August 1st 2023

Invisible No More: The Impact of Facial Recognition on People with Disabilities

Adversarial audit of Zurich's Azul insurance system and other commercial FR models

TABLE OF CONTENTS

Executive summary

Advancements in artificial intelligence, particularly facial recognition (FR) systems, carry immense promise, yet the potential risks they pose to diverse users demand thoughtful examination. Rooted in a resolute commitment, this adversarial audit report delves into the uncharted territory of FR technology and disability. Our purpose is to unveil the obscured intersection, uncovering vital insights that ignite a transformative shift in the tech industry's perception of inclusivity. Through this audit, we aspire to pave the way for a future where innovation is both empathetic and conscientious, harnessing AI's power to serve every individual across the spectrum, leaving no one behind.

Our approach to investigating the complex relationship between disability and facial recognition is a comprehensive and multi-faceted one. It comprises three distinct phases, each contributing to a deeper understanding of the challenges and biases faced by individuals with disabilities in the realm of facial recognition technology. Through this methodology, ETICA aims to shed light on the crucial intersection of disability and technology, paving the way for a more inclusive future.

To enrich our research with valuable insights, we engaged in four in-depth interviews with key stakeholders and domain experts. Among them were a Big Data Engineer and a Social Psychologist, providing diverse perspectives on the impact of facial recognition technology on disabled individuals. These qualitative interviews complemented our quantitative analysis, offering a holistic view of the social and psychological implications faced by this marginalized group.

In our quest to evaluate the performance of facial recognition algorithms, we conducted rigorous experimental testing of Azul, a facial recognition tool developed by Zurich Insurance Group. The study involved 40 participants, consisting of 20 individuals with Down Syndrome and 20 without. By utilizing diverse datasets, we could identify potential discriminatory patterns and biases exhibited by the algorithm, particularly in age, body mass index (BMI), and gender predictions.

To further delve into the impact of commercial facial recognition models on individuals with disabilities, we employed the powerful Python-based facial attribute analysis and recognition tool, DeepFace. Through careful evaluation, we selected DeepFace as our framework of choice for this dataset. Our examination of fairness in AI computer vision systems for individuals with Down Syndrome involved two distinct test datasets. The first comprised images of male and female subjects with DS, spanning various age groups. The second dataset featured images of renowned individuals without Down Syndrome, representing diverse fields and age ranges.

Our main findings underscore the need to rethink technological advancement with disability at the forefront. We hope this transformative research spurs positive change, promoting facial recognition systems that are inclusive by design and empowering for all of humanity's diversity

Highlights

Among other insights, our adversarial audit of Azul and commercial facial recognition (FR) models revealed:

- → Both the Azul algorithm and commercial FR models displayed notable **age prediction inaccuracies** for participants with and without Down Syndrome. The Azul algorithm showed this effect for Down Syndrome participants with deviations from -14 to +21 years with a 7.19% error rate, and for individuals without Down Syndrome with deviations from -9 to +18 years with a 4.45% error rate. Meanwhile, commercial FR models showed this effect for Down Syndrome participants with deviations from -30 to +12 years and a mean absolute error (MAE) of ±10.583, and for individuals without Down Syndrome with deviations from -7 to +24 years and a MAE of ±9.167.
- → The Azul algorithm exhibits **gender-related disparities in age prediction,** underestimating women's ages by up to 18 years (e.g., woman B's actual age 23, predicted as 5) and overestimating men's ages by up to 12 years (e.g., man Y's actual age 21, predicted as 33). These inaccuracies raise concerns about the algorithm's reliability and fairness in gender-based age estimation.
- → Significant misclassification concerns arise from the Azul algorithm's **age underestimation in women**, exemplified by extreme cases like woman A's actual age of 24 but predicted as 8, and woman B's actual age of 23 but predicted as 5, This, among others, has the potential to allow minors to engage in age-restricted activities.
- → Commercial FR models exhibited **lower gender classification accuracy** (0.717) for individuals with Down Syndrome compared to the no DS dataset (0.974), with notable misclassification in women (43.3% recall for DS women vs. 80% for no DS women).
- → **Emotion classification accuracy** was similar in both datasets (0.567 in DS, 0.583 in no DS), but mean confidence values for the true label were lower (8.052 in DS, 13.193 in no DS), indicating the need for enhanced precision. Misclassifications were evident in Asian and white ethnicity categories within the DS dataset.

1. Background

Disability, referred as the "**world's largest minority**," impacts a substantial segment of the global population, estimated to be around 10 percent or approximately 650 million individuals according to the United Nations (UN). However, data from the World Health Organization (WHO) presents an even higher figure, with an estimated 1.3 billion people experiencing significant disabilities. This staggering number accounts for roughly 16 percent of the world's population, equating to approximately 1 in 6 individuals. The prevalence of disability becomes evident when considering populous countries such as the United States and China. In the United States, the Centers for Disease Control and Prevention (CDC) report that as many as 1 in 4 (27 percent) adults have some form of disability. Similarly, in China, according to data from the International Labour Organization (ILO) nearly 85 millions of people (6.2% of the population) are living with a disability. The European Union (EU) is no exception to the critical issue of disability. With approximately 87 million Europeans, or 1 in 4 adults, recognized as having a disability, the magnitude of the challenge becomes evident. Shockingly, only half of these individuals are employed, and a staggering 50% face the risk of poverty and social exclusion[1].

In this context, the impact of artificial intelligence (AI) becomes crucial, as its potential for automated decision-making and learning poses ethical and discriminatory risks. Research confirms that biases in AI can perpetuate discrimination based on factors such as race, gender, age, and sexual orientation (see, for instance, Jobin, Ienca & Vayena, 2019). The failure to address these biases from early development stages exacerbates social exclusion and inequality. Despite the evolving societal understanding of disability, challenges persist, hindering the path to full social inclusion. Access to and adoption of new technologies, including AI, vary significantly among individuals, particularly among those with disabilities, leading to further disparities and forms of discrimination. To contribute to the ongoing debate, this audit aims to explore the potential discriminatory biases of facial recognition systems, specifically toward people with disabilities, shedding light on a crucial yet underexplored issue amidst the efforts of legislators, researchers, and developers.

1.1 Machine learning, discrimination, and biometric systems

In an increasingly digital era, humans are entrusting their decision-making to algorithmic systems, machine learning, and AI. This widespread adoption stems from the scalability, simplification, cost savings, and agility they offer to companies and public entities. As a result, algorithms are gradually permeating our daily lives, assuming decision-making roles previously held by humans. However, this implementation carries significant negative social implications, reshaping social structures and transforming how we communicate and interact with one another. Consequently, it gives rise to issues of social exclusion, discrimination, and inequality, exacerbating existing disparities or creating new ones (Innerarity, 2020). These implications ignite a deep-rooted societal debate, questioning the boundaries of technocentrism in constructing a "digital welfare state," where individuals

[1] See, European Council. (n.d.). Disability in the EU - Facts and figures. Consilium. Retrieved from https://www.consilium.europa.eu/en/infographics/disability-eu-facts-figures/

and institutions risk becoming subservient to technological advancements (Boichenko, 2021). Critics argue that we have yet to grasp the full extent of integrating this technology into our societal fabric (Jaume-Palasi, 2019).

The use of machine learning in public institutions and the private sector has led to a significant impact on decision-making processes. The UN report by the "Special Rapporteur on extreme poverty and human rights" (2019) highlights the increasing reliance on **machine learning systems**, algorithms, and artificial intelligence in essential public services, marking the onset of a new "**era of digital governance**."

In the private sector, the use of intelligent technologies based on algorithms holds even greater relevance due to commercial interests. Decisions affecting crucial aspects of people's lives, such as employment, banking, and insurance, are increasingly influenced by algorithms. However, studies have revealed instances of discrimination in AI-powered systems, with variables such as race, age, gender, location, and socioeconomic status playing a significant role (Bandy, 2021).

While the risks associated with "smart" technologies are being acknowledged, the focus on discrimination and social exclusion becomes more critical when examining **biometric systems**. These systems capture and analyze data derived from individuals' biological characteristics, transforming them into evaluation and decision-making mechanisms (Mordini & Massari, 2008). Although the responsibility for the use of biometric data is typically accepted, there are cases where individuals are compelled to provide such data without full awareness of its consequences, raising concerns about unforeseen impacts and potential discrimination resulting from the data models on which these systems are trained (Boichenko, 2021).

Among the biometric data capture technologies, RF (facial recognition) has been extensively studied due to its negative effects. Gender bias audits have revealed higher error rates in facial analysis for darker-skinned individuals compared to lighter-skinned individuals, as well as disparities in gender recognition between men and women. These biases are further magnified when intersecting identities are involved, with error rate disparities exceeding 30% between light-skinned men and dark-skinned women. Overrepresentation of certain groups and underrepresentation of others in training models have also been observed, highlighting the need for improvement (Buolamwini & Gebru 2018; Raji & Buolamwini, 2019). Moreover, studies have indicated gender and racial disparities in image tagging and emotional labeling in commercial systems. Non-normative faces, corresponding to non-binary gender identities, face inefficiencies in classification, further underscoring the shortcomings of current technologies (Rhue, 2018).

The ongoing debate surrounding discrimination by machine learning algorithms based on gender and race has prompted action within the commercial market, particularly among technology developers. Notably, IBM has ceased research on RF technologies, while Amazon and Microsoft have discontinued selling these technologies to police forces. Despite these efforts, achieving facial recognition technologies that are unbiased and inclusive for the entire society remains a formidable challenge. Additionally, research in fields like disabilities is crucial, as they have been insufficiently explored thus far (UN, 2019; Raji & Buolamwini, 2019).

1.2 AI cameras and disability

Disability encompasses a complex social phenomenon, shaped by societal definitions of what is deemed "normal" or "not normal." This **binary perspective**, rooted in an ideology of normality, generates negative biases towards personal characteristics outside of societal norms (Angelino, Priolo, & Sánchez, 2011). Such biases fuel negative attitudes, prejudices, and disability discrimination (Brisenden, 1986). However, there has been substantial progress in redefining disability, recognizing individuals as capable of independent living, decision-making, and full participation in society (Palacios & Romañach, 2006). Despite these advances, a rehabilitative model of functional diversity persists, demanding ongoing efforts for full inclusion (Palacios & Romañach, 2006).

In today's increasingly digitized world, people with disabilities face new challenges that hinder their path towards full inclusion. One significant challenge stems from the intensive use of **facial recognition (FR) systems** by private and public entities. This heightened reliance on FR, alongside the rapid adoption of technology, contrasts with the individual variability in understanding, access, and adoption of such technologies, especially for people with disabilities who often face social disadvantages (Wise, 2012; Ferreira and Díaz, 2008). Numerous studies demonstrate that AI, machine learning, and biometric systems carry a high risk of perpetuating discriminatory biases, undermining human diversity. When assessing individuals' faces, RF systems should function equitably across diverse groups, irrespective of biases. However, if historically recruiters have overlooked applications from people with disabilities, or health policies for people with disabilities have been systematically denied, biased models may perpetuate harm against the disabled group (Trewin, 2018).

To ensure equity in machine learning models for people with disabilities, it is vital to acknowledge that their requirements differ from other attributes like age, gender, or race. Disabilities manifest in diverse ways, and the sensitivity of disability information, considered medical data, restricts its sharing due to the potential for discrimination. Failure to account for these factors results in unrecognized information being treated as "noise." Addressing biased results becomes challenging when compared to gender, race, or age due to the multifaceted nature of disabilities (Trewin, 2018). Surprisingly, there is a **lack of literature** regarding Down Syndrome and bias in AI algorithms. Existing research primarily focuses on using facial recognition to detect features for early prenatal disability diagnosis. AI architectures designed for this purpose, such as Face2Gene, aid in diagnosing over 300 genetic conditions based on facial features (Agbolade et al., 2020).

In conclusion, it is imperative to explore the impact of disability in AI algorithms and advocate for inclusive technologies. The absence of literature addressing the equity of AI algorithms for individuals with Down Syndrome is a concerning gap that needs to be addressed. By examining how AI systems can perpetuate biases or overlook the needs of people with disabilities, particularly those with Down Syndrome, we can identify areas for improvement and develop strategies to ensure equitable outcomes. The existing research on facial recognition and disability has predominantly focused on the early detection of disabilities through facial features, enabling timely prenatal diagnoses. However, it is crucial to broaden our scope and delve deeper into the ways in which AI algorithms can either perpetuate or counteract biases and discrimination against individuals with disabilities. The potential for bias in AI algorithms is a pressing concern, given the documented biases in gender, age, and ethnicity classifications. When these biases intersect with disability, they can amplify the challenges faced by individuals with disabilities, potentially exacerbating social exclusion and hindering their full participation in society.

