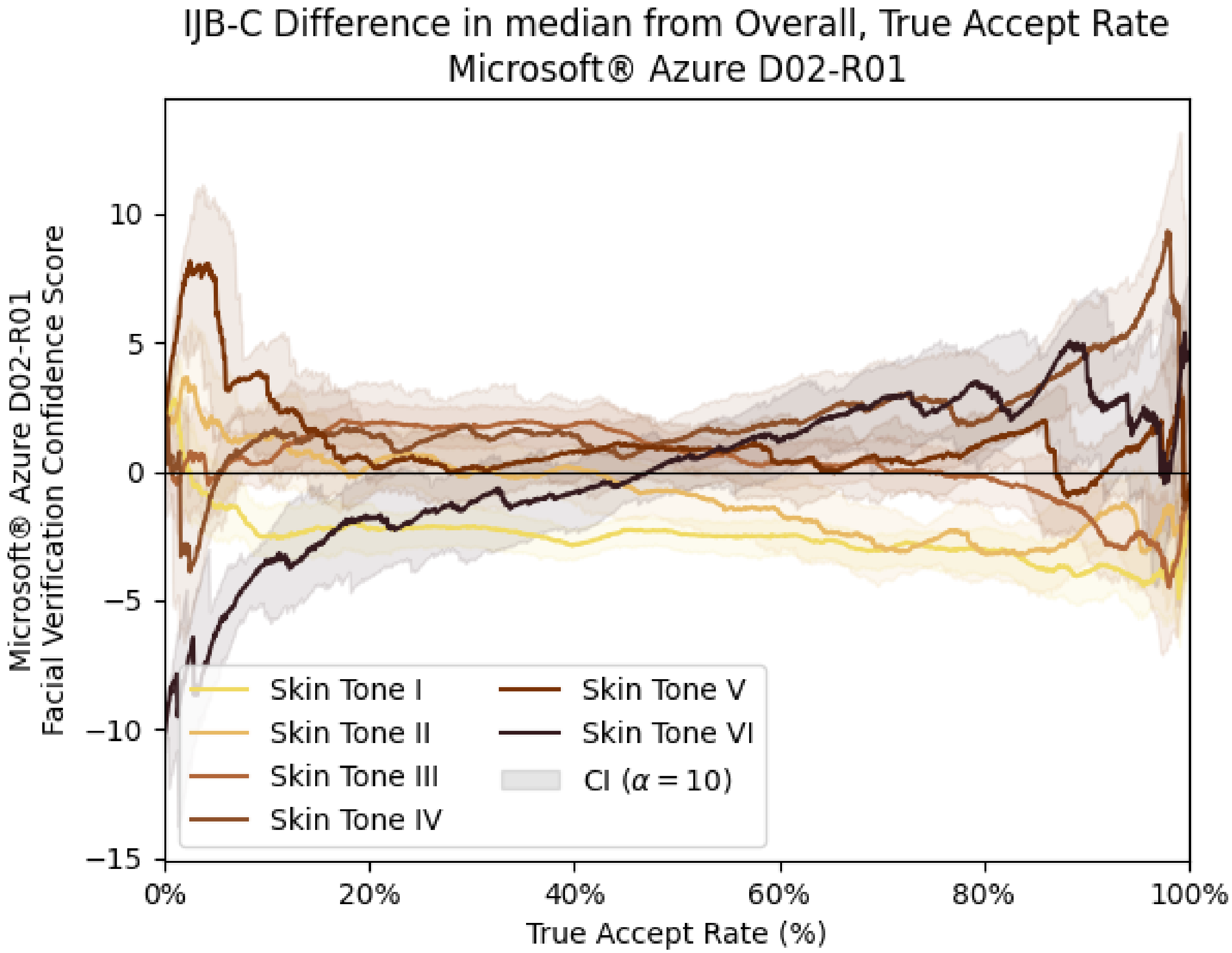


FIGURE 57 The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under Microsoft Face API *with a configuration of* detection_02 *and* recognition_01*, for the six skin tone classifications as defined by the IJB-C skin tone classification schema.*

Microsoft Face API with a configuration of detection_02 and recognition_01

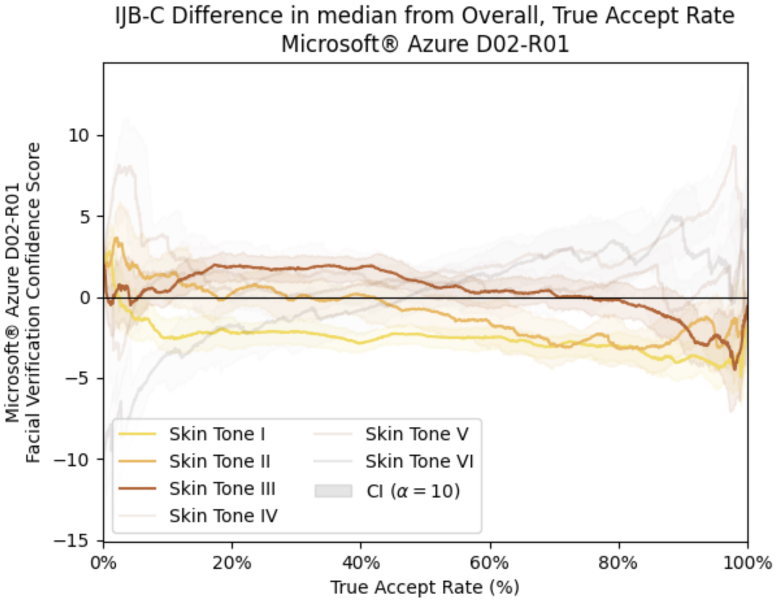


***FIGURE 58** The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under* Microsoft Face API *with a configuration of* detection_02 *and* recognition_01, *for the three light skinned skin tone classifications, collectively **Skin Tone I** (Light Pink), **Skin Tone II** (Light Yellow), and **Skin Tone III** (Medium Pink / Brown), using the IJB-C skin tone classification schema.*

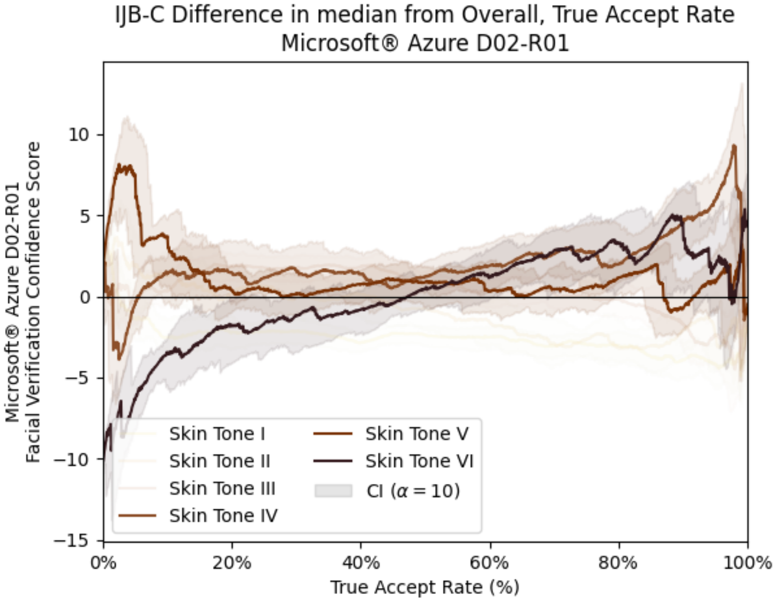


***FIGURE 59** The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under* Microsoft Face API *with a configuration of* detection_02 *and* recognition_01, *for the three dark skinned skin tone classifications, collectively **Skin Tone IV** (Medium Yellow / Brown), **Skin Tone V** (Medium-Dark Brown), and **Skin Tone VI** (Dark Brown), using the IJB-C skin tone classification schema.*

Microsoft Face API, Released 2019

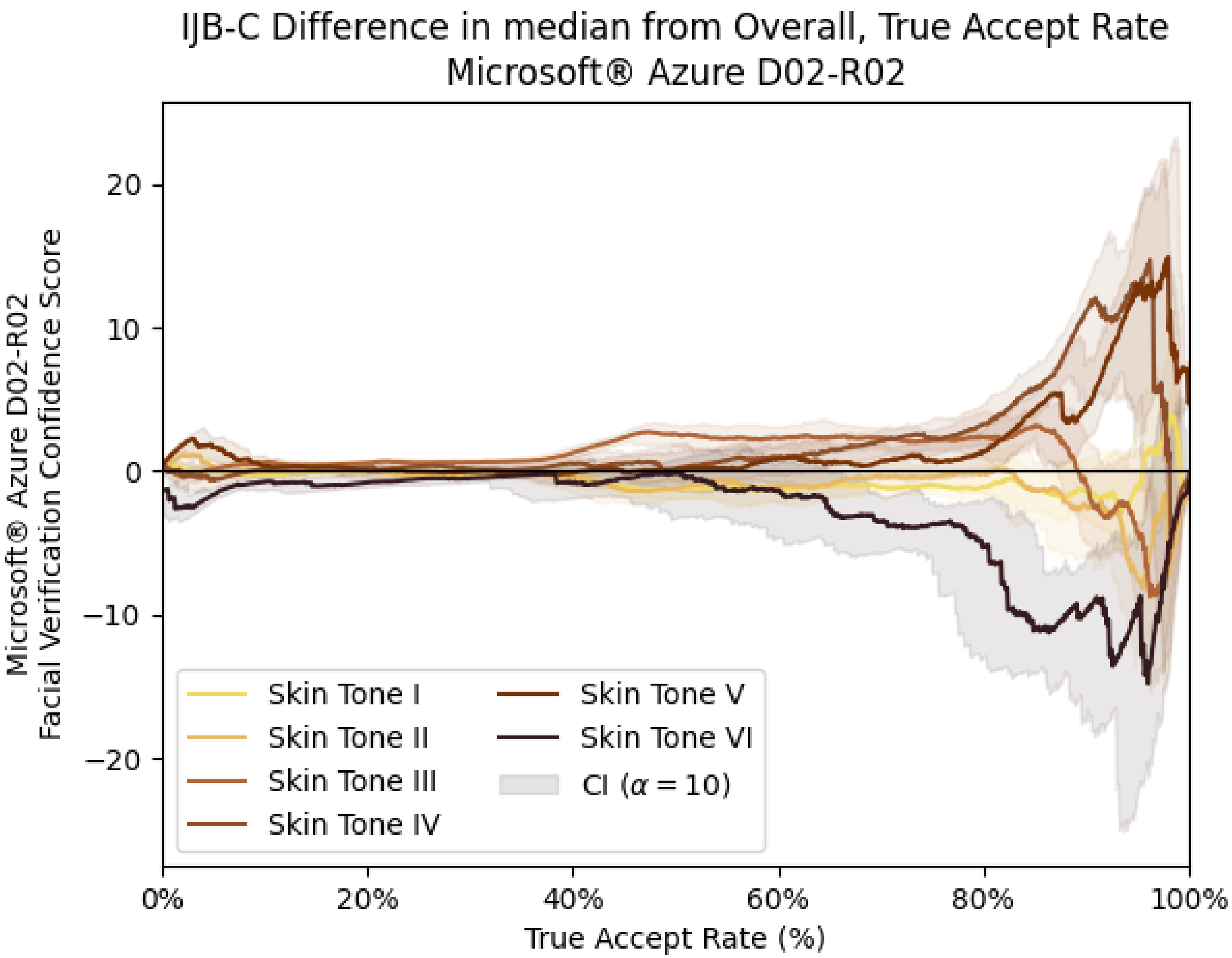


FIGURE 60 *The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under* Microsoft Face API *under its default configuration of* detection_02 *and* recognition_02, *released in 2019, for the six skin tone classifications as defined by the IJB-C skin tone classification schema.*

Microsoft Face API, Released 2019

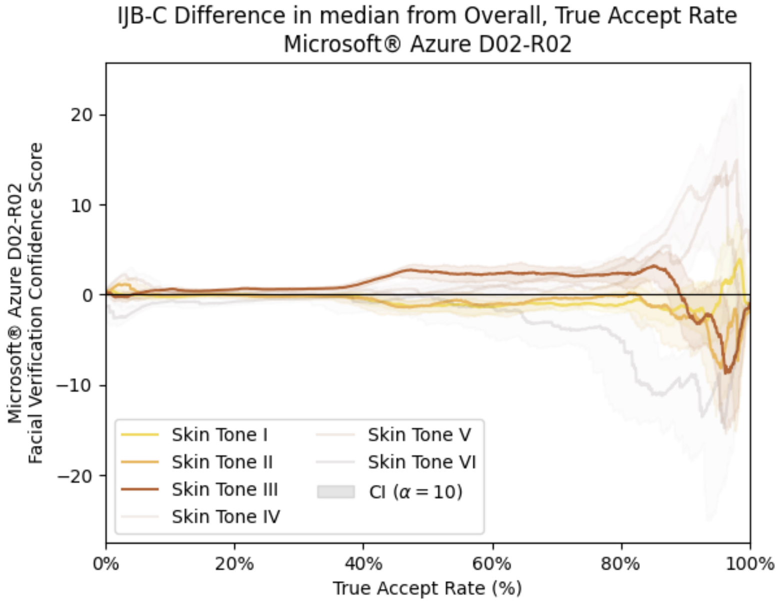


FIGURE 61 The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under Microsoft Face API *under its default configuration of* detection_02 *and* recognition_02*, released in 2019, for the three light skinned skin tone classifications, collectively **Skin Tone I** (Light Pink), **Skin Tone II** (Light Yellow), and **Skin Tone III** (Medium Pink / Brown), using the IJB-C skin tone classification schema.*

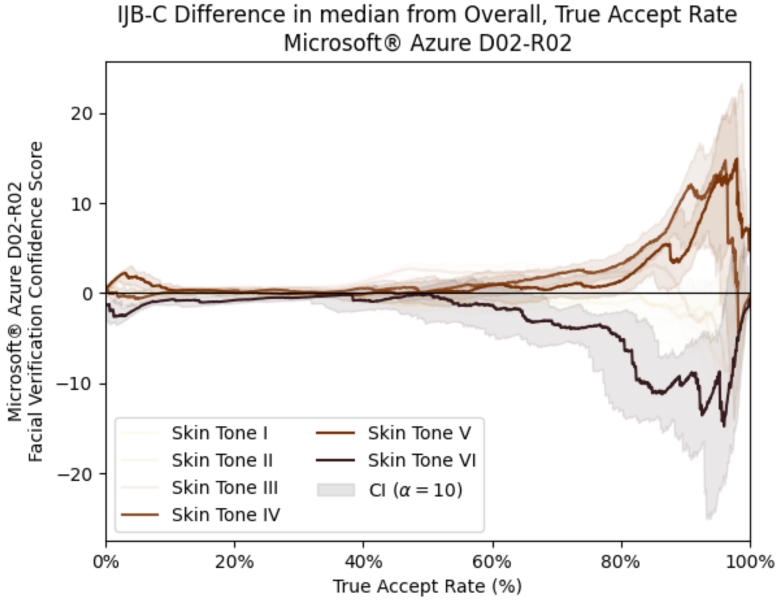


FIGURE 62 The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under Microsoft Face API *under its default configuration of* detection_02 *and* recognition_02*, released in 2019, for the three dark skinned skin tone classifications, collectively **Skin Tone IV** (Medium Yellow / Brown), **Skin Tone V** (Medium-Dark Brown), and **Skin Tone VI** (Dark Brown), using the IJB-C skin tone classification schema.*

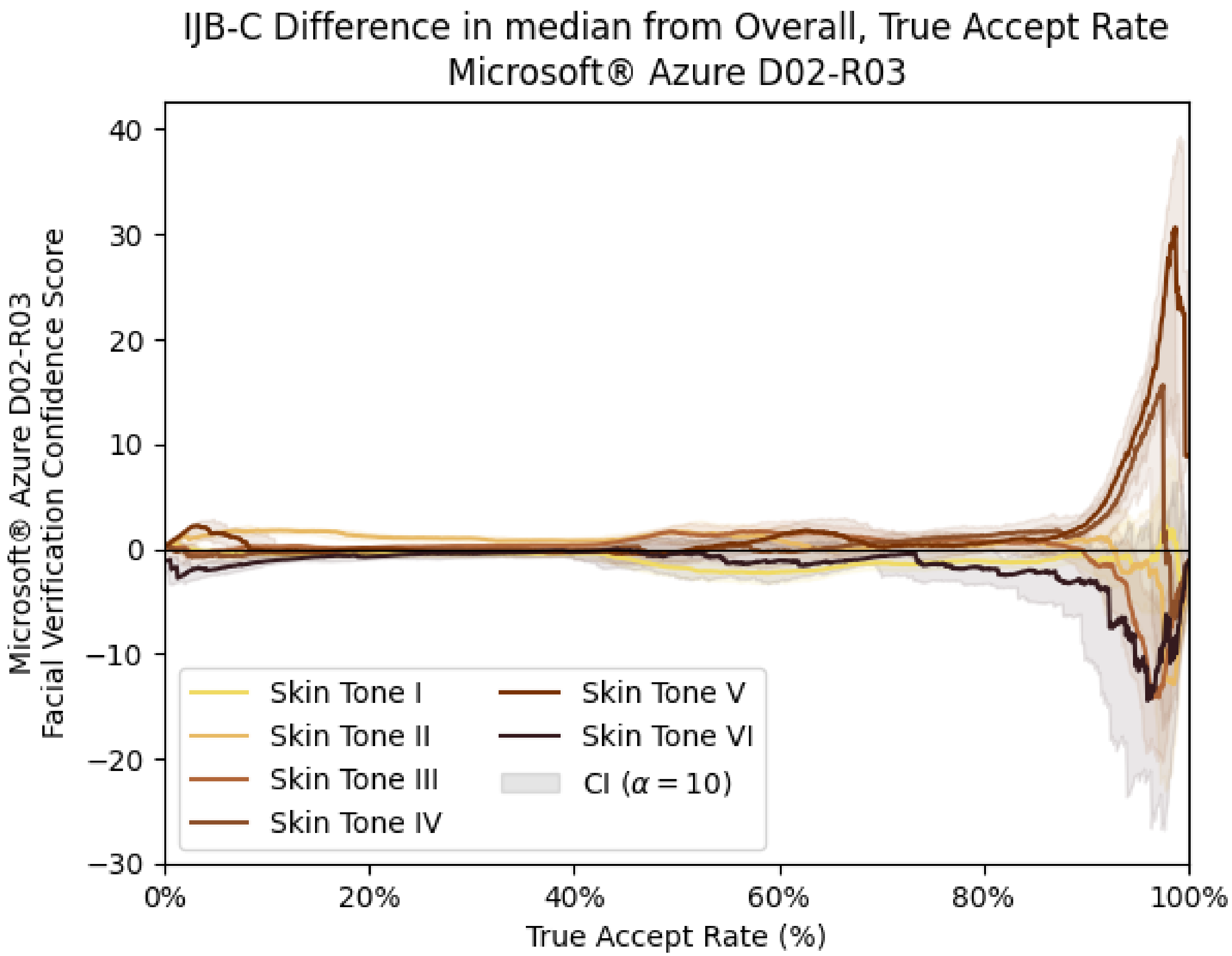


FIGURE 63 The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under Microsoft Face API *under its default configuration of* detection_02 *and* recognition_03*, released in 2020, for the six skin tone classifications as defined by the IJB-C skin tone classification schema.*

Microsoft Face API, Released 2020

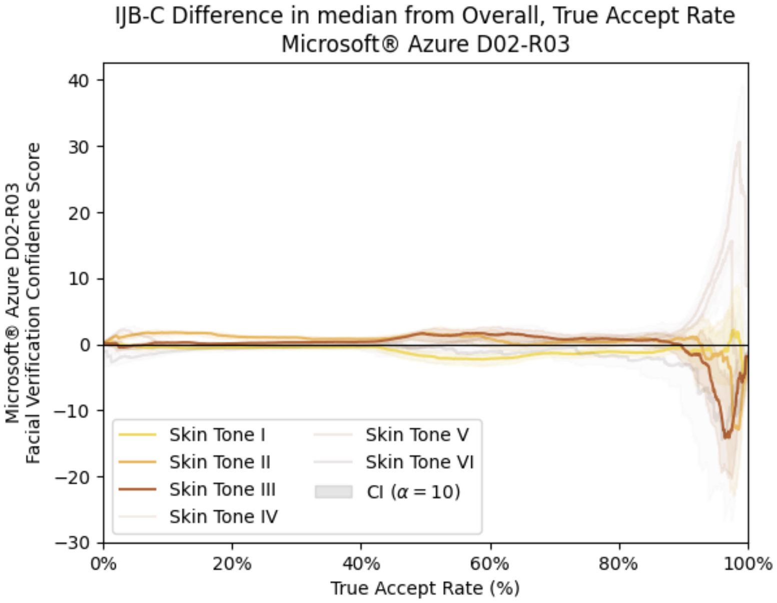


FIGURE 64 The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under Microsoft Face API *under its default configuration of* detection_02 *and* recognition_03, *released in 2020, for the three light skinned skin tone classifications, collectively **Skin Tone I** (Light Pink), **Skin Tone II** (Light Yellow), and **Skin Tone III** (Medium Pink / Brown), using the IJB-C skin tone classification schema.*

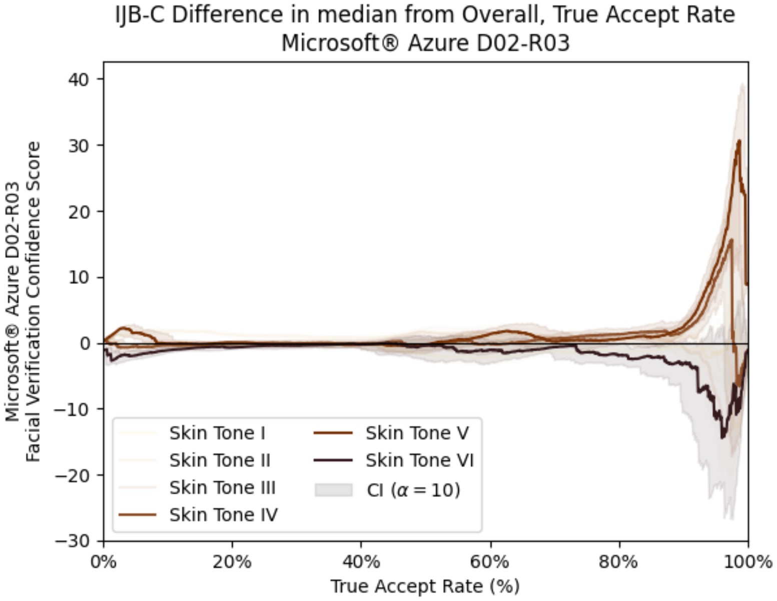


FIGURE 65 The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under Microsoft Face API *under its default configuration of* detection_02 *and* recognition_03, *released in 2020, for the three dark skinned skin tone classifications, collectively **Skin Tone IV** (Medium Yellow / Brown), **Skin Tone V** (Medium-Dark Brown), and **Skin Tone VI** (Dark Brown), using the IJB-C skin tone classification schema.*

Microsoft Face API with a configuration of detection_02 and recognition_04

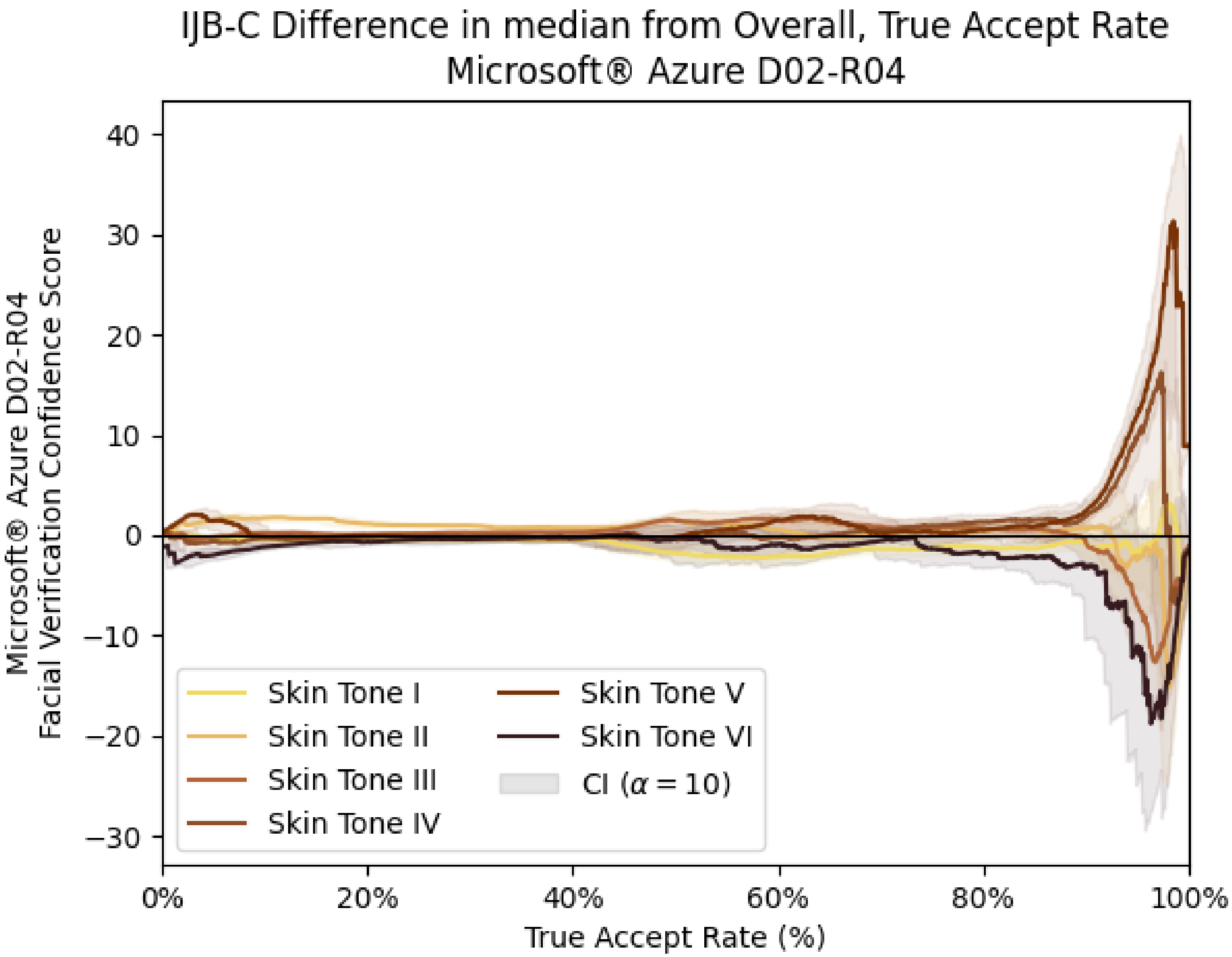


FIGURE 66 The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under Microsoft Face API *with a configuration of* detection_02 *and* recognition_04*, for the six skin tone classifications as defined by the IJB-C skin tone classification schema.*

Microsoft Face API with a configuration of detection_02 and recognition_04

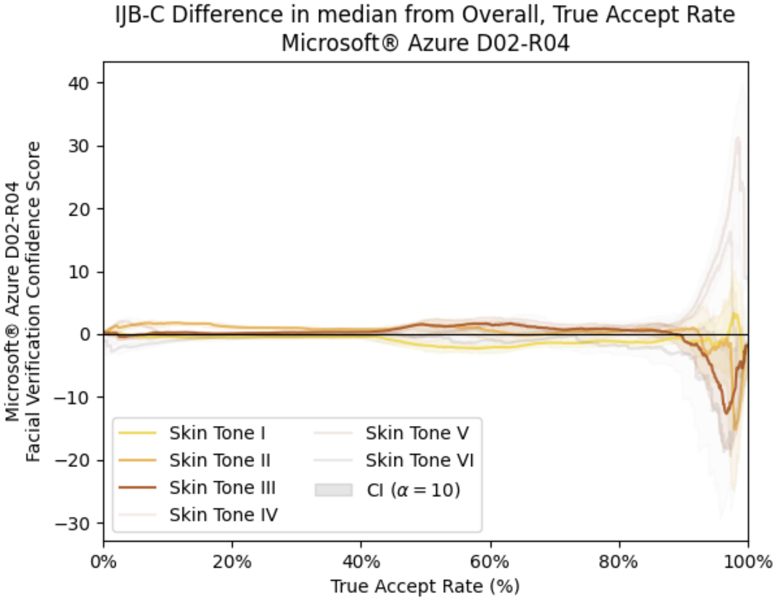


FIGURE 67 *The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under* Microsoft Face API *with a configuration of* detection_02 *and* recognition_04, *for the three light skinned skin tone classifications, collectively* ***Skin Tone I*** *(Light Pink),* ***Skin Tone II*** *(Light Yellow), and* ***Skin Tone III*** *(Medium Pink / Brown), using the IJB-C skin tone classification schema.*

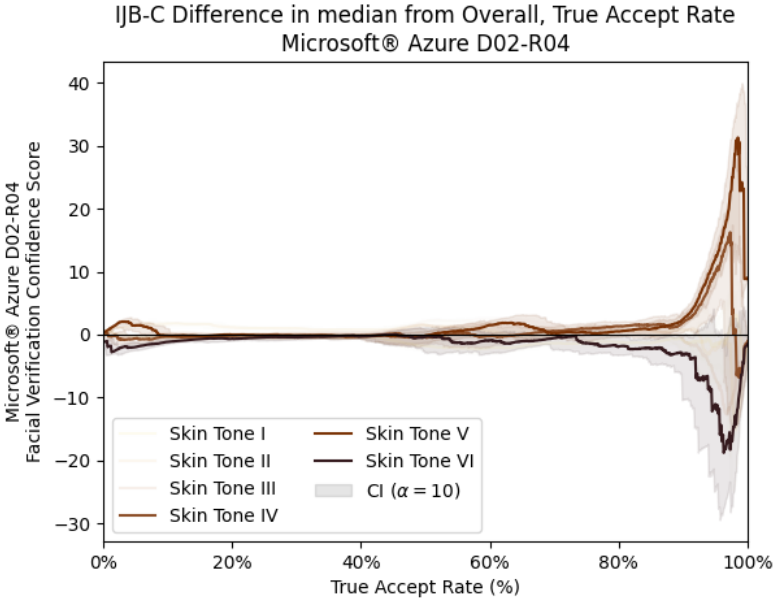


FIGURE 68 *The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under* Microsoft Face API *with a configuration of* detection_02 *and* recognition_04, *for the three dark skinned skin tone classifications, collectively* ***Skin Tone IV*** *(Medium Yellow / Brown),* ***Skin Tone V*** *(Medium-Dark Brown), and* ***Skin Tone VI*** *(Dark Brown), using the IJB-C skin tone classification schema.*

Microsoft Face API with a configuration of detection_03 and recognition_01

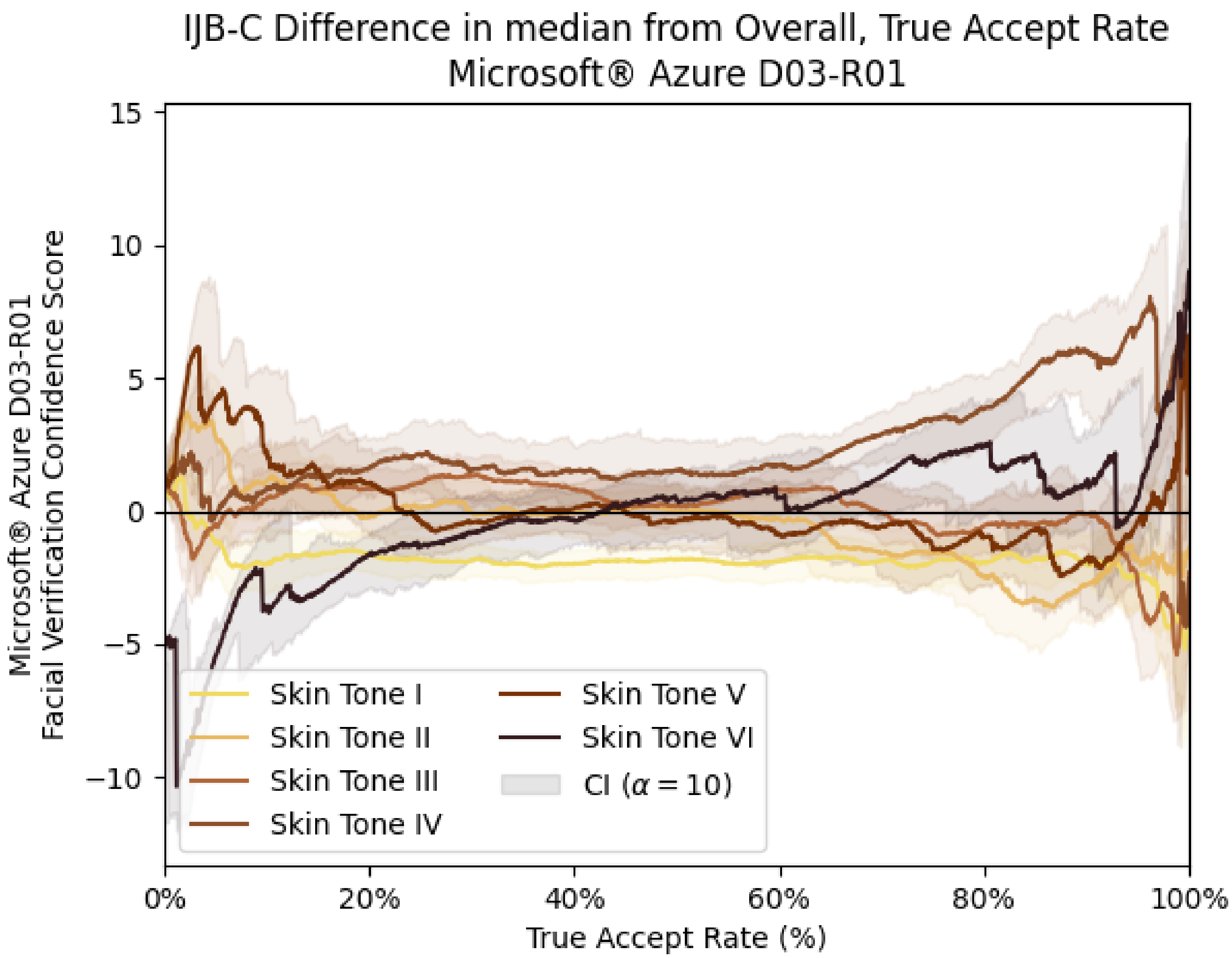


FIGURE 69 The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under Microsoft Face API *with a configuration of* detection_03 *and* recognition_01, *for the six skin tone classifications as defined by the IJB-C skin tone classification schema.*

