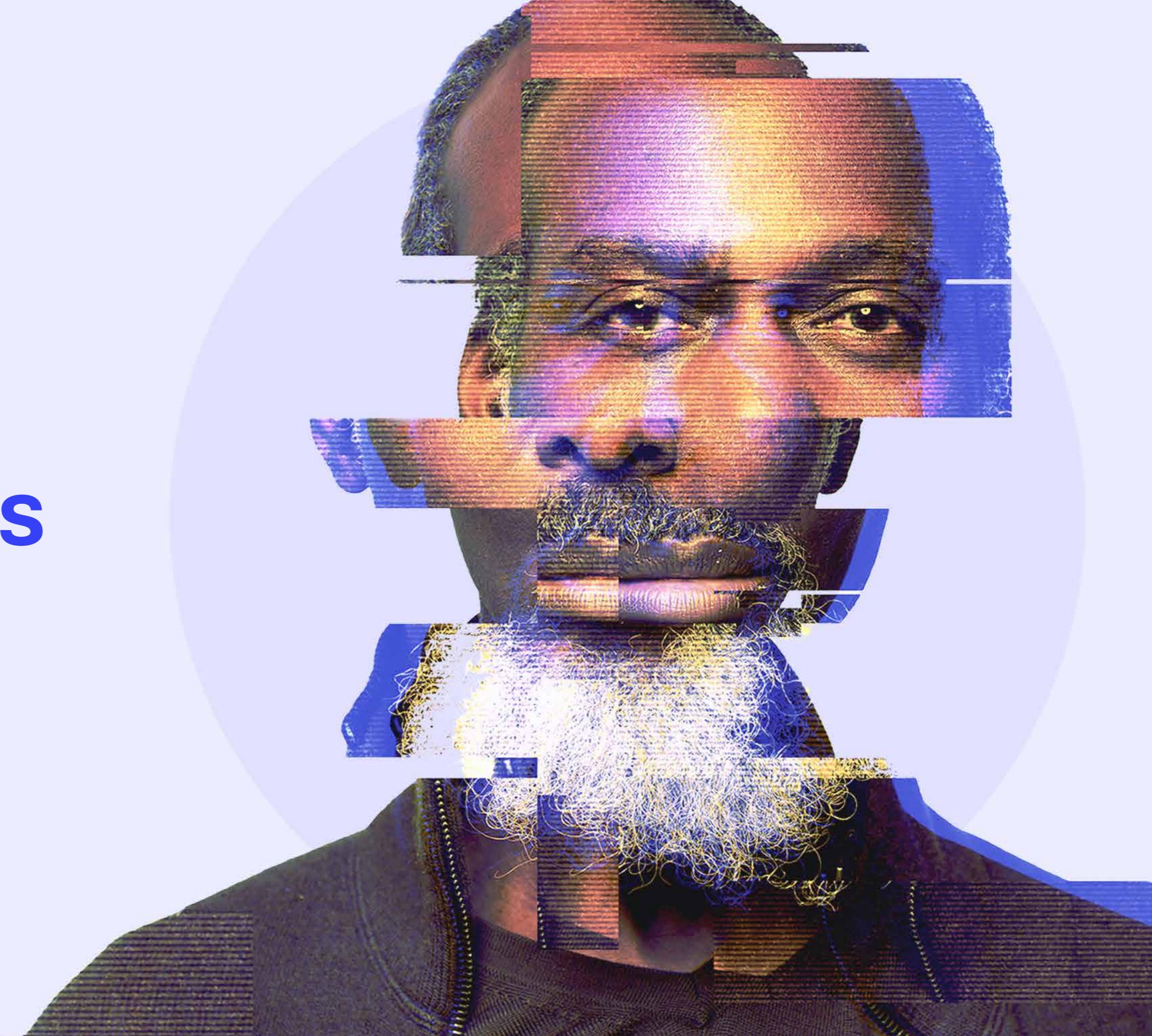


Building AI without bias

Reducing bias in biometrics



Introduction

Biometric analysis should operate to the same standard for everyone.

Artificial intelligence (AI) enables the digital age. It automates tasks at unimaginable scale, so we can have everything, at any time, lightning fast. But, AI can perform better for some than others, particularly when applied to biometrics. In other words, it can be biased.