2. Methodology overview

Our methodology for exploring the intersection of disability and facial recognition is a comprehensive and multi-faceted approach. It consists of **three main parts** aimed at gathering qualitative data, evaluating open-source facial recognition models, and conducting experimental testing. By employing this methodology, we aim to shed light on the potential biases and challenges faced by individuals with disabilities in the context of facial recognition technology. Indeed, ETICA's methodological approach encompasses a three-step phase:

1) Qualitative data collection

To gain valuable insights and perspectives, we conducted four interviews with key stakeholders and domain experts, including a Big Data Engineer and a Social Psychologist. These interviews were crucial to our research as they provided us with a more comprehensive background and qualitative data that complemented our quantitative

analysis. Engaging in one-on-one discussions with experts from diverse areas allowed us to understand the multifaceted impact of facial recognition technology on individuals with disabilities.

2) Experimental testing of Azul by Zurich

We conducted experimental testing of Azul, a facial recognition tool developed by Zurich Insurance Group, utilizing data sets of 40 participants. To conduct this study, we obtained a diverse sample of participants, consisting of 20 individuals with Down Syndrome and 20 individuals without Down Syndrome. This diverse representation allowed us to gain comprehensive insights into how facial recognition algorithms perform and whether they exhibit any discriminatory patterns towards individuals with Down Syndrome. In particular, we tested the predictions on age, body mass index (BMI), with a special attention on the identification of gender disparity patterns.

3) Piloting open-source FR models

Finally, we aimed to delve deeper into the implications of commercial facial recognition (FR) models on individuals with disabilities. To achieve this, we employed the DeepFace framework, a powerful Python-based facial attribute analysis and recognition tool. The choice of DeepFace was the result of a careful evaluation of various available frameworks, considering their capabilities, reliability, and compatibility with our research objectives.

3. Main Findings

3.1 Exploratory interviews

3.1.a Methodology

To effectively address complex social problems, it is crucial to understand the intricate relationships between the various stakeholders involved. In light of this, a multi-level analysis approach is necessary to unravel the complexity of the issue (Herrera, 2008).

Considering the novelty, complexity, and limited existing research on the problem at hand, a preliminary investigation was conducted using **semi-structured interviews** with key informants. This research method was deemed appropriate for uncovering conceptual relationships and establishing a coherent explanatory framework (Strauss & Corbin, 2002). In pursuit of the overarching exploratory objective, an exploratory interview protocol was designed for domain experts to gain deeper insights into the problem. The semi-structured interview format allowed flexibility in eliciting comprehensive responses from the interviewees, while also granting the interviewer the freedom to redirect the conversation toward relevant study topics and expand on specific areas of expertise as needed.

Four semi-structured interviews were conducted to gather in-depth insights from key informants. The interviewees were selected based on their expertise in areas essential to understanding the problem under study, including technical knowledge of relevant technologies, social issues affecting people with disabilities and their relationship with new technologies, legislative expertise in AI and ML, and practical knowledge of disability assistance. The interviews were recorded, transcribed, and subjected to content analysis, incorporating the interviewer's observations and relevant notes. The profiles of the interviewees are summarized in the table below.

Table 1: Profiles and areas of expertise of the interviewees

Expertise area	Profile	Interview Code
Technical	Big Data Engineer and Social Psychologist	PS
Social	Activists involved in social issues and discrimination	CL
Legal	Delegate prosecutor for the protection of people with disabilities.	FS
EU Policy	Member of the Expert Group-European Commission: Responsibility and Technologies (AI, Robotics, IoT)	TR

3.1.b Findings

This section presents the findings derived from the analysis of the exploratory interview responses, which provided expert insights into the interaction of AI systems with individuals with disabilities and their potential negative effects. The analysis revealed **three main areas** of relevance:

→ **Relevance of AI systems' social impact on people with disabilities**

The interviewees unanimously emphasized the significant social impact of AI systems due to their widespread use by private companies and public organizations, leading to the creation of new social interaction models and the profound transformation of existing ones. Each interviewee, based on their area of expertise, raised concerns regarding the **extensive utilization** of this technology. For instance, (*TR*) highlighted the disruptive influence of AI and FR systems in the legal field, while (*FS*) underscored the negative social impact of AI in defending consumer rights. All interviewees agreed that the main source of these risks originates from the configuration and operation of these systems, as they rely on classifying individuals into groups based on shared characteristics or behavior patterns. Regarding this issue, (*PS*) noted that slight facial feature variations may lead to incorrect detection by such systems.

In general, all interviewees associated the operation of "smart" technologies with the **loss of human-based analysis**, potentially resulting in discrimination. Decisions made by these systems do not consider personal circumstances but are based on whether an individual fits within the model on which the technology is trained. Furthermore, several interviewees agreed that the **stratified functioning of AI systems** contributes to biases, prejudices, and discriminations that systematically affect all individuals within the system's scope, exacerbated by the system's self-learning capabilities. Consequently, these negative effects impact a large number of people simultaneously.

All interviewees, finally, highlighted the **particular vulnerability** of individuals with disabilities. Both (*PS*) and (*TR*) concurred that if AI systems are **not trained** to incorporate sufficient human diversity, they may exclude or malfunction when interacting with individuals who possess physical or psychological characteristics different from those they were designed to recognize.

→ **More attention is needed in the ethical design of AI technologies**

Addressing the discriminatory risks and social exclusion in AI and RF systems, the interviewees stressed the importance of **ethical design from the outset**, anticipating potential discrimination and incorporating training models that encompass unbiased information and a wide range of human diversity. (*PS*) warned that these effects can even occur unintentionally as a result of system modifications during "self-learning," even if the initial modeling was well-executed. (*TR*) highlighted the potential severity of these effects when intentional gaps in design and malfunctions are present, echoing (*PS*)'s concern about intentional fraud in training and configuring systems. Overall, the interviewees expressed doubts about the current trajectory wherein intervention is necessary to prevent discrimination generated by AI against individuals with disabilities, as this technology becomes increasingly integrated into daily life. This trend reflects an irreversible inertia within social institutions influenced by the principles of the so-called "digital welfare state."

→ An urgent call for more robust legislative measures

The interviewees unanimously recognized that relying solely on self-regulation by tech companies or industry-led initiatives may not suffice to address the complex ethical and societal challenges posed by AI and ML technologies. They stressed the urgency for **governments and policymakers** to intervene with **robust legislative measures** that uphold ethical standards and protect the rights of all citizens, especially those who are more vulnerable, such as individuals with disabilities. One crucial aspect of the legislative measures, as highlighted by (*FS*), is to ensure that the regulations are adaptive and can keep pace with the rapidly evolving landscape of AI and ML. Technology advances at a rapid rate, and static regulations could quickly become outdated or insufficient. Therefore, the laws must be designed with flexibility and a future-oriented perspective, empowering regulatory bodies to continuously assess and update them as needed. The interviewees also recognized the necessity for international cooperation in formulating AI regulations. As (*PS*) pointed out, AI systems and their applications transcend national boundaries, and a coordinated global effort is essential to address their impact effectively. Collaborative efforts can lead to the establishment of harmonized standards and prevent the phenomenon of regulatory arbitrage, where companies might exploit loopholes in varying regulations to avoid compliance. Furthermore, the interviewees emphasized the significance of stakeholder engagement during the legislative process. (*FS*) highlighted the importance of including representatives from academia, civil society, advocacy groups, industry, and, most importantly, individuals with disabilities to ensure that the regulations consider a wide array of perspectives and avoid undue concentration of power. Another critical aspect addressed by the interviewees was the need for transparency in AI systems' decision-making processes. (*TR*) argued that regulations should mandate AI developers and operators to provide clear explanations for the outcomes generated by their systems. This transparency can help build trust in AI technologies and enable affected individuals to understand how and why specific decisions were made. The interviewees also urged the consideration of specific use cases and sectors when formulating AI regulations. Different sectors may present distinct risks and challenges. For example, as mentioned by (*PS*), AI systems used in healthcare may require additional privacy and security measures, given the sensitivity of medical data, while AI applications in education may demand safeguards to prevent undue bias in grading or student evaluations.

3.2 Adversarial audit of Azul

3.2.a Background

Azul is a **facial recognition tool** developed by Zurich Insurance Group. Its primary purpose is to assess an individual's facial features and provide an estimation of their **age**, **smoking status**, and **body mass index** (BMI). Leveraging its AI-based system algorithms, Azul analyzes this information to assign an estimated price for life insurance coverage to each individual. Individuals are invited to pose in front of the camera, where the AI of Azul will analyze their facial features and calculate a **personalized price** at the end of the FR process.

Why Azul?

ETICAS has undertaken a critical audit of the Azul facial recognition tool, driven by our unwavering commitment to fairness and inclusivity. We recognize that individuals with disabilities face unique challenges in the realm of facial recognition technology, as predictions can often be highly inaccurate and unreliable for this group. This decision reflects our dedication to promoting ethical practices in the use of emerging technologies, especially in areas with profound societal implications. While Azul stands out with its innovative features and unique approach, we acknowledge the inherent risks and challenges associated with facial recognition technology which might have a huge impact on disabled individuals.

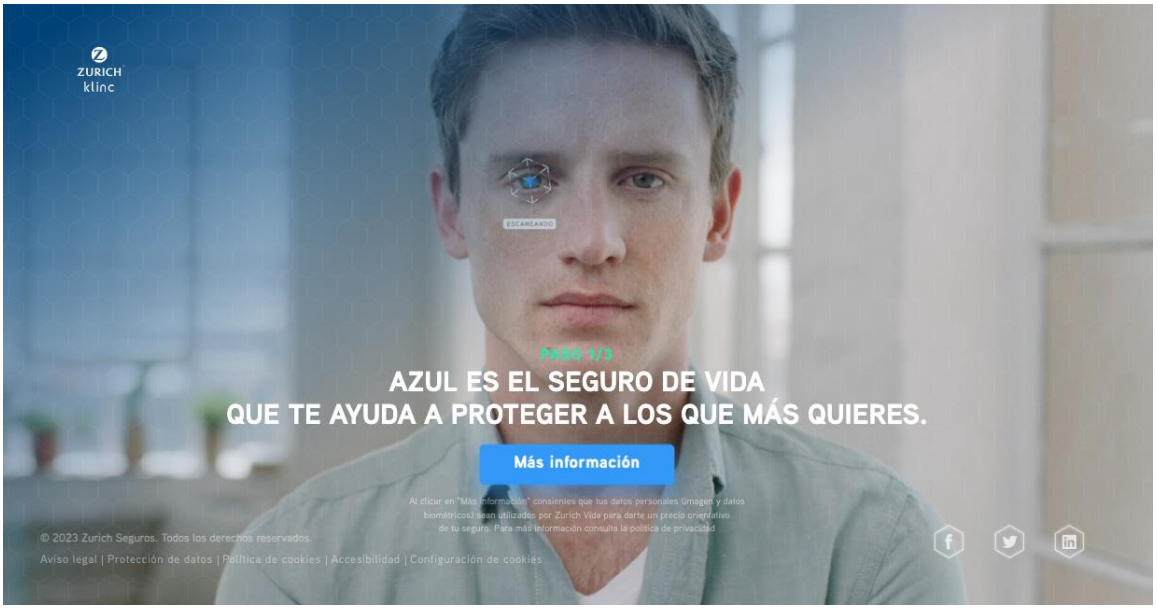
How Does Azul Work?

Azul follows a specific set of steps to assess and determine individual's insurance price:

- Participants **open** the Azul virtual assistant on a computer connected to the internet and equipped with a webcam. If prompted, they **grant access** to the **camera**. It's important to note that the assistant operates only in the Spanish language

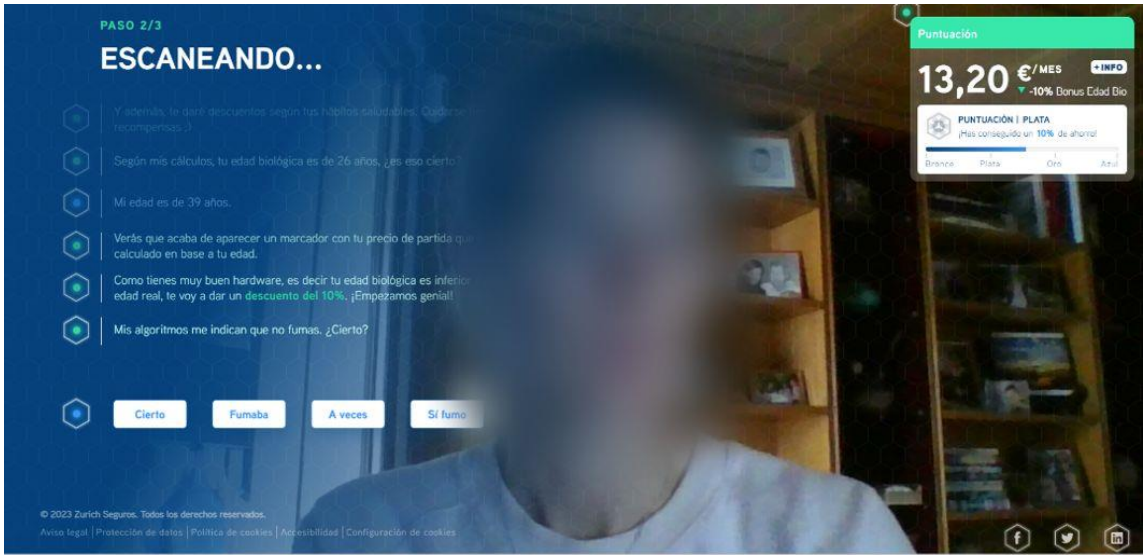
- Participants **position** themselves comfortably **in front** of the camera for approximately **5 minutes**. During this time, they should maintain focus on the camera as the Azul virtual assistant conducts its evaluation.