Microsoft Face API with a configuration of detection_03 and recognition_01

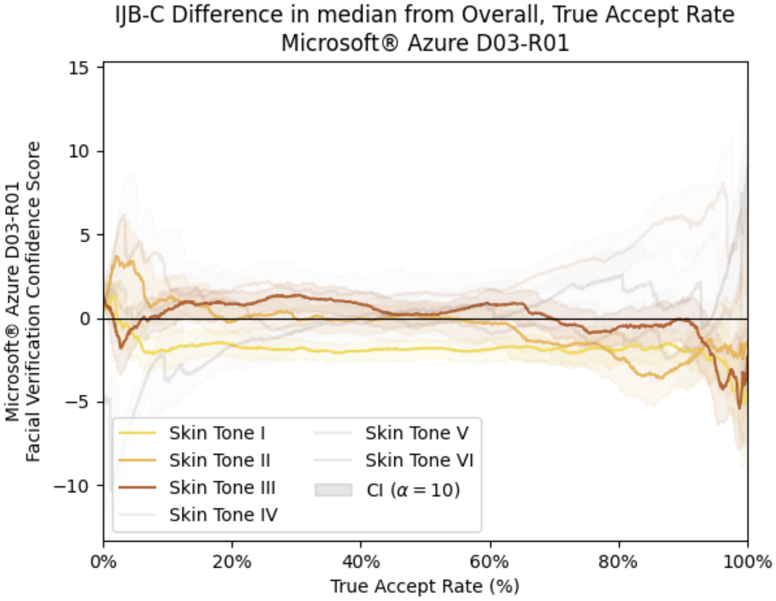


Figure 70 *The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under* Microsoft Face API *with a configuration of* detection_03 *and* recognition_01, *for the three light skinned skin tone classifications, collectively **Skin Tone I** (Light Pink), **Skin Tone II** (Light Yellow), and **Skin Tone III** (Medium Pink / Brown), using the IJB-C skin tone classification schema.*

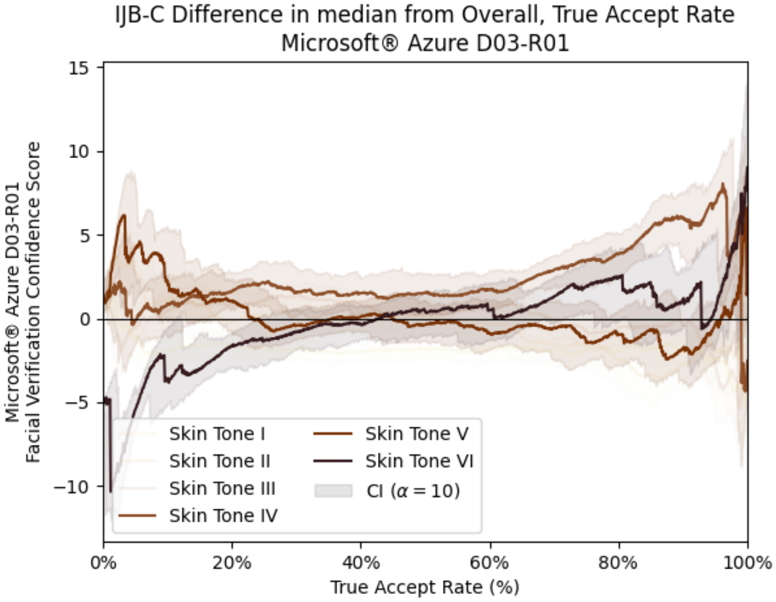


Figure 71 *The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under* Microsoft Face API *with a configuration of* detection_03 *and* recognition_01, *for the three dark skinned skin tone classifications, collectively **Skin Tone IV** (Medium Yellow / Brown), **Skin Tone V** (Medium-Dark Brown), and **Skin Tone VI** (Dark Brown), using the IJB-C skin tone classification schema.*

Microsoft Face API with a configuration of detection_03 and recognition_02

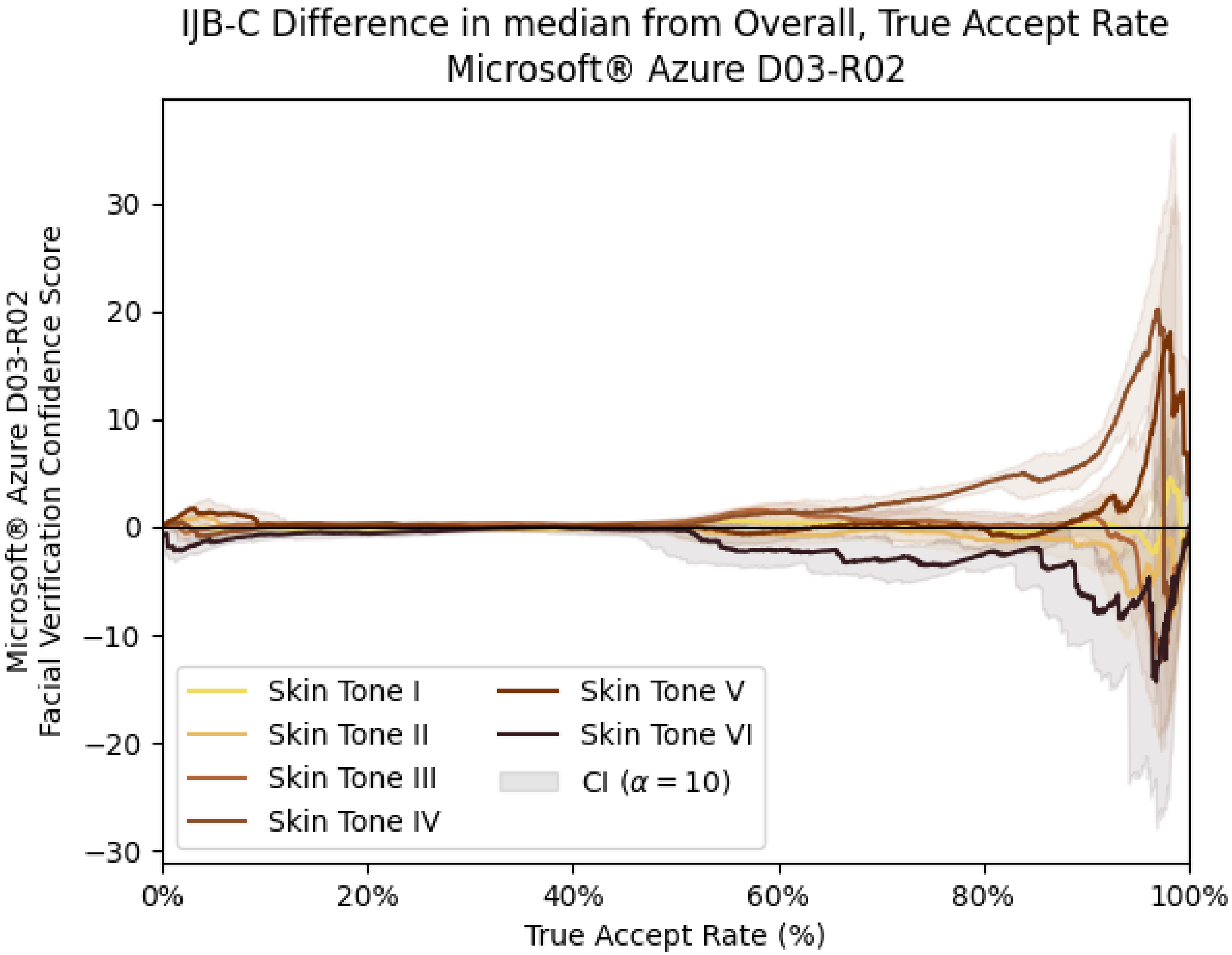


FIGURE 72 The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under Microsoft Face API *with a configuration of* detection_03 *and* recognition_02*, for the six skin tone classifications as defined by the IJB-C skin tone classification schema.*

Microsoft Face API with a configuration of detection_03 and recognition_02

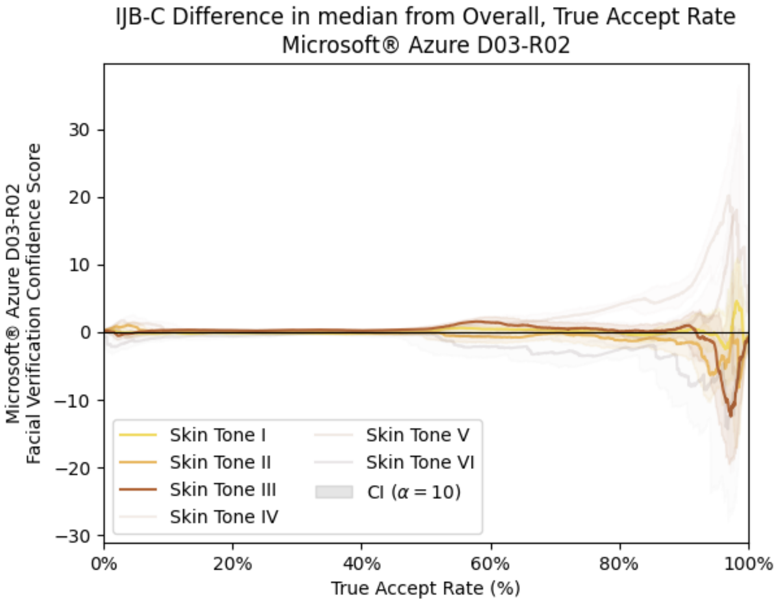


FIGURE 73 The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under Microsoft Face API *with a configuration of* detection_03 *and* recognition_02*, for the three light skinned skin tone classifications, collectively* ***Skin Tone I*** *(Light Pink),* ***Skin Tone II*** *(Light Yellow), and* ***Skin Tone III*** *(Medium Pink / Brown), using the IJB-C skin tone classification schema.*

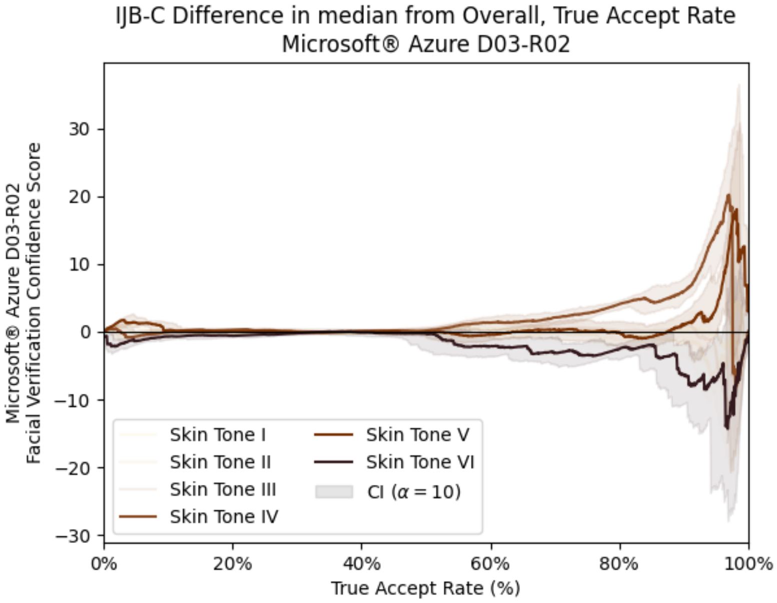


FIGURE 74 The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under Microsoft Face API *with a configuration of* detection_03 *and* recognition_02*, for the three dark skinned skin tone classifications, collectively* ***Skin Tone IV*** *(Medium Yellow / Brown),* ***Skin Tone V*** *(Medium-Dark Brown), and* ***Skin Tone VI*** *(Dark Brown), using the IJB-C skin tone classification schema.*

Microsoft Face API with a configuration of detection_03 and recognition_03

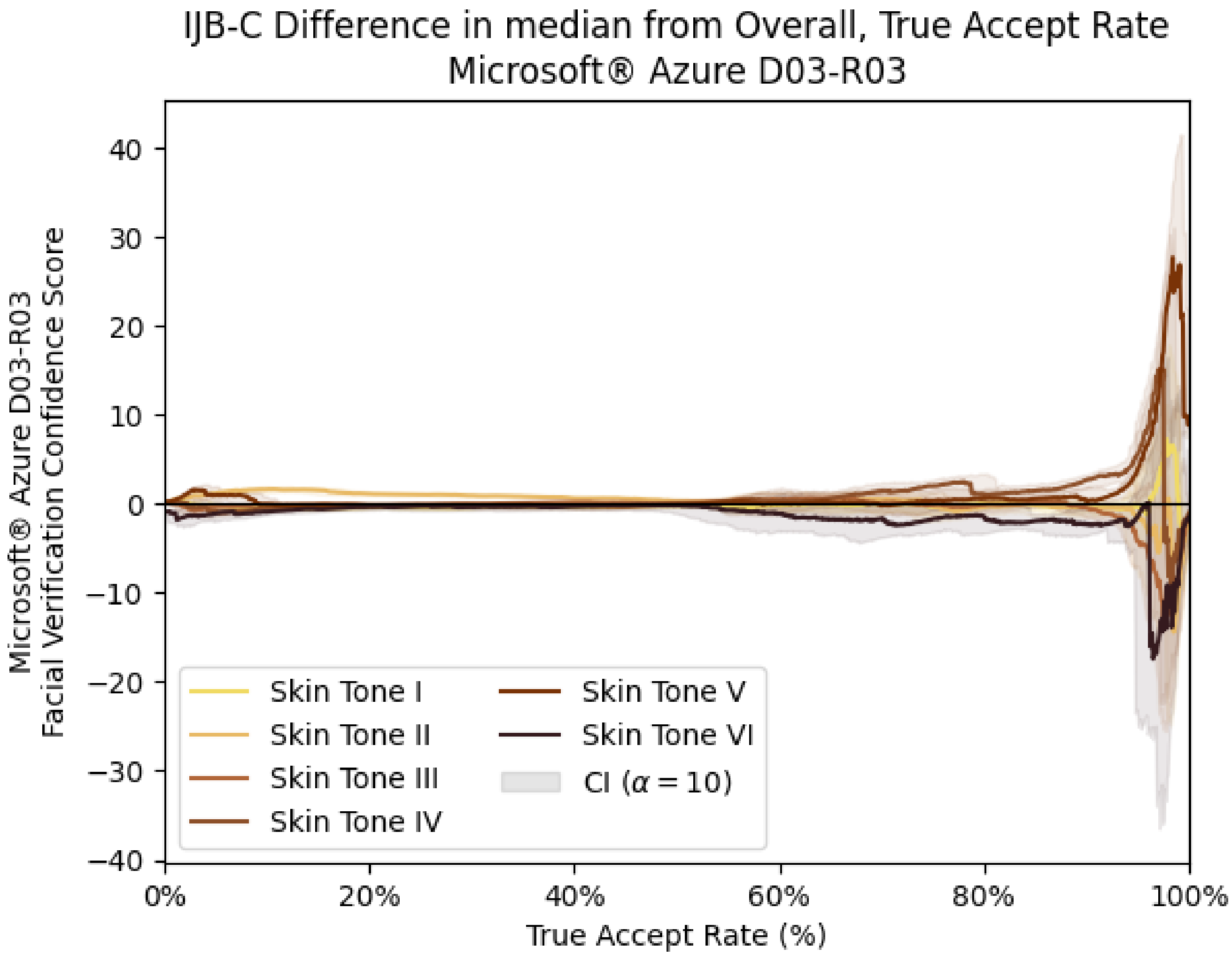


FIGURE 75 The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under Microsoft Face API *with a configuration of* detection_03 *and* recognition_03, *for the six skin tone classifications as defined by the IJB-C skin tone classification schema.*

Microsoft Face API with a configuration of detection_03 and recognition_03

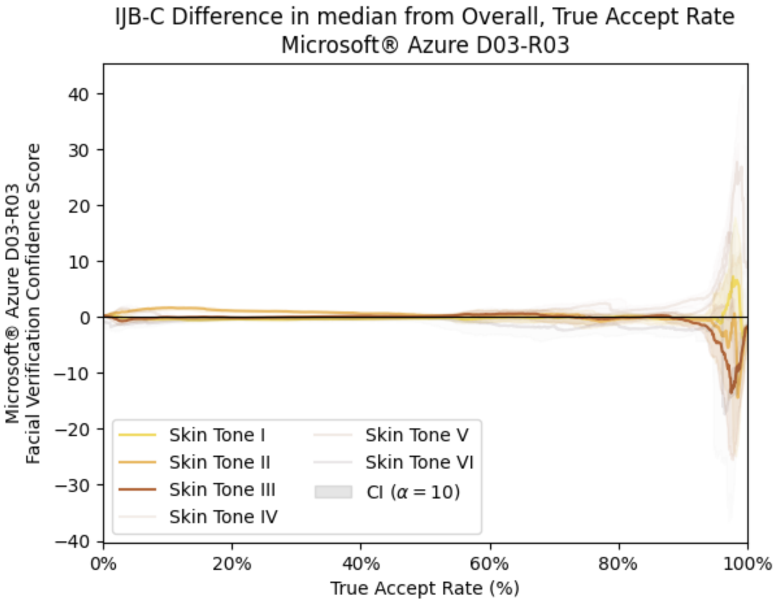


Figure 76 *The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under* Microsoft Face API *with a configuration of* detection_03 *and* recognition_03*, for the three light skinned skin tone classifications, collectively* ***Skin Tone I*** *(Light Pink),* ***Skin Tone II*** *(Light Yellow), and* ***Skin Tone III*** *(Medium Pink / Brown), using the IJB-C skin tone classification schema.*

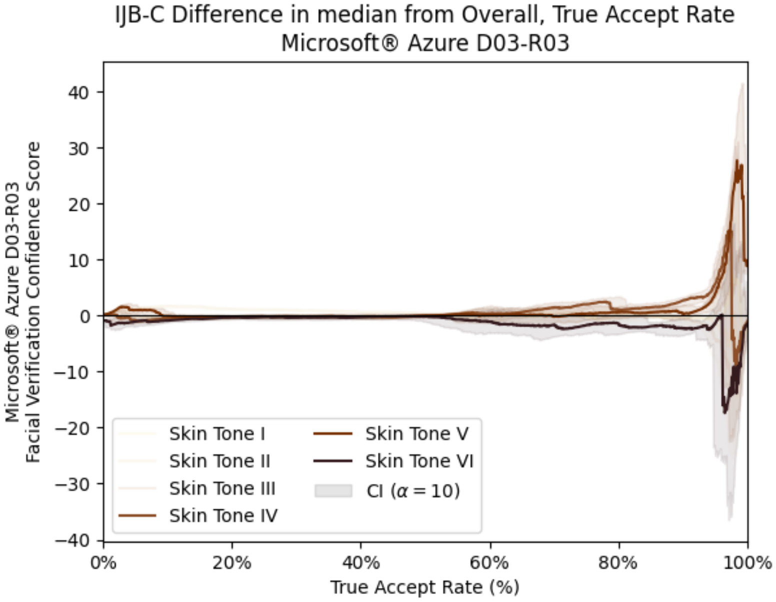


Figure 77 *The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under* Microsoft Face API *with a configuration of* detection_03 *and* recognition_03*, for the three dark skinned skin tone classifications, collectively* ***Skin Tone IV*** *(Medium Yellow / Brown),* ***Skin Tone V*** *(Medium-Dark Brown), and* ***Skin Tone VI*** *(Dark Brown), using the IJB-C skin tone classification schema.*

Microsoft Face API, Released 2021

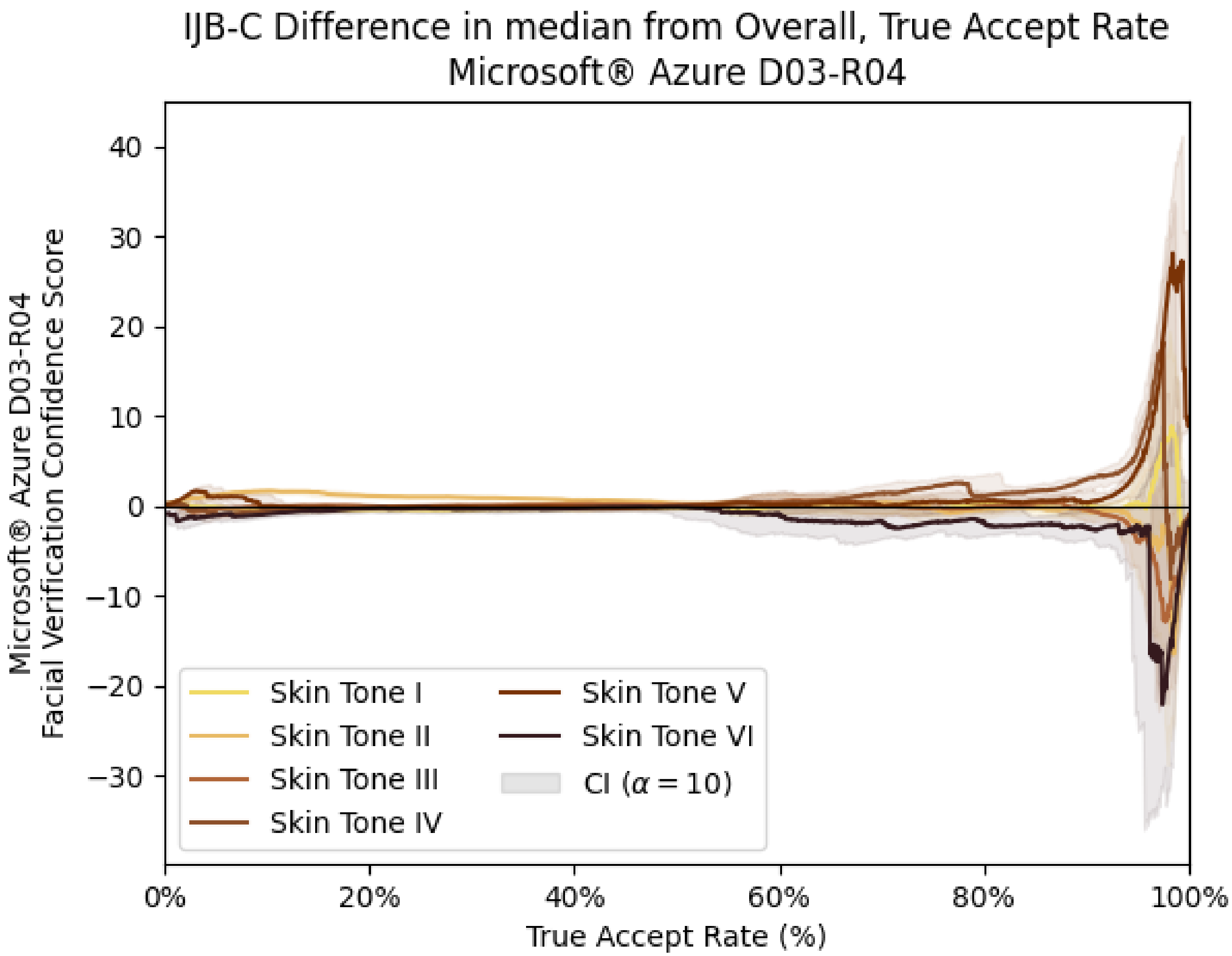


FIGURE 78 The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under Microsoft Face API *under its default configuration of* detection_03 *and* recognition_04*, released in 2021, for the six skin tone classifications as defined by the IJB-C skin tone classification schema.*

Microsoft Face API, Released 2021

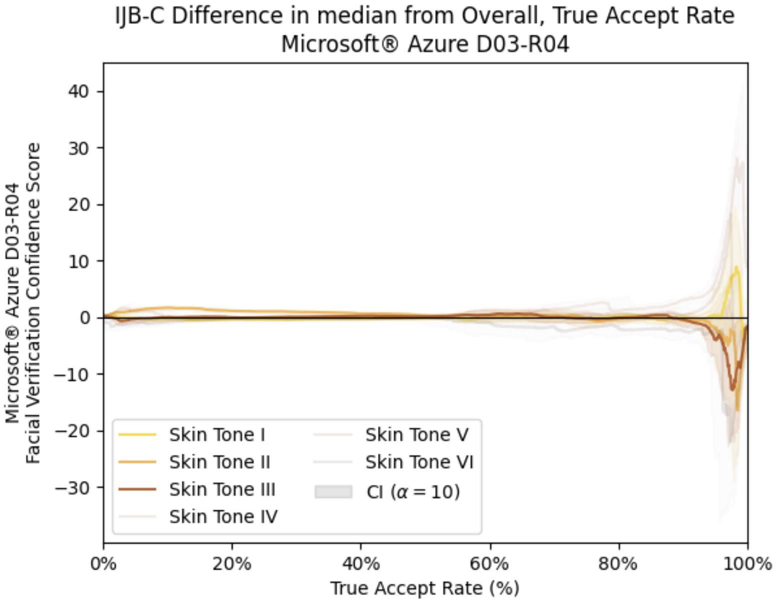


***Figure 79** The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under* Microsoft Face API *under its default configuration of* detection_03 *and* recognition_04, *released in 2021, for the three light skinned skin tone classifications, collectively **Skin Tone I** (Light Pink), **Skin Tone II** (Light Yellow), and **Skin Tone III** (Medium Pink / Brown), using the IJB-C skin tone classification schema.*

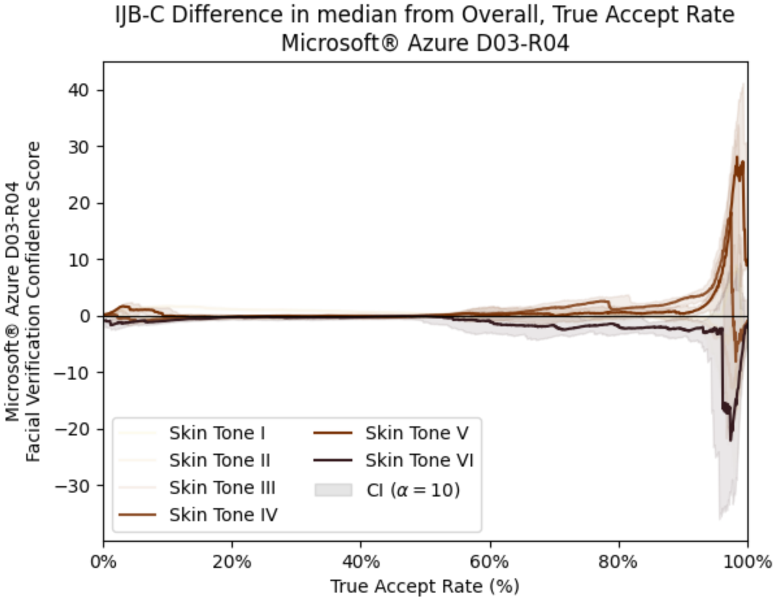


***Figure 80** The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under* Microsoft Face API *under its default configuration of* detection_03 *and* recognition_04, *released in 2021, for the three dark skinned skin tone classifications, collectively **Skin Tone IV** (Medium Yellow / Brown), **Skin Tone V** (Medium-Dark Brown), and **Skin Tone VI** (Dark Brown), using the IJB-C skin tone classification schema.*

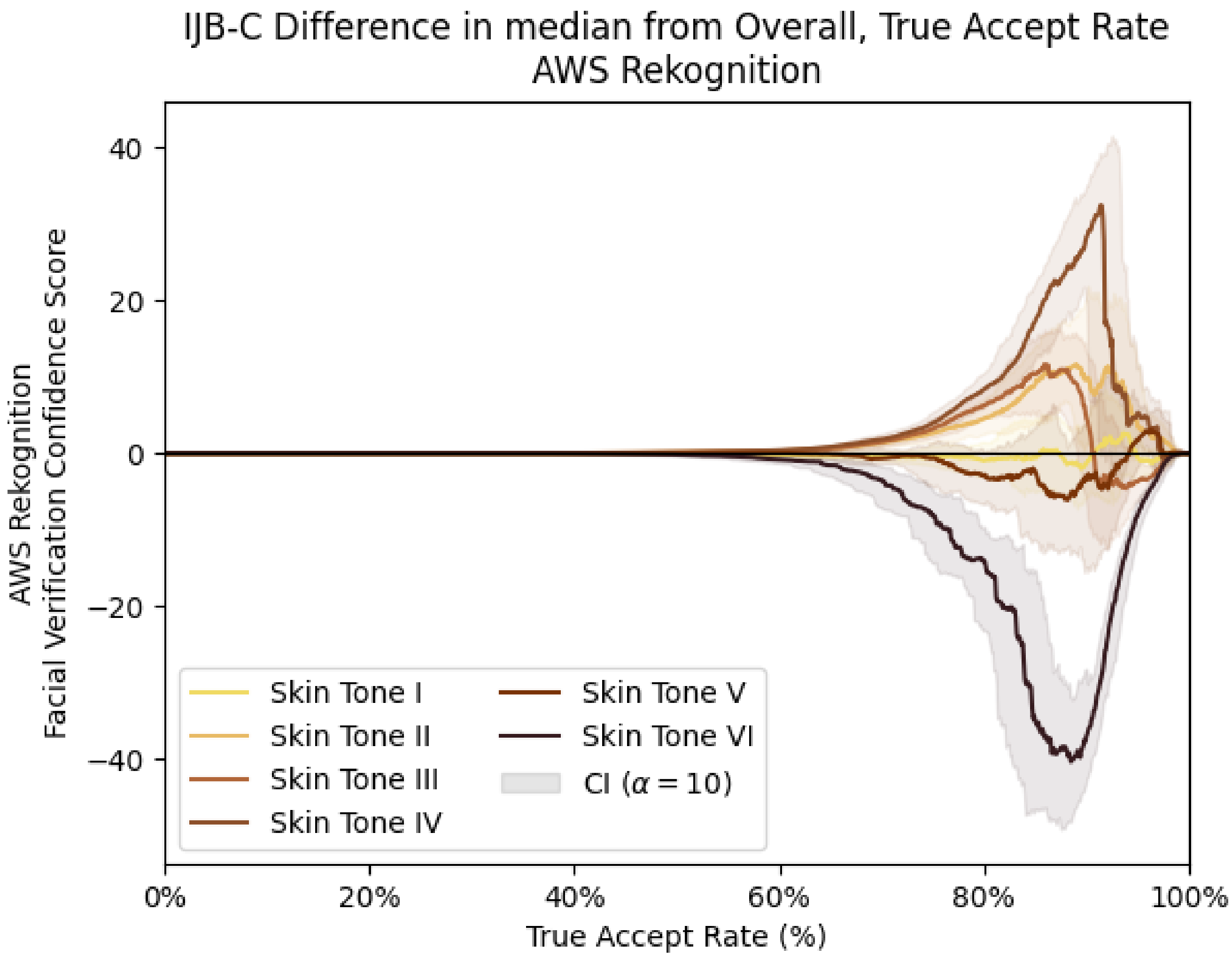


FIGURE 81 The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under AWS Rekognition *for the six skin tone classifications as defined by the IJB-C skin tone classification schema.*

AWS Rekognition, As of Fall 2021

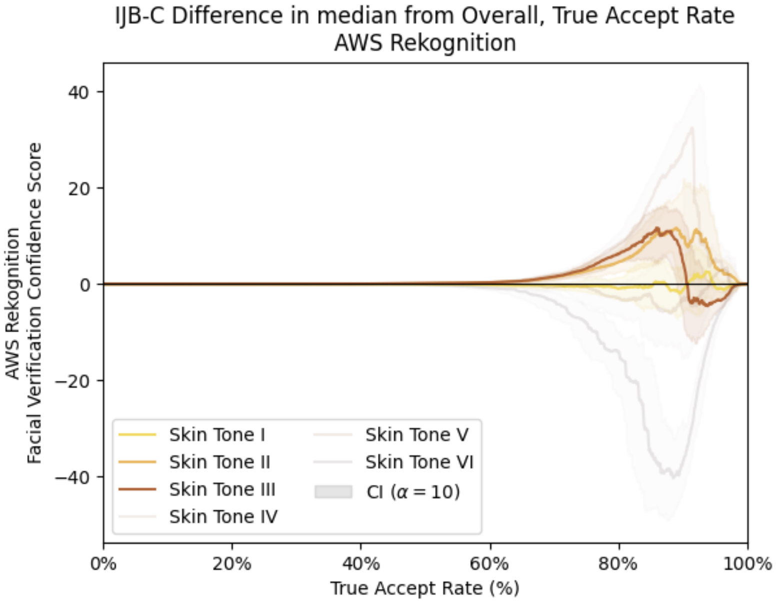


***FIGURE 82** The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under* AWS Rekognition *for the three light skinned skin tone classifications, collectively **Skin Tone I** (Light Pink), **Skin Tone II** (Light Yellow), and **Skin Tone III** (Medium Pink / Brown), using the IJB-C skin tone classification schema.*

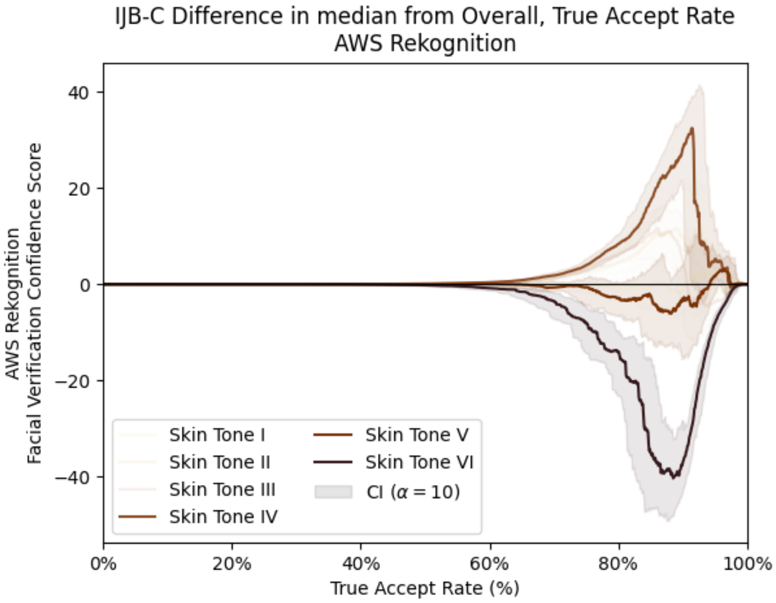


***FIGURE 83** The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under* AWS Rekognition, *for the three dark skinned skin tone classifications, collectively **Skin Tone IV** (Medium Yellow / Brown), **Skin Tone V** (Medium-Dark Brown), and **Skin Tone VI** (Dark Brown), using the IJB-C skin tone classification schema.*

Distance-based Variance Interclass Bias Metric

Subsequently it is important to transform these graphical understandings of interclass bias into the d_{l_2} measure of interclass bias between a class and the overall performance. Table 44-44 relay the maximum d_{l_2} measure for each of the six skin tone classifications for the performance of *Microsoft Face API* under all of its possible configurations, and *AWS Rekognition*. The tables encode the configuration of the Microsoft Face API as DXX-RYY where XX represents the detection model detection_XX and YY represents the recognition model recognition_YY. For example, D01-R02 would indicate *Microsoft Face API* with a configuration of detection_01 and recognition_02.