The problem of bias did not arise with AI and automation. Human processes are equally vulnerable to bias. But AI allows biases to be amplified. An individual bank employee assessing credit applications can be biased, but they are only able to process a relatively low number of applications. A biased algorithm could process thousands of times more applications, and impact thousands more lives.

At Onfido, one method we use to verify identity is AI-powered biometric

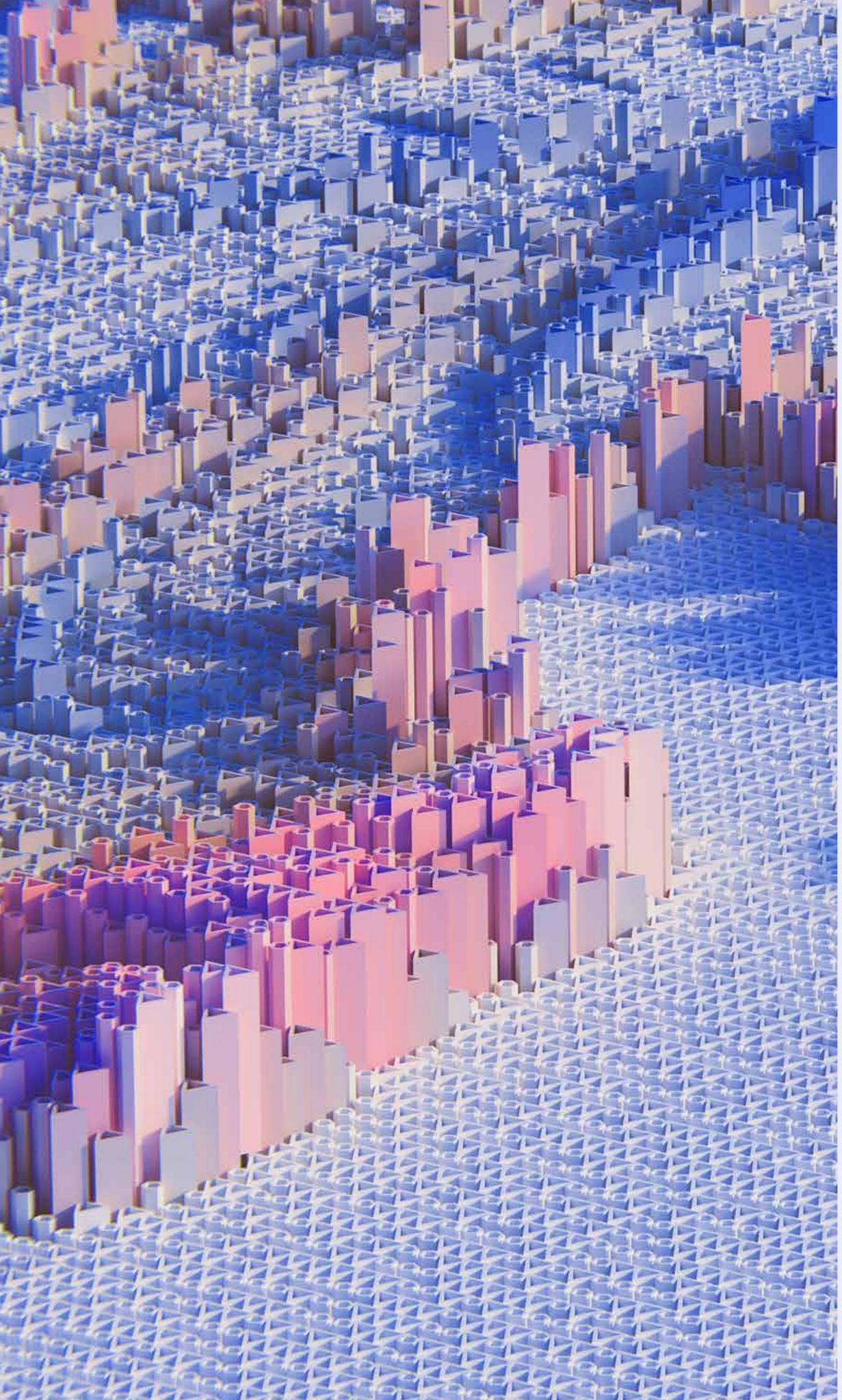
analysis. It creates trust between businesses and their customers — so they can be remotely onboarded. Biometric verification is becoming increasingly popular — 76.7% of users find it convenient and 82.8% find it secure¹. For businesses, it offers high-assurance in the face of identity fraud, which has increased 41% since 2020².