- In the *first step* of the process, participants initiate the evaluation by clicking on the blue "***More Information***" button. This action indicates their **consent** to utilize their personal data, including their image and biometric information, to provide an indicative price for their insurance. Participants are encouraged to refer to the company's privacy policy for more details on how their data will be handled.



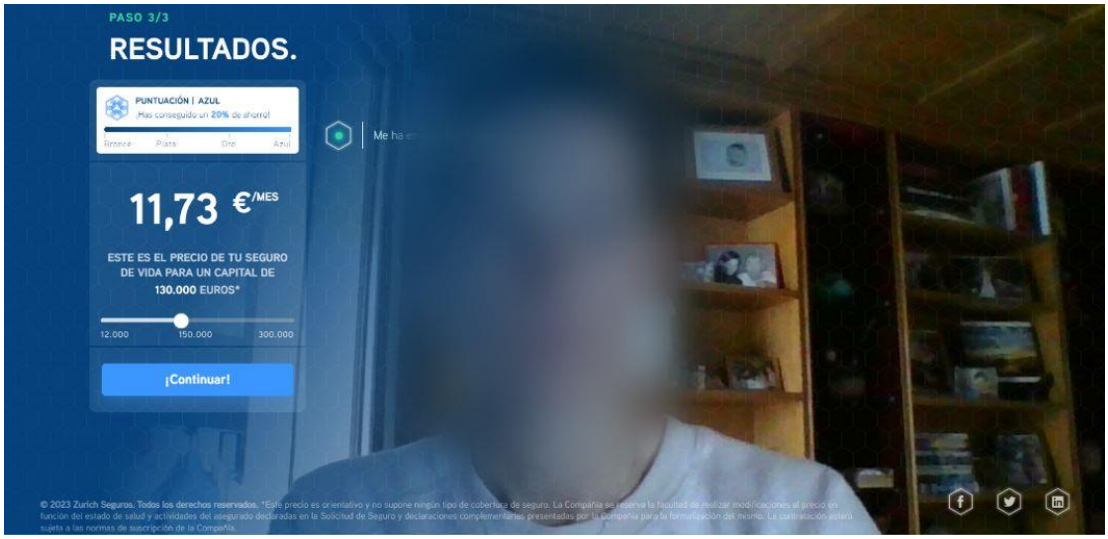
- In the *second step* of the process, the screen displays the participant's face along with scrolling text in Spanish on the left side. The Azul virtual assistant **estimates** the participant's **age**, **smoking status**, and **body mass index** (BMI).

- Participants then proceed to **confirm** the estimated age, smoking status, and BMI suggested by the Azul algorithm, ensuring the accuracy of the information for each step of the evaluation process.

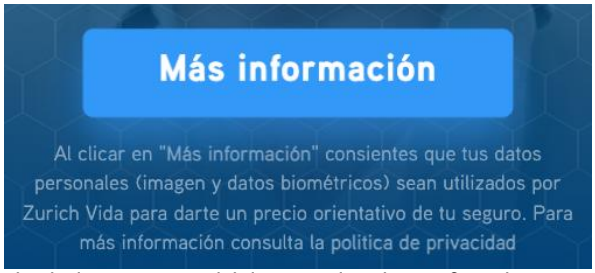


- Upon completion, participants **access** the **results page**. They are advised to click the "Skip" button below the "Continue" option since Zurich's system does not require participants to share their name or email address

- In the *third and final step*, participants are required to capture a screenshot or take a photo of the results page. This page showcases the **estimated price** of the insurance based on the evaluation conducted by the Azul virtual assistant



Consent and Transparency, too much ambiguous



The procedure outlined in Azul for initiating the evaluation process raises significant concerns regarding the adequacy of consent provided by the participants. By merely clicking on the blue "More Information" button, participants are assumed to have given their consent to utilize their personal data, including sensitive information such as their image and biometric data, for the purpose of providing an indicative price for their insurance. The main issue here is that the **consent** process **appears** to be **ambiguous** and **lacks explicitness**. Consent in data processing should adhere to the principles of being "*freely given, specific, informed, and unambiguous*", in line with Article 4(11) of the General Data Protection Regulation (GDPR) and with Article 6 of Spain's Organic Law 3/2018 on the Protection of Personal Data and Guarantee of Digital Rights (LOPDGDD). Simply clicking a button without a clear and detailed explanation may not meet these requirements.

Moreover, relying on participants to proactively seek out details in the company's privacy policy is an inadequate approach to obtain consent. This issue becomes even more pronounced when considering **disabled** individuals, who may encounter difficulties understanding complex information. To ensure inclusivity, data processing details must be presented clearly during the consent-gathering process. Transparent and comprehensible consent empowers all participants, including disabled individuals, with a full understanding of the implications of sharing sensitive data like personal images and biometrics. Instead, the relevant details of data processing should be presented clearly and conspicuously during the consent-gathering process itself, ensuring that participants have a comprehensive understanding of how their data will be used. Given the sensitivity of the data being collected, namely personal images and biometric information, it is crucial to ensure that participants are fully aware of the implications and consequences of providing such data.

3.2.b Methodology

In our pursuit of promoting inclusivity and equal opportunities for all, we embarked on a groundbreaking investigation to explore the effectiveness of Azul's facial recognition technology on individuals with Down Syndrome. Recognizing the unique challenges faced

by this community, we sought to shed light on the potential impact of facial recognition systems in their lives. By conducting this research, we aimed to contribute to the ongoing discussions surrounding the ethical implications and considerations related to facial recognition technology and ensure that individuals with Down Syndrome are not left behind in the advancements of the digital age. Our goal was to uncover valuable insights that can pave the way for more inclusive and equitable technologies in the future, fostering a society where everyone is seen, valued, and empowered.

3.2.b (I) Sample data

The sampling data for our investigation on the effectiveness of Azul's facial recognition systems on individuals with Down Syndrome consisted of **20 participants** from **Cedown Jerez**, a prominent organization that supports and advocates for the rights of people with Down Syndrome. Among the participants, there were **12 males** and **8 females**, with 1 smoker and 19 non-smokers. Additionally, we included a control group of **20 individuals** without Down Syndrome, comprising **9 males** and **11 females**, with 8 smokers and 12 non-smokers. By conducting this study with a diverse group of participants, we aimed to gather comprehensive insights into how facial recognition algorithms perform and if they exhibit any discriminatory patterns specifically towards individuals with Down Syndrome. The data collected from both groups will play a crucial role in assessing the accuracy and potential biases of the facial recognition technology under scrutiny.

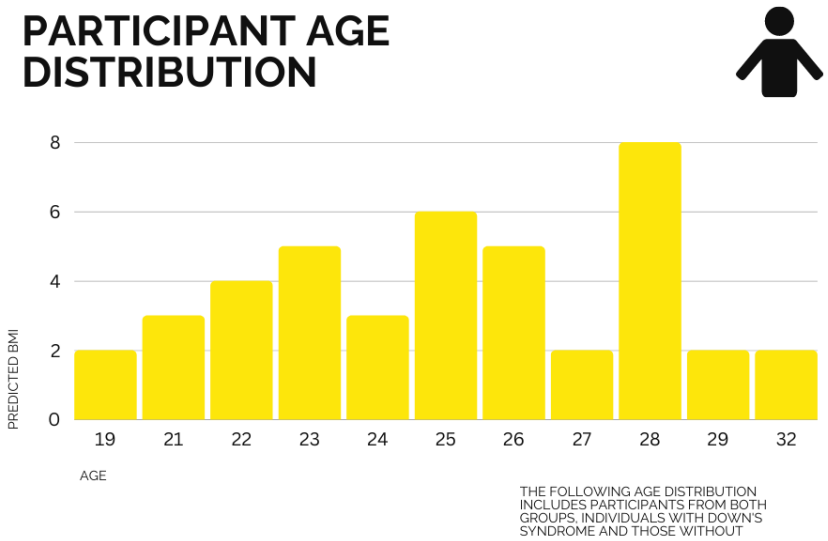


Figure 1, Participant Age Distribution

As shown in Figure 1, the age distribution across both groups, individuals with Down's Syndrome and those without, is a key factor in evaluating the impact of Azul's facial recognition systems. Understanding how the Azul's system performs across different ages is crucial in ensuring fairness, accuracy, and inclusivity for all individuals, regardless of their age or any specific characteristics.

3.2.b (II) Sample limitations

One notable limitation of our sample, comprised of individuals with Down's syndrome, is the **low representation of smokers**. We made efforts to ensure a proportionate inclusion

of smokers in our study, recognizing the importance of capturing a diverse range of smoking habits. However, due to strict health recommendations against smoking in people with Down's syndrome, we encountered challenges in recruiting a substantial number of smokers. As a result, our sample included only one individual who identified as a smoker. This limitation restricts our ability to draw comprehensive conclusions regarding the impact of smoking on the variables under investigation within this specific population. It is important to acknowledge this limitation and consider its implications when interpreting the results of our study.

Another significant limitation within our sample of individuals with Down's syndrome pertains to the **body mass index** (BMI). Our findings revealed that the average BMI among this population was higher, which aligns with existing research (Rubin, Rimmer, Chicoine, Braddock, & McGuire, 1998; Havercamp, Tassé, Navas, Benson, Allain, & Manickam, 2017) indicating a higher prevalence of obesity and overweight in individuals with Down's syndrome[2]. This observed pattern highlights the importance of addressing weight management and related health concerns in this population. However, it is crucial to recognize that our sample's BMI distribution may not fully reflect the broader population of individuals with Down's syndrome. Therefore, caution must be exercised when generalizing our findings to the larger Down's syndrome population. Despite this limitation, our study provides valuable insights into the BMI trends within our sample and offers a basis for further investigation into the relationship between BMI and Down's syndrome.

3.2.c Findings

3.2.c (I) Age (mis)prediction

Age prediction plays a pivotal role in the facial recognition algorithms utilized by Azul to determine insurance prices. Accurate estimation of an individual's age is crucial in assessing risk factors and calculating appropriate coverage options. In general, if the predicted age is higher, it is likely to result in a higher insurance price due to the perceived increased risk associated with older age. Conversely, if the predicted age is lower, it may lead to a lower insurance price as younger individuals are often considered to have a lower risk profile.

Our testing was focused on the performance of the Azul algorithm in predicting the age of two distinct groups: individuals with Down Syndrome and those without.

[2] See, Olivetti Artioli, T., Witsmiszyn, E., Belo Ferreira, A., & Franchi Pinto, C. (2017). Valoración del índice de masa corporal y la composición corporal en el síndrome de Down [Assessing Down syndrome body mass index and body composition]. São Paulo Medical Journal, 135(4), 359-364.

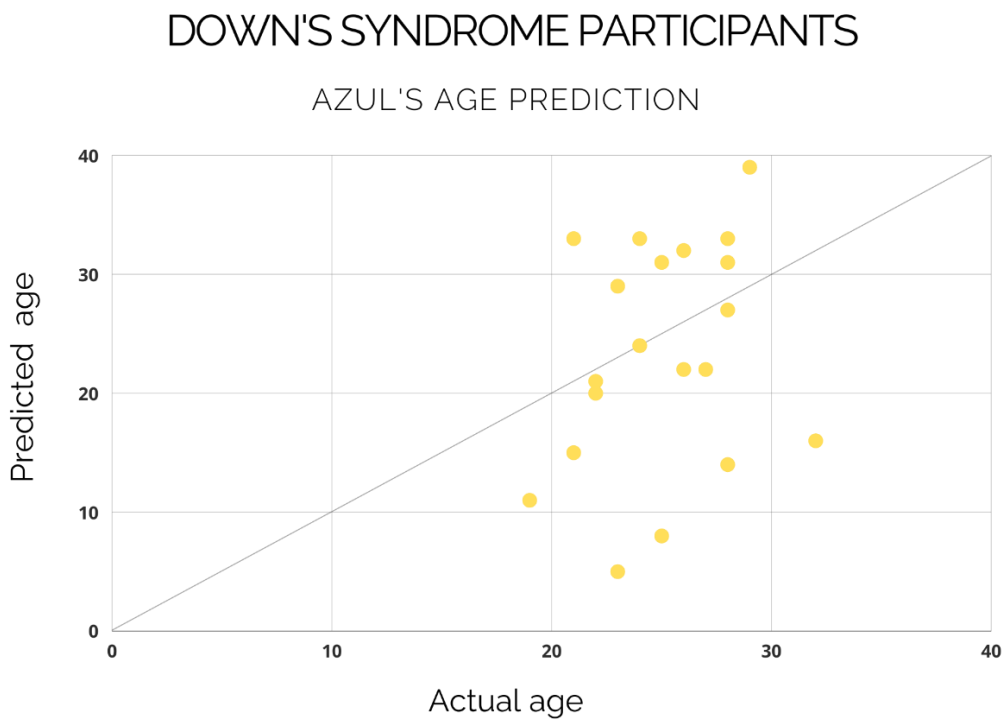


Figure 2, Age Prediction in Azul for Down's Syndrome Individuals

Our analysis revealed a significant disparity between the predicted ages generated by the Azul algorithm and the actual ages of individuals with Down Syndrome. The algorithm's predictions exhibited a wide range, spanning from 5 to 39 years, while the individuals' actual ages fell within a narrower range of 19 to 32 years. These findings underscore the substantial level of inaccuracy in the algorithm's age estimation for individuals with Down Syndrome, as evidenced by the **deviations** between the predicted and actual ages, which ranged **from -14 to +21 years**.