TABLE 4 *The maximum measure of interclass bias for each of the six skin tone classifications under* Microsoft Face API *with a configuration of* detection_01, detection_02, detection_03 *each paired with* recognition_01, *using the IJB-C skin tone classification schema.*

Class	***Microsoft Face API*** *D01-R01*	***Microsoft Face API*** *D02-R01*	***Microsoft Face API*** *D03-R01*
Skin Tone I *(Light Pink)*	2.669012	3.454001	2.599026
Skin Tone II *(Light Yellow)*	1.97817	2.307768	1.933965
Skin Tone III *(Medium Pink / Brown)*	1.207779	1.626414	1.264912
Skin Tone IV *(Medium Yellow / Brown)*	4.009657	3.495838	4.13781
Skin Tone V *(Medium-Dark Brown)*	1.892936	3.004449	1.807422
Skin Tone VI *(Dark Brown)*	2.391841	2.83183	2.326273
Total	***4.009657***	***3.495838***	***4.13781***

TABLE 5 *The maximum measure of interclass bias for each of the six skin tone classifications under* Microsoft Face API *with a configuration of* detection_01, detection_02, detection_03 *each paired with* recognition_02, *using the IJB-C skin tone classification schema.*

Class	***Microsoft Face API*** *D01-R02*	***Microsoft Face API*** *D02-R02*	***Microsoft Face API*** *D03-R02*
Skin Tone I *(Light Pink)*	0.843805	1.528638	0.817172
Skin Tone II *(Light Yellow)*	1.899399	2.366992	1.777788
Skin Tone III *(Medium Pink / Brown)*	1.181191	2.048876	1.16398
Skin Tone IV *(Medium Yellow / Brown)*	3.349543	3.606959	3.128856
Skin Tone V *(Medium-Dark Brown)*	2.079491	3.21583	2.134085
Skin Tone VI *(Dark Brown)*	3.882438	5.540444	4.303409
Total	***3.882438***	***5.540444***	***4.303409***

TABLE 6 *The maximum measure of interclass bias for each of the six skin tone classifications under* Microsoft Face API *with a configuration of* detection_01, detection_02, detection_03 *each paired with* recognition_03, *using the IJB-C skin tone classification schema.*

	Microsoft	***Microsoft***	***Microsoft***
	Face API	***Face API***	***Face API***
Class	*D01-R03*	*D02-R03*	*D03-R03*
Skin Tone I *(Light Pink)*	0.971448	1.756607	0.855982
Skin Tone II *(Light Yellow)*	1.441749	1.723081	1.526481
Skin Tone III *(Medium Pink / Brown)*	1.193055	1.374286	1.194568
Skin Tone IV *(Medium Yellow / Brown)*	1.852326	2.074765	1.883101
Skin Tone V *(Medium-Dark Brown)*	2.062876	2.6775	2.186757
Skin Tone VI *(Dark Brown)*	3.020116	3.474086	3.09069
Total	***3.020116***	***3.474086***	***3.09069***

TABLE 7 *The maximum measure of interclass bias for each of the six skin tone classifications under* Microsoft Face API *with a configuration of* detection_01, detection_02, detection_03 *each paired with* recognition_04, *using the IJB-C skin tone classification schema.*

	Microsoft	***Microsoft***	***Microsoft***
	Face API	***Face API***	***Face API***
Class	*D01-R04*	*D02-R04*	*D03-R04*
Skin Tone I *(Light Pink)*	0.9212	1.718912	0.846978
Skin Tone II *(Light Yellow)*	1.530604	1.648383	1.590849
Skin Tone III *(Medium Pink / Brown)*	1.208315	1.315521	1.29098
Skin Tone IV *(Medium Yellow / Brown)*	2.06674	2.146795	2.134128
Skin Tone V *(Medium-Dark Brown)*	2.069679	2.791353	2.222468
Skin Tone VI *(Dark Brown)*	3.242087	3.629941	3.261147
Total	***3.242087***	***3.629941***	***3.261147***

TABLE 8 *The maximum measure of interclass bias for each of the six skin tone classifications under* AWS Rekognition, *using the IJB-C skin tone classification schema.*

	AWS
Class	***Rekognition***
Skin Tone I *(Light Pink)*	1.08468
Skin Tone II *(Light Yellow)*	3.007234
Skin Tone III *(Medium Pink / Brown)*	2.396734
Skin Tone IV *(Medium Yellow / Brown)*	4.972686
Skin Tone V *(Medium-Dark Brown)*	2.186609
Skin Tone VI *(Dark Brown)*	7.837107
Total	***7.837107***

Table 9 *The 90% two-sided confidence intervals ($\alpha = 10\%$) for the maximum measure of interclass bias for the binary masculine or feminine gender presentation classification schema under* Microsoft Face API *under all of its possible configurations, and* AWS Rekognition*.*

Class	***Lower Bound***	***Upper Bound***
Microsoft Face API *D01-D02*	0.880488	4.009657
Microsoft Face API *D01-D02*	0.283942	3.882438
Microsoft Face API *D01-D03*	0.470357	3.020116
Microsoft Face API *D01-D04*	0.50061	3.242087
Microsoft Face API *D02-D01*	1.07061	3.495838
Microsoft Face API *D02-D02*	0.514375	5.540444
Microsoft Face API *D02-D03*	0.747604	3.474086
Microsoft Face API *D02-D04*	0.711602	3.629941
Microsoft Face API *D03-D01*	0.821633	4.13781
Microsoft Face API *D03-D02*	0.317373	4.303409
Microsoft Face API *D03-D03*	0.389303	3.09069
Microsoft Face API *D03-D04*	0.429666	3.261147
AWS Rekognition	0.230991	7.837107

IJB-C Gender Presentation Audit of Commercial Facial Verification Algorithms

A total of $14,436$ comparisons were submitted to each algorithm for scoring, yielding $14,436$ confidence scores for each of the commercial facial verification algorithms: (a) *Microsoft Face API* under its default configuration of detectiosn_01 and recognition_01 released in 2017, (b) *Microsoft Face API* with a configuration of detection_01 and recognition_02, (b) *Microsoft Face API* with a configuration of detection_01 and recognition_03, (c) *Microsoft Face API* with a configuration of detection_01 and recognition_04, (d) *Microsoft Face API* with a configuration of detection_02 and recognition_01, (e) *Microsoft Face API* with a configuration of detection_02 and recognition_02 released in 2019, (f) *Microsoft Face API* with a configuration of detection_02 and recognition_03 released in 2020, (h) *Microsoft Face API* with a configuration of detection_02 and recognition_04, (i) *Microsoft Face API* with a configuration of detection_03 and recognition_01, (j) *Microsoft Face API* with a configuration of detection_03 and recognition_02, (k) *Microsoft Face API* with a configuration of detection_03 and recognition_03, (l) *Microsoft Face API* with its latest released a configuration of detection_03 and recognition_04 released in 2021, and (m) *AWS Rekognition*. For each of these commercial facial verification algorithms, the confidence scores reported were partitioned into two gender presentation classifications as defined by the IJB-C classification schema.

True Accept Rates

Subsequently, confidence bounds for the true accept rate ("TAR"), calculated from the fraction of genuine comparisons that correctly exceed the threshold, for each of two gender presentation classifications, using the bootstrap method. The author set $\alpha = 10\%$ for the two-sided confidence interval, α was selected based on the expected statistical power from the aforementioned asymptotic relative efficiency adjustment to the Mann-Whitney method for determining sample size. It has been shown that for this significance level that a minimum of 599 bootstrap resamples must be conducted (Davidson and MacKinnon 2000; Wilcox 2010), as computational power and resources were freely available, the author conducted 9999 resamples in keeping with the default choice for the statistical software package used to generate the resamples (Pedregosa et al. 2011). Using these resampled confidence scores, the threshold is varied across the entire domain of the confidence scores to plot the TAR and the confidence bounds for the facial verification algorithms performance for each of the two gender presentation classifications. Figures 44-44 plot the TAR and the corresponding confidence score threshold to achieve it under the thirteen commercial facial verification algorithms.

Microsoft Face API, Released 2017

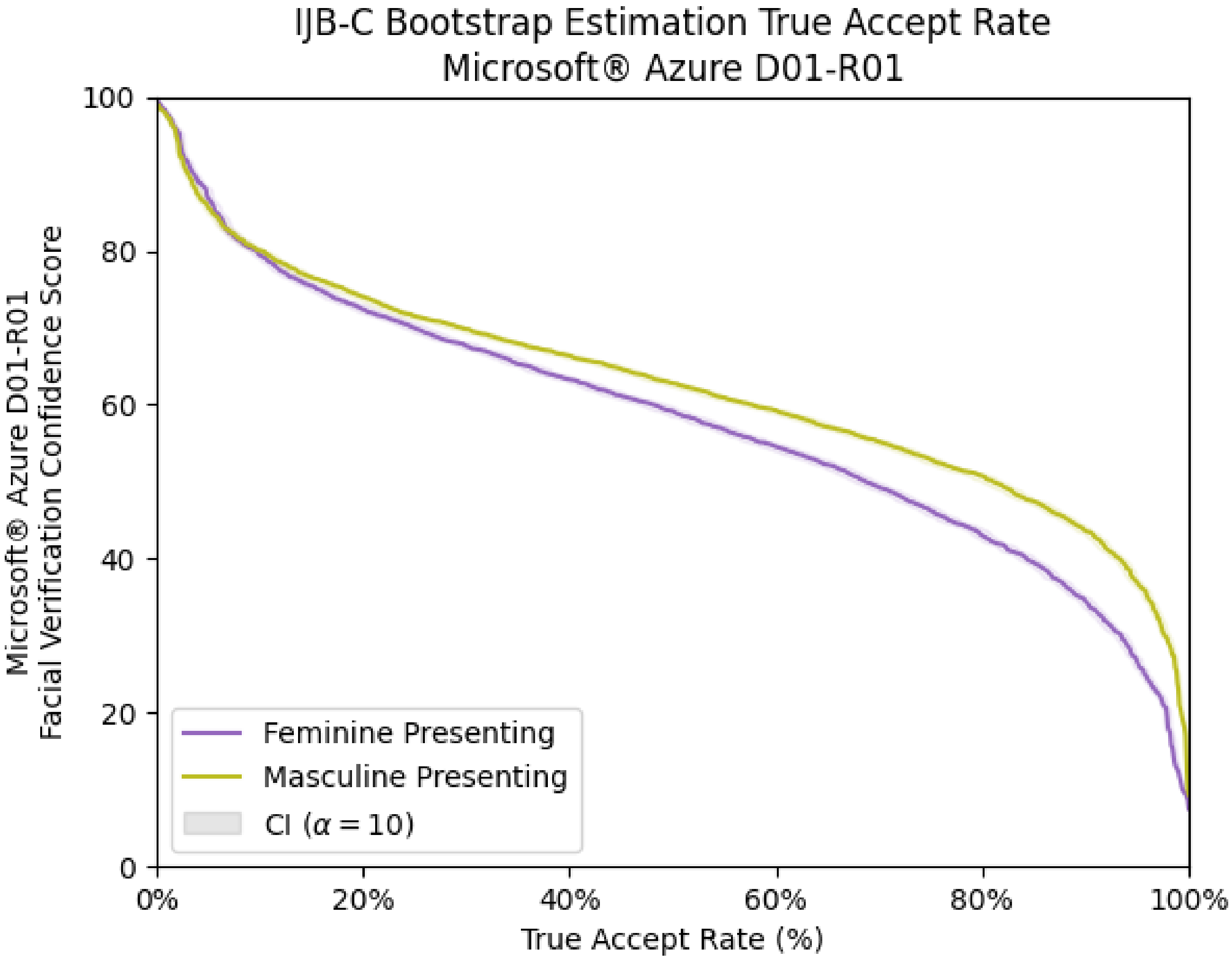


Figure 84 *The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under* Microsoft Face API *under its default configuration of* detection_01 *and* recognition_01, *released in 2017, for the binary masculine or feminine gender presentation classification schema.*

Microsoft Face API with a configuration of detection_01 and recognition_02

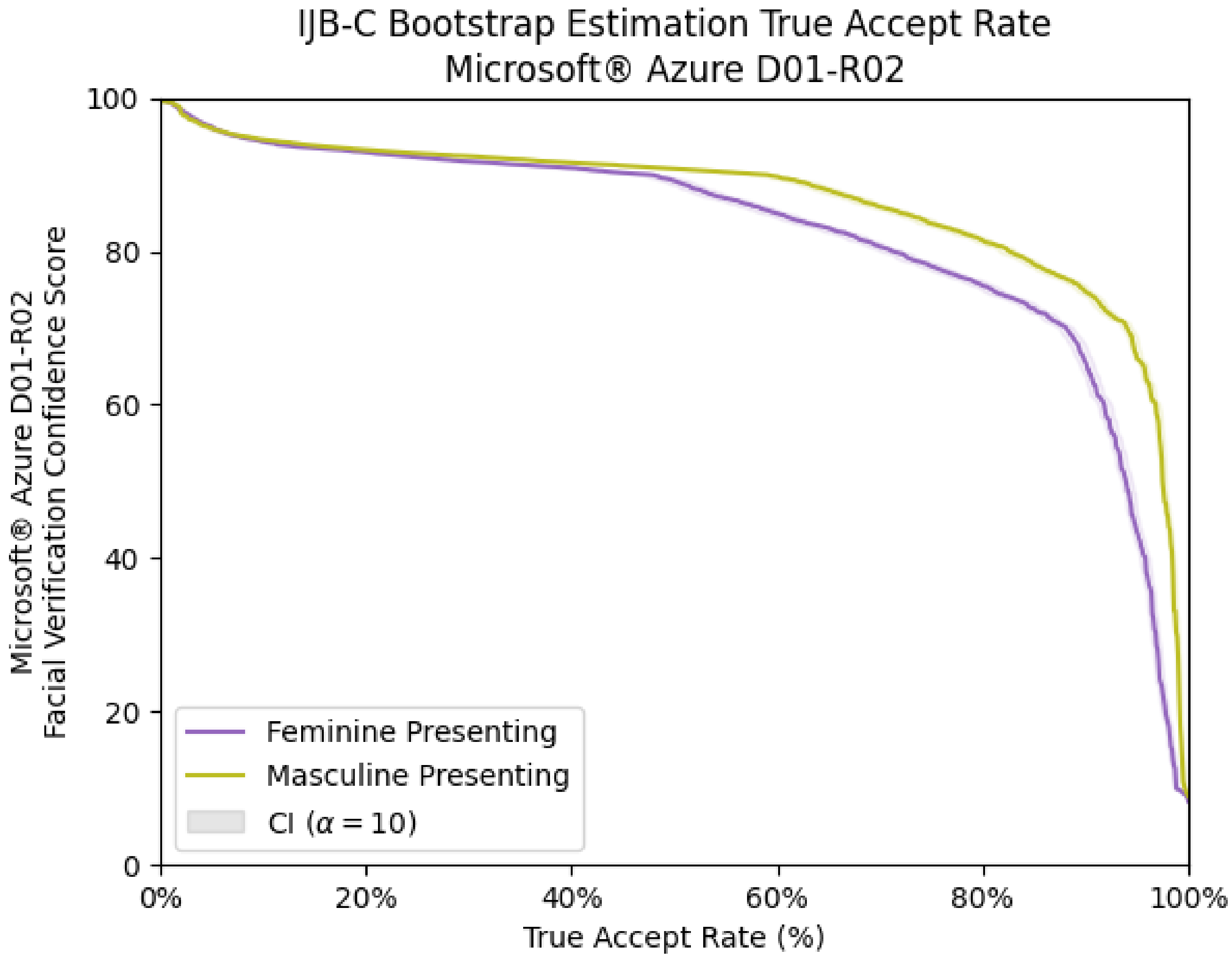


FIGURE 85 The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under Microsoft Face API *with a configuration of* detection_01 *and* recognition_02*, for the binary masculine or feminine gender presentation classification schema.*

Microsoft Face API with a configuration of detection_01 and recognition_03

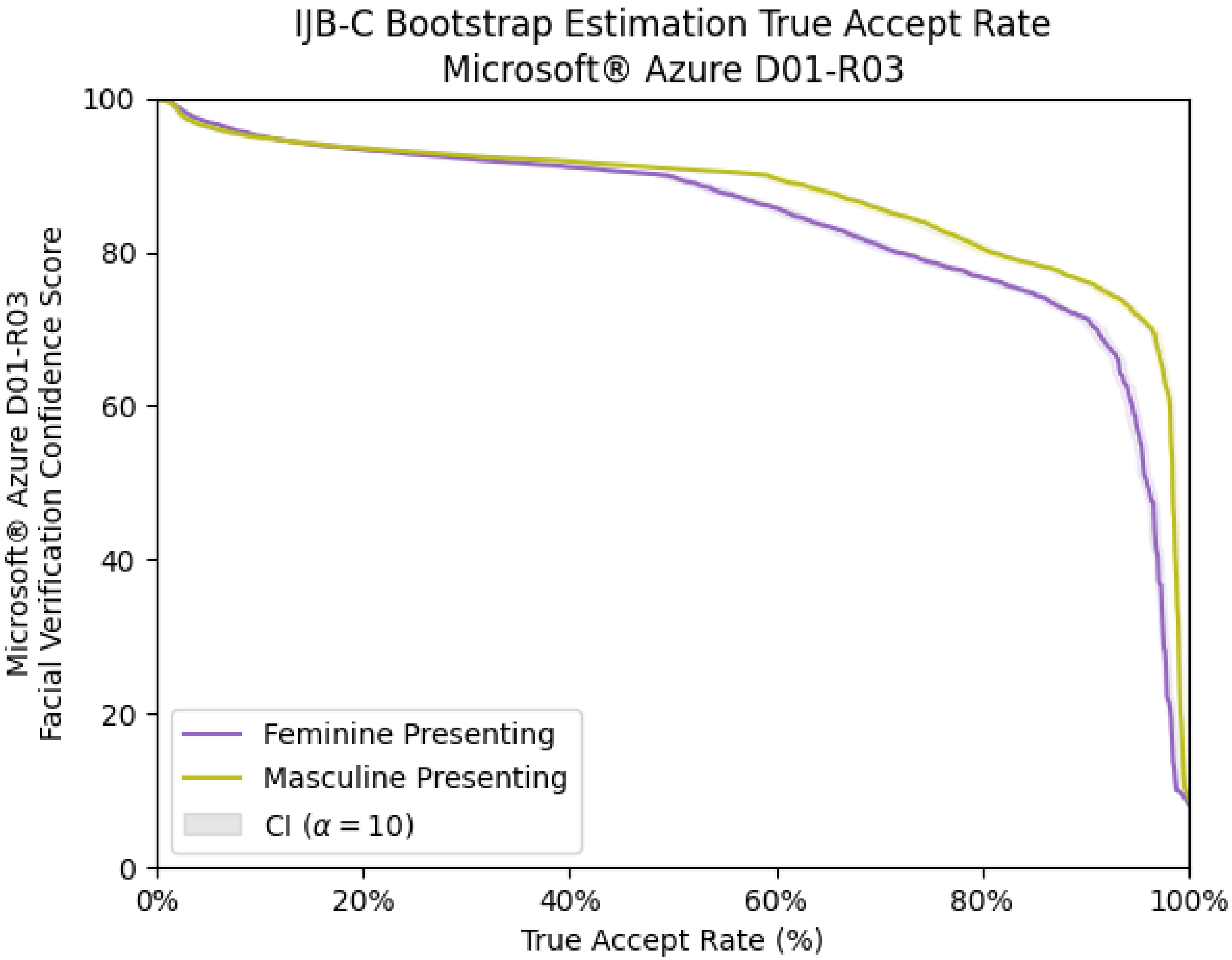


FIGURE 86 The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under Microsoft Face API *with a configuration of* detection_01 *and* recognition_03*, for the binary masculine or feminine gender presentation classification schema.*

Microsoft Face API with a configuration of detection_01 and recognition_04

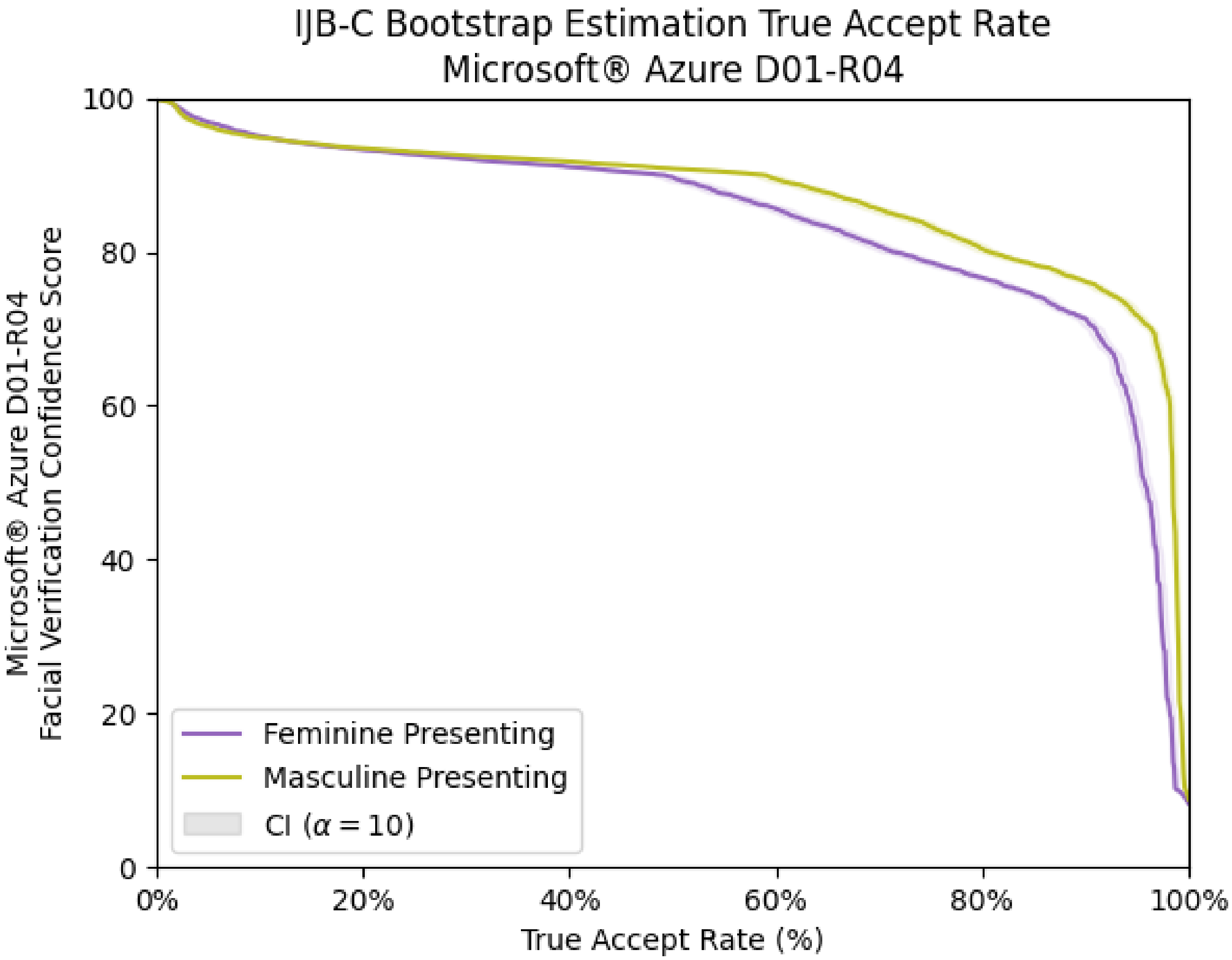


Figure 87 *The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under* Microsoft Face API *with a configuration of* detection_01 *and* recognition_04*, for the binary masculine or feminine gender presentation classification schema.*

Microsoft Face API with a configuration of detection_02 and recognition_01

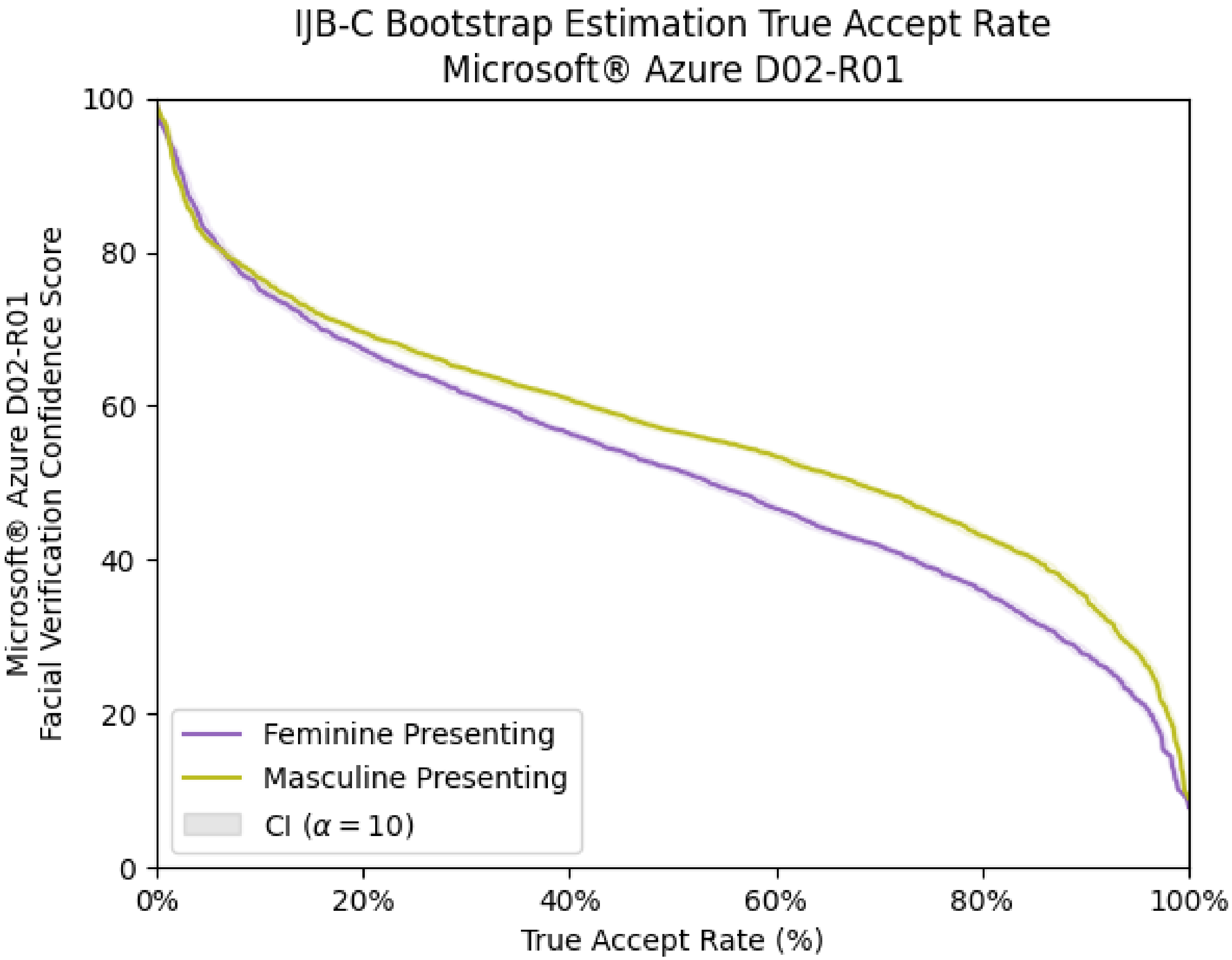


Figure 88 *The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under* Microsoft Face API *with a configuration of* detection_02 *and* recognition_01, *for the binary masculine or feminine gender presentation classification schema.*

Microsoft Face API, Released 2019

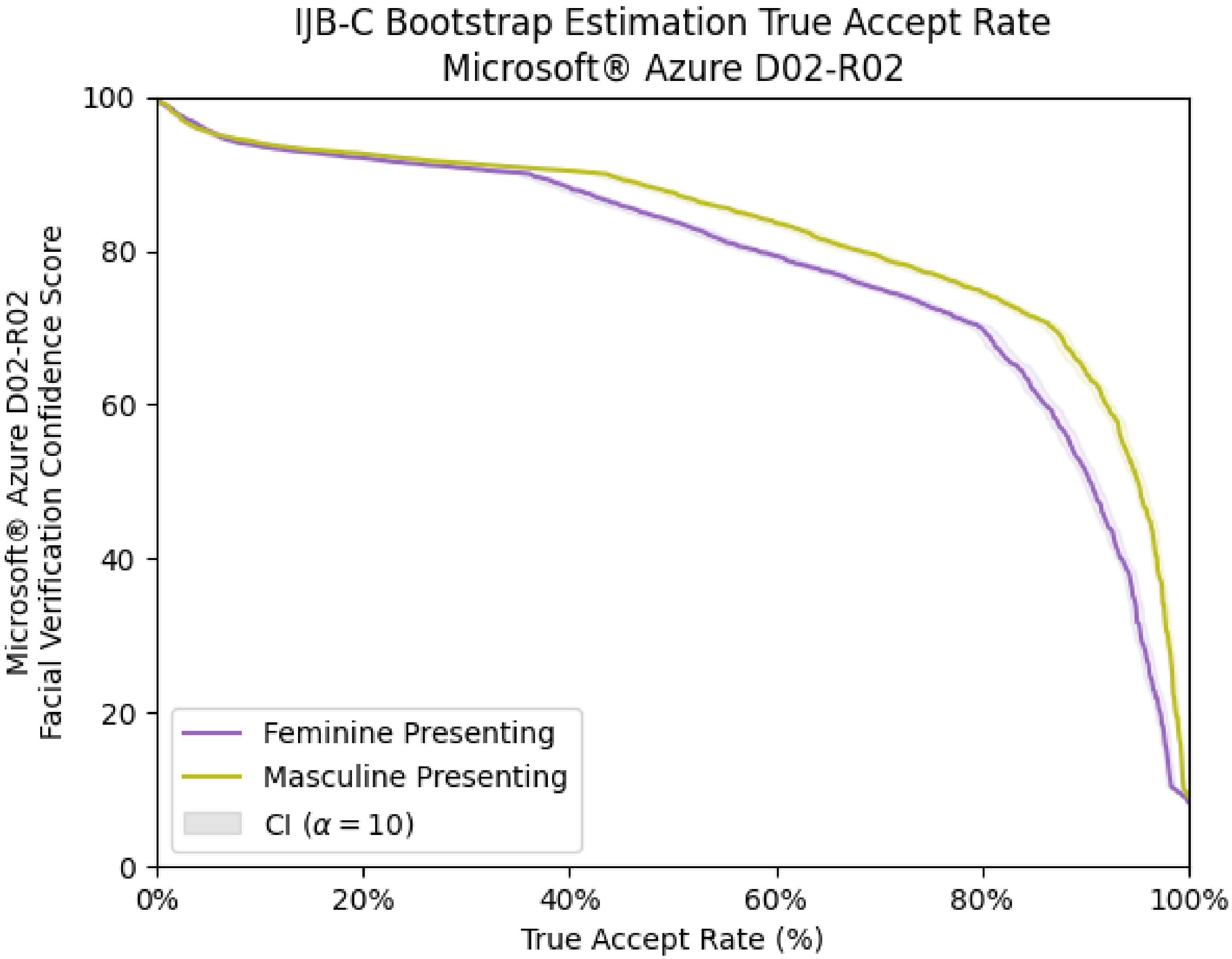


Figure 89 *The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under* Microsoft Face API *under its default configuration of* detection_02 *and* recognition_02*, released in 2019, for the binary masculine or feminine gender presentation classification schema.*

Microsoft Face API, Released 2020

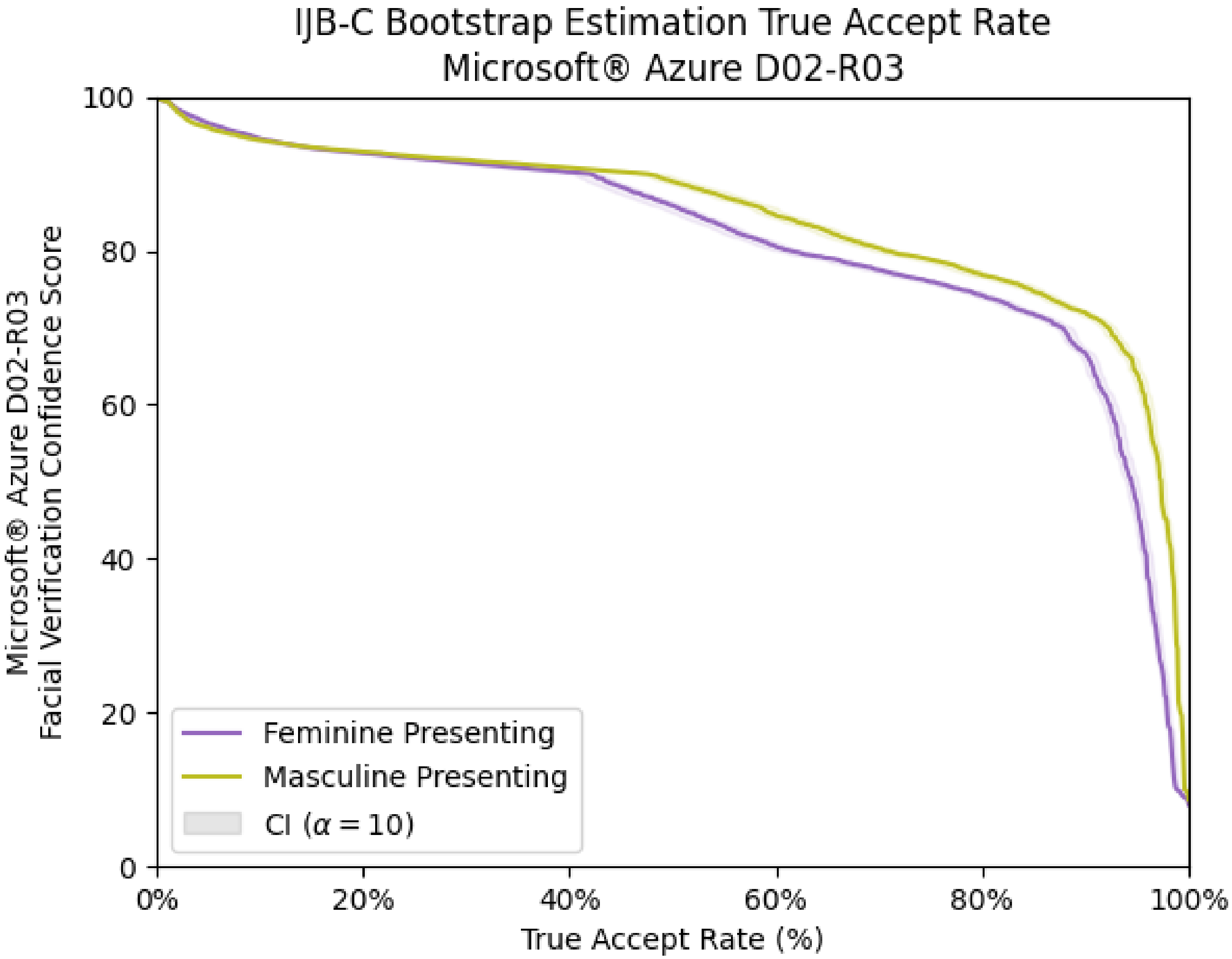


Figure 90 *The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under* Microsoft Face API *under its default configuration of* detection_02 *and* recognition_03*, released in 2020, for the binary masculine or feminine gender presentation classification schema.*