But when biometric analysis is being used to grant access to services — it should operate to the same standard for everyone.

So what are we doing to make AI ethical? This whitepaper offers guidance based on our experience of defining, measuring, and mitigating biometric bias, and describes our experience executing bias mitigation in our next-generation biometric solution, Onfido Motion.

¹ Onfido, [Identity Fraud Report 2022](#) (2022).

² Onfido, [Digital by Default](#) (2021).



Contents

01

Measuring bias in biometric products

Why defining bias, segmenting data, and defining measurements is critical.

02

Strategies for bias mitigation

How data preparation, training, and post-processing strategies can mitigate bias, and how they are used at Onfido.

03

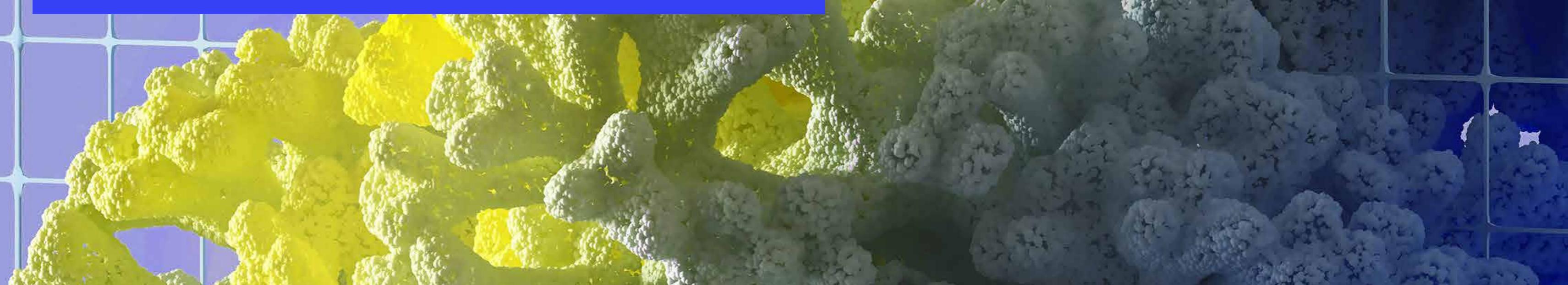
Case study: Onfido Motion

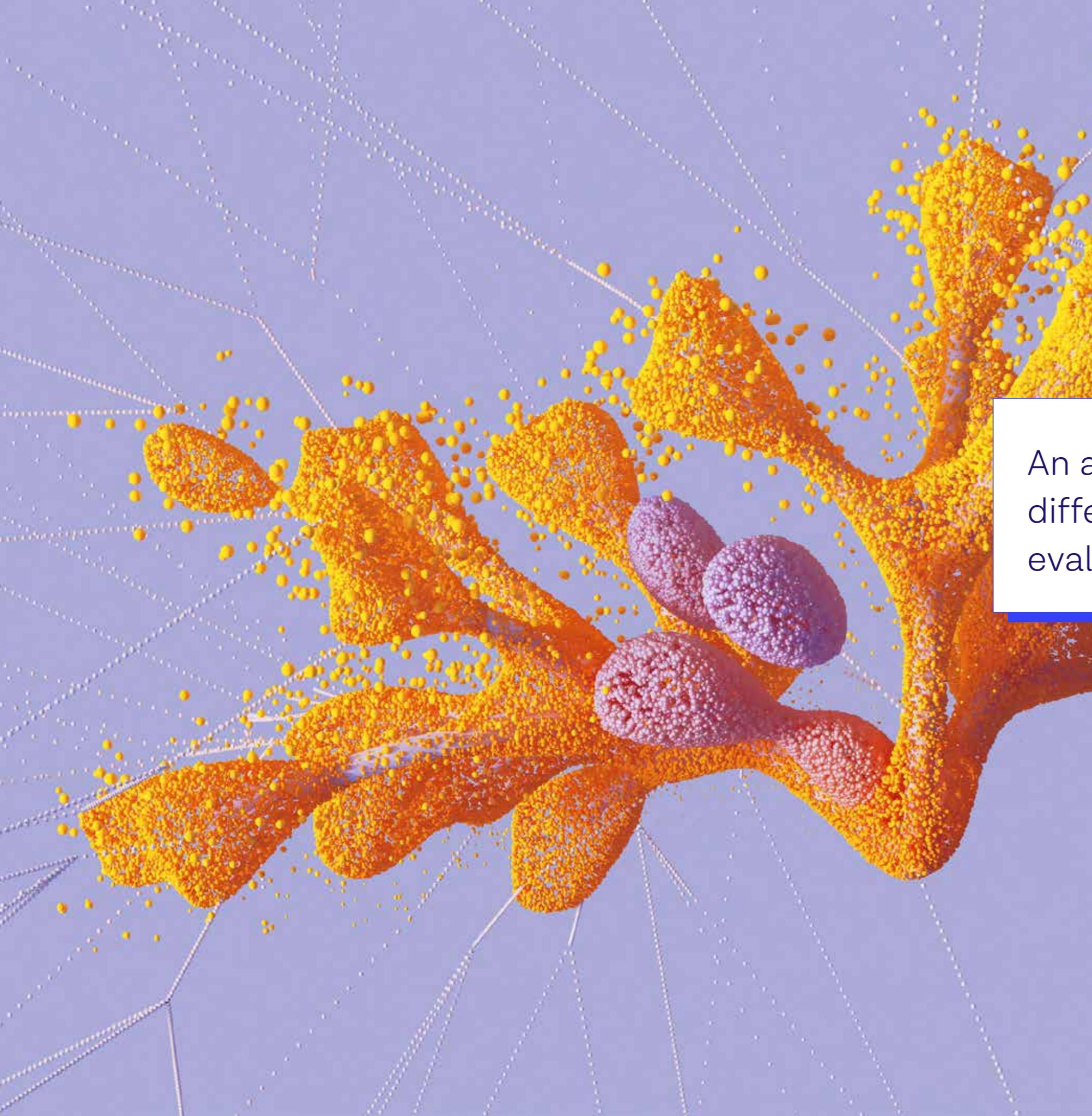
Our performance, and what's next.