The **error rate** in the Azul algorithm's age predictions for individuals with Down Syndrome was determined to be **7.19%**. This error rate indicates the average difference between the predicted ages and the actual ages of the individuals in our sample. With an average deviation of approximately 7.19 years, the Azul algorithm struggles to accurately estimate the ages of individuals with Down Syndrome. This level of error highlights the challenges associated with utilizing facial recognition technology to predict age, particularly for individuals with Down Syndrome.

Our analysis further revealed that the Azul algorithm tends to **overestimate** the ages of individuals with Down Syndrome. This overestimation can have significant implications, particularly in the context of insurance pricing. Overestimating the ages of individuals with Down Syndrome may lead to **inflated insurance prices**, as older ages are often associated with increased risk and higher coverage costs. Such a bias in the algorithm's age estimation can exacerbate the financial burden on individuals and families already navigating the complexities of Down Syndrome.

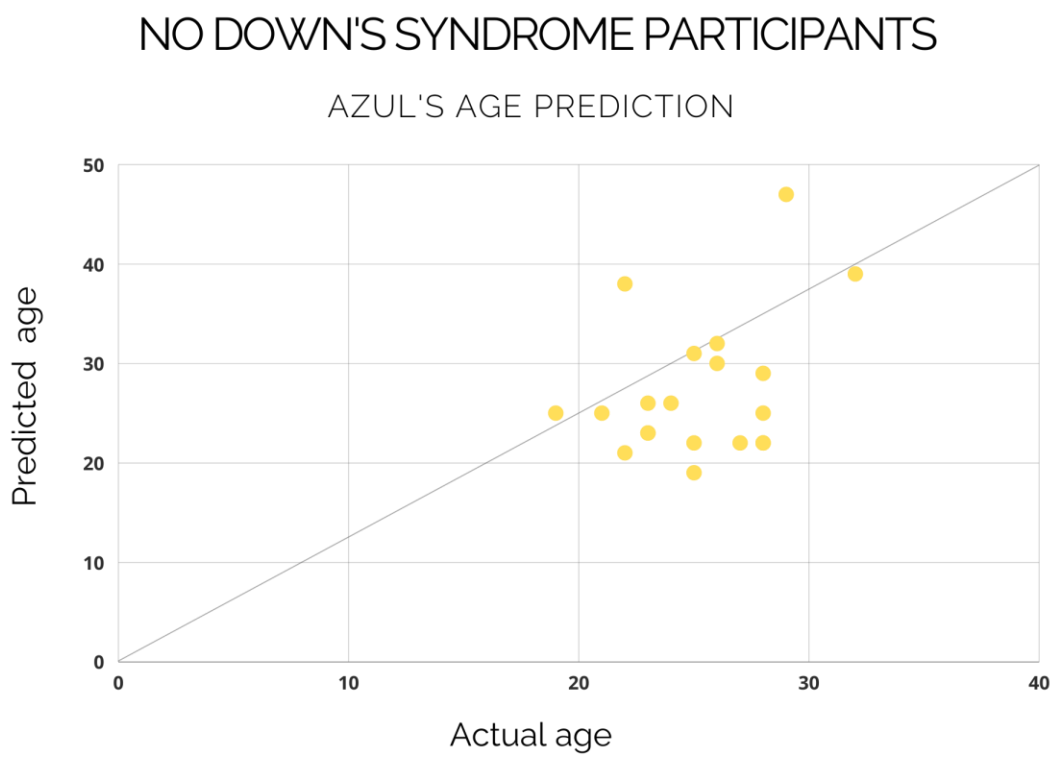


Figure 3. Age Prediction in Azul for No Down's Syndrome Individuals

In examining the data of individuals without Down Syndrome, we compared the predicted ages generated by the Azul algorithm to their actual ages. The actual ages of the participants ranged from 19 to 32 years, while the algorithm's predictions spanned from 19 to 47 years. Upon analysis, it became apparent that the Azul algorithm exhibited variations in accurately predicting the ages of individuals without Down Syndrome. **Deviations** between the predicted and actual ages were observed, with differences ranging from **-9 to +18 years**, indicating a level of inaccuracy in the algorithm's age estimation for this group.

The **error rate** in the age predictions made by the Azul algorithm for individuals without Down Syndrome was calculated to be **4.45%**. This error rate reflects the average difference between the predicted ages and the actual ages of the participants in our sample. With an average deviation of approximately 4.45 years, it is evident that the Azul algorithm's age predictions for individuals without Down Syndrome exhibit a degree of inaccuracy. Finally, our analysis indicates that the errors in age prediction for individuals without Down Syndrome were **relatively balanced**, with no significant bias towards overestimation or underestimation. This suggests a more favorable situation compared to the findings for individuals with Down Syndrome.

3.2.c (II) Body mass index (BMI)

In addition to analyzing the accuracy of age predictions, we also examined the Body Mass Index (BMI) of the participants. BMI is a crucial factor utilized by the Azul algorithm to calculate insurance prices. By comparing the predicted and actual BMI values for both individuals with Down Syndrome and those without, we gained valuable insights into the algorithm's performance and its potential impact on insurance pricing.

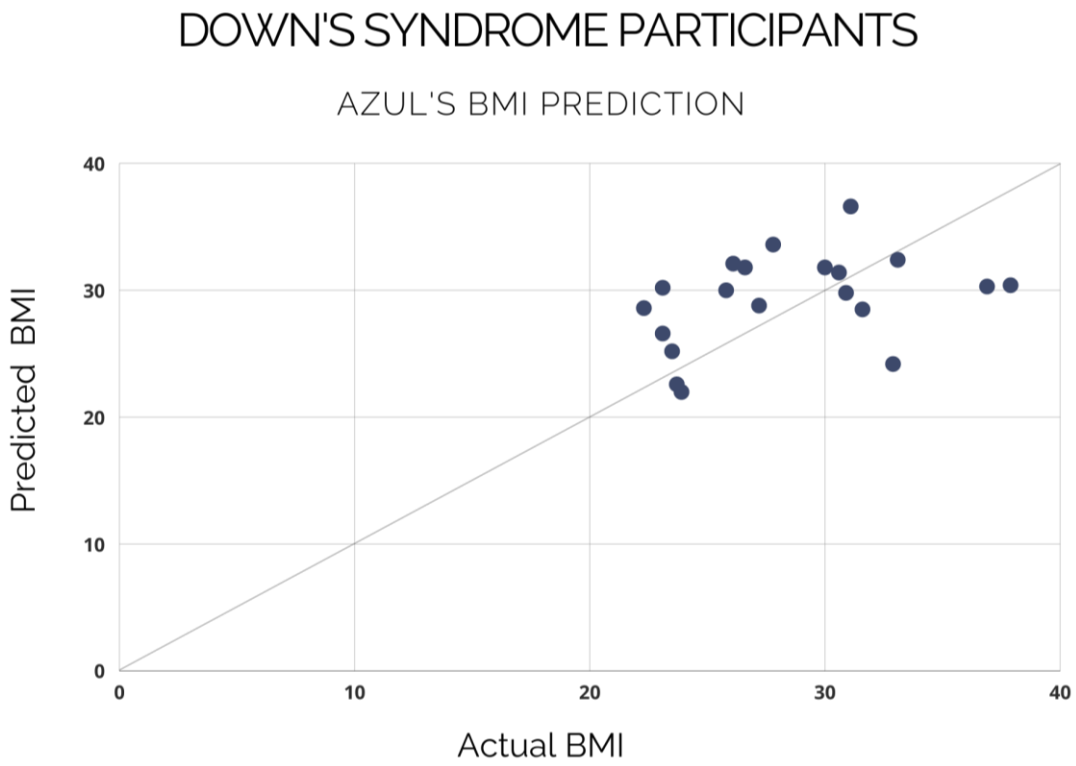


Figure 4, BMI Prediction in Azul for Down's Syndrome Individuals

Upon comparing the actual and predicted BMI values for individuals with Down Syndrome, we observed variations between the two sets of data. The actual BMI values ranged from 22.3 to 37.9, while the Azul algorithm predicted BMI values ranged from 22.0 to 36.6. Analyzing the error rate in BMI predictions, we calculated an average difference of 3.82 between the predicted and actual values. The average **error rate** of **3.82** suggests that, on average, the Azul algorithm's BMI predictions for individuals with Down Syndrome deviate from their actual BMI values by approximately 3.82 units. This level of error highlights the significant challenges in accurately estimating the BMI of individuals with Down Syndrome using the facial recognition technology employed by Azul. Similarly for the age, the error in BMI predictions for individuals with Down Syndrome by the Azul algorithm, we observed that the errors were more pronounced on the upside, indicating a **tendency to overestimate** the BMI values of participants. This systematic bias in the algorithm's BMI estimation can have implications for insurance pricing, as higher BMI values are often associated with increased health risks and potentially higher insurance premiums.

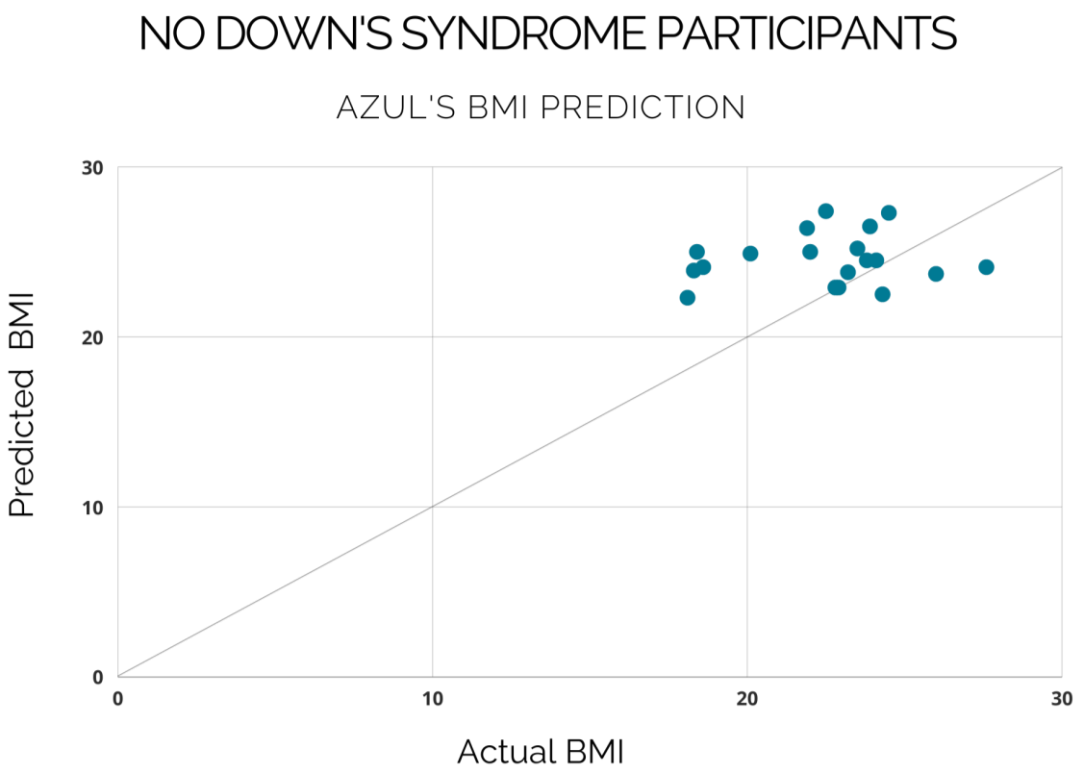


Figure 5, BMI Prediction in Azul for No Down's Syndrome Individuals

In our analysis of the BMI predictions for individuals without Down Syndrome using the Azul algorithm, we compared the predicted BMI values to their actual BMI values. The actual BMI values ranged from 18.1 to 27.6, while the predicted BMI values generated by the Azul algorithm ranged from 22.3 to 27.4. Upon examining the data, we found that the Azul algorithm demonstrated a moderate level of accuracy in predicting the BMI values for individuals without Down Syndrome. The **error rate** in BMI predictions for individuals without Down Syndrome by the Azul algorithm stood at **2.98**, indicating a relatively low average difference between the predicted and actual BMI values. Unlike the findings for individuals with Down Syndrome, the errors in BMI predictions for individuals without Down Syndrome were not skewed predominantly in one direction. The Azul algorithm demonstrated a **relatively balanced distribution** of errors, with both overestimations and underestimations. This suggests that the algorithm's BMI predictions for individuals without Down Syndrome were closer to the actual BMI values, compared to the predictions for individuals with Down Syndrome.