Microsoft Face API with a configuration of detection_02 and recognition_04

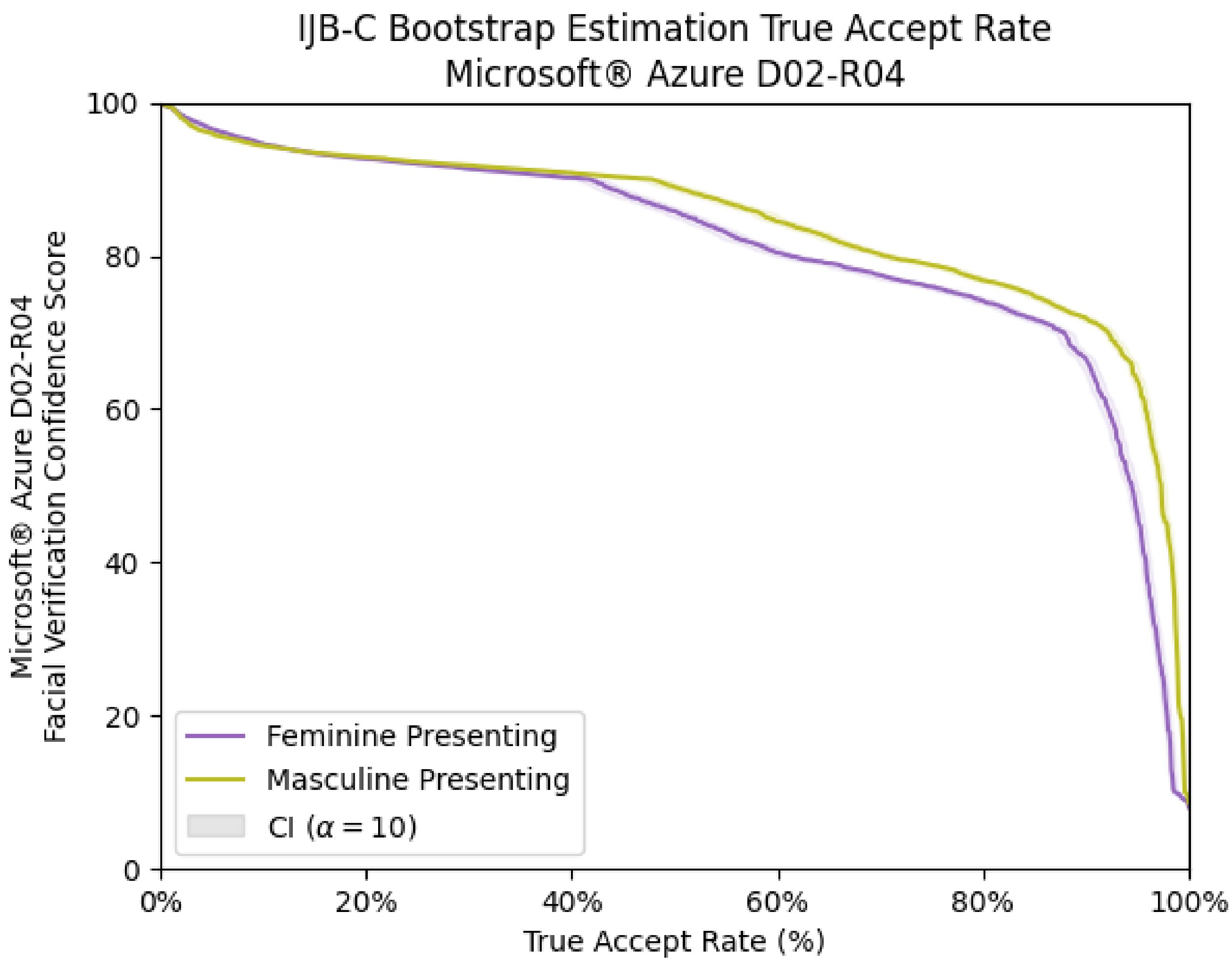


***Figure 91** The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under* Microsoft Face API *with a configuration of* detection_02 *and* recognition_04*, for the binary masculine or feminine gender presentation classification schema.*

Microsoft Face API with a configuration of detection_03 and recognition_01

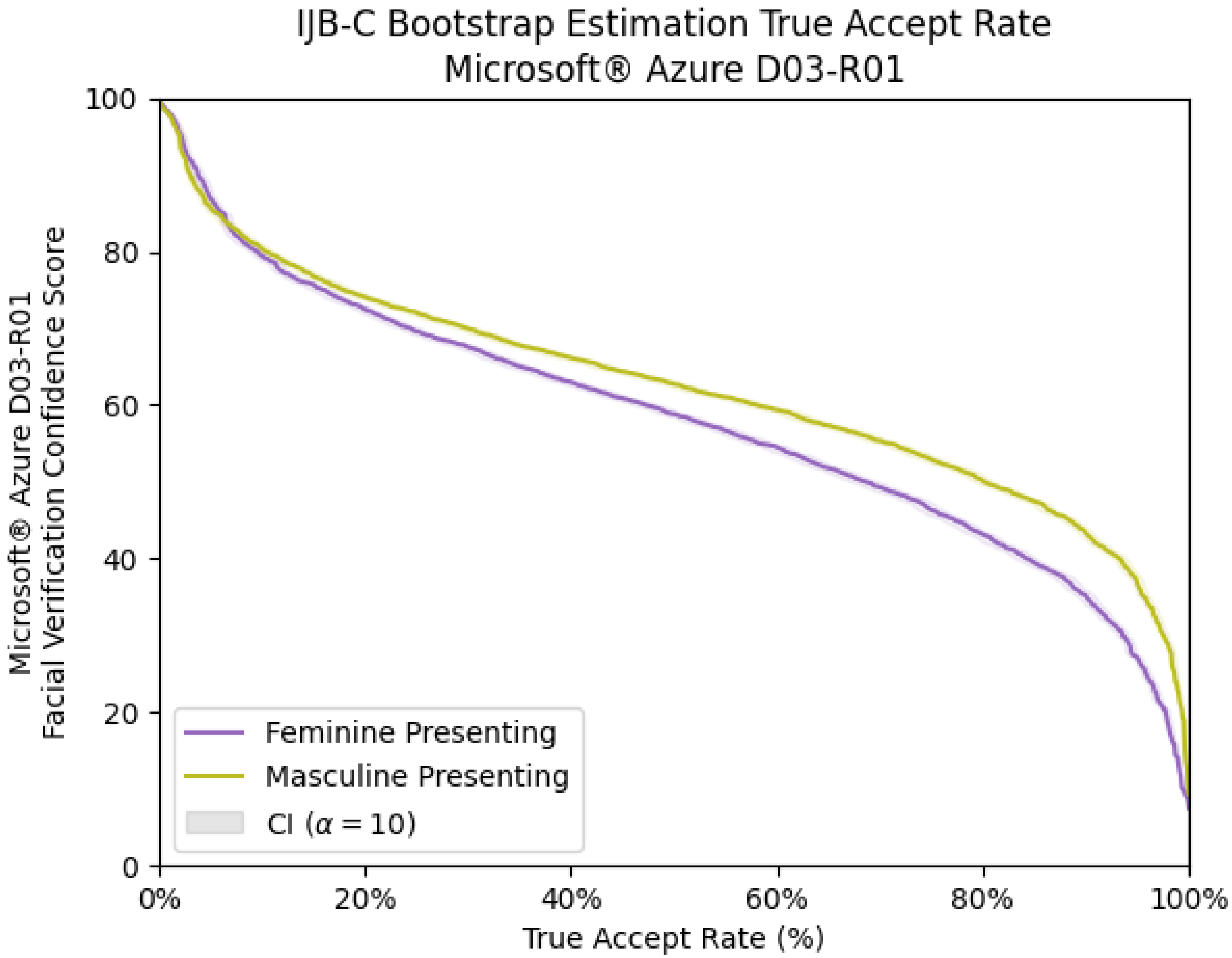


FIGURE 92 *The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under* Microsoft Face API *with a configuration of* detection_03 *and* recognition_01, *for the binary masculine or feminine gender presentation classification schema.*

Microsoft Face API with a configuration of detection_03 and recognition_02

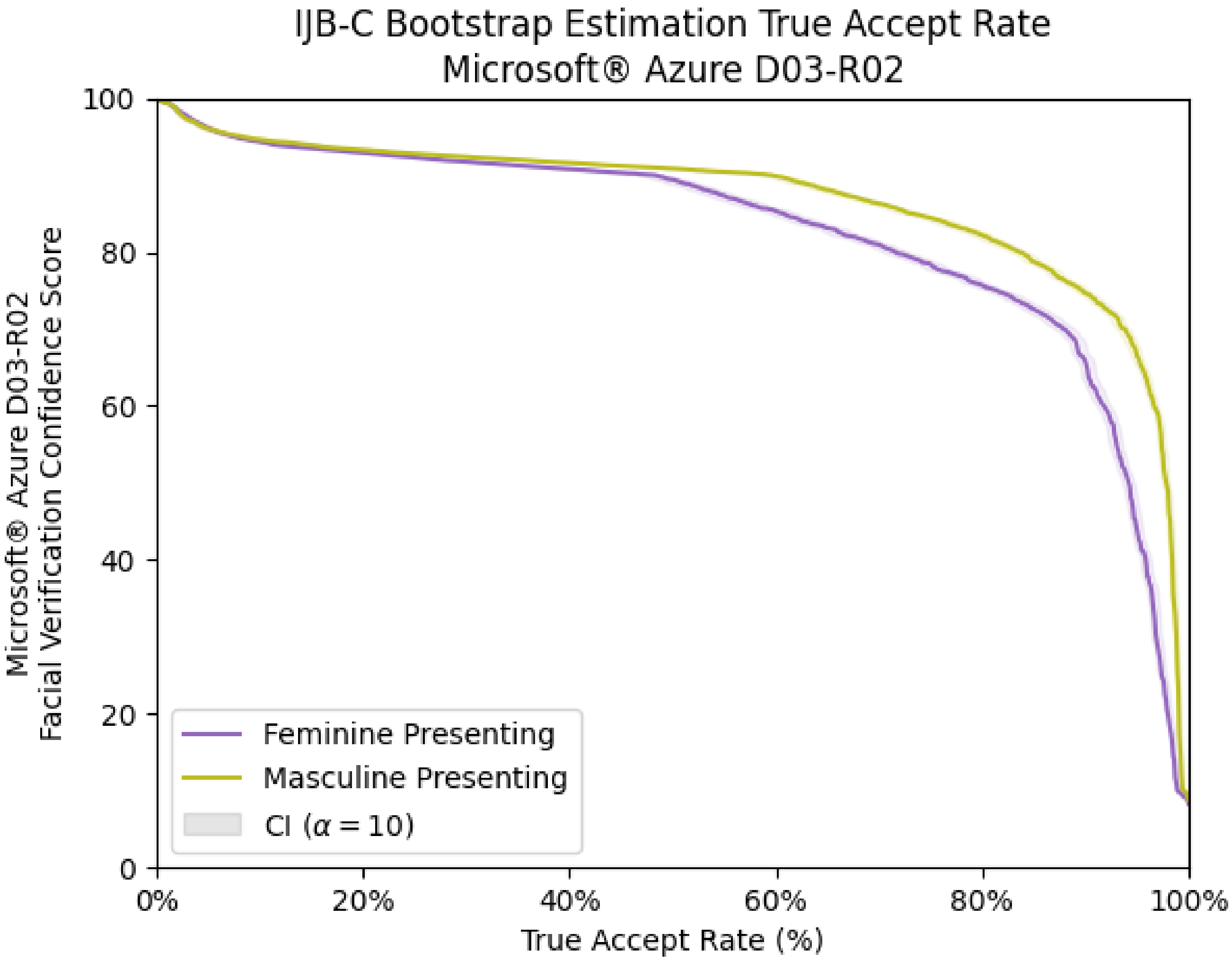
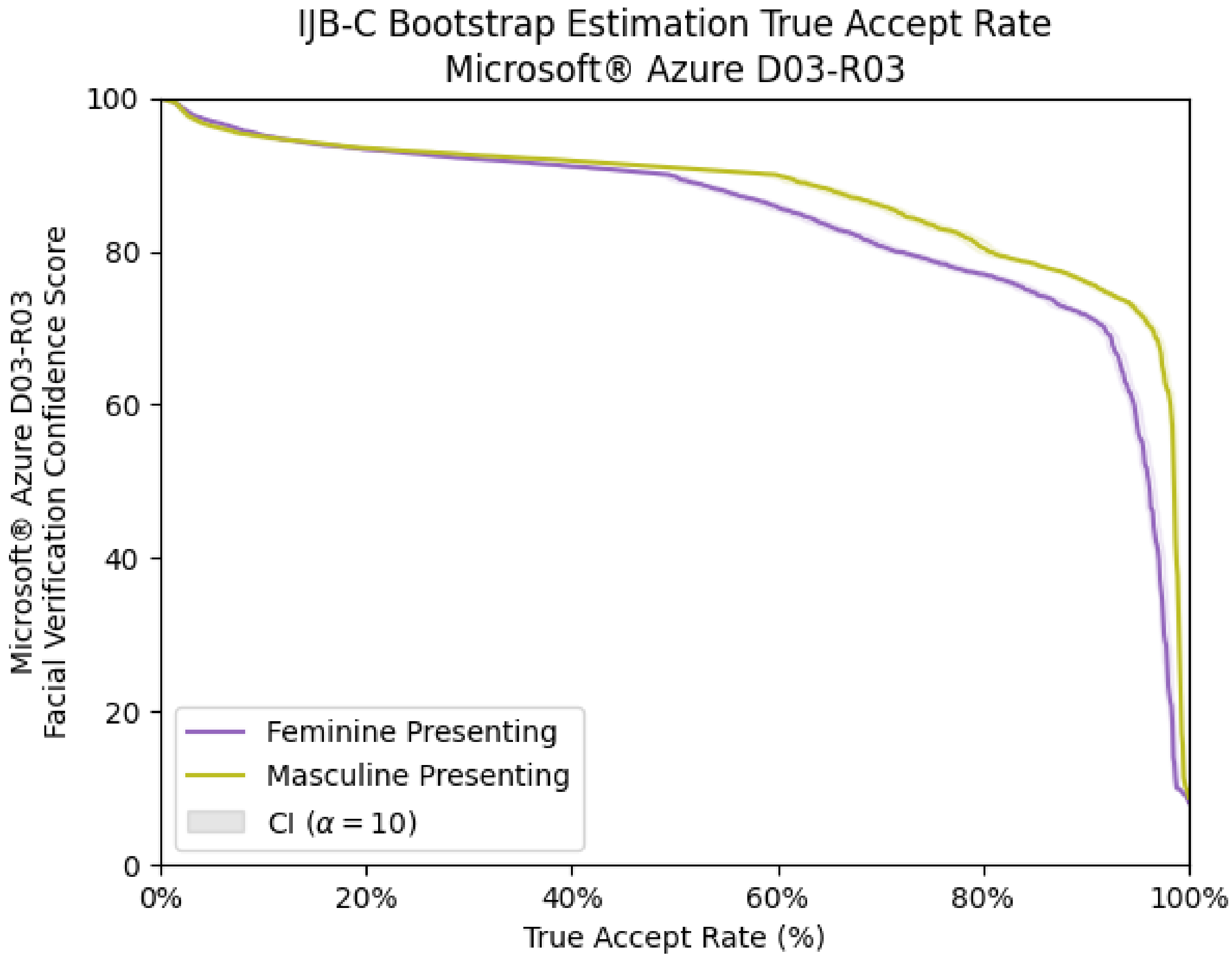


Figure 93 *The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under* Microsoft Face API *with a configuration of* detection_03 *and* recognition_02, *for the binary masculine or feminine gender presentation classification schema.*

Microsoft Face API with a configuration of detection_03 and recognition_03



*F*IGURE *94 The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under* Microsoft Face API *with a configuration of* detection_03 *and* recognition_03*, for the binary masculine or feminine gender presentation classification schema.*

Microsoft Face API, Released 2021

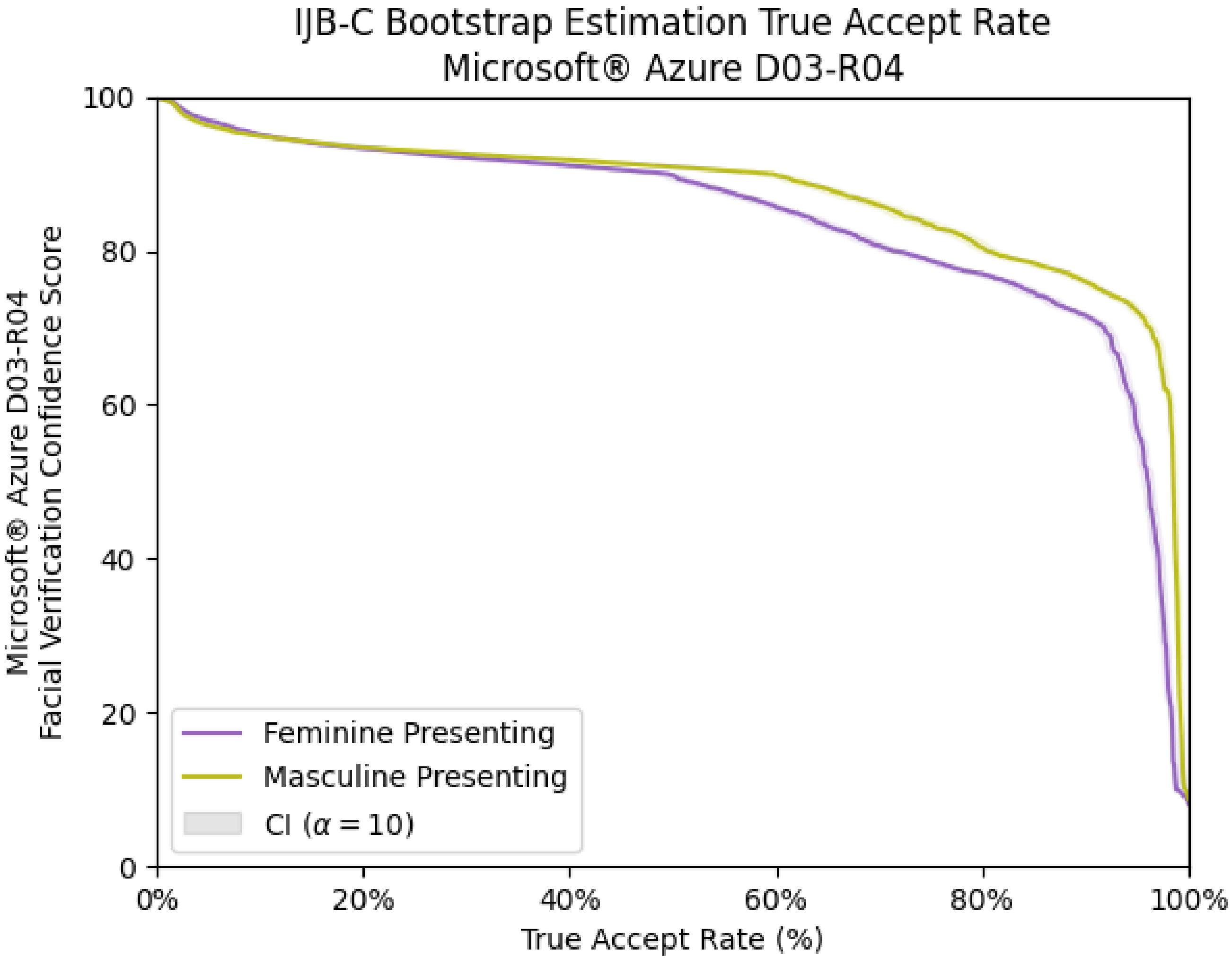


Figure 95 *The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under* Microsoft Face API *under its default configuration of* detection_03 *and* recognition_04*, released in 2021, for the binary masculine or feminine gender presentation classification schema.*

AWS Rekognition, As of Fall 2021

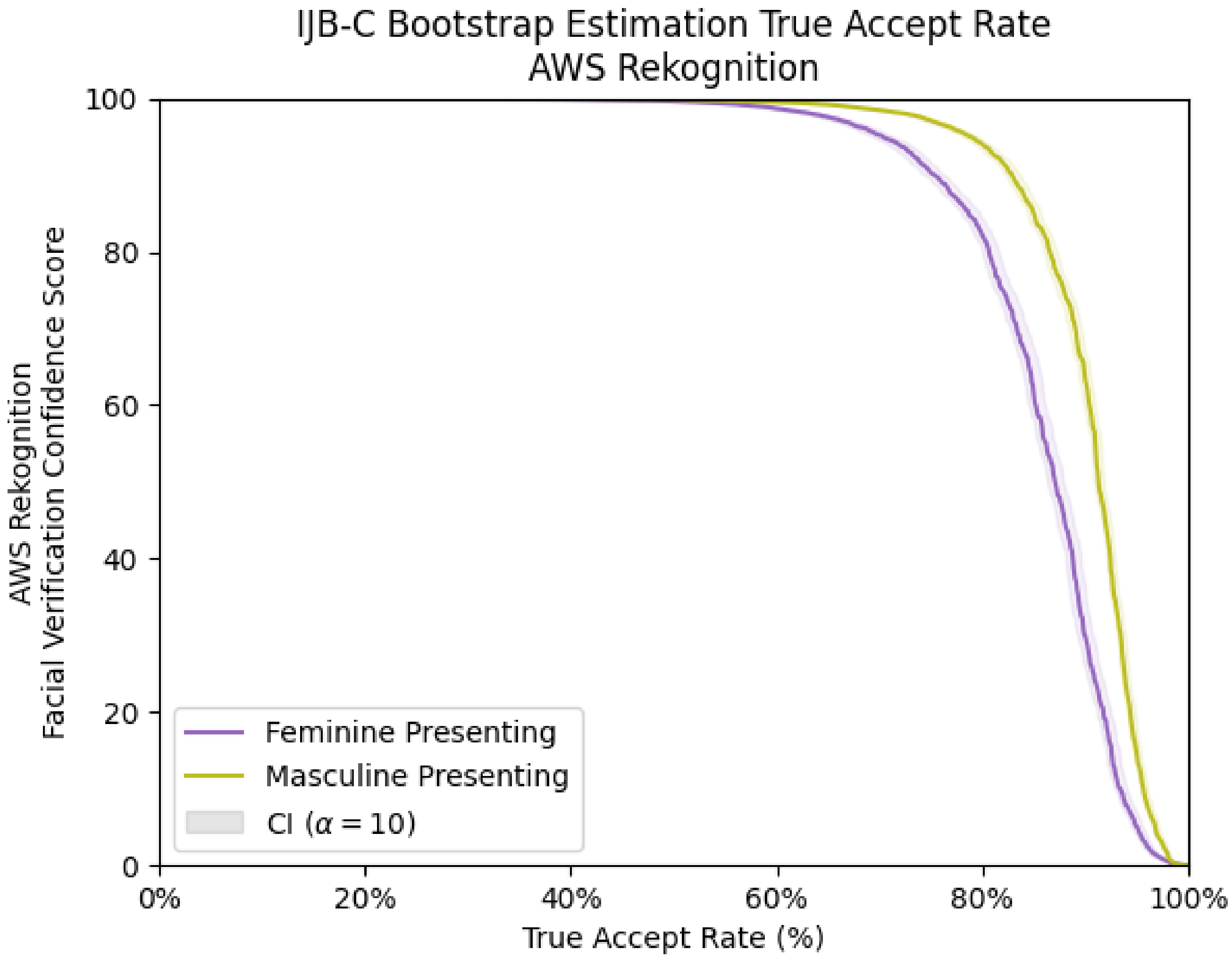
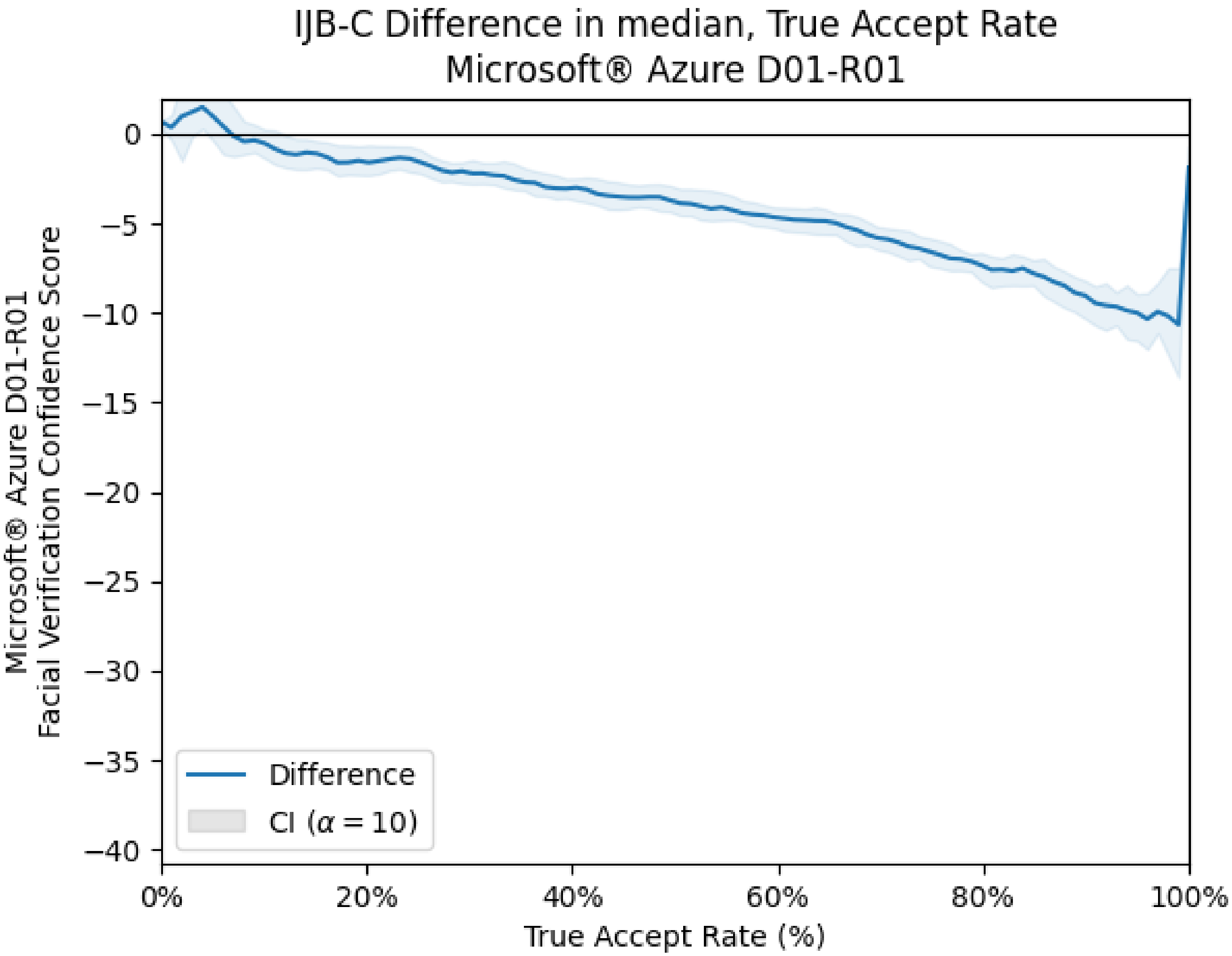


***FIGURE 96** The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under* AWS Rekognition *for the binary masculine or feminine gender presentation classification schema.*

Difference in Medians from Median True Accept Rate

These insights can be furthered, through calculation of a score detailing the interclass bias between the two gender presentation classifications. Following the procedure outline earlier, for each of the bootstrap samples generated, the difference between the two gender presentation classifications was measured. These difference in performance curves are shown Figures 44-44 to provide a clearer understanding of the interclass bias for the two gender presentation classifications.