The abstract images in this report were sourced from [Visualising AI](#), a collection of open source imagery created by a diverse range of artists, that builds a multi-dimensional picture of how AI can impact society.

01

Measuring bias in biometric products





Defining bias

What does ‘bias’ mean in the context of biometric products?

The term is used in different contexts to express a number of meanings. At Onfido, we recognize a specific meaning:

An algorithm is biased if it exhibits different levels of performance when evaluated on different subsets of data.

These ‘subsets’ refer to human characteristics, for example gender and age. Our definition of bias isn’t too different from the dictionary definition that states ‘bias is an inclination or prejudice for or against one person or group especially in a way considered to be unfair’. We measure the ‘inclination or prejudice’ of an algorithm by looking at its accuracy, define a group as a subset of data, and see unfairness as different levels of accuracy for different groups.

Two things are critical to mitigating bias.

1.

First, identifying bias requires labeled data with categories relevant to bias such as gender, age and race. Without labels relating to relevant sub-groups, it is not possible to measure how an AI performs for each specific sub-group — and so not possible to measure bias between them.

2.

Secondly, bias should be examined through statistical analysis of the finished product — to ensure analysis is representative of a real user experience. Bias cannot be observed by looking at individual cases; instead we have to look at aggregate performance statistics over a group of people. And rather than looking at constituent parts, we need to consider the whole finished product. With machine learning models, our ability to reason from first principles is limited. For example, while training a model on an imbalanced data set is likely to lead to a biased model, the converse is not true: training a model on a balanced dataset will not, by itself, necessarily lead to an unbiased model. All claims about bias must be supported by empirical

measurements of the product rather than anecdotal evidence or abstract reasoning.

We will measure and compare error rates across different subgroups of the overall population. But the full picture of bias is larger than that. As described by National Institute of Standards and Technology (NIST),³ besides the statistical aspects described here, bias is also affected by systemic and human factors. Systemic biases are reflected in the datasets we have available for product development. For example, the male-female imbalance among users with Indian documents is significantly larger compared to other countries. And human biases affect the way the results of the AI system are used and interpreted. AI systems are often used to inform human decision-making and, as such, are inextricably tied to human social behavior. As AI developers, we need to anticipate how human operators will interact with and use the AI system and design it in a way that reduces potential human biases.

³ Reva Schwartz, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, Patrick Hall, [Towards a Standard for Identifying and Managing Bias in Artificial Intelligence](#) (NIST Special Publication 1270, 2022).