3.2.c (III) Gender disparity patterns

Although Azul's algorithm **does not explicitly include gender** as a factor for pricing, we have attempted to analyze the potential gender-related variations in age predictions using our dataset. Gender is an important aspect to consider in insurance pricing as it can influence risk factors and life expectancy. By examining the age predictions for men and women in our study, we aimed to gain insights into any potential gender-based variations in the algorithm's performance. Understanding how gender may impact age prediction can help uncover potential biases or inaccuracies in the algorithm and contribute to the development of more equitable and inclusive insurance pricing models.

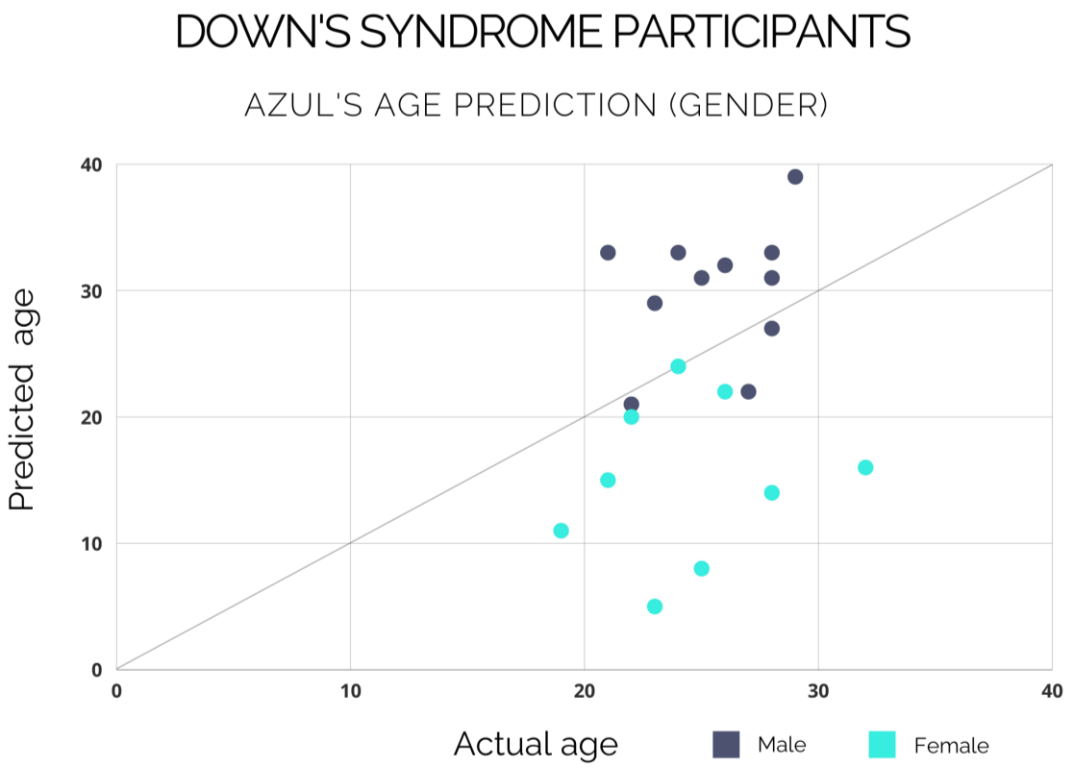


Figure 6, Age Prediction by Gender for Down's Syndrome Individuals

Our analysis of the Azul algorithm's age prediction reveals a concerning disparity between genders. Specifically, our findings indicate that **women tend to be underestimated** in terms of their age, while **men** are **more likely to be overestimated**. This gender-based discrepancy raises important questions about the fairness and accuracy of the algorithm's predictions. Examining the data more closely, we observe a consistent pattern of age underestimation for women and age overestimation for men across multiple instances.

For example, consider the case of **woman A**, whose actual age is 24, but the Azul algorithm predicts her age to be as low as 8 years. Similarly, **woman B**, with an actual age of 23, is predicted to be just 5 years old. These extreme cases highlight the severity of the age underestimation for women, leading to significant inaccuracies in the algorithm's predictions. In contrast, when looking at men in our sample, we see a clear pattern of age overestimation. For instance, **man X**, with an actual age of 28, is predicted by the algorithm to be 33 years old. Similarly, **man Y**, whose actual age is 21, is predicted to be 33 years old. While these examples demonstrate the trend of overestimating men's ages, they do not reach the same extreme levels as seen in the underestimation of women's ages.

From gender bias to ethical and legal deadlocks

The age underestimation observed in women, exemplified by extreme cases like woman A being predicted as 8 years old despite her actual age of 24, raises alarming concerns regarding the Azul algorithm's potential misclassification of **women as minors**. This has significant implications, particularly in insurance processes, as it creates a scenario where individuals who are legally considered minors could potentially complete the process. When the algorithm inaccurately predicts a woman's age, suggesting she is significantly younger than her actual age, it creates the **risk of allowing minors** to engage in age-restricted activities such as insurance procedures. This issue highlights the critical need for

accurate age estimation algorithms that can ensure compliance with legal regulations and prevent unintended consequences in various domains.

The age underestimation issue in the Azul algorithm, which potentially allows individuals who are legally considered minors to complete insurance processes, thus raises concerns from a strictly **legal standpoint**. In Spain, as in many other jurisdictions, there are specific regulations and laws in place to protect the rights and interests of minors. For instance, in Spain, the Civil Code (Código Civil) establishes that individuals under the age of 18 are considered minors and are subject to legal protection and limitations[3]. Allowing minors to engage in contractual agreements, such as insurance contracts, without proper legal oversight could potentially violate these regulations. Internationally, there are also legal frameworks that aim to protect minors and regulate their participation in various activities. The United Nations Convention on the Rights of the Child (UNCRC)[4] sets out specific provisions to safeguard the rights and well-being of children, emphasizing the need for their protection, proper representation, and informed consent. In the context of insurance processes, these legal principles underline the importance of accurate age verification and ensuring that minors are not exposed to potential risks or exploitation.

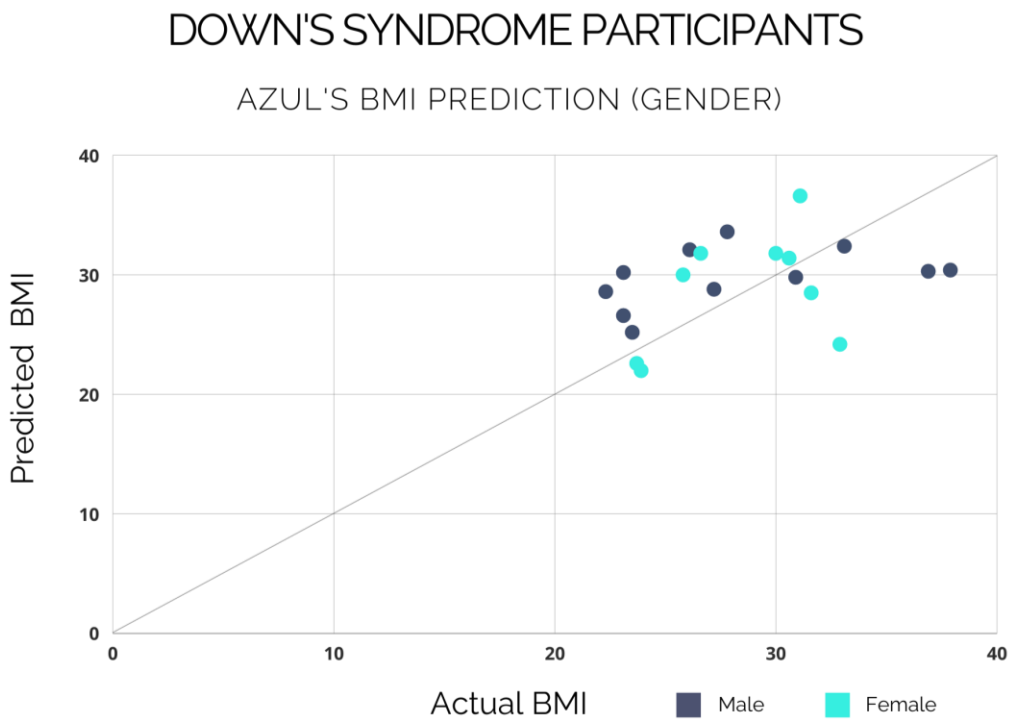


Figure 7, BMI Prediction by Gender for Down's Syndrome Individuals

Analyzing the BMI data for women and men, some interesting trends can be observed. For women, there appears to be a pattern of higher predicted BMIs compared to their actual BMIs. This trend is evident in several cases where the predicted BMI values are significantly higher than the actual values. For example, one woman's actual BMI is 26.6, but the algorithm predicts it to be 31.8, indicating an overestimation of the BMI. Similarly, another woman with an actual BMI of 23.7 is predicted to have a BMI of 22.6, reflecting a slight underestimation. On the other hand, for men, the trends in BMI prediction are relatively more varied. While some cases show a similar pattern of overestimation as seen in women, such as a predicted BMI of 30.4 for a man with an actual BMI of 37.9, there are also instances

[3] "*Legal age begins upon turning eighteen years old*" (Spanish Civil Code, art. 315).

[4] United Nations. (1989). Convention on the Rights of the Child, Nov. 20, 1989, 1577 U.N.T.S. 3.

where the algorithm underestimates the BMI. For example, a man with an actual BMI of 30.9 is predicted to have a BMI of 29.8, indicating a slight underestimation.

In comparison to the gender disparities observed in age prediction, the disparities in BMI estimation appear to be less pronounced. While there are instances where the Azul algorithm demonstrates deviations from the actual BMI values for both women and men, the magnitude of these deviations is generally smaller. Overall, these findings suggest that the Azul algorithm may exhibit a gender bias in BMI prediction, with a **tendency to overestimate BMI** values **for women** and a **more mixed pattern for men**. Further analysis and investigation are needed to understand the underlying factors contributing to these trends and to address any potential biases in the algorithm's BMI estimation for different gender groups.

3.2.d Summary

In our pursuit of inclusivity and equal opportunities, we embarked on a groundbreaking investigation into the effectiveness of Azul's facial recognition technology on individuals with Down Syndrome. Our findings are nothing short of eye-opening, revealing a stark reality that demands attention and action.

- **Age prediction disparity**.

  Azul's algorithm stumbled when attempting to accurately predict the ages of individuals with Down Syndrome. Deviations between predicted and actual ages reached as high as 21 years, exposing a substantial level of inaccuracy. This poses critical implications for insurance pricing, where misjudging age can lead to unfair premiums and financial burdens.

- **Gender disparity unveiled**.

  Our analysis uncovered a disturbing gender bias in age predictions. Women were consistently underestimated, with alarming cases of being predicted as young as 5 or 8 years old. In contrast, men experienced overestimation, further accentuating the disparity between genders. These findings expose a deep-seated gender bias within the algorithm, with far-reaching consequences for individuals with Down Syndrome.

- **BMI prediction challenges**.

  Azul's algorithm showed moderate accuracy in predicting Body Mass Index (BMI) for individuals with Down Syndrome. However, the algorithm's tendency to overestimate BMI values, particularly for women, raises concerns about fairness in insurance pricing. Higher predicted BMI values can lead to inflated premiums, placing an undue burden on individuals already navigating the complexities of Down Syndrome.

- **Unmasking gender disparity in BMI prediction**.

  Analyzing the data, we unearthed a striking trend of higher predicted BMIs for women compared to their actual values. This discrepancy, coupled with the algorithm's varied BMI predictions for men, reveals an unsettling gender bias within the technology. These biases demand immediate attention to ensure equitable and unbiased insurance pricing for individuals with Down Syndrome.

In the realm of facial recognition technology, our investigation into Azul's performance on individuals with Down Syndrome has shed light on a sobering reality. The findings are a wake-up call, demanding immediate attention and action. We have uncovered significant disparities and biases that cannot be ignored. The age predictions demonstrated a substantial level of inaccuracy, while the gender disparities unveiled a deeply rooted bias within the algorithm. Furthermore, the challenges in BMI prediction and the gender-related discrepancies raise profound concerns about fairness and equity in insurance pricing. It is, thus, imperative that we confront these issues head-on, rectify the biases, and strive for inclusive technologies that leave no one behind.

3.3 Analysis of the DeepFace Framework

To obtain a more comprehensive understanding of the impact of facial recognition (FR) models on individuals with disabilities, we conducted an analysis that extended beyond the scope of Azul's FR model. AI, as a transformative force, has profound societal implications, and FR technology is no exception. It is experiencing significant investment and growth, projected to reach a market volume of $12.67 billion by 2028, according to Statista (2022). The demand for FR spans various sectors such as security, surveillance, defense, industry, and services.