Microsoft Face API, Released 2017



The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under Microsoft Face API *under its default configuration of* detection_01 *and* recognition_01, *released in 2017, for the binary masculine or feminine gender presentation classification schema.*

Microsoft Face API with a configuration of detection_01 and recognition_02

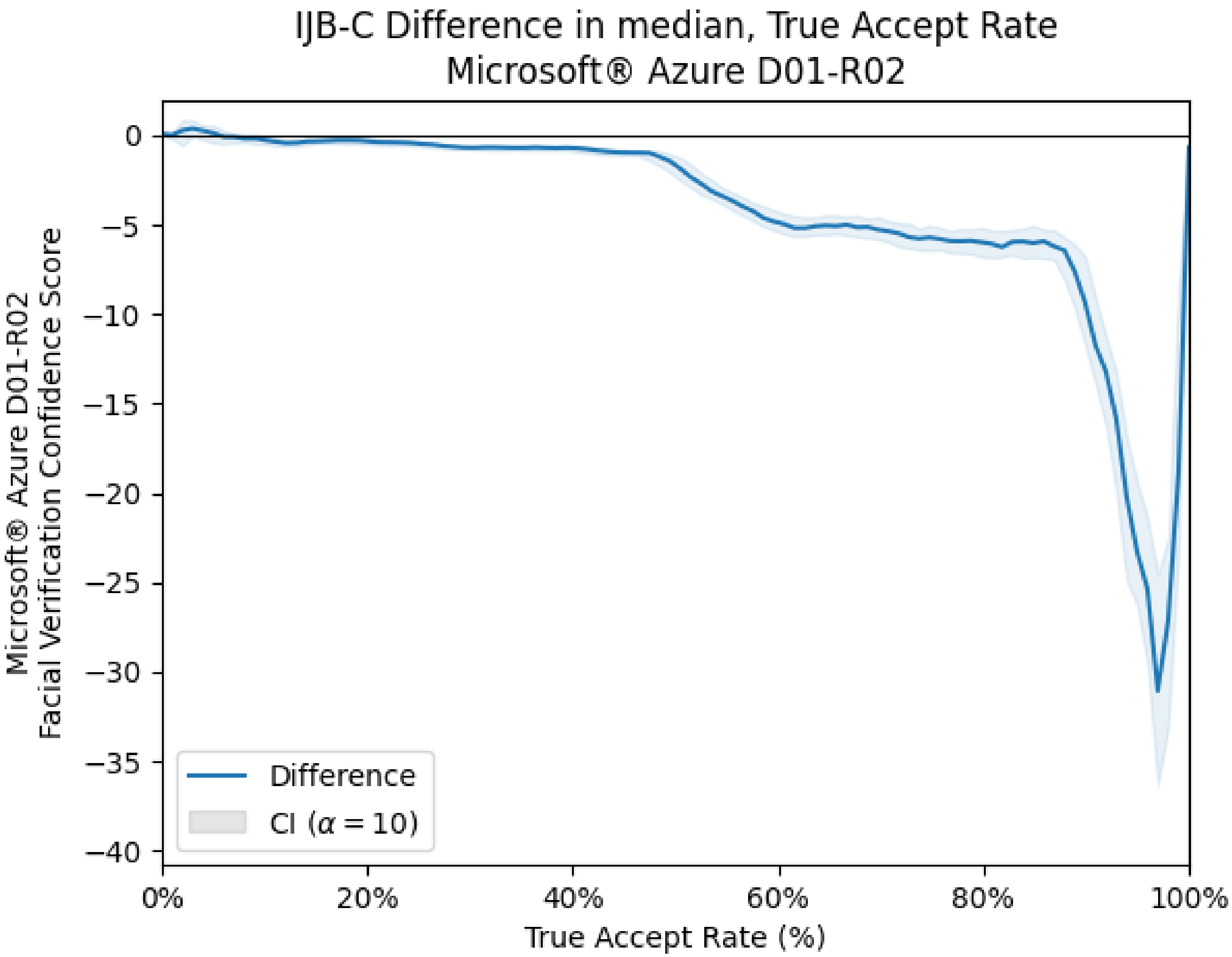


FIGURE 97 The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under Microsoft Face API *with a configuration of* detection_01 *and* recognition_02*, for the binary masculine or feminine gender presentation classification schema.*

Microsoft Face API with a configuration of detection_01 and recognition_03

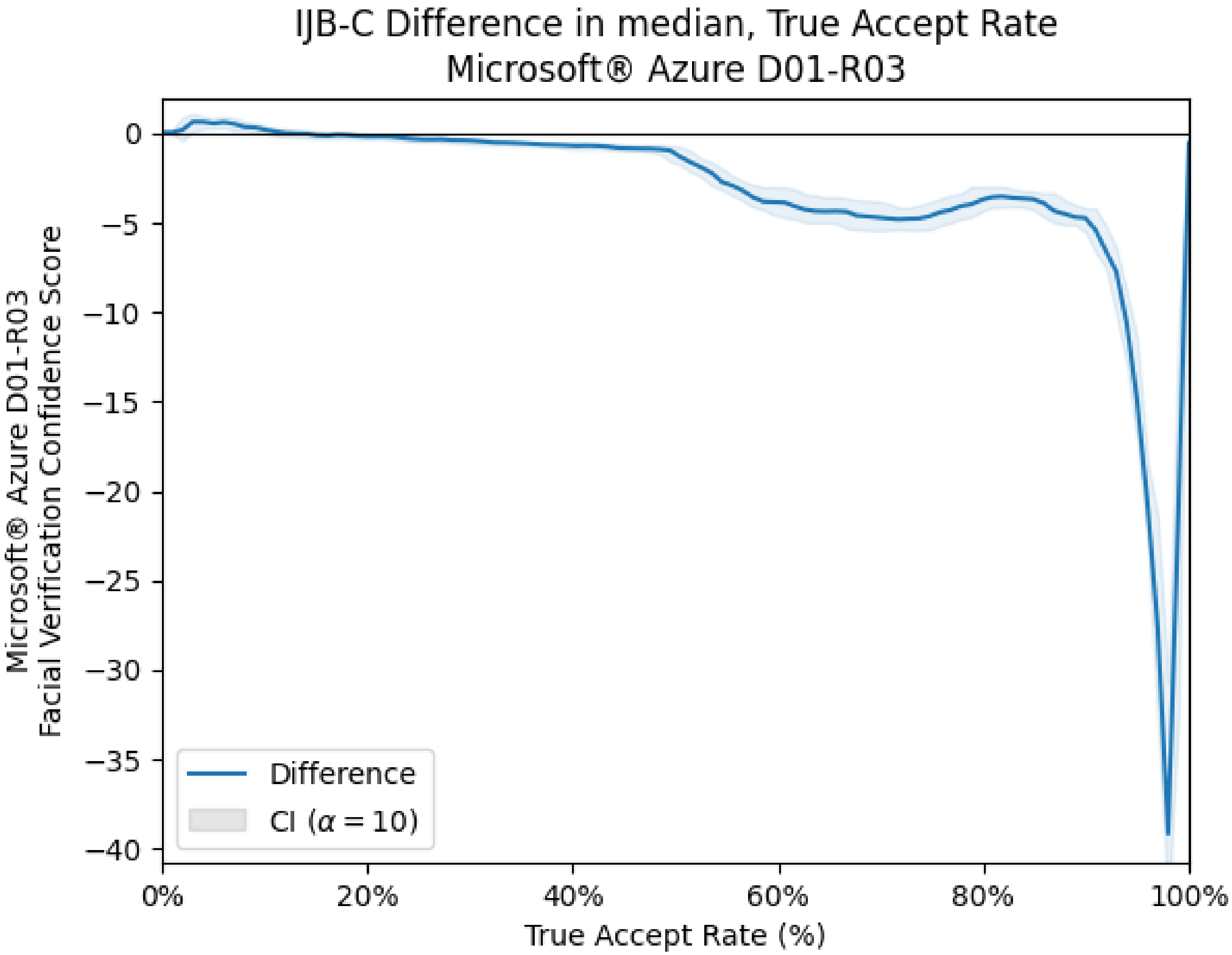


FIGURE 98 *The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under* Microsoft Face API *with a configuration of* detection_01 *and* recognition_03*, for the binary masculine or feminine gender presentation classification schema.*

Microsoft Face API with a configuration of detection_01 and recognition_04

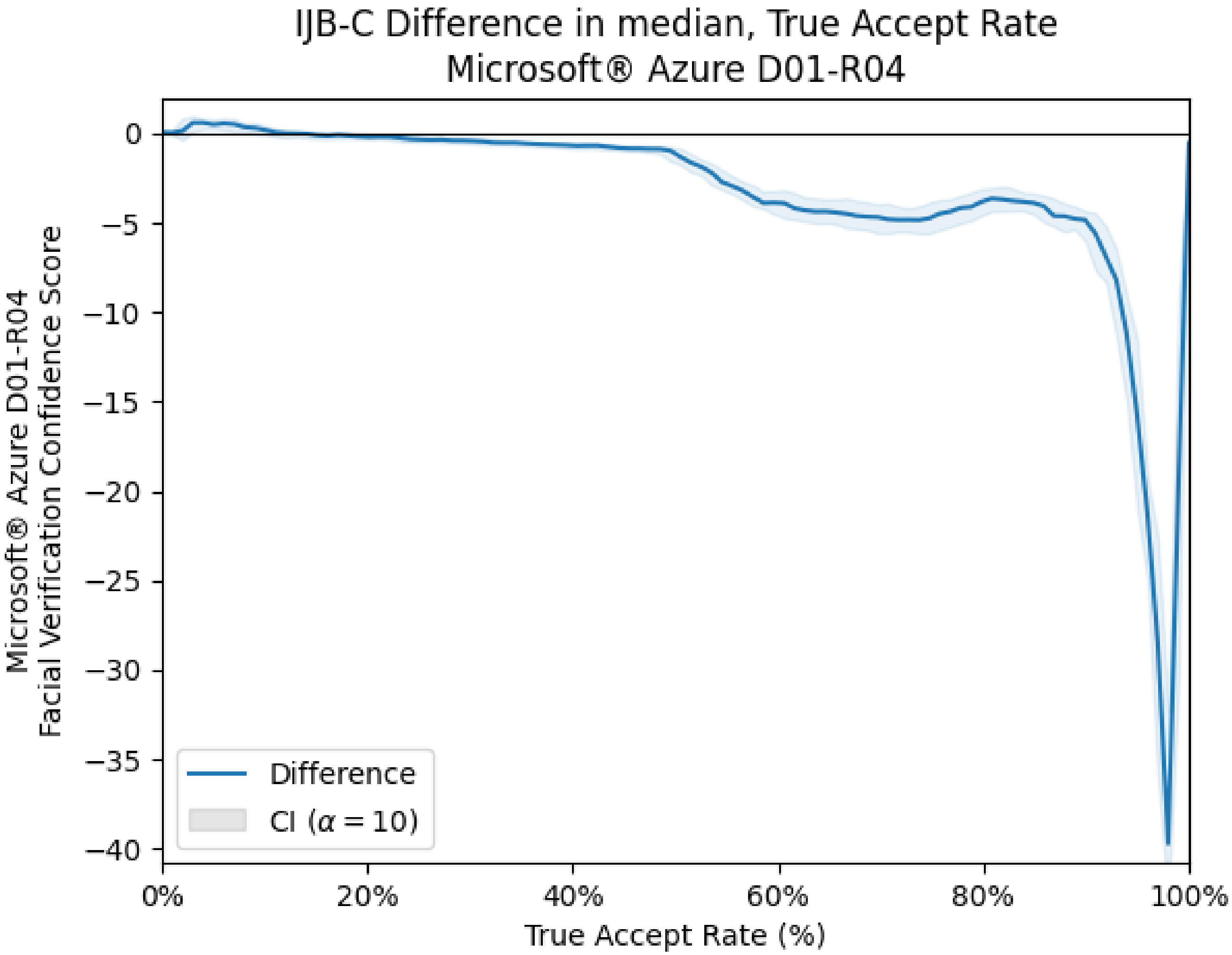


FIGURE 99 The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under Microsoft Face API *with a configuration of* detection_01 *and* recognition_04, *for the binary masculine or feminine gender presentation classification schema.*

Microsoft Face API with a configuration of detection_02 and recognition_01

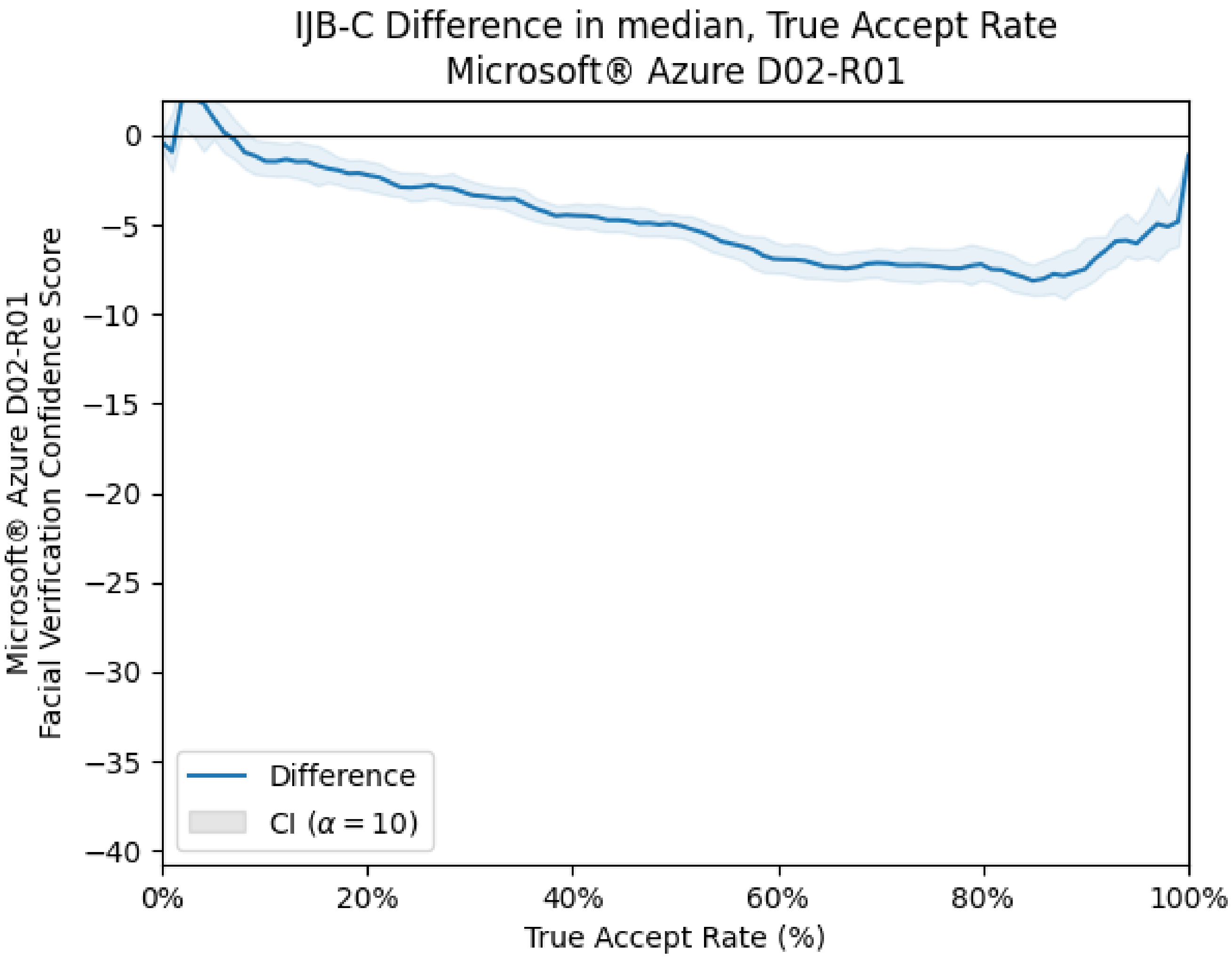


***Figure 100** The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under* Microsoft Face API *with a configuration of* detection_02 *and* recognition_01*, for the binary masculine or feminine gender presentation classification schema.*

Microsoft Face API, Released 2019

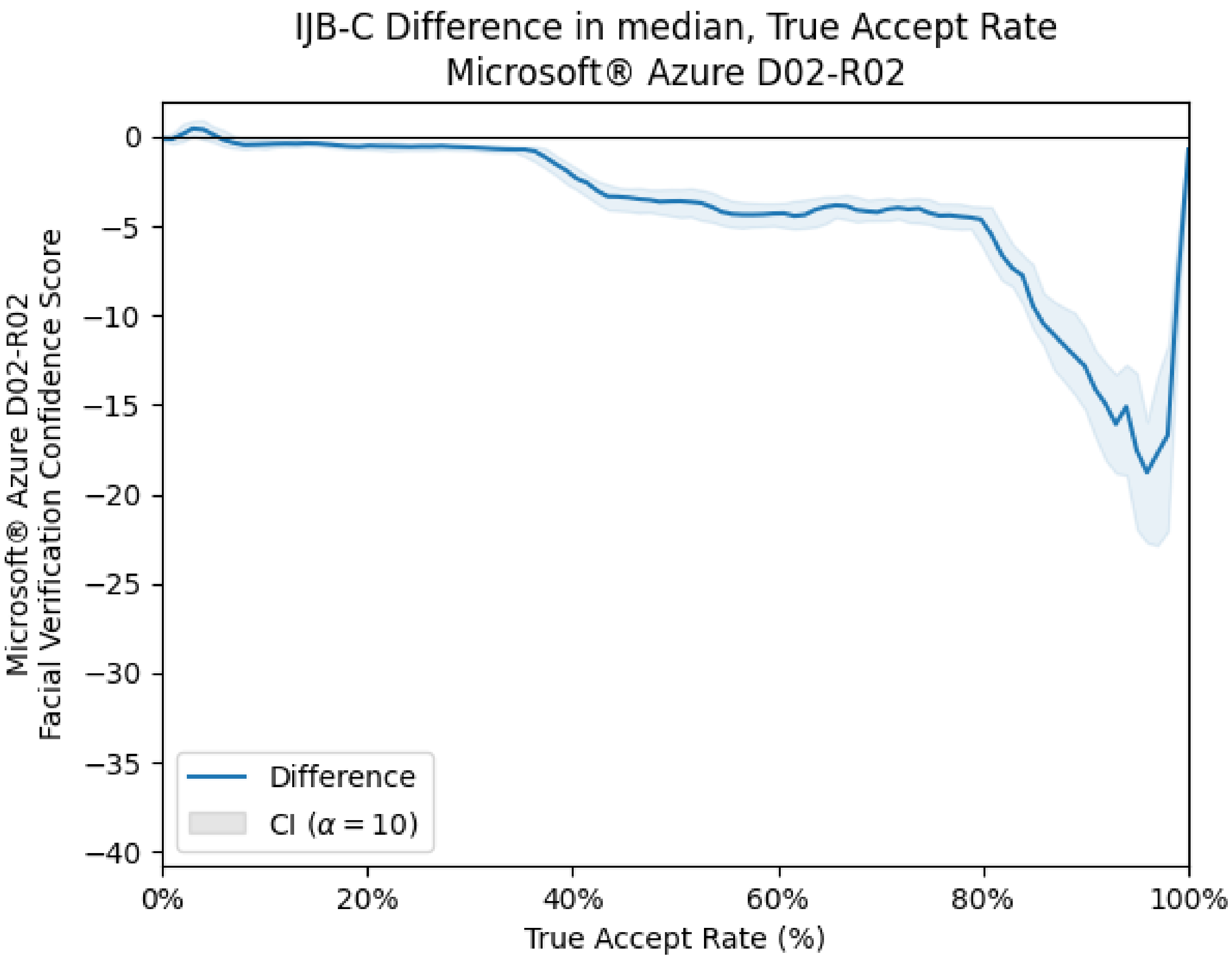


FIGURE 101 *The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under* Microsoft Face API *under its default configuration of* detection_02 *and* recognition_02, *released in 2019, for the binary masculine or feminine gender presentation classification schema.*

Microsoft Face API, Released 2020

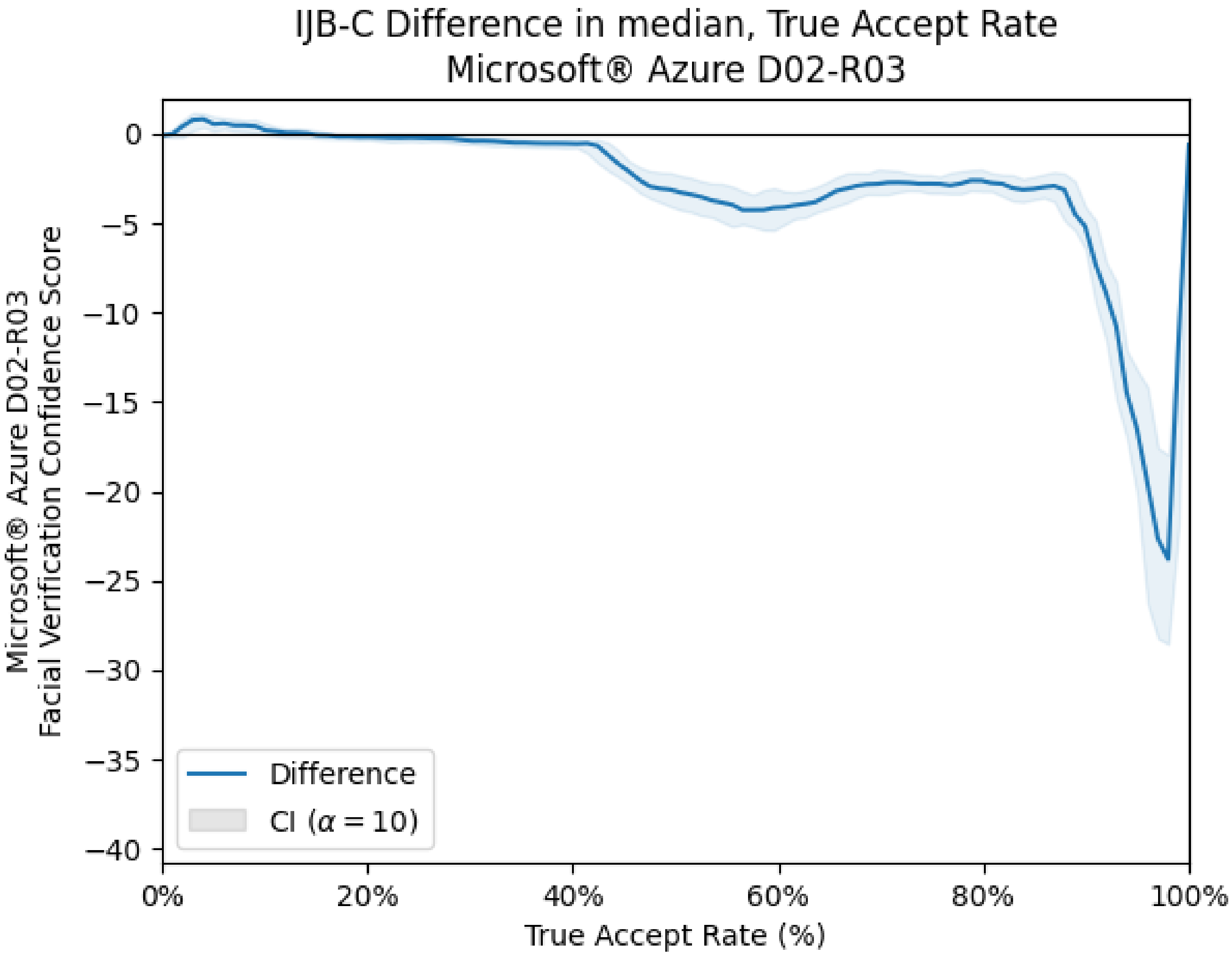


FIGURE 102 The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under Microsoft Face API *under its default configuration of* detection_02 *and* recognition_03*, released in 2020, for the binary masculine or feminine gender presentation classification schema.*

Microsoft Face API with a configuration of detection_02 and recognition_04

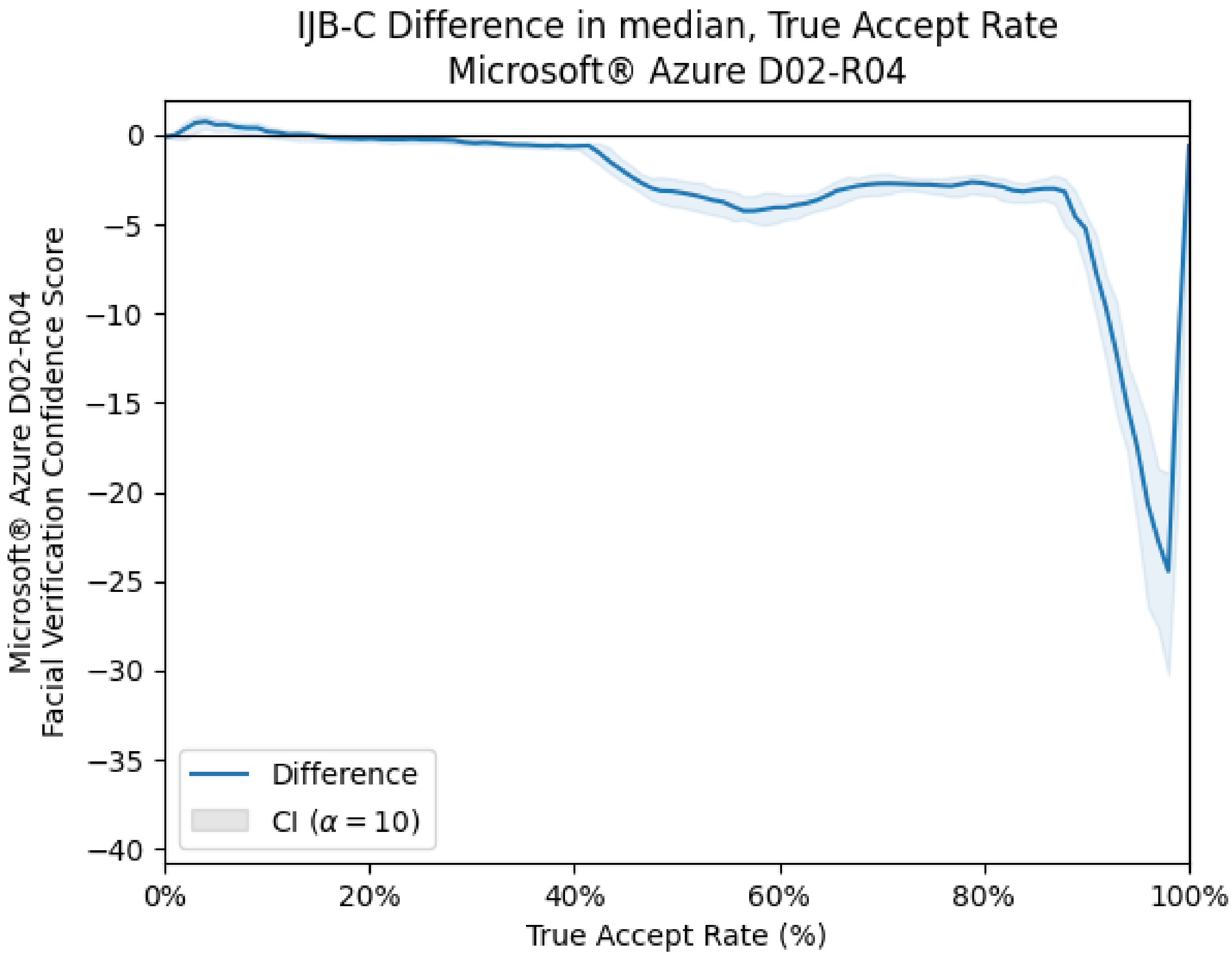


FIGURE 103 *The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under* Microsoft Face API *with a configuration of* detection_02 *and* recognition_04*, for the binary masculine or feminine gender presentation classification schema.*

Microsoft Face API with a configuration of detection_03 and recognition_01

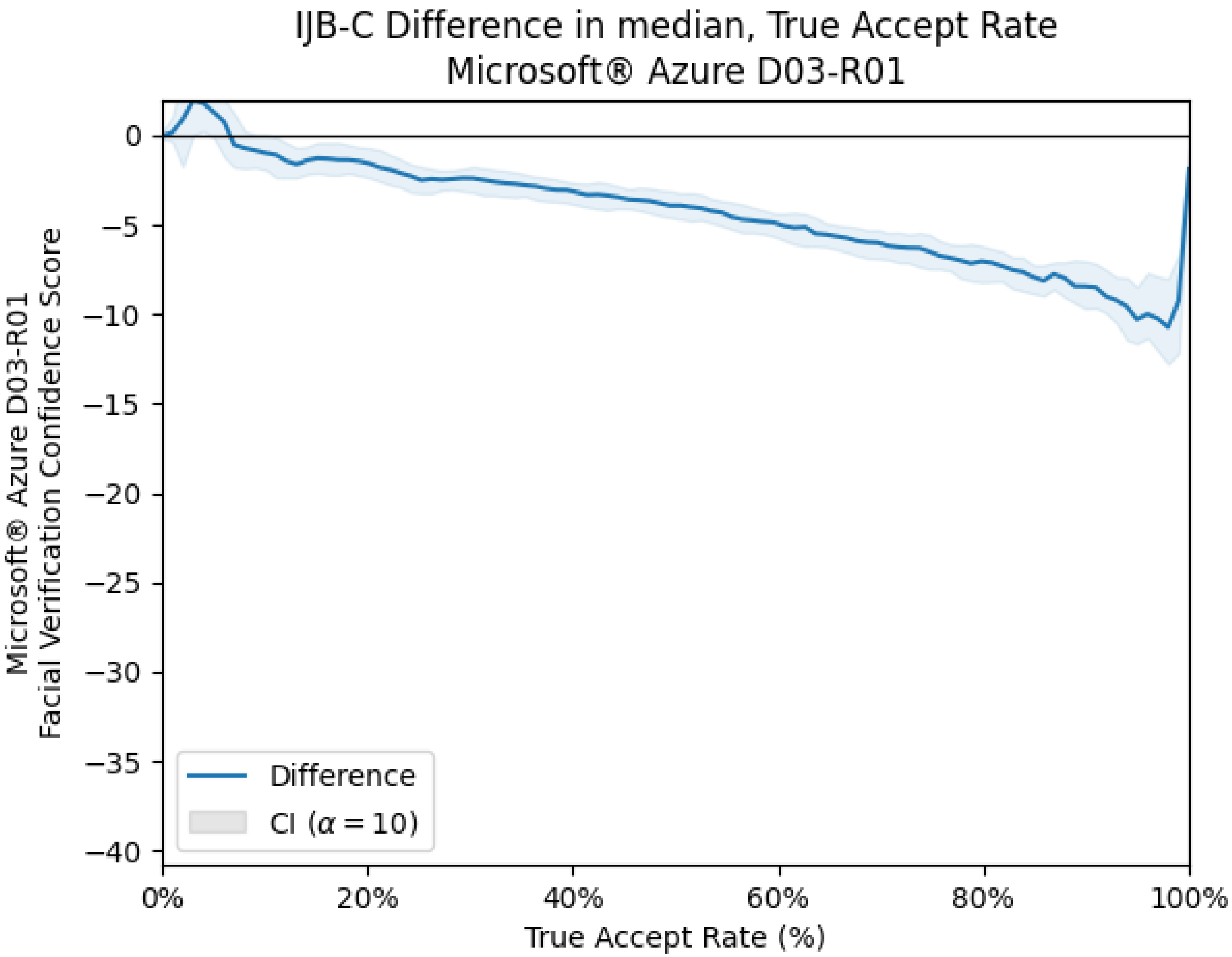


Figure 104 *The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under* Microsoft Face API *with a configuration of* detection_03 *and* recognition_01, *for the binary masculine or feminine gender presentation classification schema.*

Microsoft Face API with a configuration of detection_03 and recognition_02

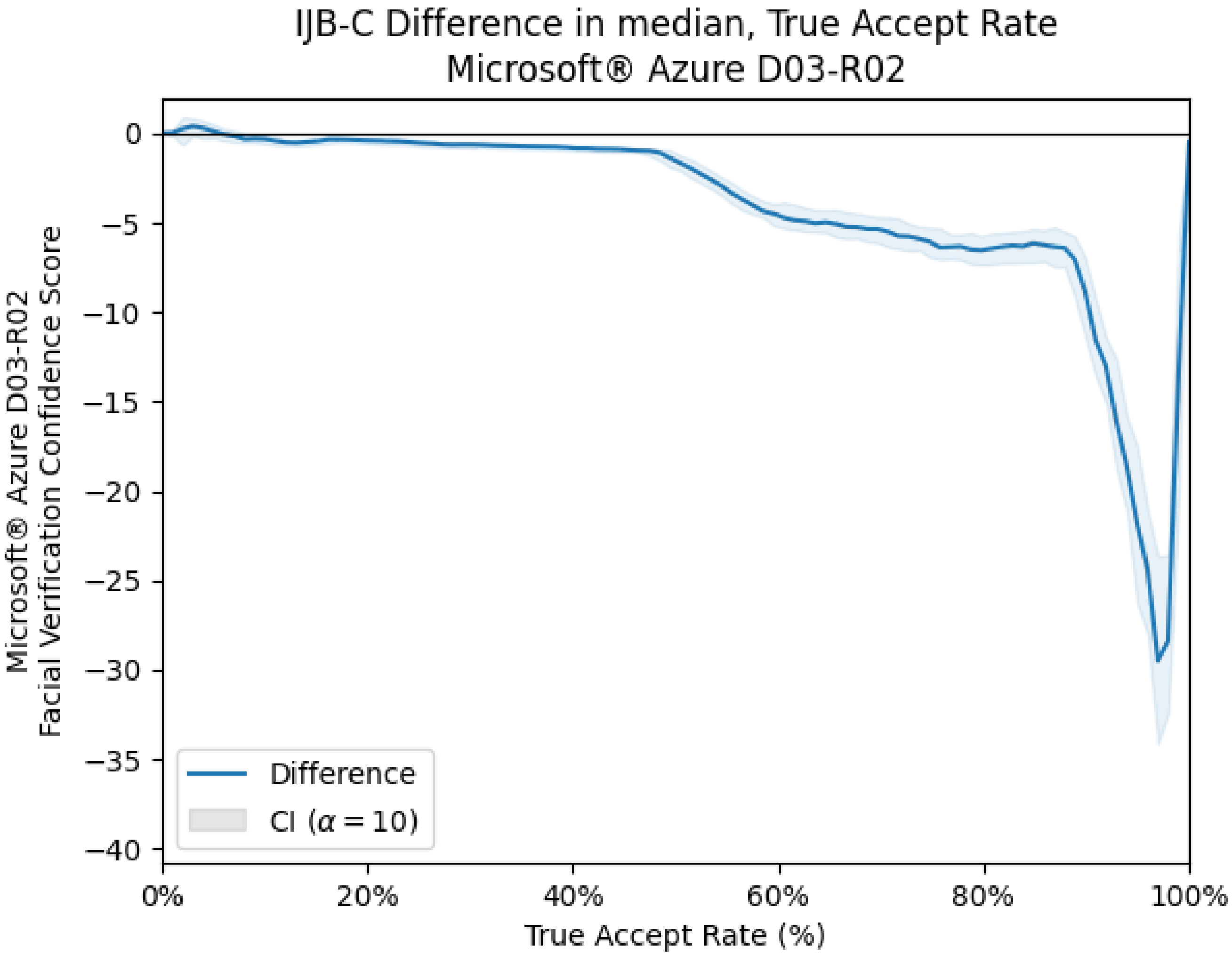


Figure 105 *The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under* Microsoft Face API *with a configuration of* detection_03 *and* recognition_02*, for the binary masculine or feminine gender presentation classification schema.*

Microsoft Face API with a configuration of detection_03 and recognition_03

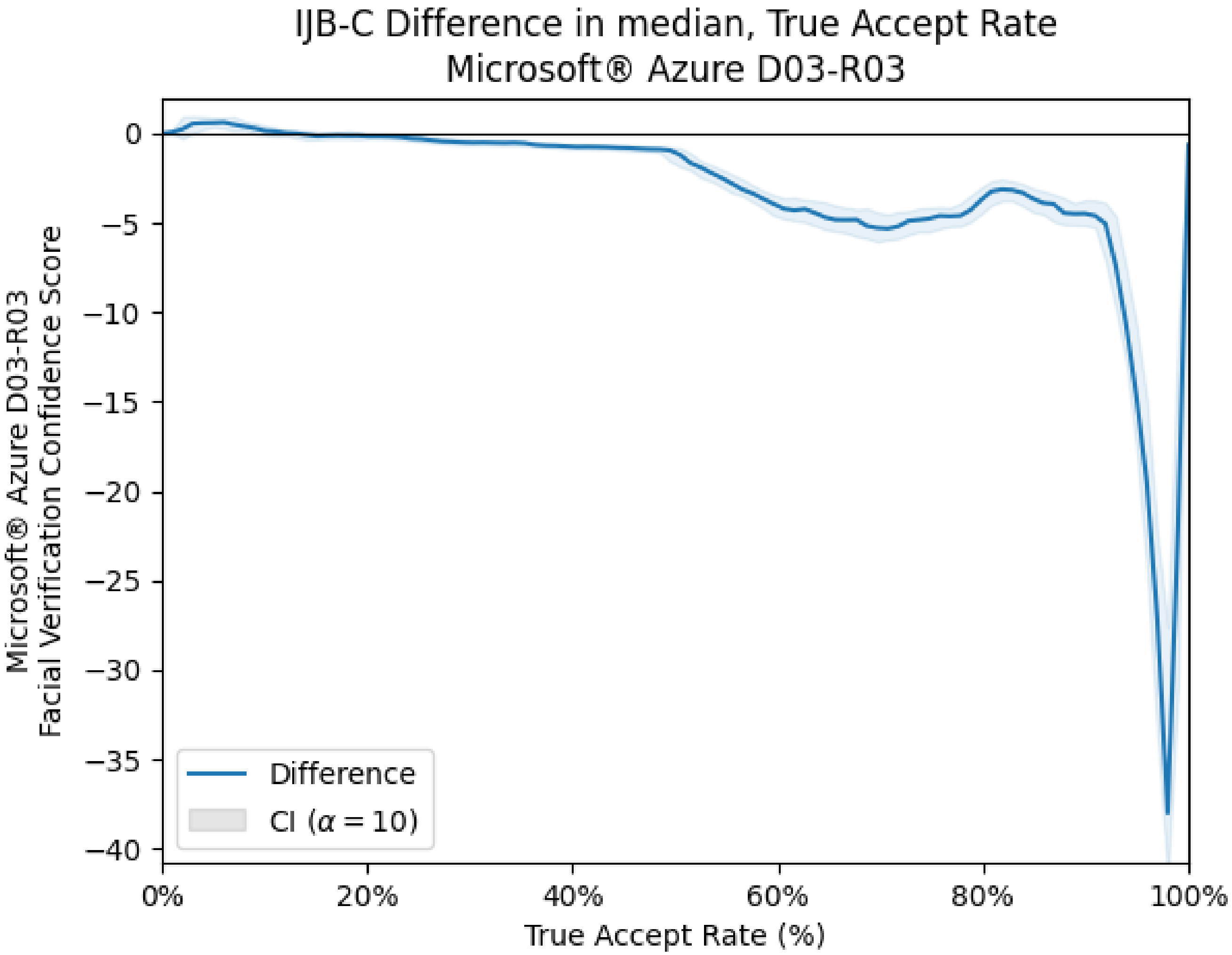


Figure 106 *The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under* Microsoft Face API *with a configuration of* detection_03 *and* recognition_03*, for the binary masculine or feminine gender presentation classification schema.*