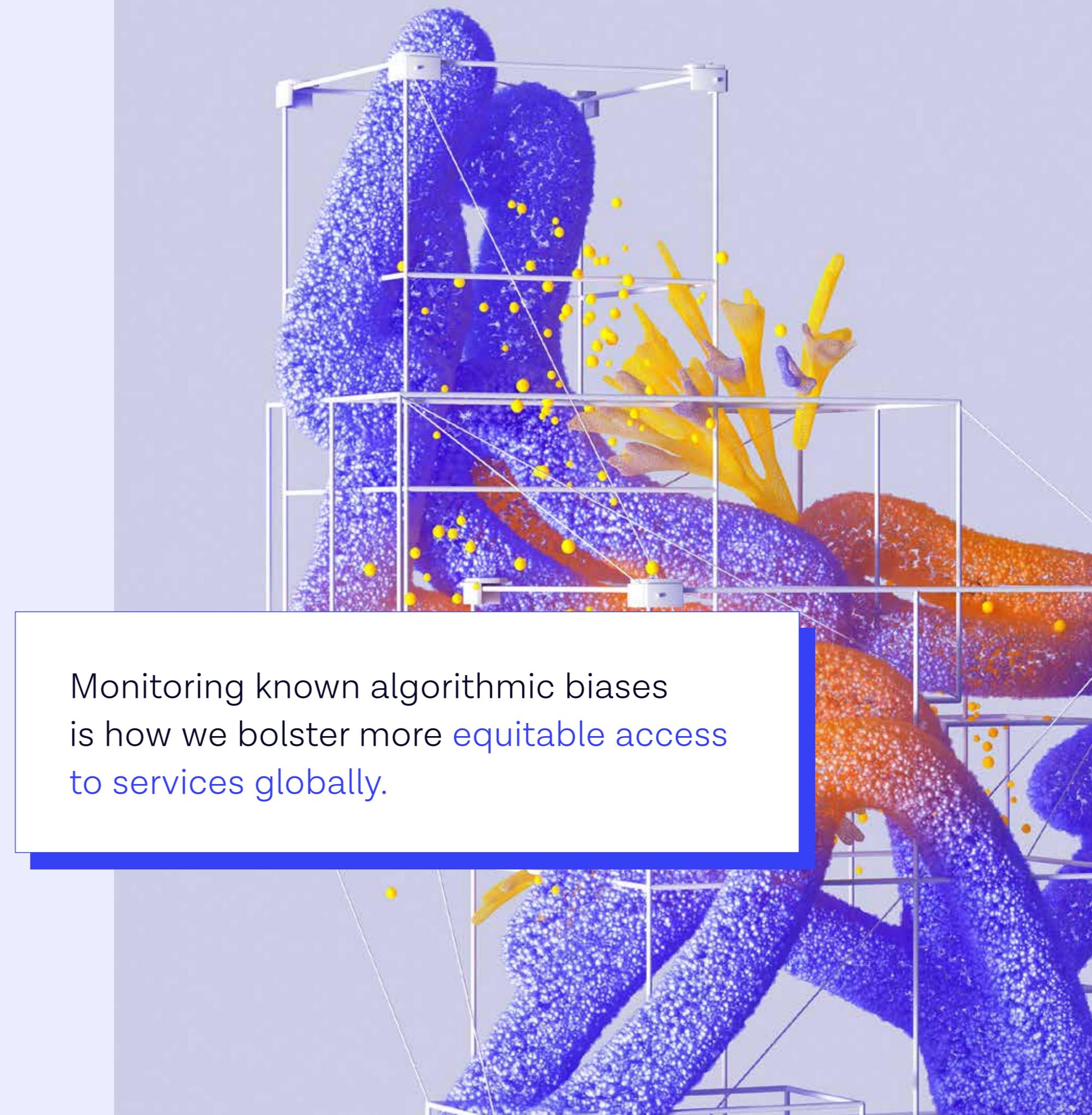


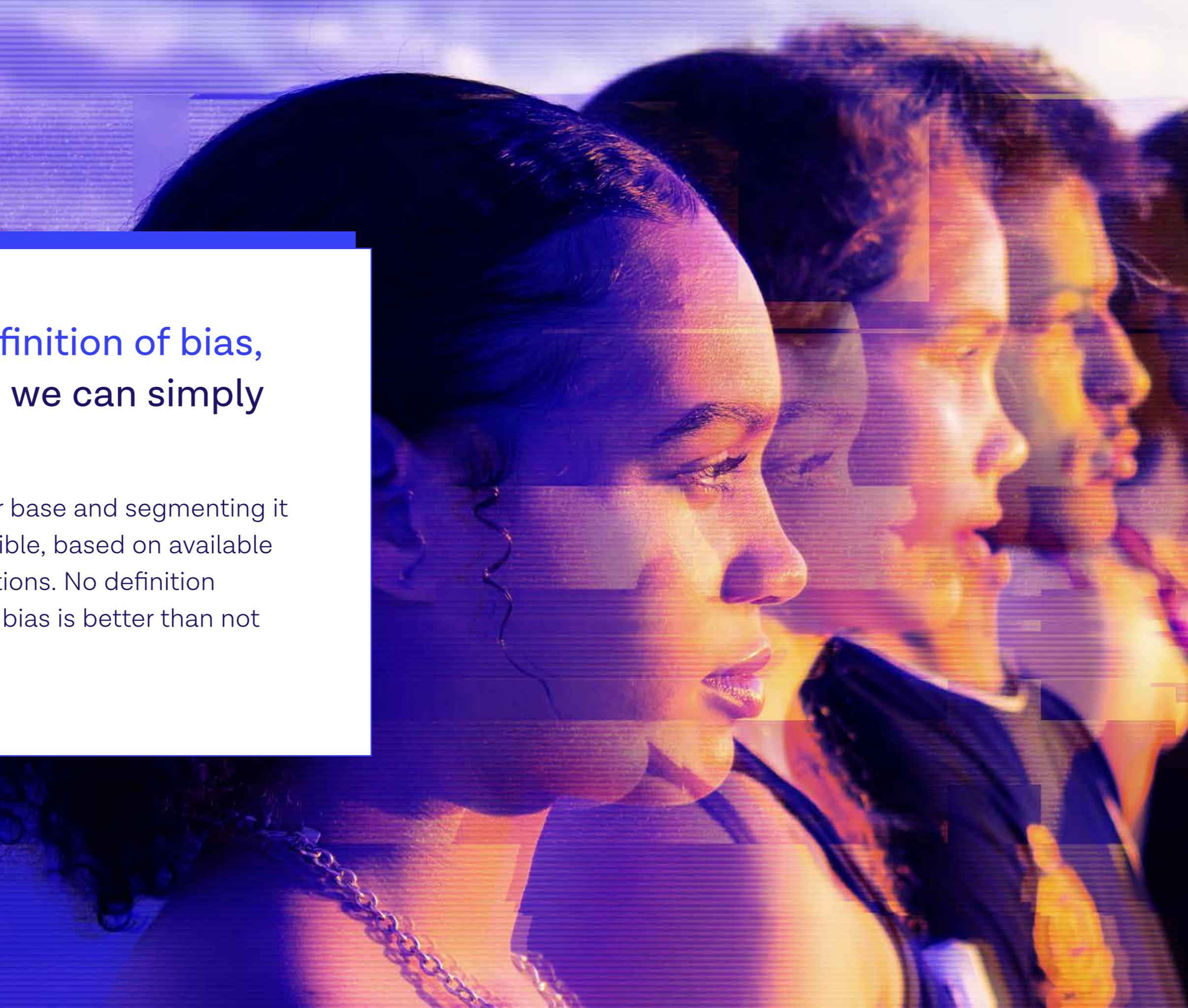
Segmenting data

To meaningfully measure bias, we need to specify the subsets of data to consider. We are not interested in bias with respect to arbitrary subsets of data, but only those defined by societally meaningful categories. Without imposing any restrictions, every machine learning model can appear biased: it is always possible to manufacture an artificial dataset split that causes a machine learning model to show different levels of accuracy.

The known natural biases of biometric algorithms can help us choose which segments of data are the most important to monitor. Fingerprint recognition is prone to age bias because fingerprints become less pronounced with age due to factors like chemical exposure and wear through manual labor⁴. Face recognition algorithms are inclined toward gender and racial biases – although there is not yet an accepted consensus on the reason why. As developers of face recognition algorithms, this means we must always be working to explicitly measure and mitigate the bias of our algorithms. We cannot hide behind ignorance. Instead, we must assume that every face recognition algorithm we train is biased until we prove otherwise. In Onfido's case, monitoring these known algorithmic biases is how we bolster more equitable access to services globally.