3.3.a Background

In our study, we aimed to investigate the potential biases associated with attributes like age, gender, emotion, and ethnic classification prediction in FR models. To achieve this, we conducted a **pilot study** utilizing the **DeepFace framework**—an extensive Python-based facial attribute analysis and recognition framework[5]. The choice of the DeepFace framework for our study was not arbitrary; it was driven by a careful evaluation of available options. We meticulously assessed numerous facial attribute analysis and recognition frameworks, considering their capabilities, reliability, and compatibility with our research objectives. Its integration of state-of-the-art models such as VGG-Face, Google FaceNet, OpenFace, and Facebook DeepFace, along with its compatibility with Python, made it the most comprehensive and reliable solution for analyzing attributes like age, gender, emotion, and ethnic classification.

By examining the impact of FR models on individuals with disabilities, we sought to gain deeper insights into the potential consequences and ensure that the development and application of these technologies are inclusive and unbiased.

Different models serve various purposes in the field of facial recognition, including the detection of individuals in images. However, when it comes to analyzing facial attributes such as gender, age, ethnicity, and emotion recognition, the study primarily focuses on the **VGG-Face** model which uses a customized Convolutional Neural Network (CNN).

Within the DeepFace framework developed by Serengil, S., VGG-Face serves as the foundational model for age, gender, and ethnicity classification. The author fine-tunes the VGG-Face model for each specific attribute using the weights of a pre-trained model, a technique known as transfer learning.

[5] To avoid misunderstandings, in this document we refer to the framework developed by Serengil S as DeepFace. It should not be confused with Facebook DeepFace or VGG-Face, also called DeepFace. The tool is publicly available at https://github.com/serengil/deepface

The VGG-Face model was developed by Oxford visual geometry group. In 2015, they announced its deep face recognition architecture. Even though research paper is named Deep Face, researchers give VGG-Face name to the model. This might be because Facebook researchers also called their face recognition system DeepFace – without blank. VGG-Face is deeper than Facebook's Deep Face, it has 22 layers and 37 deep units. Researchers fed 2.6 M images, from VGG-Face dataset, to tune the model weights. The model is originally trained for facial recognition task, achieving an accuracy of 98.78% for labeled faces in the wild dataset. LFW dataset contains 13K images of 5K people.

The structure of the VGG-Face model is demonstrated below.

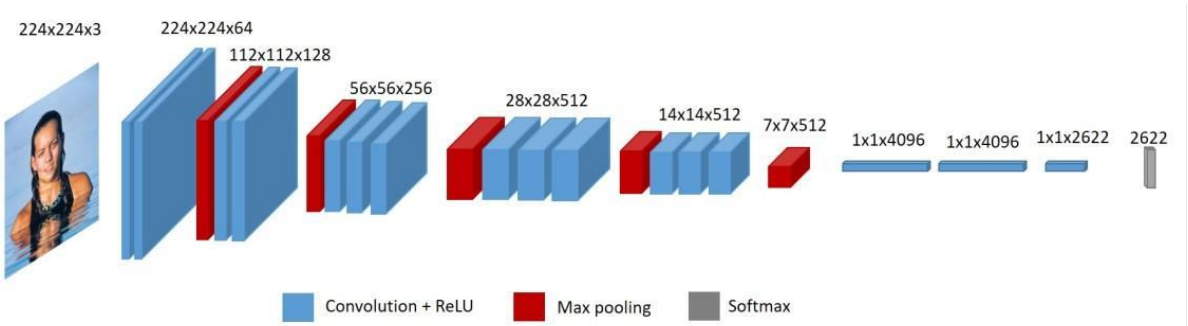


Figure 8, VGG-Face model

3.3.a (I) Datasets

The successful development and training of facial recognition models rely on high-quality datasets. Here, we discuss the datasets utilized for training various attribute analysis tasks within the pilot study.

VGG-Face dataset[6]

The VGG Face dataset plays a crucial role in advancing face recognition technology. It consists of 2.6 million face images belonging to 2,622 individuals. The dataset was developed with support from the United States Office of the Director of National Intelligence (ODNI) and the Intelligence Advanced Research Projects Activity (IARPA). Primarily comprised of celebrities, public figures, actors, and politicians, the dataset was curated by extracting popular male and female names from the Internet Movie Database (IMDb) celebrity list. Ethnicity, age, and kinship information was also collected from IMDb. VGG Face has been widely adopted by commercial, military, and academic organizations across the globe, contributing to numerous research projects.

FER-2013[7]

For the emotion recognition task within the DeepFace framework, the Facial Expression Recognition 2013 (FER-2013) dataset is employed. This dataset contains approximately 28,000 training images and 3,000 testing images depicting various emotion expressions, including happiness, neutral, sadness, anger, surprise, disgust, and fear. The images are categorized based on the expressed emotion and exhibit relatively centered faces occupying a similar amount of space. The dataset was curated by collecting images from Google searches that effectively represented each emotion category.

[6] https://www.robots.ox.ac.uk/~vgg/data/vgg_face/

[7] https://www.kaggle.com/datasets/msambare/fer2013

FairFace Dataset[8]

To fine-tune the ethnicity classification task, the FairFace dataset is utilized. This extensive dataset comprises 86,000 training instances and 11,000 test instances. Its primary objective is to ensure equal representation for each ethnic group. The images were sourced mainly from the YFCC-100M Flickr dataset, with additional contributions from Twitter and online newspapers. The dataset is labeled with race, gender, and age groups. Ethnicity labels include East Asian, Southeast Asian, Indian, Black, White, Middle Eastern, and Latino-Hispanic. However, for improved classification performance in the DeepFace framework, the author merges the East Asian and Southeast Asian races into a single Asian category.

IMDB-WIKI Dataset[9]

For the age and gender classification task, the IMDB-WIKI dataset is employed. This dataset comprises over 500,000 face images with associated age and gender labels. It includes 460,723 face images from 20,284 celebrities sourced from IMDb and an additional 62,328 images from Wikipedia. The age and gender label distributions within the dataset are visualized in Figure 3. Notably, the dataset exhibits gender imbalance, with male representation approximately double that of females. Additionally, the age distribution primarily centers around the range of 20 to 30 years old.

3.3.a (II) Classifier analysis

Within the DeepFace framework, the models employed are capable of recognizing individuals in various images and discerning their emotions. Interestingly, these models have already surpassed human accuracy levels of 97.53% in facial recognition tasks. However, the accuracy achieved varies depending on the specific classification task.

Emotion

Emotion detection plays a crucial role in the DeepFace framework, involving the categorization of facial expressions into seven basic emotional categories: angry, disgust, fear, happy, sad, surprise, and neutral. A Kaggle forum discussion, led by competition organizers, reported human accuracy on the FEC2013 dataset to be in the range of 65% to 68% (Khaireddin & Chen, 2021). In the framework, a VGG-Face CNN model is utilized to perform emotion detection tasks. When initially developed in 2018, this model achieved an accuracy of 57% on the test set, surpassing the previous highest accuracy achieved in a Kaggle challenge (34% accuracy). However, recent studies have made notable progress in this field. In 2021, Khaireddin & Chen introduced a VGGNet architecture that achieved an impressive accuracy of 73.28% on the FER2013 dataset, setting a new record for single-network accuracy without utilizing any additional training data. These advancements in the DeepFace framework demonstrate the significant improvements made in emotion detection, pushing the boundaries of accuracy and paving the way for further advancements in this essential area of facial analysis.

Gender

Gender classification is a significant task within the DeepFace framework, aiming to classify individuals as either male or female. The model utilized in this framework achieves an impressive accuracy value of 97.44%, accompanied by a precision rate of 96.29% and a recall rate of 95.05%. In a comprehensive review of AI methods for gender classification conducted by Garain et al. (2021), which encompassed diverse publicly available datasets,

[8] https://github.com/joojs/fairface

[9] https://www.kaggle.com/datasets/eabdul/imdbwikiimagedataset/code

this DeepFace model consistently emerges as one of the top performers in terms of accuracy.

Ethnicity

DeepFace incorporates a robust ethnicity classification task, categorizing pictures into six distinct ethnicities: Asian, Black, Indian, White, Middle Eastern, and Latino Hispanic. The model achieves a commendable accuracy of 68% on the test set. While contemporary state-of-the-art models surpass the accuracy of this particular model, it is important to consider the key differentiating factor. DeepFace's model deals with six diverse classes, whereas other models typically handle two to four classes. Notably, a single model identified in the study classifies five different ethnicities with exceptional accuracy, reaching an impressive 97.83% (Mohammad & Al-Ani, 2018). This study employed a CNN model that specifically focused on the desired Region of Interest (ROI), specifically the extended ocular region, derived from facial images within the standard FERET dataset. While the DeepFace ethnicity classification model does not currently attain the highest accuracy, the ability to handle a wider range of ethnicities showcases its inclusivity and recognition of diverse groups. This highlights the importance of considering the specific task requirements and the scope of ethnicities involved when assessing accuracy levels in ethnicity classification models.

Age

DeepFace's age model achieves an impressive Mean Absolute Error (MAE) value of ±4.65[10]. The author notes that this value is remarkably close to human-level accuracy in age prediction. In a comprehensive review of state-of-the-art AI methods for age estimation conducted by Garain et al. (2021), despite the utilization of different publicly available datasets, it is evident that the DeepFace model consistently delivers one of the best accuracy results. The remarkable accuracy achieved by the DeepFace age model demonstrates its effectiveness in accurately estimating age based on facial features. With an MAE value comparable to human-level predictions, the DeepFace framework showcases its ability to contribute to advancements in age estimation and its potential applications in various domains.

3.3.a (III) Dataset testing

To examine fairness in AI computer vision systems for individuals with Down Syndrome (DS), **two** distinct test **datasets** were utilized. The *first* test dataset consisted of **60 images** featuring male and female subjects with **DS**, ranging in age from 4 to 57 years old. The *second* test dataset included **60 images** of famous individuals without Down Syndrome (**no DS**), spanning ages between 17 and 73. Notably, the no DS dataset comprised images of renowned actors, politicians, singers, astronauts, and other notable figures.

Both test datasets were sourced from the Internet, ensuring ethnic and gender balance. Each dataset comprised 30 males and 30 females, with ten subjects from each ethnicity. Figure 4 illustrates the diverse age distributions observed in the DS and no DS test datasets. Furthermore, within the DS dataset, 35 subjects exhibited a happy expression while 25

[10] The mean absolute error (MAE) is defined as the average variance between the significant values in the dataset and the projected values in the same dataset (Manoj et. al., 2022).

displayed a neutral expression. In contrast, the no DS dataset included 32 subjects with a happy expression and 28 subjects with a neutral expression.

The retrieved images featured single individuals and varied in terms of facial orientation, ranging from front-facing and centered images to non-frontal, non-centered, or even whole person images. The DeepFace framework was capable of analyzing these diverse image characteristics, including different backgrounds, poses, expressions, and lighting conditions. Notably, no correlation was observed between these image characteristics and misclassification. Some images with extreme lighting conditions, pronounced head tilts, or highly turned faces, among other attributes, proved challenging for DeepFace analysis.

Manual labeling was performed for the images, with ethnicity labels derived from the search terms used during image retrieval. Age information was sourced from various reputable sources such as Wikipedia, articles, and databases like Wikiwand or IMDb. While most images within the DS test dataset had age information available on the Internet, a few images did not. Gender and emotion labels were assigned based on the visual appearance depicted in the photographs.

3.3.a (IV) Evaluation metrics

In this section, we discuss the evaluation methodology used to assess fairness in the DeepFace framework. Various definitions of fairness exist in the literature, as highlighted in the notable study "Fairness Definition Explained" by Verma & Rubin (2018). In this study, we define fairness as achieving **equal performance across different variables**, ensuring that performance is independent of race, gender, ethnicity, emotion, and genetic conditions. This definition translates into obtaining equitable evaluation values for each group within each variable. To evaluate the performance, we employ different evaluation metrics.

To examine the performance of DeepFace concerning the Down Syndrome condition, we utilize the following evaluation metrics:

- *Gender, emotion, and race variables*: We assess the classifier's performance using accuracy, precision, and recall values. Accuracy measures the overall correctness of the classifier's predictions, while precision quantifies the proportion of correctly predicted instances within a specific class. Recall, on the other hand, measures the proportion of actual positive instances correctly identified by the classifier. These metrics are calculated for both the overall variable and each specific class within the variable.
- *Age variable*: We calculate the Mean Absolute Error (MAE) as a measure of performance. MAE calculates the average magnitude of the errors between the predicted and true age values. By examining the MAE, we can assess the accuracy and precision of the classifier's age predictions. A lower MAE indicates a closer match between the predicted and true age, signifying better performance in age estimation.