Microsoft Face API, Released 2021

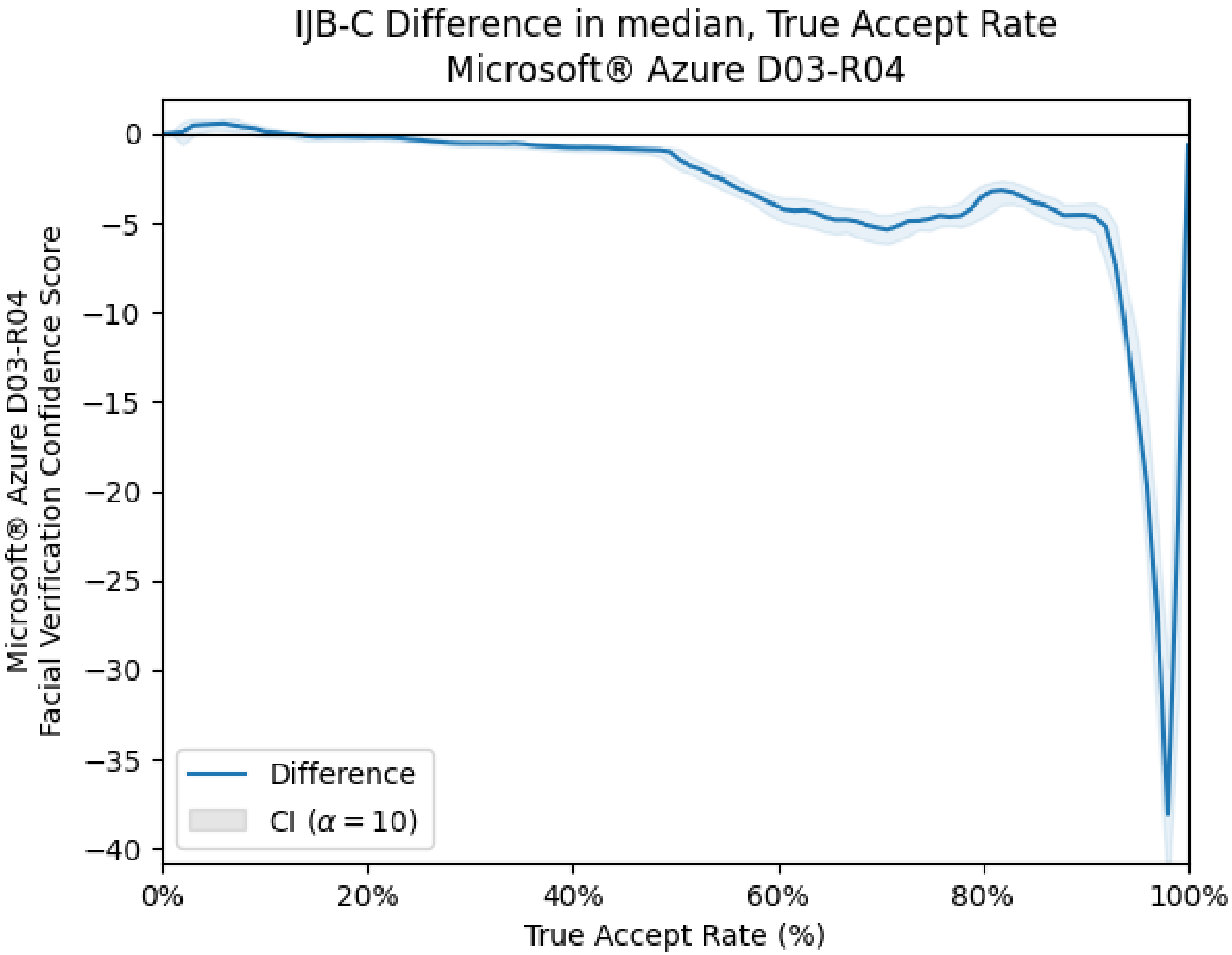


Figure 107 *The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under* Microsoft Face API *under its default configuration of* detection_03 *and* recognition_04*, released in 2021, for the binary masculine or feminine gender presentation classification schema.*

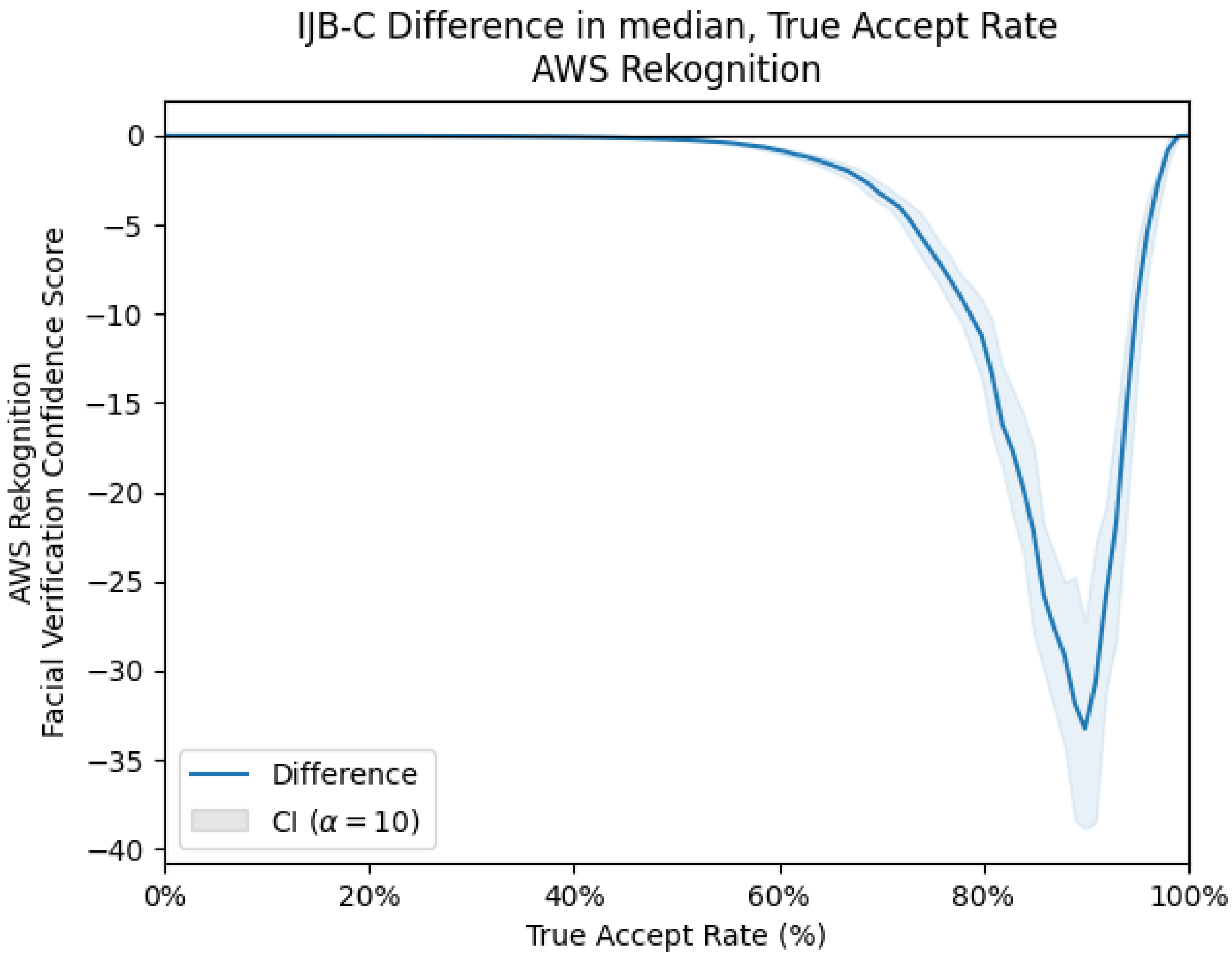


Figure 108 *The difference between the mean overall performance true accept rate (TAR) and each skin tone against the corresponding confidence score threshold under* AWS Rekognition *for the binary masculine or feminine gender presentation classification schema.*

Distance-based Variance Interclass Bias Metric

Subsequently it is important to transform these graphical understandings of interclass bias into the d_{l_2} measure of interclass bias between a class and the overall performance. Table 44-44 relay the maximum d_{l_2} measure for the two gender presentation classifications for the performance of *Microsoft Face API* under all of its possible configurations, and *AWS Rekognition*.

Table 10 *A matrix representing the maximum measure of interclass bias for the binary masculine or feminine gender presentation classification schema under Microsoft Face API with all possible configurations it's detection and recognition models.*

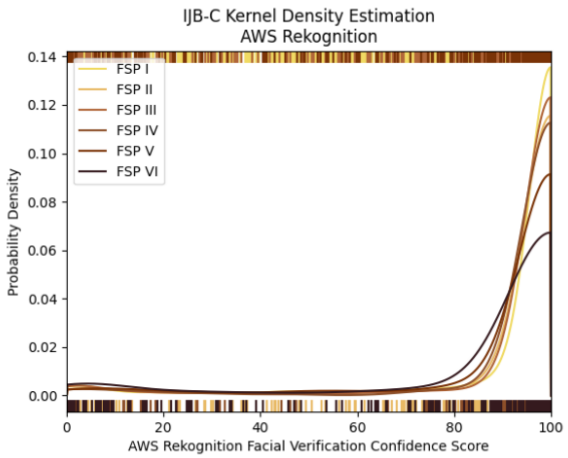
Microsoft Face API	*detection_01*	*detection_02*	*detection_03*
recognition_01	5.048655	5.551684	5.242811
recognition_02	5.021978	5.073328	4.925848
recognition_03	3.969225	3.881258	3.975921
recognition_04	4.132754	3.896087	4.037939

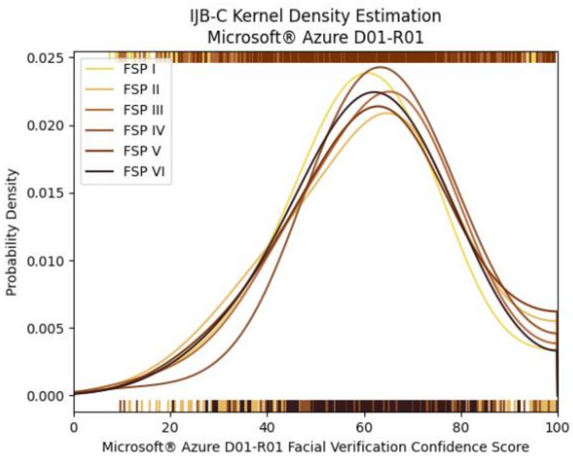
Table 11 *The 90% two-sided confidence intervals ($\alpha = 10\%$) for the maximum measure of interclass bias for the binary masculine or feminine gender presentation classification schema under* Microsoft Face API *under all of its possible configurations, and* AWS Rekognition*. The table encodes the configuration of the* Microsoft Face API *as* DXX-RYY *where* XX *represents the detection model* detection_XX *and* YY *represents the recognition model* recognition_YY*. For example,* D01-R02 *would indicate* Microsoft Face API *with a configuration of* detection_01 *and* recognition_02*.*

Class	***Lower Bound***	***Upper Bound***
Microsoft Face API *D01-D02*	3.679845	5.048655
Microsoft Face API *D01-D02*	3.524707	5.021978
Microsoft Face API *D01-D03*	2.659155	3.969225
Microsoft Face API *D01-D04*	2.724611	4.132754
Microsoft Face API *D02-D01*	2.523772	3.896087
Microsoft Face API *D02-D02*	2.343196	3.881258
Microsoft Face API *D02-D03*	3.369334	5.073328
Microsoft Face API *D02-D04*	4.050693	5.551684
Microsoft Face API *D03-D01*	3.731456	5.242811
Microsoft Face API *D03-D02*	3.440768	4.925848
Microsoft Face API *D03-D03*	2.632185	3.975921
Microsoft Face API *D03-D04*	2.674466	4.037939
AWS Rekognition	3.555945	5.454197

Discussion

Before exploring the proposed novel metric for interclass bias in a classification schema, the author validates an assumption previously stated in the methodology to ensure that subsequent analysis holds. The confidence scores, reported from the commercial facial verification algorithms, were partitioned into six skin tone classifications as defined by the IJB-C skin tone classification schema. The author then extracted a probabilistic density function from the partitioned confidence scores. The proposed methodology ignores parametric approaches, as they require assumptions on the underlying distribution of the data or the prior model and an incorrect model can greatly affect the predictive power. Instead, the author shifted their attention to nonparametric methods, which avoid the need to make *a priori* assumptions on the sample distribution. One such method is the kernel density estimation ("KDE") method is able to directly estimate the probability density function which simplifies further analysis. A critical parameter in any kernel-based estimator is the search bandwidth. A small bandwidth will detect a density surface with small, spiky event hotspots, while larger bandwidths return density surfaces with smoother and larger event clusters. Various methods have been developed to aid the selection of an appropriate bandwidth, such as the rule-of-thumb (Silverman 1986) and plug-in (Scott 2014). It is widely accepted in the literature that the choice of bandwidth is more important than the choice of kernel; as such the author used the *Improved Sheather-Jones algorithm*, an improvement to the Scott (2014) plug-in selector, to select the optimal bandwidth (Botev et al. 2010). This method was selected as this bandwidth selection algorithm performs better for data that is far from normal, or multimodal, which holds true for the reported confidence scores as measured by a combined omnibus test of normality (D'Agostino and Pearson 1973; Oliphant 2007).





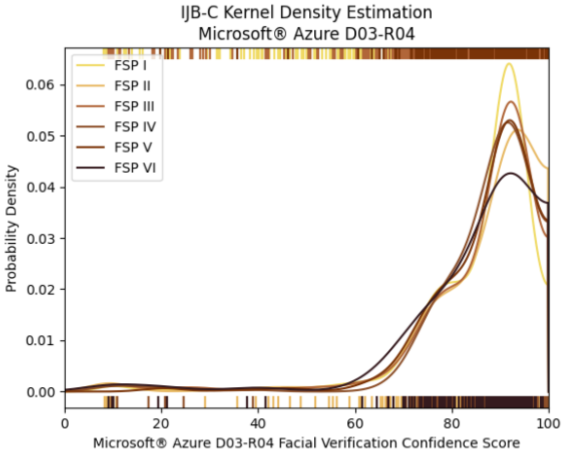


***Figure 109(a, b, c)** A kernel density estimation of the probability density function for various commercial facial verification algorithms confidence scores for six skin tone classifications: (a)* AWS Rekognition *(b)* Microsoft Face API *under its default configuration of* detection_01 *and* recognition_01, *released in 2017, and (c) its latest released configurations* detection_03 *and* recognition_04, *released in 2021; using the IJB-C skin tone classification schema.*

FIGURE 109(A, B, C), above, show the kernel density estimates for the performance of *Microsoft Face API*, under its default configuration of detection_01 and recognition_01, and its latest released configurations detection_03 and recognition_04; and *AWS Rekognition* for each of the six skin tone classifications. In **FIGURE 109(A)** one can see how the confidence scores for *AWS Rekognition* distribution of each of the six skin tones classifications is highly skewed near the recommended threshold of 99%. Furthermore, we observe that the darker skin tone (i.e., **Skin Tone IV** (Medium Yellow / Brown), **Skin Tone V** (Medium-Dark Brown), and **Skin Tone VI** (Dark Brown)) the likelihood of obtaining a high confidence score, above the recommended threshold, decreases. In contrast, the *Microsoft Face API*, under its default configuration, in **FIGURE 109(b)**, the confidence scores for the distribution of the six skin tones classifications are not only less skewed towards the top of the confidence score values, but also appears more symmetric and normal; with the notable exception of the boundary conditions imposed at 0 and 100. Additionally, under this facial verification algorithm the different skin tone classifications are tightly clustered and there is no clear skin tone classification that has dominate the likelihood of obtaining a high confidence score. In **FIGURE 109(C)** the latest released configuration of the *Microsoft Face API* has improved its algorithm as the confidence scores are more skewed than the default configuration and is tightly clustered near the upper limit of the confidence scores. Taken together these kernel density estimations of the probability density function for the facial verification algorithms confidence scores take many forms due to both the design of the algorithm but also due to the bounds of the confidence scores themselves. This supports the need to use analysis methods, as described in the methodology above, that do not require any model knowledge or assumptions about the underlying distribution.

The Investigative Power of the True Accept Rate

IJB-C Six Skin Tone Audit of Commercial Facial Verification Algorithms

We therefore turn our attention to computing confidence bounds for the true accept rate (“TAR”), calculated from the fraction of genuine comparisons that correctly exceed the threshold, for each class in the classification schema using the bootstrap method.

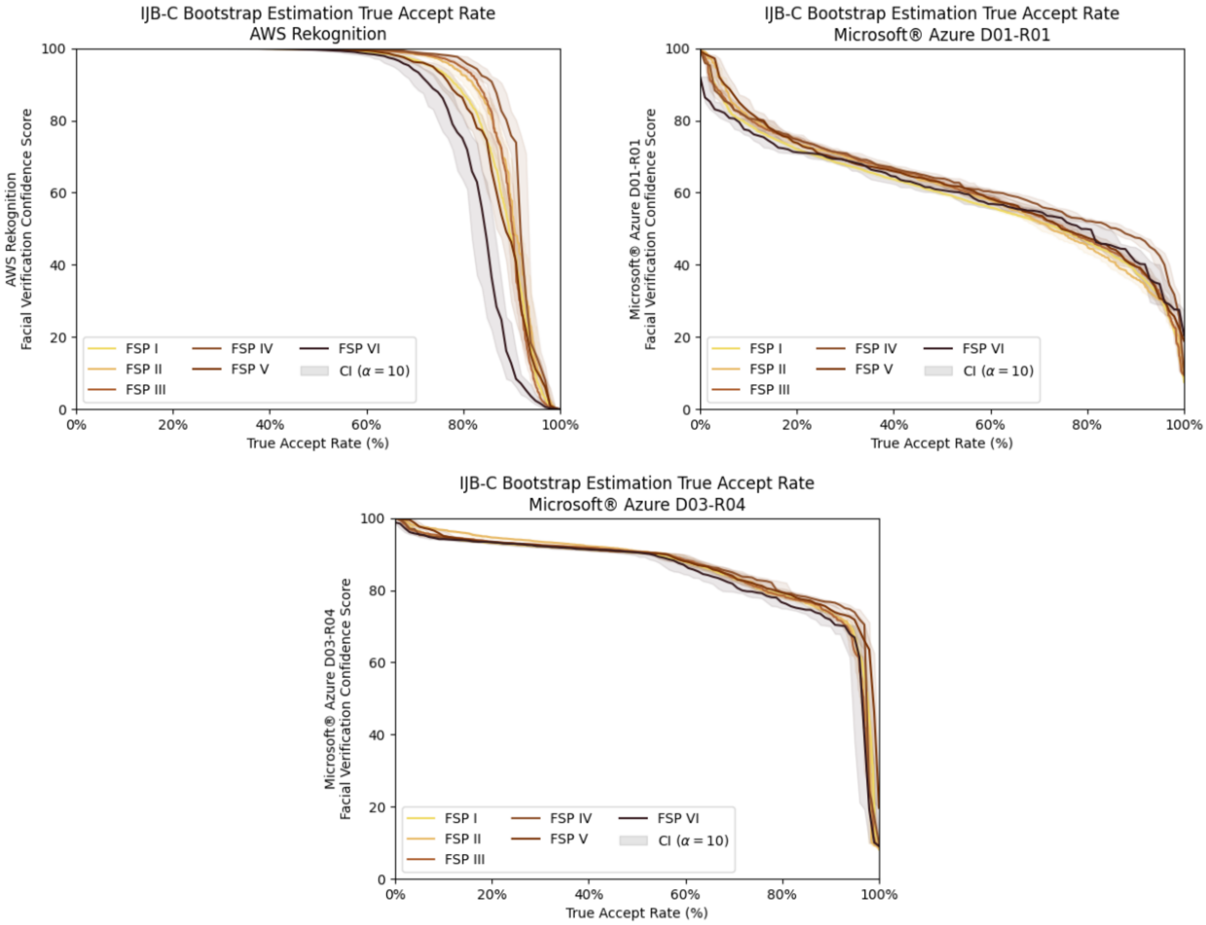


FIGURE 110(A, B, C) *The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under various commercial facial verification algorithms: (a)* AWS Rekognition *(b)* Microsoft Face API *under its default configuration of* detection_01 *and* recognition_01, *released in 2017, and (c) its latest released configurations* detection_03 *and* recognition_04, *released in 2021; using the IJB-C skin tone classification schema.*

As mentioned in ***Findings***, the two-sided confidence interval ($\alpha = 10\%$) for the TAR for each class in the classification schema is evaluated using the bootstrap method. Using these resampled confidence scores, the threshold is varied across the entire domain of the confidence scores to plot the TAR and the confidence bounds for the facial verification algorithms performance for each of the six skin tone classifications.

Figure 110**(a, b, c)** on page 138, shows the TAR and confidence bounds for the performance of *Microsoft Face API* under its default configuration, and its latest released, and *AWS Rekognition* for each of the six skin tone classifications. Taken all together, the six curves represent the facial verification algorithms TAR for individuals across of all skin tone classifications (as classified under the IJB-C schema). The TAR curves are more illuminating than the density estimates and communicate the differing performance between classes than the skewed nature of the probability density function estimates in **FIGURE 109(A, B, C)**. In **FIGURE 110(A)**, it is evident that there is considerable overlap between all six skin tone classifications when the *AWS Rekognition* facial verification algorithm is extremely confident (i.e., confidence scores greater than 99%) but as the algorithm is less confident there is a large spread across skin tone classifications. In particular, **Skin Tone V** (Medium-Dark Brown) and **Skin Tone VI** (Dark Brown) exhibit a significant departure from the other classes *AWS Rekognition* performance. In contrast the *Microsoft Face API*, under its default configuration, is more tightly clustered, shown in **FIGURE 110(B)**, with the confidence bounds for each of the skin tone classifications overlapping with another class for most of the bounds of the threshold. Notably, the darkest skin tone classification **Skin Tone VI** (Dark Brown) does not reach the same maximum confidence score as the other classes, and maintains lower confidence scores than other classes for the first quartile of the TAR. Additionally, it is interesting that in this facial verification algorithm the second darkest skin tone **Skin Tone V** (Medium-Dark Brown) maintains higher confidence scores than the other classes across the entire range. This is surprising as many other examples of bias in the larger space of facial recognition systems show lighter skin tones performing better, so seeing one of the darkest skin tone classes outperforming others is unexpected. That is to say, it is surprising to see a bias towards darker skinned individuals. Lastly in **FIGURE 110(C)** the latest released configuration of the *Microsoft Face API* shows significant improvement in confidence scores reported, but more importantly it is evident that this iteration of the algorithm the bias has been eliminated. All classes in the schema are tightly following each other, in some areas the confidence bands both narrow and overlap for most confidence score values. This visual representation of the interclass bias between classes in the six skin tone classification has provided additional detail to the biases exhibited by the various facial verification algorithms, and provided new insights not visible in a probability density function, further emphasizing the investigative power of the true accept rate and the novel methodology for measuring the interclass bias in a classification schema.

IJB-C Gender Presentation Audit of Commercial Facial Verification Algorithms

Following the same methodology, outlined in ***Findings*** and used in the six skin tone audit, the two-sided confidence interval ($\alpha = 10\%$) for the TAR for each of the two gender presentation classes is evaluated using the bootstrap method. Using these resampled confidence scores, the threshold is varied across the entire domain of the confidence scores to plot the TAR and the confidence bounds for the facial verification algorithms performance for both of the gender presentation classifications.

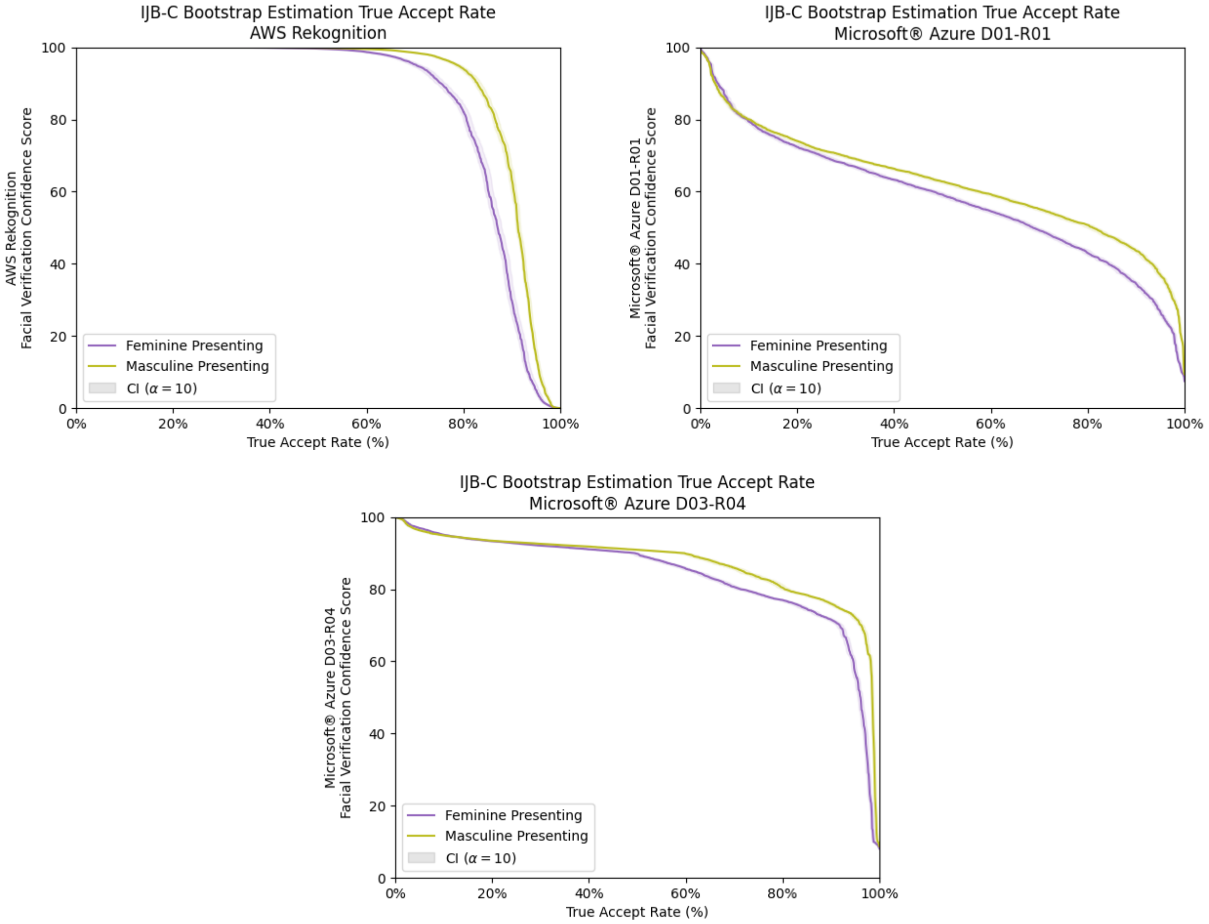


FIGURE 111(A, B, C) The true accept rate (TAR) and the corresponding confidence score threshold to achieve it under various commercial facial verification algorithms: (a) AWS Rekognition *(b)* Microsoft Face API *under its default configuration of* detection_01 *and* recognition_01*, released in 2017, and (c) its latest released configurations* detection_03 *and* recognition_04*, released in 2021; using the IJB-C gender presentation classification schema.*

FIGURE 111(A, B, C) on page 138, shows the TAR and confidence bounds for the performance of *Microsoft Face API* under its default configuration, and its latest released, and *AWS Rekognition* for both masculine and feminine gender presentation classifications. The most prominent feature in these commercial facial verification algorithms is how the masculine gender presentation classification performs better than the feminine gender presentation classification, almost entirely, across the full threshold range. This is surprising, especially when compared to the results of the six skin tone audit, as the bias between the two gender presentation classifications remains present in all of the plotted commercial facial verification algorithms.

Similar to the findings from the skin tone audit, there remains little spread across the gender classifications when the *AWS Rekognition* facial verification algorithm is extremely confident (i.e., confidence scores greater than 99%). As the confidence in the match decreases, *AWS Rekognition* exhibits a widening gap between the masculine and feminine presenting classifications. This trend holds for the *Microsoft Face API*, under its default and latest configurations, as shown in **FIGURE 111(B, C)**, with an interval in both commercial facial verification algorithms where there remains little spread in the confidence scores across the gender classifications. These confidence score intervals are distinct to each commercial facial verification algorithm, greater than ~80% and 90% for *Microsoft Face API*, under its default and latest configurations respectively. Yet these confidence score intervals operate similar to the extremely confident confidence score interval for *AWS Rekognition*, perhaps this interval functions in a similar manner and provides some insights on how to interpret the confidence scores from *Microsoft Face API*, which doesn't provide any documentation to assist in that effort.

Distance-based Variance Interclass Bias Metric

IJB-C Six Skin Tone Audit of Commercial Facial Verification Algorithms

These insights can be furthered, through calculation of a score detailing the interclass bias between one of the six skin tone classifications and the overall performance. Following the procedure outline earlier, the author adopts an overall performance TAR curve that assumes that all classes should perform equally. Therefore, for each of the bootstrap samples generated, the overall performance curve was calculated for that resample as the mean of the six classes thresholds at the TAR; and the difference between each skin tone class and the overall performance curve was measured. Furthermore the d_{l_2} measure of interclass bias between a class and the overall performance can be calculated for each of the six skin tone classifications. **TABLE 12** relays the maximum d_{l_2} measure for each of the six skin tone classifications for the performance of *Microsoft Face API* under its default configuration, its latest released configuration, and *AWS Rekognition*.

TABLE 12 *A measure of interclass bias for each of the six skin tone classifications under various commercial facial verification algorithms:* AWS Rekognition*,* Microsoft Face API *under its default configuration of detection_01 and recognition_01, and its latest released configurations detection_03 and recognition_04.*

Class	AWS Rekognition	Microsoft Face API D01-R01 (Default)	Microsoft Face API D03-R04 (Latest)
Skin Tone I (Light Pink)	1.001005	2.569466	0.812459
Skin Tone II (Light Yellow)	2.846956	1.899081	1.547066
Skin Tone III (Medium Pink / Brown)	2.241796	1.171121	1.243549
Skin Tone IV (Medium Yellow / Brown)	5.042035	4.014602	1.956673
Skin Tone V (Medium-Dark Brown)	1.838562	1.945524	2.151646
Skin Tone VI (Dark Brown)	7.28833	2.577463	3.280281
Total	*7.28833*	*4.014602*	*3.280281*

When analyzed under the d_{l_2} measure of interclass bias between a class and the overall performance, the graphical interpretations delineated earlier are made obvious: the latest released configuration of the *Microsoft Face API* facial verification algorithm reduced the amount of interclass bias from its predecessor, under the default configuration; and outperforms *AWS*

Rekognition due to tighter confidence intervals and more even algorithmic performance. This relationship is also seen again when looking at the sample standard deviation of the interclass bias between a class and the overall performance. The latest released configuration of the *Microsoft Face API* not only avoids the extreme outliers but also maintains a smaller spread of interclass bias for all classes in the skin tone classification schema. One interesting observation across these three commercial facial verification algorithms, **Skin Tone III** (Medium Pink / Brown) outperforms **Skin Tone II** (Light Yellow). This is mildly surprising as overall darker skin tones perform worse regardless of reported commercial facial verification algorithms.

IJB-C Gender Presentation Audit of Commercial Facial Verification Algorithms

Similarly following the same methodology, outlined in ***Findings*** and used in the six skin tone audit, the score detailing the interclass bias between the two gender presentation classification can be calculated. **TABLE 13** relays the maximum d_{l_2} measure for the binary gender presentation classification for the performance of *Microsoft Face API* under its default configuration, its latest released configuration, and *AWS Rekognition*. This metric illuminates that though bias persists in all three tabulated commercial facial verification algorithms, the latest released configuration of the *Microsoft Face API* facial verification algorithm reduced the amount of interclass bias from its predecessor, under the default configuration; and outperforms *AWS Rekognition* due to tighter confidence intervals and more even algorithmic performance.

TABLE 13 *A measure of interclass bias for each of the binary gender presentation classifications under various commercial facial verification algorithms:* AWS Rekognition*,* Microsoft Face API *under its default configuration of detection_01 and recognition_01, and its latest released configurations detection_03 and recognition_04.*

Class	***AWS Rekognition***	***Microsoft Face API*** *D01-R01 (Default)*	***Microsoft Face API*** *D03-R04 (Latest)*
Gender Presentation	5.454197	5.048655	4.037939

The Importance of Building in Public

Microsoft Face API provides end-users with a choice of three detection models (i.e., detection_01, detection_02, detection_03), used to detect faces in a submitted image, and four recognition models (i.e., recognition_01, recognition_02, recognition_03, recognition_04), used to extract face features to facilitate comparisons. These models are continually supported by Microsoft to ensure backwards compatibility. One interesting consequence of this decision is that by submitting the same set of comparisons to each of the various commercial facial verification algorithms, a histology of bias within *Microsoft Face API*'s development efforts is revealed. In particular, this section is concerned with the following commercial facial verification algorithms (a) *Microsoft Face API under* its default configuration of detection_01 and recognition_01 released in 2017, (b) *Microsoft Face API* with a configuration of detection_02 and recognition_02 released in 2019, (c) *Microsoft Face API* with a configuration of detection_02 and recognition_03 released in 2020, and (d) *Microsoft Face API* with its latest released configuration of detection_03 and recognition_04 released in 2021. For ease of understanding, and to underscore the impact of chronology, through the rest of this section the author will refer to each algorithm by the year in which it was released (i.e., 2019 Microsoft Face API will refer to the *Microsoft Face API* with a configuration of detection_02 and recognition_02).

Whenever Microsoft releases a new version of their commercial facial verification algorithm, they typically provide a short description documenting the important changes made to the algorithm. In 2019, Microsoft updated its commercial facial verification algorithm to improve the accuracy on small, side-view, and blurry faces. In 2020, Microsoft introduced updates to its recognition models to improve recognition for facial imagery containing face covers (e.g., surgical masks, N95 respirators, and cloth masks). In 2021, Microsoft further improved the accuracy, especially on smaller faces and rotated face orientations (Microsoft 2022).

Analyzing the same $14,436$ comparisons submitted to each of the four discussed releases of the *Microsoft Face API* commercial facial verification algorithm for scoring, shows a general trend towards improvements to bias in both the six skin tone and binary gender presentation classification metrics. **FIGURE 112** and **FIGURE 113** show the d_{l_2} measure of interclass bias for skin tone and gender presentation, respectively, for the four discussed releases of the *Microsoft Face API* commercial facial verification algorithm.

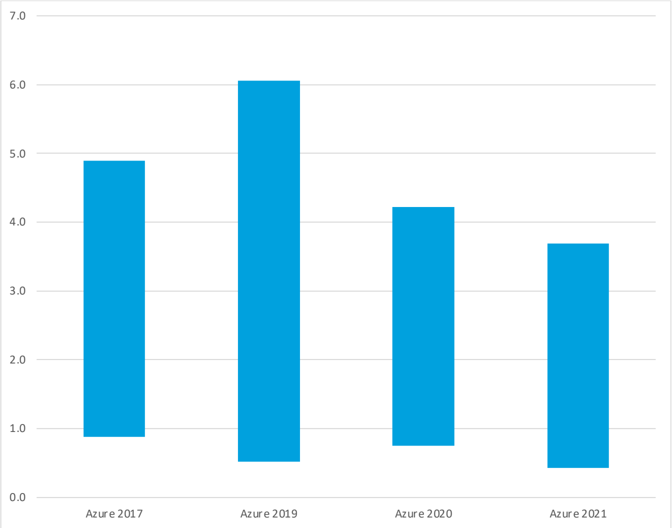


***FIGURE 112** The 90% two-sided confidence intervals (α = 10%) for the maximum measure of interclass bias for the six skin tone classifications schema under: (a)* Microsoft Face API *under its default configuration of detection_01 and recognition_01 released in 2017, (b)* Microsoft Face API *with a configuration of detection_02 and recognition_02 released in 2019, (c)* Microsoft Face API *with a configuration of detection_02 and recognition_03 released in 2020, and (d)* Microsoft Face API *with its latest released a configuration of detection_03 and recognition_04 released in 2021; using the IJB-C skin tone classification schema.*

When analyzing **FIGURE 112**, it's clear that 2021 *Microsoft Face API*, the latest algorithm, performs better than any of its predecessors, as the biases are reduced. However, it's clear that changes to the algorithm in 2019 to improve the algorithms accuracy on small, side-view, and blurry faces, caused the interclass bias for the six skin tone classification to increase dramatically. This was subsequently corrected in the 2020 and 2021 releases.

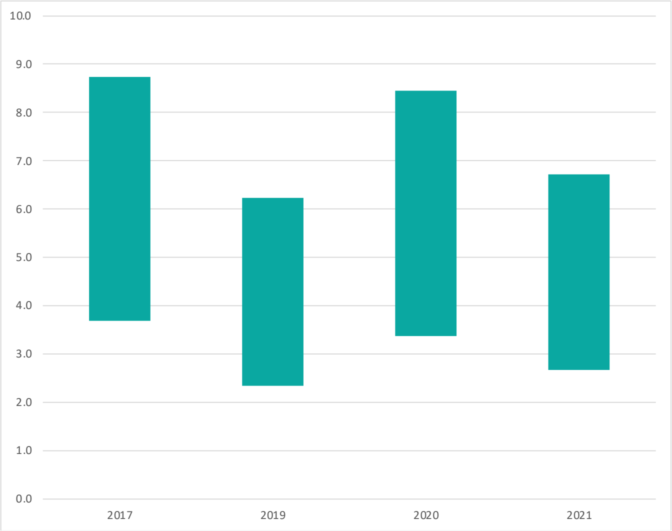


FIGURE 113 *The 90% two-sided confidence intervals (α = 10%) for the measure of interclass bias for the binary masculine or feminine gender presentation classification schema under: (a)* Microsoft Face API *under its default configuration of detection_01 and recognition_01 released in 2017, (b)* Microsoft Face API *with a configuration of detection_02 and recognition_02 released in 2019, (c)* Microsoft Face API *with a configuration of detection_02 and recognition_03 released in 2020, and (d)* Microsoft Face API *with its latest released a configuration of detection_03 and recognition_04 released in 2021; using the IJB-C gender presentation classification schema.*

When the confidence intervals for the measure of interclass, bias are plotted in **FIGURE 113**, it is evident that every algorithm has improved from the original 2017 *Microsoft Face API*. Major improvements were made to the bias metric when Microsoft released an update in 2019, yet when the algorithm was updated again in 2020 to improve recognition for imagery with face covers, Microsoft unexpectedly exaggerated the bias metric similar to its original 2017 *Microsoft Face API*. Even overall improvements to the algorithm in 2021, were unable to reduce the bias to the lowest intervals recorded in 2019.

It is interesting to note that improvements made to the algorithm that were designed to target a specific type of facial imagery can affect the metric of interclass bias so unpredictably. The same improvements in 2019 drastically improved the gender presentation bias, but also at the same time exaggerated the skin tone classification bias. These unintuitive consequences highlight the need for performing continuous measure of interclass bias within a classification schema throughout the design and commissioning of a facial verification algorithm. Additionally, these consequences also provide an important recommendation for developers of commercial facial verification algorithms.

The histology of bias allows potential facial recognition system owner’s the opportunity to not only evaluate a potential commercial facial verification algorithm at a specific point in time, but also by evaluating a set of algorithms, the owner could evaluate its developer's priorities regarding bias mitigation during the bidding process. This increased insight into a particular developer's priorities, may help gain the trust necessary from the end-users (i.e., the individuals submitting their facial imagery to the facial verification algorithm) of the facial recognition system as it shows the system's inequalities are known and under active improvement. This is strengthened by the design of the proposed metric for interclass bias within a classification schema, which requires no access to the underlying properties, configuration, or architecture of the underlying facial verification system, only it's outputs. This preserves any patents, copyright, trade secrets or other intellectual property rights that the commercial facial verification algorithm's developer may lay claim to.

While individual system owners may decide to scrutinize a potential facial verification algorithm with sample facial imagery relevant to their use case, developers need not shift the responsibility for this evaluation to their clients. The author believes that developers can and should release audits about the biases present in their algorithm to help owners during the competitive bidding process.

Important Assumptions

Assumptions About the Use of Unconstrained Images

Underpinning this work is an assumption shared by much of world of facial verification algorithms. This section lays this assumption out, the problems associated with it, and why the author has decided to retain the assumption in its proposal of a measurement of interclass bias in a classification schema.

This work assumes that the images that comprise the dataset used for generating the testing protocol, that is to say the individual images with facial imagery, are drawn from the same distributions and populations as the world in which the algorithm will be used. This assumption is similar to the assumptions made by facial verification algorithms when they are designed. Most existing facial verification algorithms rely on data prepared ahead of time. Typical facial recognition systems use some form of statistical or machine learning that are trained from a pre-existing dataset of facial images. An assumption employed by these systems is that the facial verification algorithm can be trained to a high accuracy given a sufficiently large training dataset (Hewitt and Belongie 2006). This theoretical accuracy is based on an assumption that the training data are drawn from the same distribution as exists in the actual application of this technology. These classifiers are vulnerable as they have no means to adapt or correct themselves if the facial images encountered are not similar to the images it was trained on.

***TABLE 14** A selection of relevant variables various variables to consider or control when considering a relatively unconstrained facial recognition system. This list is presented in alphabetical order and is incomplete. It was generated by looking at variables described by various datasets designed for facial recognition algorithm training (Berg et al. 2004; Hewitt and Belongie 2006; Huang et al. 2007, 2008; Ramanan et al. 2007; Zhao et al. 2018).*

- accessories
- age
- background
- background clutter
- subject's facial hair
- camera quality
- changes in appearance through time
- clothing
- color images
- color saturation
- candid photography
- diffuseness of lighting
- distance to subject
- facial expressions
- ethnicity
- expressions
- using vision eyewear
- facial expression
- focus
- gender presentation
- hairstyles
- head rotation
- illumination
- lighting angle
- lighting intensity
- lighting spectrum
- non-visible light modalities
- occlusions
- orientation
- pose
- race
- represent a broad range of individuals
- head rotation
- scale
- using scarves
- variation of time
- viewing angle

This assumption may not be valid in practice. When these facial verification algorithms are used by private corporations or governments, they are used in often uncontrolled environments. In these environments, the verification algorithms can expect to encounter immense variation: image sensor size, viewing angles, distance to the subjects, intensity, direction, color, and form of lighting, color saturation, focus, just to name a few of the variables related to the technical setup of the camera and lighting. **TABLE 14** above above, shows a short list of the various variables the author was able to find described by various datasets designed for facial recognition. With so many variables to consider, there can be no assurances that the training dataset has taken all the relevant variables into account, or that the distributions in the training dataset are the same as what can be expected in the actual application of the verification algorithm.

There are two ways that facial verification algorithms have attempted to square this assumption. The first are innovations to the statistical classifiers that verification algorithms are based on. Newer algorithms often employ convolutional neural networks, often referred to as CNNs, because they excel at this unconstrained task (Balaban 2015; Springenberg et al. 2015; Taigman et al. 2014b; Yi et al. 2014). One important feature of many CNN architectures is *pooling* (or sometimes referred to as *subsampling*), a mathematical operation that reduces the input over a certain area into a single value. Simply put, pooling consists of stepping a small window across an image and taking the maximum value from the window at each step. Pooling provides basic invariance to rotations and translations of a feature in an image, is considered to improve the object detection capability of convolutional networks (e.g., a face in an image is not in the center of the image but located off center (a slight translation) can be detected by the convolutional filters because the valuable information is provided to the correct neurons by the pooling operation) (Liu et al. 2016). Considered slightly differently, pooling can be considered a way to take large images and shrink them down while preserving the most important information in them. After a typical pooling operation an image has about a quarter as many pixels as it started with.

The second way facial verification algorithms have attempted to improve their performance on the actual applications, is to merely expand their training datasets to ones that better capture the diversity of the images the verification algorithms can expect in the actual application. An early database designed to help train and produce facial verification algorithms, Labeled Faces in the Wild, categorized their effort as providing a large set of "relatively unconstrained face images." They go on to define unconstrained as imagery that provides variation evident in everyday life: "variation in pose, lighting, expression, background, race, ethnicity, age, gender, clothing, hairstyles, camera quality, color saturation, focus, and other parameters." The motivation behind the unconstrained facial verification algorithm, was developing an algorithm that could recognize people in images the algorithm had no control over (i.e., pre-existing face images) (Huang et al. 2007). This trend of unconstrained facial imagery continued into other popular datasets (Klare et al. 2015; Maze et al. 2018; Wang et al. 2018b; Whitelam et al. 2017).

These innovations to the verification algorithm architecture and contributions to training datasets have resulted in the *in the wild* design paradigm adopted by the Labelled Faces in the Wild dataset becoming a defining feature of the problem facial verification algorithms aim to solve. Indeed, large datasets that have become popular benchmarks used to rank and compare facial verification algorithms from different vendors like *Labelled Faces in the Wild*, the *IJB-C* and *Adience*, do not report metrics about the distribution of the individual images in a dataset across any of the variables mentioned earlier (Eidinger et al. 2014; Huang et al. 2007; Maze et al. 2018). Instead, these datasets only guarantee a minimum facial size, or in some cases guarantee that each image can be detected by a simple facial detection algorithm. Furthermore, these datasets are often generated by scraping publicly available imagery from the internet using automated processes. The bulk (sometimes unpermitted) image collection using search queries provides a baseline randomness that ensures the desired variations evident in everyday life are visible in the generated dataset.

However, this in the wild framing introduces some problematic sources of bias. One important note inherent in the unconstrained problem as formulated above, is that under the best-case scenarios the datasets replicate the world around us. That is to say that they replicate the biases inherent in the world around us. Identities that are less frequently photographed are necessarily less represented in any dataset collected from existing images. As such, people who hold systematically marginalized and excluded identities will most likely be represented less (if at all) in the generated datasets, and as such the algorithms will be at a disadvantage when they encounter them in the actual application. In a manner of speaking, the algorithms learn the biases of the (sampled) photographers. However, there is an additional source of bias incorporated, from the designers of the camera equipment the photographers employ. This can take primarily two forms. The first is that the type of camera available to a photographer is a function of their access to the technology and their means to purchase or use it. This means that objects, vistas, or people of interest to photographers with less access to capital or in countries with restrictions to the import or distribution of modern cameras are less likely to photographed. Furthermore, images from photographers with intermittent or no internet connectivity will most likely be represented less (or not represented in the case of no internet connectivity) in the generated datasets, and as such the algorithms will be at a disadvantage when they encounter them in the actual application. The second form of camera bias comes from the design of the camera itself. Researchers and historians have established that photography cameras and equipment were designed to better represent white skin tones and do not equivalently represent people with darker skin tones. As such, datasets that feature people with darker skin tones might have images that are less bright or sharp than images of people with lighter skin tone. These darker images could affect facial verification algorithms that are sensitive to changes in the intensity, direction, color, and form of lighting or color saturation, or other variables mentioned in **TABLE 14** on page 146 above.

In fact, looking at the dataset generated for the commercial facial verification algorithm audit conducted in this dissertation, it is clear that there are biases potentially attributable to the camera. *AWS Rekognition* provides access to an API that provides a measure of the brightness of a face, to help end-users gauge if a face is bright enough to be enrolled into its facial recognition system. This API is part of AWS Rekogniton's image quality measures. The API reports a value representing the brightness of the face between 0 and 100, where a higher value indicates a brighter face image. When the individual image quality of each image in the dataset is measured by this API is aggregated by the six skin tone classification schema, it is evident that the reported brightness decreases with each darker skin tone. **FIGURE 114** below illustrates this phenomenon.

The author believes that because of the nature of the unconstrained problem that these algorithms are trying to solve, and the datasets these algorithms are trained on, the algorithms performance and measures of interclass bias in a classification schema must also be gauged on the unconstrained problem. As evidenced above, the number of variables to control in a relatively unconstrained problem is only constrained by the taxonomy with which one decides to employ to classify it. Some of these variables are directly correlated or associated with biases baked into or part of other systems, or even society at large. Discerning which variables are fair to control for in the testing protocol, and which the facial verification algorithm should be impervious to, are questions that are difficult to control for in the current paradigm for data collection. Controlling for some but not all of these variables could yield disparate results that damn or vindicate a particular algorithm on the preferences of the researchers, facial verification algorithm providers, end-users, activists, auditors, or governments who commissioned the analysis. In the end the most important consideration is that each image incorporated into the testing protocol comes from the same distribution as is expected in the actual application.

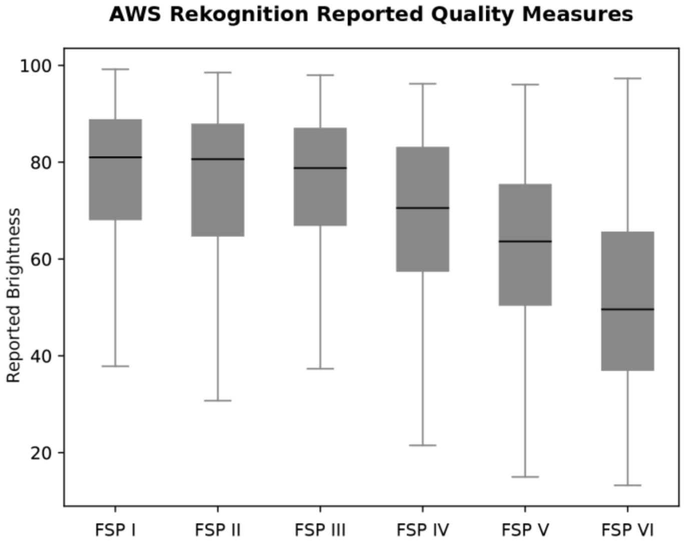


***FIGURE 114** A box plot of reported brightness values reported by* AWS Rekognition*'s image quality measurements by skin tone classification. Within each box, horizontal white lines denote median values; boxes extend from the 25th to the 75th percentile of each group's distribution of values; vertical extending lines denote adjacent values (i.e., the most extreme values within 1.5 interquartile range of the 25th and 75th percentile of each group); dots denote observations outside the range of adjacent values.*

In the end, some novel research techniques in training facial verification algorithms that utilize machine generated datasets (either through computer-generated three-dimensional imagery, or using generative adversarial networks) seem to provide respite for the concerns associated with the practicality of generating the large and varied dataset required to provide the variation evident in everyday life. Yet even as it becomes easier to build these better datasets and facial verification algorithms the problem will not disappear. The authors of the Labeled Faces in the Wild dataset provide a valuable insight on the drawbacks of this panacea:

> *On the other hand, in order to study more general face recognition problems, in which faces are drawn from a very broad distribution, one may wish to train and test face recognition algorithms on highly diverse sets of faces. While it is possible to manipulate a large number of variables in the laboratory in an attempt to make such a database, there are two drawbacks to this approach. The first is that it is extremely labor intensive. The second is that it is difficult to gauge exactly which distributions of various parameters one should use in order to make the most useful database. What percentage of subjects should wear sunglasses? What percentage should have beards? How many should be smiling? How many backgrounds should contain cars, boats, grass, deserts, or basketball courts?*
>
> *One possible solution to this problem is simply to measure a "natural" distribution of faces. Of course, no single canonical distribution of faces can*

> *capture a natural distribution of faces that is valid across all possible application domains (Huang et al. 2007).*

Assumptions in the Commercial Facial Verification Algorithm Audit

This next section refers to specific supporting statements that outline how these assumptions continue to hold in the commercial facial verification algorithm audit.

The commercial facial verification algorithms, analyzed in the audit, adopt the "*in the wild"* framing of facial verification. These providers market their algorithms to a whole variety of end-users who have just as many use-cases for the technology. Furthermore, *AWS Rekognition* and *Microsoft Face API* follow the convention employed by many in the wild datasets, and only require a minimum template size. However, both algorithms provide additional recommendations on the types of images that are likely to perform better on the algorithm.

AWS Recognition suggests the following recommendations for facial comparison input images (Amazon Web Services, Inc n.d.):

> 1. *Ensure that images are sufficiently large in terms of resolution. Amazon Rekognition can recognize faces as small as 50 x 50 pixels in image resolutions up to 1920 x 1080. Higher-resolution images require a larger minimum face size. Faces larger than the minimum size provide [sic] a more accurate set of facial comparison results.*
>
> 2. *Use images that are bright and sharp. Avoid using images that may be blurry due to subject and camera motion as much as possible. DetectFaces can be used to determine the brightness and sharpness of a face.*

Microsoft Face API states the following requirements for its facial recognition systems' detection algorithm, a preprocessing step before it can be submitted to its facial verification algorithm (Microsoft n.d.):

> — *The minimum detectable face size is 36x36 pixels in an image no larger than 1920x1080 pixels. Images with dimensions higher than 1920x1080 pixels will need a proportionally larger minimum face size.*
>
> — *For optimal results when querying Face - Identify, Face - Verify, and Face - Find Similar ('returnFaceId' is true), please use faces that are: frontal, clear, and with a minimum size of 200x200 pixels (100 pixels between eyes).*

The author read these requirements and infers that the facial verification algorithms use a convolutional neural network designed to handle high definition 1920×1080 pixel images and uses a series of convolutions to look for faces of the minimum pixel size anywhere within it. This seems likely to follow the same in the wild paradigm mentioned earlier. So, while it is expected that algorithms are able to perform better with larger faces, larger or brighter images are not required by the commercial facial verification algorithm providers and as such not guaranteed by end users.

Additionally, the IJB-C is a common benchmarking dataset used to evaluate performance across many different academic and commercial facial verification algorithms. The dataset follows the *in the wild* design in its sampling of images for inclusion into the dataset; as such, this dataset exhibits relatively unconstrained variation. Taken together, the author believes that the algorithms performance and measures of interclass bias in a classification schema must also be gauged on the unconstrained problem.

Conclusion

This dissertation reviews current metrics of success for facial verification algorithms, proposes a new realistic bias model that incorporates algorithmic bias (and fairness), and constructs a novel measure of interclass bias within a classification schema. This measure is constructed to be continuous, differentiable, monotonic and does not require access to the underlying properties, configuration, or architecture of the underlying facial recognition system.

The author develops a case study based on the audit of two commercial facial verification algorithm providers from Microsoft Corporation and Amazon Web Services, Inc. to evaluate the efficacy of this proposed interclass bias measure, the author utilized a subset of the IARPA Janus Benchmark C dataset, and it's 1:1 verification protocol. To the best of the author's knowledge, this is the first analysis of commercial facial recognition systems bias for "in the wild" facial imagery. The computation of this interclass bias metric shows that darker-skinned people have the least accurate verification matches, with an interclass bias measure of up to 7.2 times higher than lighter-skinned people. Additionally, the results show that one of Microsoft's commercial facial verification algorithms statistically eliminates the interclass bias for skin tone. Yet, all thirteen commercial facial verification algorithms evaluated experienced worse performance for feminine presenting persons compared to masculine presenting persons. The substantial disparities in the accuracy of classifying darker-skinned and feminine presenting people require urgent attention, if commercial companies are to build genuinely equal, transparent, and accountable facial verification algorithms. This case study shows how the measure of interclass bias can engender comprehensive analyses of facial verification algorithms biases without violating the facial recognition system's developer's intellectual property protections.

The construction of this measure of interclass bias can also be used by commercial facial verification algorithm developers to eliminate biases that can be incorporated into an algorithm's design, implementation, or training processes. The case study shows that efforts with Microsoft's commercial facial verification algorithm to improve accuracy with small, blurry, and side faces improved the equitable performance for skin tone but exaggerated the bias for gender presentation. These unintended consequences reiterate the need for careful monitoring throughout a facial verification algorithm training process. The author postulates that the metric for interclass bias can be incorporated directly into an algorithms training loss function because of its continuous, differentiable, and monotonic construction, but leaves this question for future researchers.

This work and the disparities revealed in these measures of bias show the utility of continuing to measure bias in facial verification algorithms, affirming earlier findings that point to the importance of using the true accept rate (also known as the sensitivity or recall), calculated from the fraction of genuine comparisons that correctly exceeds a given threshold, to expose underlying distributions that may be obfuscated by aggregated accuracy measures. Measures like this interclass bias metric can help ensure that regardless of the threshold utilized by a facial recognition system owner, the facial verification algorithm performs fairly for people of all classification schemas; especially as AI agents are increasingly involved as an integrated technology deployed in a menagerie of sectors. The author hopes that this metric assists researchers, policy makers and industry practitioners to develop better supervision throughout an AI agent’s development and production lifecycle to prevent disparate impacts, and better construct goals to ensure AI agents are solving the intended problem.

Thesis Contributions

This dissertation contributes to the applied uses of artificial intelligence (i.e., machine learning operations) in the area of facial recognition. Specifically, by focusing on feature bias in commercial facial recognition systems, it introduces novel thinking and techniques to measuring algorithmic fairness. In particular, it expands the evaluations of a facial recognition system's performance and success. These contributions are detailed below:

Introduction of a Realistic Bias Model

The author proposes a bias model that supports a measure of a facial recognition system's interclass bias. This bias model incorporates, (i) the system operators' limited understanding of the system's confidence scores; (ii) evaluations based on the true accept rate across the entire range of the threshold; (iii) the developer's judgement on the perceived impact of Type I Error*s* and Type II Errors; and (iv) an understanding of bias rooted in the jurisprudence of anti-discrimination laws and the theory of disparate impact (i.e., the discriminatory consequences of a neutral policy). To the best of the author's knowledge, this is the first investigation that adopts such a realistic model of bias.

A Novel Methodology to Measure Facial Recognition Systems' Interclass Bias within a Classification Schema

The author provides a novel methodology, metric and calculation of a facial recognition system's interclass bias within a classification schema. It assumes no access to the underlying properties, configuration, or architecture of the underlying facial recognition system, only it's outputs. It provides a generalized metric of "in the wild" facial recognition systems' performance for an entire class of people. This metric is shown to be a continuous, differentiable, and monotonic metric, that can be incorporated into a facial recognition systems' design and implementation, and/or a facial recognition system owner's testing and commissioning processes.

Testing this Methodology on Two Commercial off the Shelf Facial Recognition Systems

The method is tested on two widely used commercial off the shelf facial recognition systems. The author presents the results of experiments carried out using a commonly accepted benchmark to identify the efficacy of this proposed interclass bias measure. To the best of the author's knowledge, this is the first analysis of commercial facial recognition systems bias for "in the wild" facial imagery.

It is important to acknowledge that much of the inspiration and motivation for this work is derived from debates around the disparities in accuracy between different groups in facial recognition systems. In the end, it is this vision that provided the guiding framework for this investigation. However, it is equally important to understand that many of the results and techniques developed in this work are not limited to facial recognition systems. For example, the bias model described is likely to be relevant to most binary classification tasks based on computer vision. Thus, the impact of the interclass bias within a classification schema is likely to extend beyond facial recognition systems.

References

Abadi, M., A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. n.d. "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems." 19.

Allyn, B. 2020. "'The Computer Got It Wrong': How Facial Recognition Led To False Arrest Of Black Man." *NPR*, June 24, 2020.

Amazon Web Services, Inc. 2021. "AWS Service Terms." *Amazon Web Services, Inc.* Accessed September 22, 2021. https://aws.amazon.com/service-terms/.

Amazon Web Services, Inc. n.d. "Comparing faces in images - Amazon Rekognition." Accessed March 7, 2022. https://docs.aws.amazon.com/rekognition/latest/dg/faces-comparefaces.html.

Amazon Web Services, Inc. n.d. "The Facts on Facial Recognition with Artificial Intelligence." Accessed March 7, 2022. https://aws.amazon.com/rekognition/the-facts-on-facial-recognition-with-artificial-intelligence/.

Andriotis, A. 2019. "Freddie Mac Tests Underwriting Software That Could Boost Mortgage Approvals." *The Wall Street Journal*, September 24, 2019.

Apple. 2019. "Apple introduces dual camera iPhone 11." *Apple Newsroom*. Accessed August 11, 2021. https://www.apple.com/newsroom/2019/09/apple-introduces-dual-camera-iphone-11/.

Apple. 2020. "Use fall detection with Apple Watch." *Apple Support*. Accessed August 11, 2021. https://support.apple.com/en-us/HT208944.

Bainbridge, W. A., P. Isola, and A. Oliva. 2013. "The intrinsic memorability of face photographs." *Journal of Experimental Psychology: General*, 142 (4): 1323–1334. https://doi.org/10.1037/a0033872.

Balaban, S. 2015. "Deep learning and face recognition: the state of the art." *arXiv:1902.03524 [cs]*, 94570B. https://doi.org/10.1117/12.2181526.

Bartrip, P. W. J. 1980. "The State and the Steam-Boiler in Nineteenth-Century Britain." *International Review of Social History*, 25 (1): 77–105. Cambridge University Press.

BBC. 2019. "Chaayos cafe: Indian cafe's facial recognition use sparks anger." *BBC News*, November 22, 2019.

Beal, J., and A. Jehring. 2017. "Facebook shuts off AI experiment after two robots begin speaking in their OWN language only they understand." *The Sun*, July 31, 2017.

Belhumeur, P. N., J. P. Hespanha, and D. J. Kriegman. 1997. "Eigenfaces vs. Fisherfaces: recognition using class specific linear projection." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19 (7): 711–720. https://doi.org/10.1109/34.598228.

Bellamy, R. K. E., K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang. 2018. "AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias." *arXiv:1810.01943 [cs]*.

Bellemare, M., S. Srinivasan, G. Ostrovski, T. Schaul, D. Saxton, and R. Munos. 2016. "Unifying Count-Based Exploration and Intrinsic Motivation." *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Berg, T. L., A. C. Berg, J. Edwards, M. Maire, R. White, Yee-Whye Teh, E. Learned-Miller, and D. A. Forsyth. 2004. "Names and faces in the news." *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, 848–854. Washington, DC, USA: IEEE.

Blackburn, D. M., M. Bone, and P. J. Phillips. 2001. *Face Recognition Vendor Test 2000: Evaluation Report*. Department of Defense Counterdrug Technology Development Program Agency.

Bolger, R. 2018. "Cafe app that knows how you take your coffee sparks security concerns." *SBS News*. Accessed April 23, 2021. https://www.sbs.com.au/news/cafe-app-that-knows-how-you-take-your-coffee-sparks-security-concerns/b3364005-f866-4789-a735-40e71a84ea00.

Botev, Z. I., J. F. Grotowski, and D. P. Kroese. 2010. "Kernel density estimation via diffusion." *Ann. Statist.*, 38 (5). https://doi.org/10.1214/10-AOS799.

Brown, R. 1991. *Society and Economy in Modern Britain 1700-1850*. London: Routledge.

Buolamwini, J., and T. Gebru. 2018. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." *Conference on Fairness, Accountability and Transparency*, 77–91. PMLR.

Burt, C. 2018. "Australia considers deploying Unisys facial recognition technology for social media checks at airports | Biometric Update." Accessed April 23, 2021. https://www.biometricupdate.com/201809/australia-considers-deploying-unisys-facial-recognition-technology-for-social-media-checks-at-airports.

Caliskan, A., J. J. Bryson, and A. Narayanan. 2017. "Semantics derived automatically from language corpora contain human-like biases." *Science*, 356 (6334): 183–186. American Association for the Advancement of Science. https://doi.org/10.1126/science.aal4230.

Calsamiglia, C. 2005. "Decentralizing Equality of Opportunity and Issues Concerning the Equality of Educational Opportunity." PhD diss. New Haven, CT: Yale University.

Cameron, R. E., and A. J. Millard. 1985. *Technology Assessment: A Historical Approach*. Kendall/Hunt Publishing Company.

Cao, J., Y. Li, and Z. Zhang. 2018a. "Celeb-500K: A Large Training Dataset for Face Recognition." *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2406–2410.

Cao, Q., L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. 2018b. "VGGFace2: A Dataset for Recognising Faces across Pose and Age." *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, 67–74.

Carnot, N. L. S. 1897. *Reflections on the Motive Power of Heat and on Machines Fitted to Develop that Power*. (R. H. Thurston, ed. & tran.). New York: John Wiley & Sons.

Cavazos, J. G., P. J. Phillips, C. D. Castillo, and A. J. O'Toole. 2020. "Accuracy comparison across face recognition algorithms: Where are we on measuring race bias?" *arXiv:1912.07398 [cs]*.

Chen, B.-C., C.-S. Chen, and W. H. Hsu. 2014. "Cross-Age Reference Coding for Age-Invariant Face Recognition and Retrieval." *Computer Vision – ECCV 2014,* Lecture Notes in Computer Science, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, eds., 768–783. Cham: Springer International Publishing.

Chen, C., A. Dantcheva, T. Swearingen, and A. Ross. 2017. "Spoofing faces using makeup: An investigative study." *2017 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*, 1–8.

Chen, S., Y. Liu, X. Gao, and Z. Han. 2018. "MobileFaceNets: Efficient CNNs for Accurate Real-Time Face Verification on Mobile Devices." *arXiv:1804.07573 [cs]*.

Conner, C. D. 2017. *A People's History of Science: Miners, Midwives, and Low Mechanicks*.

Conover, W. J. 1971. *Practical Nonparametric Statistics*. Wiley.

Cook, C. M., J. J. Howard, Y. B. Sirotin, J. L. Tipton, and A. R. Vemury. 2019. "Demographic Effects in Facial Recognition and Their Dependence on Image Acquisition: An Evaluation of Eleven Commercial Systems." *IEEE Trans. Biom. Behav. Identity Sci.*, 1 (1): 32–41. https://doi.org/10.1109/TBIOM.2019.2897801.

Costa-jussà, M. R., C. Escolano, C. Basta, J. Ferrando, R. Batlle, and K. Kharitonova. 2020. "Gender Bias in Multilingual Neural Machine Translation: The Architecture Matters." *arXiv:2012.13176 [cs]*.

Crawford, K., and T. Paglen. 2019. "Excavating AI." -. Accessed April 23, 2021. https://excavating.ai.

Crumpler, W., and J. A. Lewis. 2021. *How Does Facial Recognition Work? A Primer*. 16. Center for Strategic and International Studies.

D'Agostino, R., and E. S. Pearson. 1973. "Tests for Departure from Normality. Empirical Results for the Distributions of b2 and √ b1." *Biometrika*, 60 (3): 613–622. [Oxford University Press, Biometrika Trust]. https://doi.org/10.2307/2335012.

Dantcheva, A., C. Chen, and A. Ross. 2012. "Can facial cosmetics affect the matching accuracy of face recognition systems?" *2012 IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 391–398.

Dastin, J. 2018. "Amazon scraps secret AI recruiting tool that showed bias against women." *Reuters*, October 10, 2018.

Davidson, R., and J. G. MacKinnon. 2000. “Bootstrap tests: how many bootstraps?” *Econometric Reviews*, 19 (1): 55–68. Taylor & Francis. https://doi.org/10.1080/07474930008800459.

Davison, A. C., and D. V. Hinkley. 1997. *Bootstrap Methods and their Application*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.

Deb, D., L. Best-Rowden, and A. K. Jain. 2017. “Face Recognition Performance under Aging.” *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 548–556.

Delta Airlines. 2021. “Delta launches first domestic digital identity test in U.S., providing touchless curb-to-gate experience.” *Delta News Hub*. Accessed April 23, 2021. https://news.delta.com/delta-launches-first-domestic-digital-identity-test-us-providing-touchless-curb-gate-experience.

Dwork, C., M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. 2011. “Fairness Through Awareness.” *arXiv:1104.3913 [cs]*.

ecobee. 2020. “Power your home with affordable, clean energy in one step.” Accessed August 11, 2021. https://www.ecobee.com/en-us/citizen/eco-plus-energy-efficiency-home/.

Efron, B. 1979. “Bootstrap Methods: Another Look at the Jackknife.” *The Annals of Statistics*, 7 (1): 1–26. Institute of Mathematical Statistics. https://doi.org/10.1214/aos/1176344552.

Efron, B., and R. J. Tibshirani. 1994. *An Introduction to the Bootstrap*. New York: Chapman and Hall/CRC.

Eidinger, E., R. Enbar, and T. Hassner. 2014. “Age and Gender Estimation of Unfiltered Faces.” *IEEE Transactions on Information Forensics and Security*, 9 (12): 2170–2179. https://doi.org/10.1109/TIFS.2014.2359646.