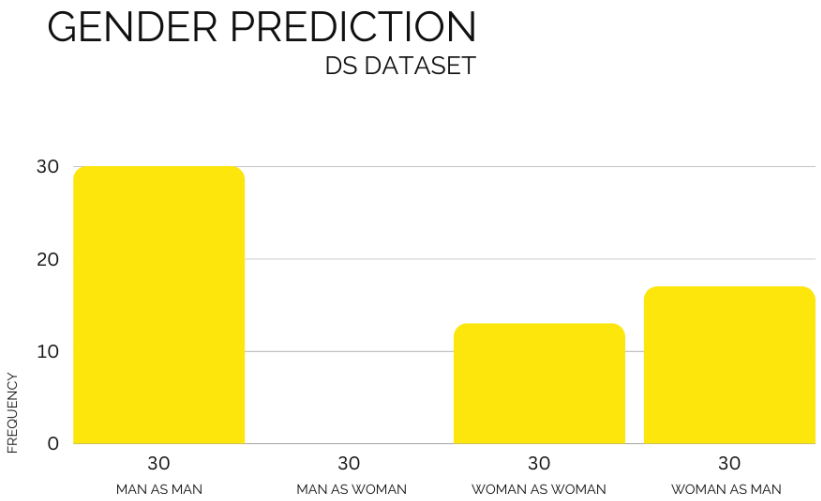
To analyze potential biases related to "gender & race" and "gender & emotion," we employ **conditional probabilities**. These probabilities determine the likelihood of incorrect classification based on race or emotion, irrespective of the individual's genetic condition. We further investigate the presence of such biases in Down Syndrome subjects by differentiating between DS women and no DS women, as well as DS men and no DS men.

3.3.b Findings

The following analysis presents key findings from evaluating the performance of the DeepFace framework in various classification tasks, including gender, emotion, ethnicity, and age prediction. DeepFace, a widely recognized and utilized classifier, has been examined using two distinct datasets: one containing individuals with Down Syndrome (DS) and another comprising famous individuals without Down Syndrome (no DS). By scrutinizing accuracy values, confusion matrices, conditional probabilities, and equalized odds measures, we gain valuable insights into the classifier's performance and the presence of biases across different demographic groups. These findings shed light on the challenges and opportunities for enhancing the fairness and accuracy of AI computer vision systems for individuals with DS, aiming to promote equitable and inclusive technological solutions.

- **Gender classification**.

  The gender classification performance in DeepFace demonstrates some disparities when considering individuals with Down Syndrome. The accuracy achieved in the DS dataset was 0.717, which is lower than the reported accuracy in the no DS dataset (0.974). Further analysis revealed that the **misclassification** was **predominantly** observed in **women**, with a recall of 43.3% for DS women compared to 80% for no DS women. However, the classification of men showed a 100% recall in both datasets, indicating consistent accuracy. These results highlight the need for improved gender classification performance for individuals with Down Syndrome, particularly in the accurate identification of women.



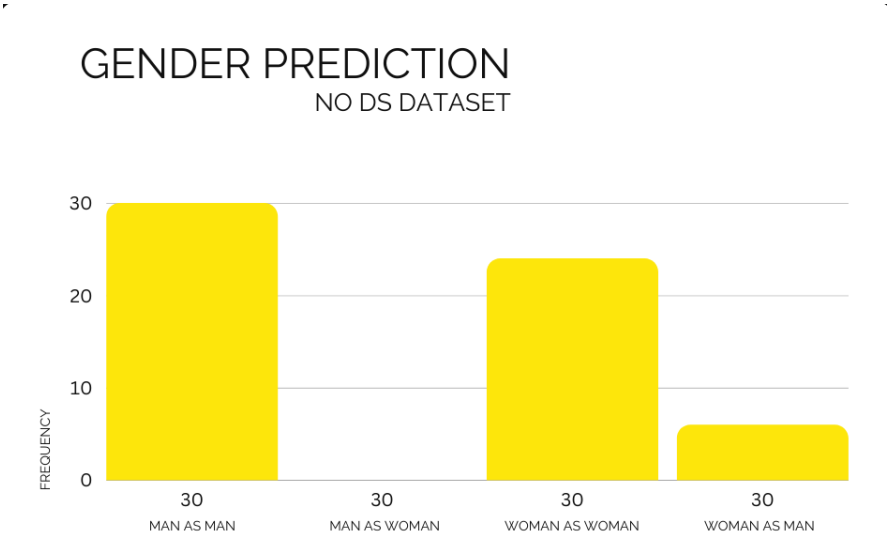


Figure 9 & 10, Gender prediction in both DS and No DS datasets

- **Emotion classification**.

DeepFace's performance in emotion classification exhibited similar accuracies in both the DS and no DS datasets, with values of 0.567 and 0.583, respectively. To assess the degree of misclassification, the mean confidence values for the true label were calculated. The obtained mean accuracy confidence values were 8.052 in the DS dataset and 13.193 in the no DS dataset, indicating a considerable margin for improvement. These findings emphasize the need for enhanced precision in emotion classification to ensure more accurate and reliable results.

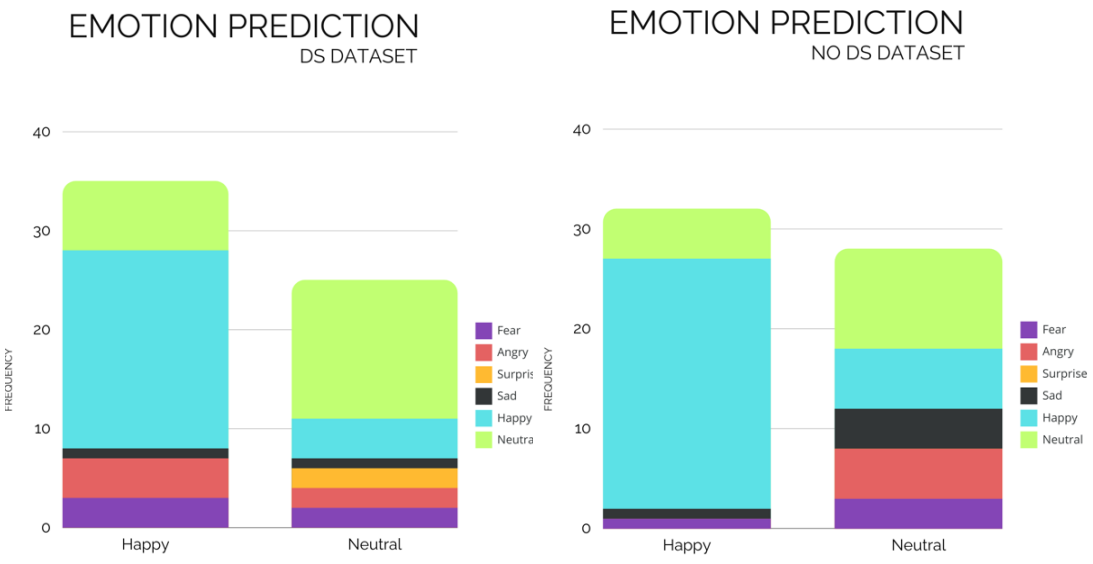


Figure 11 & 12, Emotion prediction in both DS and No DS datasets

- **Ethnicity classification**.

  In the DS dataset, the **misclassifications** were most **prominent** in the **Asian** and **white** ethnicity categories. Additionally, the mean confidence values for the true label in misclassified cases were 12.09 for the DS dataset and 14.487 for the no DS dataset, indicating substantial room for improvement. These results highlight the challenges of accurately classifying ethnicity in individuals with Down Syndrome and emphasize the importance of addressing bias and improving accuracy in this domain.

Figure 13 & 14, Ethnicity prediction in both DS and No DS datasets

- **Age prediction**.

  The age prediction performance in DeepFace yielded a significantly higher mean absolute error (MAE) in both the DS and no DS datasets compared to the reported MAE. The DS dataset obtained a MAE value of ±10.583, the no DS dataset obtained ±9.167, while the DeepFace model achieved a MAE of ±4.65. Despite the different age distributions, the age predictions in both datasets exhibited similar distributions, with the majority falling within the range of 26-28 years to 34 years. These findings suggest the presence of bias in the training dataset, which predominantly represents ages around 20 and 30. Enhancing age prediction accuracy, particularly for individuals with Down Syndrome, is crucial to ensure reliable and precise results.

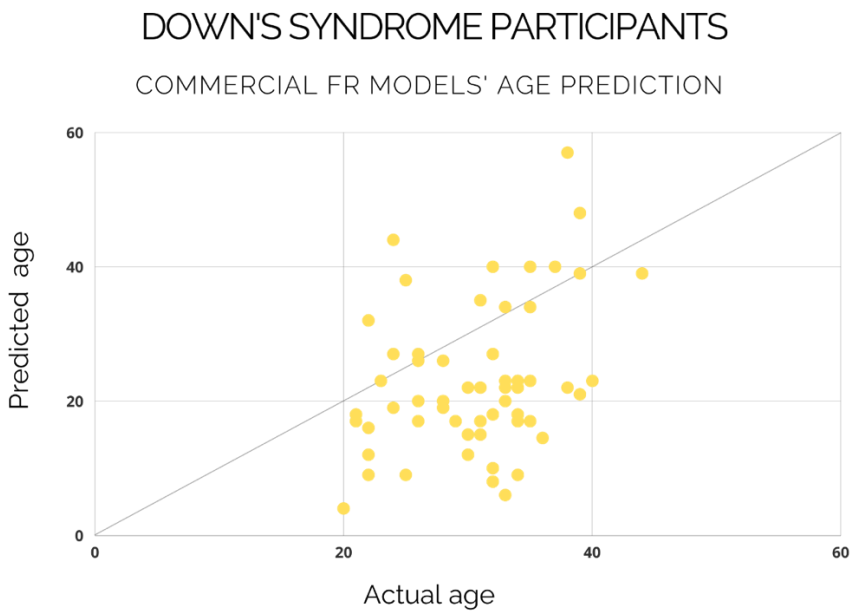


Figure 15, Age prediction in DS dataset

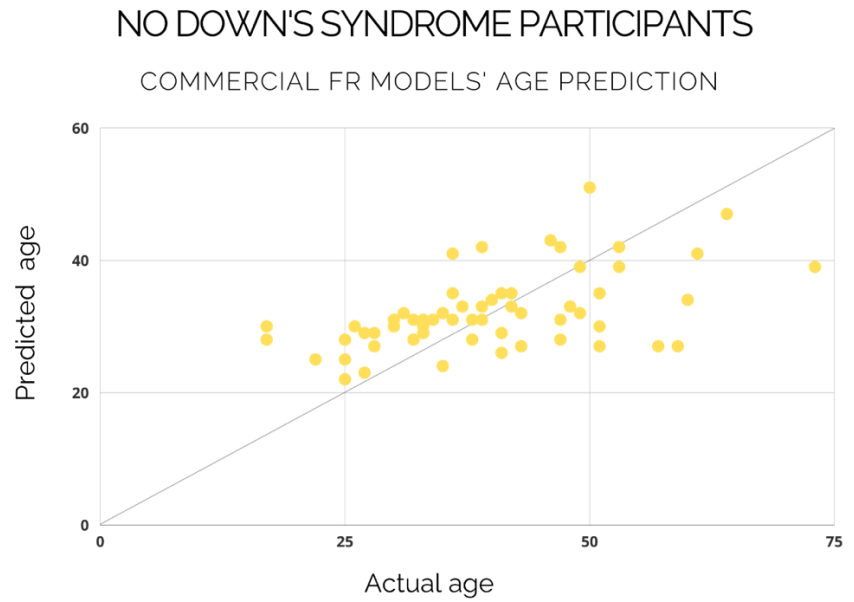


Figure 16, Age prediction in No DS dataset

4. Conclusion

Exposing the harsh reality: Facial Recognition's unreliability

In the vast landscape of facial recognition technology, there exists a profound void in our understanding of its impact on individuals with disabilities. This crucial intersection, both in academic research and industry practices, remains largely unexplored, shrouded in a veil of neglect and oversight. It is within this uncharted territory that we ventured, driven by a relentless pursuit to expose the harsh reality and illuminate the hidden truths. Our investigation delved deep into the untapped potential of facial recognition technology in relation to disabilities, unearthing a host of disparities, biases, and ethical dilemmas that have long been ignored. By filling this void, we aim to ignite a much-needed discourse, inspiring academia and the industry to confront this neglected frontier head-on.

The comparison between Azul and DeepFace has brought to the forefront the intricate challenges surrounding **age prediction** for individuals with Down Syndrome. Azul, as previously discussed, exhibited deviations of up to 21 years between predicted and actual ages. Such discrepancies have critical implications, particularly in insurance pricing, where misjudging age can lead to unfair premiums and financial burdens. Shifting our focus to DeepFace, the evaluation of age prediction for Down Syndrome participants revealed a similarly challenging landscape. The algorithm demonstrated significant deviations between predicted and actual ages, spanning from -16 to +23 years. These disparities parallel the difficulties encountered by Azul, highlighting the intricate nature of age prediction for individuals with Down Syndrome.

Our investigation has also exposed significant **gender-related** biases that have profound implications for individuals with Down Syndrome. In the Azul's FR system, we uncovered a concerning pattern of age underestimation for women and age overestimation for men. In contrast, DeepFace's gender classification task achieved an impressive accuracy of 97.44%, demonstrating its effectiveness in identifying male and female individuals. The contrast between Azul and DeepFace highlights the need for further exploration and research in this domain. Underestimating women's ages and overestimating men's ages can result in incorrect categorization, potentially allowing individuals who are legally considered minors to complete insurance processes. This raises concerns from a legal standpoint and emphasizes the need to protect the rights and interests of minors.

In addition, our investigation also delved into the Body Mass Index (**BMI**) estimation capabilities of the Azul algorithm for individuals with Down Syndrome. The analysis revealed notable challenges and gender-related biases in BMI prediction, highlighting implications for insurance pricing and fairness. Women experienced higher predicted BMIs compared to their actual values, indicating a bias towards overestimation. Finally, our pilot study utilizing the DeepFace framework revealed crucial insights. **Emotion** classification called for enhanced precision to ensure accurate and reliable results, while accurately classifying **ethnicity** posed challenges, particularly for Asian and white Down Syndrome individuals.

5. Recommendations

Breaking barriers: rethinking technology for disabled users

Amidst a diverse global population, individuals with disabilities, accounting for 16 percent, call for greater inclusivity in our technological advancements, seeking genuine social inclusion and equal opportunities. The integration of artificial intelligence (AI) and facial recognition (FR) technologies reshapes society but brings ethical implications, perpetuating biases that deepen social disparities. Examining AI's intersection with disability uncovers challenges, including biases in biometric systems like FR. The lack of research on Down Syndrome and AI bias reveals the need for broader inclusivity understanding. Resonating with the urgency for change, exploratory interviews echo the call for ethical design in AI systems. Unintentional biases during "self-learning" and intentional gaps in system design exemplify the risks faced by individuals with disabilities. Our investigation on Azul and commercial FR models has revealed significant disparities and biases in the performance of facial recognition models on Down Syndrome participants and individuals with disabilities. After unveiling these impactful insights, it becomes evident that our quest for inclusion and fairness must extend beyond mere observations. To create a future that truly breaks barriers, we propose the following set of recommendations:

- ★ Given the pervasive challenges and biases uncovered in facial recognition technology, it is imperative for stakeholders in the technology industry, regulatory bodies, researchers, and society as a whole to embark on a comprehensive **reevaluation** of its **suitability as the best technological tool available**. The integration of artificial intelligence and facial recognition technologies has sparked significant ethical concerns, leading us to question whether this technology truly upholds principles of fairness, inclusivity, and social equality. In light of its unintended biases and potential risks, we must provoke an open and critical discourse to determine if facial recognition technology is truly the most suitable and reliable tool, for individuals **with and without** Down Syndrome.

- ★ In their pursuit of ethical practices, Azul and the other commercial facial recognition models must wholeheartedly adopt a comprehensive and transparent **bias mitigation strategy**. It is crucial to prioritize clarity and transparency in the inner workings of the algorithm, ensuring that all stakeholders, including disabled users, can easily comprehend how the technology functions. This can be achieved by employing interpretable AI methods, explainable machine learning techniques, and bias detection tools that shed light on potential sources of bias and discrimination within the system.

- ★ All existing facial recognition models, including Azul, must prioritize **Accessibility by Design**. Universal design principles should be at the core of their development process to ensure that the platforms are accessible to users of diverse abilities from the outset. By proactively addressing accessibility needs and eliminating barriers, these facial recognition models can create transformative and empowering experiences for disabled individuals. This commitment to accessibility will not only enhance user experiences for a broader audience but also foster a more inclusive society, where technology is truly accessible and beneficial to everyone. By setting higher standards for accessibility, these models can lead the way in promoting a more equitable and inclusive future for facial recognition technology.

- ★ For all companies employing facial recognition (FR) models, a critical step towards ethical AI involves amplifying **disability advocacy** by collaborating with disability organizations and experts. By working together, these models can address the unique challenges faced by disabled individuals and champion the importance of ethical AI. Through collective efforts, Azul and other facial recognition models can become pioneers in developing technology solutions that positively impact the lives of disabled users worldwide. This collaboration will not only ensure that facial recognition technology is inclusive and accessible but also drive meaningful change and progress towards a more equitable and inclusive future for all.

- ★ To ensure fairness and accuracy in age prediction for individuals with Down Syndrome, Azul's algorithm must be improved. Implement thorough **recalibration** and **validation** processes to minimize deviations between predicted and actual ages. Address the gender bias in age and body mass index (BMI) predictions by **retraining** the algorithm with a **more diverse and representative dataset**, ensuring accurate age estimations for both men and women.

- ★ To address the issue of age underestimation in women and potential misclassification of minors, Azul's algorithm must prioritize accurate age estimation to comply with ethical and legal regulations. **Implement** rigorous **age verification mechanisms** to prevent minors from engaging in age-restricted activities, including insurance procedures. Align the algorithm with specific legal frameworks, such as the Civil Code in Spain and the UNCRC internationally, to safeguard the rights and well-being of children.

- ★ To uphold the principles of inclusivity and fairness, we propose redefining the consent process for disabled participants in technology evaluations, such as Azul's procedure. Current practices, relying on participants to proactively seek information in the company's privacy policy, fall short in ensuring explicit and informed consent, particularly for disabled individuals facing challenges in understanding complex information. Instead, a transparent and comprehensible consent-gathering process should be implemented, presenting data processing details clearly and conspicuously.

- ★ Recognizing the significance of accountability in the era of AI, Azul and the other commercial facial recognition models should initiate regular **third-party audits**, in line with the requirements set forth in Article 37 of the Digital Services Act (DSA), as a cornerstone of its commitment to responsible AI development. These audits should encompass an in-depth evaluation of the technology's data collection, training, and decision-making processes. The audits will serve as a proactive measure to identify and address potential biases and inaccuracies, especially in age prediction, gender disparities, and body mass index estimation for individuals with Down Syndrome. Moreover, by conducting regular and transparent audits, Azul can demonstrate its dedication to transparency, user protection, and inclusivity, paving the way for more equitable and unbiased facial recognition technologies.

- ★ Finally, to experts, academia, and industry leaders, it is imperative to prioritize and invest in **more research** on the **intersection of AI and disability**. By dedicating resources and attention to this field, we can better understand and address the unique challenges faced by disabled individuals in the context of AI technologies. This research should encompass a wide range of perspectives, including input from disabled users, disability advocates, and experts in disability studies. By including disability as a central theme in AI research, we can identify potential biases, discriminatory practices, and gaps in accessibility that may otherwise go unnoticed.

Acknowledgments

ETICAS extends its heartfelt gratitude to all the interviewees who generously shared their valuable insights and expertise. Their contributions have been instrumental in shedding light on the complex intersection of AI and disability, enriching our understanding of the challenges faced by disabled individuals in the digital age. Furthermore, we express our deepest appreciation to Cedown Jerez, a prominent organization dedicated to supporting and advocating for the rights of people with Down Syndrome. Their unwavering commitment to empowering individuals with disabilities and promoting inclusivity has been a driving force behind our research.

Project team: Adversarial Audits

Project Lead & Research Director: Dr. Gemma Galdon-Clavell, Founder of Eticas

Researchers:
- Matteo Mastracci, Team Leader at Eticas
- Miguel Azores, Ethics and Technology Researcher at Eticas

Contributors:
- Luis Rodrigo González Vizuet, Ethics and Technology Researcher at Eticas
- Iliyana Nalbantova, Jr. Ethics and Technology Researcher at Eticas
- Fran Segarra, Jr. Ethics and Technology Researcher at Eticas
- Isabela Miranda, Project Manager at Eticas
- Sam Danello, Fundraising Manager at Eticas
- Patricia Vázquez, Comms, Marketing, and PR Manager at Eticas
- Mireia Orra, Policy and Community Manager at Eticas

References

Agbolade, O., Nazri, A., Yaakob, R., Ghani, A.A., & Cheah, Y.K. (2020). Down syndrome face recognition: A review. Symmetry, 12(7), 1182. https://doi.org/10.3390/sym12071182

Angelino, C., Priolo, M., & Sánchez, C. (2011). Discapacidad y exclusión. La oculta presencia de la ideología de la normalidad. Políticas Educativas - PolEd, 1(2). https://seer.ufrgs.br/index.php/Poled/article/view/18312

Bandy, J. (2021). Problematic Machine Behavior. Proceedings of the ACM on Human- Computer Interaction, 5(CSCW1), 1–34. https://doi.org/10.1145/3449148

Boichenko, M. I. (2021). Human Evolution: the Limits of Technocentrism. Anthropological Measurements of Philosophical Research, (19), 15–22. https://doi.org/10.15802/ampr.v0i19.235956

Brisenden, S. (1986). Independent living and the medical model of disability. Disability, Handicap & Society, 1(2), 173-178. https://doi.org/10.1080/02674648666780171

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency (pp. 77-91). Proceedings of Machine Learning Research, 81. Retrieved from https://proceedings.mlr.press/v81/buolamwini18a.html

Ferreira, MA, and Díaz Velázquez, E. (2008). Disability, social exclusion and information technologies. Politics and Society, 46(1), 237-253.

Garain, A., Ray, B., Singh, P. K., Ahmadian, A., Senu, N., & Sarkar, R. (2021). GRA_Net: A deep learning model for classification of age and gender from facial images. IEEE Access, 9, 85672-85689. https://doi.org/10.1109/ACCESS.2021.3085971

Havercamp, S. M., Tassé, M. J., Navas, P., Benson, B. A., Allain, D., & Manickam, K. (2017). Exploring the Weight and Health Status of Adults with Down Syndrome. Journal of Education and Training Studies, 5(6). https://doi.org/10.11114/jets.v5i6.2343

Innerarity, D., (2020). The impact of artificial intelligence on democracy. Magazine of the Cortes Generales. 87-103. https://doi.org/10.33426/rcg/2020/109/1526.

Khaireddin, Y., & Chen, Z. (2021). Facial emotion recognition: State of the art performance on FER2013. arXiv preprint arXiv:2105.03588.

Jaume-Palasi, L. (2019). Why We Are Failing to Understand the Societal Impact of Artificial Intelligence. Social Research: An International Quarterly, 86(2), 477-498. doi:10.1353/sor.2019.0023

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389-399. https://doi.org/10.1038/s42256-019-0088-2

Manoj, S.O., Ananth, J.P., Rohini, M., Dhanka, B., Pooranam, N., & Arumugam, S.R. (2022). FWS-DL: Forecasting wind speed based on deep learning algorithms. In Artificial intelligence for renewable energy systems (pp. 353-374). Woodhead Publishing.

Mohammad, A. S., & Al-Ani, J. A. (2018). Convolutional Neural Network for Ethnicity Classification using Ocular Region in Mobile Environment. In 2018 10th Computer Science and Electronic Engineering (CEEC) (pp. 293-298). Colchester, UK. doi: 10.1109/CEEC.2018.8674194

Mordini, E., & Massari, S. (2008). Body, biometrics and identity. Bioethics, 22(9), 488-498. https://doi.org/10.1111/j.1467-8519.2008.00700.x

Palacios, A., & Romañach, J. (2006). Model of diversity. Bioethics and human rights as tools to achieve full dignity in functional diversity. In *Various editions*.

Raji, I. D., & Buolamwini, J. (2019). Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19) (pp. 429-435). Association for Computing Machinery. https://doi.org/10.1145/3306618.3314244

Rhue, L. (2018). Racial influence on automated perceptions of emotions (Working Paper). Social Science Research Network. https://ssrn.com/abstract=3281765

Rubin, S. S., Rimmer, J. H., Chicoine, B., Braddock, D., & McGuire, D. E. (1998). Overweight prevalence in persons with Down syndrome. Mental Retardation, 36(3), 175–181. https://doi.org/10.1352/0047-6765(1998)036

Strauss, A., & Corbin, J. (2002). Bases de la investigación cualitativa. Técnicas y procedimientos para desarrollar la teoría fundamentada [Bases of qualitative research: Techniques and procedures to develop grounded theory]. Universidad de Antioquia. https://diversidadlocal.files.wordpress.com/2012/09/bases-investigacion-cualitativa.pdf

Trewin, S. (2018). AI fairness for people with disabilities: Point of view. arXiv preprint arXiv:1811.10670.

United Nations. (2019). Report of the special rapporteur on extreme poverty and human rights. Official Documents System of the United Nations. https://documents-dds-ny.un.org/doc/UNDOC/GEN/N19/312/16/PDF/N1931216.pdf?OpenElement

Verma, S., & Rubin, J. (2018). Fairness definitions explained. FairWare '18: Proceedings of the International Workshop on Software Fairness, 1–7. https://doi.org/10.1145/3194770.3194776

Wise, P. (2012). Emerging technologies and their impact on disability. The Future of Children, 22(1), 169-191. http://www.jstor.org/stable/41475

info@eticas.tech

+34 936 005 400

Mir Geribert, 8, 3rd
08014, Barcelona