El Khiyari, H., and H. Wechsler. 2016. “Face Verification Subject to Varying (Age, Ethnicity, and Gender) Demographics Using Deep Learning.” *J Biom Biostat*, 07 (04). https://doi.org/10.4172/2155-6180.1000323.

Faulkner v. Super Valu Stores, Inc. 1993. *F.3d*, 1419.

Feldman, M., S. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. 2015. “Certifying and removing disparate impact.” *arXiv:1412.3756 [cs, stat]*.

Field, M. 2017. “Facebook shuts down robots after they invent their own language.” *The Telegraph*, August 1, 2017.

Findley, B. 2020. “Why Racial Bias is Prevalent in Facial Recognition Technology.” *Harvard Journal of Law & Technology*, (M. Bellamoroso, ed.).

Forczmański, P., and M. Furman. 2012. “Comparative Analysis of Benchmark Datasets for Face Recognition Algorithms Verification.” *Computer Vision and Graphics*, Lecture Notes in Computer Science, L. Bolc, R. Tadeusiewicz, L. J. Chmielewski, and K. Wojciechowski, eds., 354–362. Berlin, Heidelberg: Springer.

Foucault, M. 1973. *The order of things; an archaeology of the human sciences.* New York: Vintage Books.

Freenome. 2021. “Freenome Presents Data Demonstrating the Importance of Clinical Test Performance and Screening Adherence on Colorectal Cancer Outcomes [Press release].” *Business Wire*, May 21, 2021.

Furl, N., P. J. Phillips, and A. J. O’Toole. 2002. “Face recognition algorithms and the other-race effect: computational mechanisms for a developmental contact hypothesis.” *Cognitive Science*, 26 (6): 797–815. https://doi.org/10.1207/s15516709cog2606_4.

Fussell, S. 2020. “A Flawed Facial-Recognition System Sent This Man to Jail | WIRED.” Accessed March 6, 2022. https://www.wired.com/story/flawed-facial-recognition-system-sent-man-jail/.

Garvie, C., A. Bedoya, and J. Frankle. 2016. “The Perpetual Line-up: Unregulated Police Face Recognition In America.” *Perpetual Line Up*. Accessed April 23, 2021. https://www.perpetuallineup.org/.

Gentzel, M. 2021. “Biased Face Recognition Technology Used by Government: A Problem for Liberal Democracy.” *Philos. Technol.*, 34 (4): 1639–1663. https://doi.org/10.1007/s13347-021-00478-z.

Gillespie, E. 2019. “Are you being scanned? How facial recognition technology follows you, even as you shop.” *the Guardian*. Accessed April 23, 2021. http://www.theguardian.com/technology/2019/feb/24/are-you-being-scanned-how-facial-recognition-technology-follows-you-even-as-you-shop.

Goldstein, A. J., L. D. Harmon, and A. B. Lesk. 1971. “Identification of human faces.” *Proceedings of the IEEE*, 59 (5): 748–760. https://doi.org/10.1109/PROC.1971.8254.

Golino, C. L. 1966. *Galileo Reappraised*. UCLA Center for Medieval and Renaissance Studies. Contributions. Berkeley: University of California Press.

Google. 2017. “Save time with Smart Reply in Gmail.” *Google*. Accessed August 11, 2021. https://blog.google/products/gmail/save-time-with-smart-reply-in-gmail/.

Gracia, J. J. E. 2001. “Are Categories Invented or Discovered? A Response to Foucault.” *The Review of Metaphysics*, 55 (1): 3–20. Philosophy Education Society Inc.

Gross, R. 2005. “Face Databases.” *Handbook of Face Recognition*, A. J. S.Li, ed. New York: Springer.

Grother, P. J., G. W. Quinn, and P. J. Phillips. 2010. *Report on the Evaluation of 2D Still-Image Face Recognition Algorithms*. NIST Interagency/Internal Report (NISTIR).

Grother, P., and M. Ngan. 2014. *Face Recognition Vendor Test (FRVT) Performance of Face Identification Algorithms NIST IR 8009*. NIST Interagency/Internal Report (NISTIR).

Grother, P., M. Ngan, and K. Hanaoka. 2017. *Ongoing Face Recognition Vendor Test (FRVT) Part 1: Verification*. 76. NIST Interagency/Internal Report (NISTIR).

Grother, P., M. Ngan, and K. Hanaoka. 2019. *Face recognition vendor test part 3:: demographic effects*. NIST IR 8280. Gaithersburg, MD: National Institute of Standards and Technology.

Guo, G., S. Z. Li, and K. Chan. 2000. "Face recognition by support vector machines." *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580),* 196–201.

Guo, G., and N. Zhang. 2019. "A survey on deep learning based face recognition." *Computer Vision and Image Understanding*, 189: 102805. https://doi.org/10.1016/j.cviu.2019.102805.

Guo, G.-D., and H.-J. Zhang. 2001. "Boosting for fast face recognition." *Proceedings IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, 96–100.

Guo, Y., L. Zhang, Y. Hu, X. He, and J. Gao. 2016. "MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition." *arXiv:1607.08221 [cs]*.

Hall, P., and M. A. Martin. 1988. "On Bootstrap Resampling and Iteration." *Biometrika*, 75 (4): 661–671. [Oxford University Press, Biometrika Trust]. https://doi.org/10.2307/2336307.

Han, H., and A. Jain. 2014. "Age, Gender and Race Estimation from Unconstrained Face Images." *MSU Technical Report*, 2014.

Han, H., A. K. Jain, F. Wang, S. Shan, and X. Chen. 2017. "Heterogeneous Face Attribute Estimation: A Deep Multi-Task Learning Approach." *arXiv:1706.00906 [cs]*.

Hardt, M., E. Price, and N. Srebro. 2016. "Equality of Opportunity in Supervised Learning." *arXiv:1610.02413 [cs]*.

Harwell, D. 2021. "Civil rights groups ask Biden administration to oppose facial recognition." *Washington Post,* February 17, 2021.

Heidari, H., and A. Krause. 2018. "Preventing Disparate Treatment in Sequential Decision Making." *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence,* 2248–2254. Stockholm, Sweden: International Joint Conferences on Artificial Intelligence Organization.

Hero. 1851. *The pneumatics of Hero of Alexandria: from the original Greek*. (B. Woodcroft, ed., J. G. Greenwood, tran.). London: Taylor, Walton and Maberly.

Hewitt, R., and S. Belongie. 2006. "Active Learning in Face Recognition: Using Tracking to Build a Face Model." *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, 157–157. New York, NY, USA: IEEE.

Hill, K. 2020a. "Facial Recognition Tool Led to Black Man's Arrest. It Was Wrong." *The New York Times*, June 25, 2020.

Hill, K. 2020b. "Facial Match And Arrest Go Wrong." *The New York Times,* December 29, 2020.

Hogg, R. V., E. A. Tanis, and D. L. Zimmerman. 2015. *Probability and Statistical Inference*. Boston: Pearson.

Howard, J. J., L. R. Rabbitt, and Y. B. Sirotin. 2020. "Human-algorithm teaming in face recognition: How algorithm outcomes cognitively bias human decision-making." *PLoS One,* 15 (8): e0237855. https://doi.org/10.1371/journal.pone.0237855.

Howard, J. J., Y. B. Sirotin, and A. R. Vemury. 2019. "The Effect of Broad and Specific Demographic Homogeneity on the Imposter Distributions and False Match Rates in Face Recognition Algorithm Performance." *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 1–8. Tampa, FL, USA: IEEE.

Hsu, J. 2020. "AI Recruiting Tools Aim to Reduce Bias in the Hiring Process." *IEEE Spectrum*, July 29, 2020.

Huang, G. B., M. Narayana, and E. Learned-Miller. 2008. "Towards unconstrained face recognition." *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 1–8. Anchorage, AK, USA: IEEE.

Huang, G. B., M. Ramesh, T. Berg, and E. Learned-Miller. 2007. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*. University of Massachusetts, Amherst.

Hunt, E. 2016. "Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter." *The Guardian*, March 24, 2016.

IBM Corporation. 2022. "True positive rate (TPR)." *IBM Cloud Pak for Data*. Accessed March 30, 2022. https://prod.ibmdocs-production-dal-6099123ce774e592a519d7c33db8265e-0000.us-south.containers.appdomain.cloud/docs/en/cloud-paks/cp-data/4.0?topic=overview-true-positive-rate-tpr.

Ives, N. 2019. "JPMorgan Chase Taps AI to Make Marketing Messages More Powerful." *Wall Street Journal*, July 30, 2019.

Jain, V., and E. Learned-Miller. 2010. *FDDB: A Benchmark for Face Detection in Unconstrained Settings*. 11. UMass Amherst Technical Report.

Jia, Y., E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. 2014. "Caffe: Convolutional Architecture for Fast Feature Embedding." *arXiv:1408.5093 [cs]*.

Johnson, M., M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, M. Hughes, and J. Dean. 2017. "Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation." *arXiv:1611.04558 [cs]*.

Kamishima, T., S. Akaho, H. Asoh, and J. Sakuma. 2012. "Fairness-Aware Classifier with Prejudice Remover Regularizer." *Machine Learning and Knowledge Discovery in Databases*, Lecture Notes in Computer Science, P. A. Flach, T. De Bie, and N. Cristianini, eds., 35–50. Berlin, Heidelberg: Springer.

Kemelmacher-Shlizerman, I., S. M. Seitz, D. Miller, and E. Brossard. 2016. "The MegaFace Benchmark: 1 Million Faces for Recognition at Scale." *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4873–4882. Las Vegas, NV, USA: IEEE.

Kenna, S. 2017. "Facebook Shuts Down AI Robot After It Creates Its Own Language." *Huffington Post*, August 2, 2017.

Klare, B. F., M. J. Burge, J. C. Klontz, R. W. Vorder Bruegge, and A. K. Jain. 2012. "Face Recognition Performance: Role of Demographic Information." *IEEE Transactions on*

Information Forensics and Security, 7 (6): 1789–1801. https://doi.org/10.1109/TIFS.2012.2214212.

Klare, B. F., B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain. 2015. "Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A." *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1931–1939.

Kleinberg, J., S. Mullainathan, and M. Raghavan. 2016. "Inherent Trade-Offs in the Fair Determination of Risk Scores." *arXiv:1609.05807 [cs, stat]*.

Knight, W. 2020. "Companies Are Rushing to Use AI—but Few See a Payoff." *WIRED*, October 20, 2020.

Knopper, S. 2018. "Why Taylor Swift Is Using Facial Recognition at Concerts." *Rolling Stone*, December 13, 2018.

Krishnapriya, K. S., K. Vangara, M. C. King, V. Albiero, and K. Bowyer. 2019. "Characterizing the Variability in Face Recognition Accuracy Relative to Race." *arXiv:1904.07325 [cs]*.

Kumar, N., A. Berg, P. N. Belhumeur, and S. Nayar. 2011. "Describable Visual Attributes for Face Verification and Image Search." *IEEE Trans. Pattern Anal. Mach. Intell.*, 33 (10): 1962–1977. https://doi.org/10.1109/TPAMI.2011.48.

Lash Group. 2018. "Lash Group Launches Artificial Intelligence-Powered Electronic Benefit Verification Solution [Press release]." Accessed July 29, 2021. https://www.lashgroup.com/news-and-events/news/2018-lash-group-launches-artificial-intelligence-powered-electronic-benefit-verification-solution.

Leoforce. 2019. "Powered by even more advanced AI, Leoforce Announces Biggest Release Ever with Arya 3.0." *Cision PRWeb*, February 25, 2019.

Leveson, N. G. 2016. *Engineering a Safer World.* Cambridge: The MIT Press.

Lewis, M., D. Yarats, Y. N. Dauphin, D. Parikh, and D. Batra. 2017. "Deal or No Deal? End-to-End Learning for Negotiation Dialogues." *arXiv:1706.05125 [cs]*.

Li, S. Z., and A. K. Jain. 2011. "Introduction." *Handbook of Face Recognition*, S. Z. Li and A. K. Jain, eds., 1–15. London: Springer.

Lipton, Z. C., K. Azizzadenesheli, A. Kumar, L. Li, J. Gao, and L. Deng. 2018. "Combating Reinforcement Learning's Sisyphean Curse with Intrinsic Fear." *arXiv:1611.01211 [cs, stat]*.

Liu, S., J. McGree, Z. Ge, and Y. Xie. 2016. "4 - Computer vision in big data applications." *Computational and Statistical Methods for Analysing Big Data with Applications*, S. Liu, J. McGree, Z. Ge, and Y. Xie, eds., 57–85. San Diego: Academic Press.

Liu, Z., P. Luo, X. Wang, and X. Tang. 2015. "Deep Learning Face Attributes in the Wild." *arXiv:1411.7766 [cs]*.

Lohr, S. 2018. "Facial Recognition Is Accurate, if You're a White Guy." *The New York Times*, February 9, 2018.

Luong, B. T., S. Ruggieri, and F. Turini. 2011. "k-NN as an implementation of situation testing for discrimination discovery and prevention." *Proceedings of the 17th ACM SIGKDD* *international conference on Knowledge discovery and data mining*, KDD ’11, 502–510. New York, NY, USA: Association for Computing Machinery.

Main, F. 2014. “Armed robber, identified by facial recognition technology, gets 22 years.” *The Chicago Sun-Times*, June 5, 2014.

Mann, H. B., and D. R. Whitney. 1947. “On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other.” *Ann. Math. Statist.*, 18 (1): 50–60. Institute of Mathematical Statistics. https://doi.org/10.1214/aoms/1177730491.

Maze, B., J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney, and P. Grother. 2018. “IARPA Janus Benchmark - C: Face Dataset and Protocol.” *2018 International Conference on Biometrics (ICB)*, 158–165.

Menegus, B. 2019. “Amazon’s Defense of Rekognition Undermined by Police Client.” *Gizmodo*. Accessed April 27, 2022. https://gizmodo.com/defense-of-amazons-face-recognition-tool-undermined-by-1832238149.

Merler, M., N. Ratha, R. S. Feris, and J. R. Smith. 2019. “Diversity in Faces.” *arXiv:1901.10436 [cs]*.

Microsoft. 2016. “Meet Tay - Microsoft A.I. chatbot with zero chill [Press release].” Accessed August 17, 2021. https:/www.tay.ai/ archived at https://web.archive.org/web/20160414074049/https:/www.tay.ai/.

Microsoft. 2022. “How to specify a detection model - Face.” *Microsoft Azure Cognitive Services*. Accessed March 7, 2022. https://docs.microsoft.com/en-us/azure/cognitive-services/face/face-api-how-to-topics/specify-detection-model.

Microsoft. n.d. “Face - Verify Face To Face - REST API (Azure Cognitive Services - Face).” Accessed March 7, 2022. https://docs.microsoft.com/en-us/rest/api/faceapi/face/verify-face-to-face.

Moghaddam, B., T. Jebara, and A. Pentland. 2000. “Bayesian face recognition.” *Pattern Recognition*, 33 (11): 1771–1782. https://doi.org/10.1016/S0031-3203(99)00179-X.

Moschoglou, S., A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou. 2017. “AgeDB: The First Manually Collected, In-the-Wild Age Database.” *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1997–2005. Honolulu, HI, USA: IEEE.

Narayanan, A. 2018. “Translation tutorial: 21 fairness definitions and their politics.” New York University, NYC.

Nest Labs. 2020. “Behind the scenes with the new Nest Thermostat.” *Google*. Accessed August 11, 2021. https://blog.google/products/google-nest/behind-scenes-new-nest-thermostat/.

Ngan, M. L., and P. J. Grother. 2015. “Face Recognition Vendor Test (FRVT) - Performance of Automated Gender Classification Algorithms.”

Noble, S. U. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.

Norval, A., and E. Prasopoulou. 2017. “Public faces? A critical exploration of the diffusion of face recognition technologies in online social networks.” *New Media & Society*, 19 (4): 637–654. SAGE Publications. https://doi.org/10.1177/1461444816688896.

Ohlheiser, A. 2016. “Trolls turned Tay, Microsoft’s fun millennial AI bot, into a genocidal maniac.” *The Washington Post*, March 25, 2016.

Oliphant, T. E. 2007. “Python for Scientific Computing.” *Computing in Science Engineering*, 9 (3): 10–20. https://doi.org/10.1109/MCSE.2007.58.

O’Neil, C. 2017. *Weapons of math destruction: how big data increases inequality and threatens democracy*.

Oracle Staff. 1803. “[A dreadful accident happened at a steam engine].” *Oracle and the Daily Advertiser*, September 14, 1803.

O’Toole, A. J., P. J. Phillips, X. An, and J. Dunlop. 2011. “Demographic Effects on Estimates of Automatic Face Recognition Performance.” *The Ninth IEEE International Conference on Automatic Face and Gesture Recognition (FG 2011)*. Santa Barbara, CA, US.

Parge, R., N. Patil, P. Yelve, and Prof. V. Gavali. 2021. “Efficient Face Recognition System for Identifying Lost People.” *IJARSCT*, 328–330. https://doi.org/10.48175/IJARSCT-1142.

Paszke, A., S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. 2019. “PyTorch: An Imperative Style, High-Performance Deep Learning Library.” *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Payne, J. 2017. “A new angle on your favorite moments with Google Clips.” Accessed August 11, 2021. https://blog.google/products/devices-services/google-clips/.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay. 2011. “Scikit-learn: Machine Learning in Python.” *Journal of Machine Learning Research*, 12 (85): 2825–2830.

Persado. 2019. “JPMorgan Chase Announces Five-Year Deal with Persado For AI-Powered Marketing Capabilities [Press release].” *Business Wire*, July 30, 2019.

Phillips, P. J., K. W. Bowyer, P. J. Flynn, A. J. O’Toole, W. T. Scruggs, C. L. Schott, and M. Sharpe. 2007. *Face Recognition Vendor Test 2006 and Iris Challenge Evaluation 2006 Large-Scale Results*. NIST Interagency/Internal Report (NISTIR).

Phillips, P. J., P. J. Grother, R. J. Micheals, D. M. Blackburn, E. Tabassi, and M. Bone. 2003. *Face Recognition Vendor Test 2002: Evaluation Report*. 56. NIST Interagency/Internal Report (NISTIR).

Phillips, P. J., H. Moon, S. A. Rizvi, and P. J. Rauss. 2000. “The FERET evaluation methodology for face-recognition algorithms.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22 (10): 1090–1104. https://doi.org/10.1109/34.879790.

Phillips, P. J., and A. J. O'Toole. 2014. "Comparison of human and computer performance across face recognition experiments." *Image and Vision Computing*, 32 (1): 74–85. https://doi.org/10.1016/j.imavis.2013.12.002.

Phillips, P. J., A. J. O'Toole, F. Jiang, A. Narvekar, and J. Ayadd. 2009. "An Other-Race Effect for Face Recognition Algorithms."

Phrasee. 2021. "Phrasee enters $8.5 Billion Customer Experience (CX) Market with First Technology to Optimize Brand Language in Real-Time [Press release]." *PRNewswire*, April 15, 2021.

Planck Re. 2017. "Planck Re Launches AI-powered Platform for Insurers to Onboard New Businesses and Streamline Renewals With a Single Click of a Button [Press release]." *PRNewswire*, October 4, 2017.

Puntoni, S., R. W. Reczek, M. Giesler, and S. Botti. 2021. "Consumers and Artificial Intelligence: An Experiential Perspective." *Journal of Marketing*, 85 (1): 131–151. SAGE Publications Inc. https://doi.org/10.1177/0022242920953847.

Purdy, M., and P. Daugherty. 2017. *How AI Boosts Industry Profits and Innovation*. 28. New York: Accenture.

Putcha, G., T.-Y. Liu, E. Ariazi, M. Bertin, A. Drake, M. Dzamba, G. Hogan, S. Kothen-Hill, J. Liao, K. Li, S. Mahajan, K. Palaniappan, P. Sansanwal, J. St John, P. Ulz, N. Wan, H. Warsinske, D. Weinberg, R. Yang, and J. Lin. 2020. "Blood-based detection of early-stage colorectal cancer using multiomics and machine learning." *JCO*, 38 (4_suppl): 66–66. Wolters Kluwer. https://doi.org/10.1200/JCO.2020.38.4_suppl.66.

pymetrics. 2018. "pymetrics Raises $40 Million in Series B Funding Led by General Atlantic [Press release]." *Business Wire*, September 27, 2018.

Raji, I. D., and G. Fried. 2021. "About Face: A Survey of Facial Recognition Evaluation." *arXiv:2102.00813 [cs]*.

Ramanan, D., S. Baker, and S. Kakade. 2007. "Leveraging archival video for building face datasets." *2007 IEEE 11th International Conference on Computer Vision*, 1–8. Rio de Janeiro, Brazil: IEEE.

Rao, A. S., and G. Verweij. 2017. *PwC's Global Artificial Intelligence Study: Sizing the prize*. PricewaterhouseCoopers.

Reuters Staff. 2011. "U.S. tests bin Laden's DNA, used facial ID: official." *Reuters*, May 2, 2011.

Reynolds, M. 2017. "Fatal AI mistakes could be prevented by having human teachers." *New Scientist*. Accessed August 10, 2021. https://www.newscientist.com/article/2145838-fatal-ai-mistakes-could-be-prevented-by-having-human-teachers/.

Rhue, L. 2018. *Racial Influence on Automated Perceptions of Emotions*. SSRN Scholarly Paper. Rochester, NY: Social Science Research Network.

Ribeiro, M. T., S. Singh, and C. Guestrin. 2016. "'Why Should I Trust You?': Explaining the Predictions of Any Classifier." *arXiv:1602.04938 [cs, stat]*.

Robertson, D. J., E. Noyes, A. J. Dowsett, R. Jenkins, and A. M. Burton. 2016. "Face Recognition by Metropolitan Police Super-Recognisers." *PLOS ONE*, 11 (2): e0150036. Public Library of Science. https://doi.org/10.1371/journal.pone.0150036.

Roemer, J. E., and A. Trannoy. 2013. *Equality of Opportunity*. SSRN Scholarly Paper. Rochester, NY: Social Science Research Network.

Ryan-Mosley, T. 2021. "The new lawsuit that shows facial recognition is officially a civil rights issue | MIT Technology Review." April 14, 2021.

Sagan, C. 1997. *Pale blue dot: a vision of the human future in space*. New York: Ballantine Books.

Salian, I. 2019. "UnitedHealth Group Infuses AI into Healthcare Services." *The Official NVIDIA Blog*. Accessed July 29, 2021. https://blogs.nvidia.com/blog/2019/04/17/unitedhealth-deep-learning-healthcare-services/.

Samsung. 2020. "Use your Galaxy Watch3 to detect falls." *Samsung Electronics America*. Accessed August 11, 2021. https://www.samsung.com/us/support/answer/ANS00087244/.

Sassoon, L. 2017. "Expert's warning after Facebook robots 'develop their own language.'" *Mirror*, July 31, 2017.

Scheeres, J. 2002. "Nuke Reactor: Show Me Your Face." *WIRED*, August 9, 2002.

Schroff, F., D. Kalenichenko, and J. Philbin. 2015. "FaceNet: A Unified Embedding for Face Recognition and Clustering." *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 815–823. https://doi.org/10.1109/CVPR.2015.7298682.

Schwartz, O. 2019. "In 2016, Microsoft's Racist Chatbot Revealed the Dangers of Online Conversation." *IEEE Spectrum*, November 25, 2019.

Scott, D. W. 2014. *Multivariate density estimation: theory, practice, and visualization*. Hoboken, New Jersey: Wiley.

Silverman, B. W. 1986. *Density estimation for statistics and data analysis*. London; New York: Chapman and Hall.

Simonite, T. 2019. "A Sobering Message About the Future at AI's Biggest Party." *WIRED*, December 13, 2019.

Sirovich, L., and M. Kirby. 1987. "Low-dimensional procedure for the characterization of human faces." *J. Opt. Soc. Am. A, JOSAA*, 4 (3): 519–524. Optica Publishing Group. https://doi.org/10.1364/JOSAA.4.000519.

Smith, D. F., A. Wiliem, and B. C. Lovell. 2015. "Face Recognition on Consumer Devices: Reflections on Replay Attacks." *IEEE Transactions on Information Forensics and Security*, 10 (4): 736–745. https://doi.org/10.1109/TIFS.2015.2398819.

Springenberg, J. T., A. Dosovitskiy, T. Brox, and M. Riedmiller. 2015. "Striving for Simplicity: The All Convolutional Net." *arXiv:1412.6806 [cs]*.

Steed, R., and A. Caliskan. 2021. "Image Representations Learned With Unsupervised Pre-Training Contain Human-like Biases." *arXiv:2010.15052 [cs]*. https://doi.org/10.1145/3442188.3445932.

Sun, Y., X. Wang, and X. Tang. 2013. “Hybrid Deep Learning for Face Verification.” *2013 IEEE International Conference on Computer Vision*, 1489–1496.

Taigman, Y., M. Yang, M. Ranzato, and L. Wolf. 2014a. “DeepFace: Closing the Gap to Human-Level Performance in Face Verification.” *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 1701–1708.

Taigman, Y., M. Yang, M. Ranzato, and L. Wolf. 2014b. “DeepFace: Closing the Gap to Human-Level Performance in Face Verification.” *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 1701–1708. Columbus, OH, USA: IEEE.

The Trade Desk. 2018. “The Trade Desk Ushers in the Next Wave of Digital Advertising featuring Koa^{TM}, Artificial Intelligence (AI) for Advertisers [Press release].” *GlobeNewswire*, June 26, 2018.

Timekettle. 2020. “Never Be Lost In Translation Again: Timekettle’s Revolutionary World First Offline Translation Feature for WT2 Plus Translator Earbud Redefines International Communication [Press release].” *PR Newswire*. Accessed August 11, 2021. https://www.prnewswire.com/news-releases/never-be-lost-in-translation-again-timekettles-revolutionary-world-first-offline-translation-feature-for-wt2-plus-translator-earbud-redefines-international-communication-301088051.html.

Turk, M., and A. Pentland. 1991. “Eigenfaces for Recognition.” *Journal of Cognitive Neuroscience*, 3 (1): 71–86. https://doi.org/10.1162/jocn.1991.3.1.71.

Upstart. 2021. “Apple Bank Launches Personal Loans Powered By Upstart [Press release].” *Business Wire*, March 3, 2021.

Valentino-DeVries, J. 2020. “How the Police Use Facial Recognition, and Where It Falls Short.” *The New York Times*, January 12, 2020.

Vitruvius. 1999. *Vitruvius: “Ten Books on Architecture.”* (I. D. Rowland and T. N. Howe, eds.). Cambridge: Cambridge University Press.

Waddell, K. 2019. “Insurers battle California’s wildfire crisis with artificial intelligence.” *Axios*, September 28, 2019.

Wang, F., L. Chen, C. Li, S. Huang, Y. Chen, C. Qian, and C. C. Loy. 2018a. “The Devil of Face Recognition is in the Noise.” *arXiv:1807.11649 [cs]*.

Wang, M., and W. Deng. 2020. “Deep Face Recognition: A Survey.” *Neurocomputing*, 429: 215–244. https://doi.org/10.1016/j.neucom.2020.10.081.

Wang, M., W. Deng, J. Hu, X. Tao, and Y. Huang. 2018b. “Racial Faces in-the-Wild: Reducing Racial Bias by Information Maximization Adaptation Network.” https://doi.org/10.48550/arXiv.1812.00194.

Waverly Labs. 2019. “Meet Ambassador, your professional interpreter.” Accessed August 11, 2021. https://www.waverlylabs.com/blog/meet-ambassador-your-professional-interpreter.

West, D. M., and J. R. Allen. 2018. *How artificial intelligence is transforming the world*. Center for Technology Innovation.

Whitelam, C., E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, N. Kalka, A. K. Jain, J. A. Duncan, K. Allen, J. Cheney, and P. Grother. 2017. “IARPA Janus Benchmark-B Face

Dataset." *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 592–600.

Wilcox, R. R. 2010. "Inferences About Robust Measures of Location." *Fundamentals of Modern Statistical Methods: Substantially Improving Power and Accuracy*, R. R. Wilcox, ed., 147–167. New York, NY: Springer.

Wilhelm, A. 2021. "Writing helper Copy.ai raises $2.9M in a round led by Craft Ventures." *TechCrunch*, March 17, 2021.

Wolf, L., T. Hassner, and I. Maoz. 2011. "Face recognition in unconstrained videos with matched background similarity." *in Proc. IEEE Conf. Comput. Vision Pattern Recognition*.

Yang, M., L. Zhang, J. Yang, and D. Zhang. 2010. "Metaface learning for sparse representation based face recognition." *2010 IEEE International Conference on Image Processing*, 1601–1604.

Yang, M.-H., D. J. Kriegman, and N. Ahuja. 2002. "Detecting faces in images: a survey." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24 (1): 34–58. https://doi.org/10.1109/34.982883.

Yi, D., Z. Lei, S. Liao, and S. Z. Li. 2014. "Learning Face Representation from Scratch." *arXiv:1411.7923 [cs]*.

Young, H. P. 1994. "Preface." *Equity: In Theory and Practice*, xi–xiv. Princeton University Press.

Zafar, M. B., I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. 2017. "Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment." *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, 1171–1180. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee.

Zafeiriou, S., C. Zhang, and Z. Zhang. 2015. "A survey on face detection in the wild: Past, present and future." *Computer Vision and Image Understanding*, 138: 1–24. https://doi.org/10.1016/j.cviu.2015.03.015.

Zebra Medical Vision. 2021. "In an up-to $200M Acquisition by Nanox, Zebra Medical Vision Brings Its AI to Reimagine Radiology Globally [Press release]." *Business Wire*, August 10, 2021.

Zemel, R., Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. 2013. "Learning Fair Representations." *Proceedings of the 30th International Conference on Machine Learning*, 325–333. PMLR.

Zhang, Z., Y. Song, and H. Qi. 2017. "Age Progression/Regression by Conditional Adversarial Autoencoder." *arXiv:1702.08423 [cs]*.

Zhao, J., Y. Cheng, Y. Xu, L. Xiong, J. Li, F. Zhao, K. Jayashree, S. Pranata, S. Shen, J. Xing, S. Yan, and J. Feng. 2018. "Towards Pose Invariant Face Recognition in the Wild." *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2207–2216. Salt Lake City, UT: IEEE.

Zhou, E., Z. Cao, and Q. Yin. 2015. “Naive-Deep Face Recognition: Touching the Limit of LFW Benchmark or Not?” *arXiv:1501.04690 [cs]*.

Appendix

Table of Appendix Contents

A Brief Aside on the Nature of Risk in the Adoption of Technology

Throughout history, technology has often gotten ahead of society's understanding of its ethical consequences. That is to say, we harness technology because of its utility and then we seek to understand it. This has resulted in increased risk or serious accidents. For example, consider the simple case of the steam engine, a symbol of industrialization and the progress of technology. In the words of Nicolas Léonard Sadi Carnot, who laid the foundations of the discipline of thermodynamics:

> *To take away today from England her steam-engines would be to take away at the same time her coal and iron. It would be to dry up all her sources of wealth, to ruin all on which her prosperity depends, in short, to annihilate that colossal power. The destruction of her navy, which she considers her strongest defence, would perhaps be less fatal. (Carnot 1897)*

Though steam driven devices were known since the first century AD (Hero 1851; Vitruvius 1999), the first successful attempts to utilize steam for practical purposes occurred in the seventeenth century. These first engines designed by Thomas Savery, and later Thomas Newcomen enjoyed considerable success, and several hundred engines were constructed to pump water from coal mines, raising water to supply cities and water mills (Brown 1991). A blacksmith, Newcomen impressed by the difficulties and expense of pumping water, engineered an engine to solve this problem that used atmospheric pressure to force a piston into a partial vacuum created by condensing steam in a cylinder. Today we understand that Newcomen's engine relied on the dissolution of air in steam; but scientists in his day, were not aware that air dissolves in water (Golino 1966). That is to say, that Newcomen's engine was not based on prevailing scientific theory, but rather improvements to previous attempts (Conner 2017). These engines belched flames, smoke, and steam, as the result of inferior materials and poor fitting boiler plates and joints; that all resulted in frequent and disastrous boiler explosions. Yet these engines proved incredibly useful, upstarted companies that commercialized newer engines that could provide greater power with better efficiency; and their demand far outstripped supply (Cameron and Millard 1985). These better engines also increased the risk of explosions, one of the first of these engines exploded in September 1803 killing five workmen (Oracle Staff 1803). The engines were further commercialized as passenger steamboat engines where frequent and disastrous explosions were commonplace. A select committee from the UK Parliament found in 1870 that there were about 50 steam boiler explosions a year in England, which claimed on average 75 to 100 lives (Bartrip 1980).

Engineers quickly collected information about thermodynamics, the action of steam in a cylinder, the strength of materials in the engines, but little was known about how steam built up in the boiler, what effect corrosion and decay had on the boiler, and the causes of the boiler explosions. The race towards stronger and better engines made many boilers' designs obsolete, and produced unmanageable strain on boilers through excessive steam pressure and weaknesses in the materials and construction. Early technological innovations in boiler safety failed, because engineers did not understand what went on in steam boilers, and this knowledge was not available until more than half a century later. Together with regulatory reforms requiring frequent inspection and improved maintenance, the diffusion of the scientific knowledge of the actions of steam in a boiler lessens the risk of boiler explosions (Cameron and Millard 1985).

This extended example, serves to clarify that humans have propensity to take large risks and use a technology for a very long time before they understand it. The author believes that facial recognition is clearly useful. It has been useful in the private sector to secure a nuclear research facility (Scheeres 2002), to protect an artist from stalkers at her shows (Knopper 2018), and in the public sector to confirm the identity of Osama bin Laden (Reuters Staff 2011) and convicting an armed thief in Chicago (Main 2014). There are some that argue, quite poignantly, that the dangers of facial recognition systems outweigh their possible contributions. The dangers they point out, as well as many more dangers that are not yet fully understood, are real and must be addressed. However, the author finds these arguments unpersuasive, mostly because of the utility facial recognition systems have provided. As illustrated by the steam engine, the dangers behind a technology do not slow its adoption. Attempts to shut down facial recognition systems’ uses entirely, are like putting shaving cream back into a spent bottle.

The responsible move forward is not only to continue to develop the technology, but to create safety systems to mitigate its dangers and introduce regulatory frameworks that curb its abuses. The author believes that there are significant improvements that must be made to reduce the dangers associated with these technologies. Furthermore, the author hopes that this work furthers this important field of work by contributing a general measure of bias that can begin to quantify one aspect of the danger currently inherent in facial recognition systems, such that future researchers are able to reduce it.