4 Andrea Rosales, Mireia Fernández-Ardèvol, [Ageism in the era of digital platforms](#) (Convergence, 2020).





There is no universal definition of bias, but that does not mean we can simply ignore the problem.

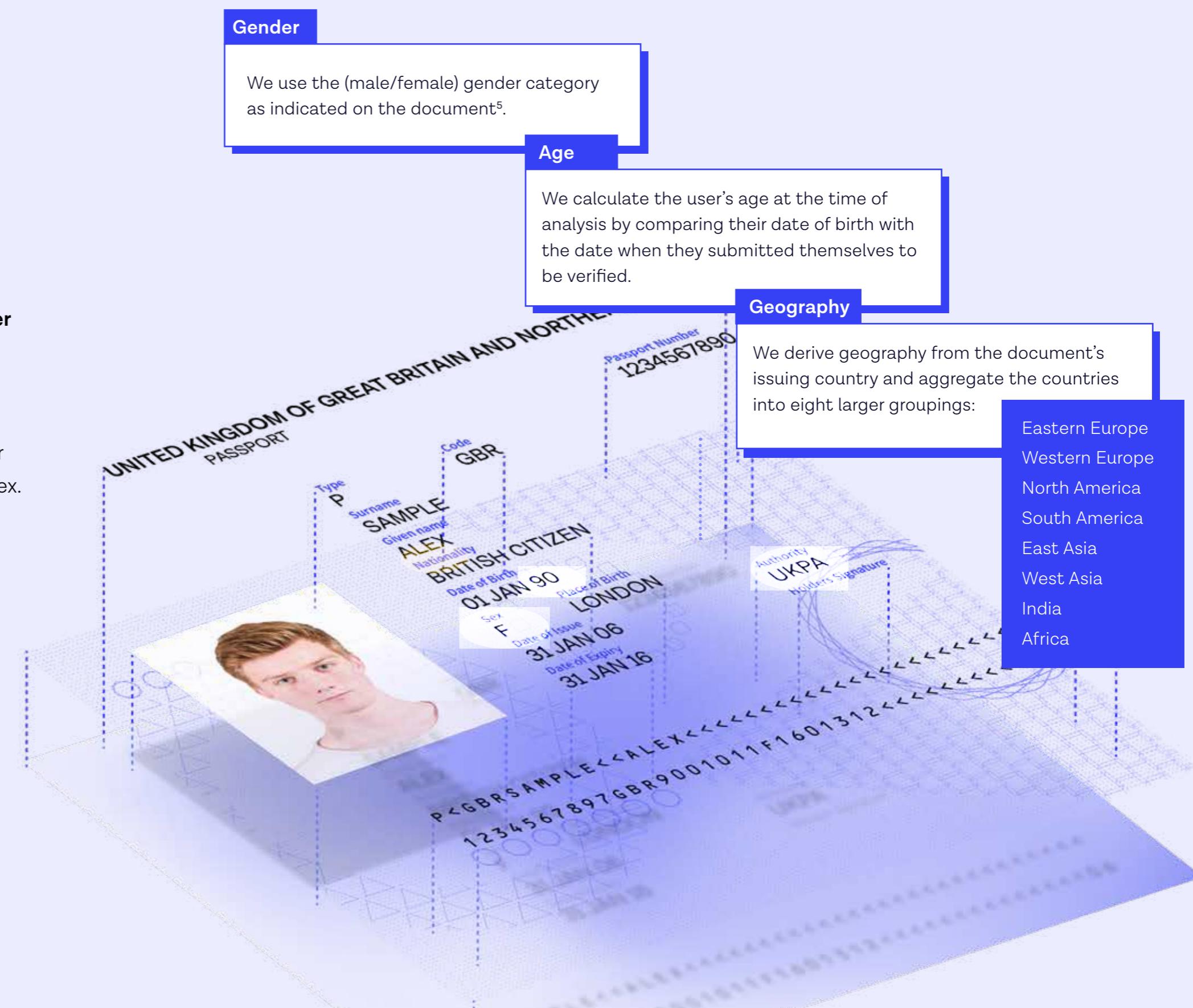
We recommend looking at your user base and segmenting it using the best approximations possible, based on available data and relevance for your applications. No definition will be perfect, but any definition of bias is better than not looking at bias at all.

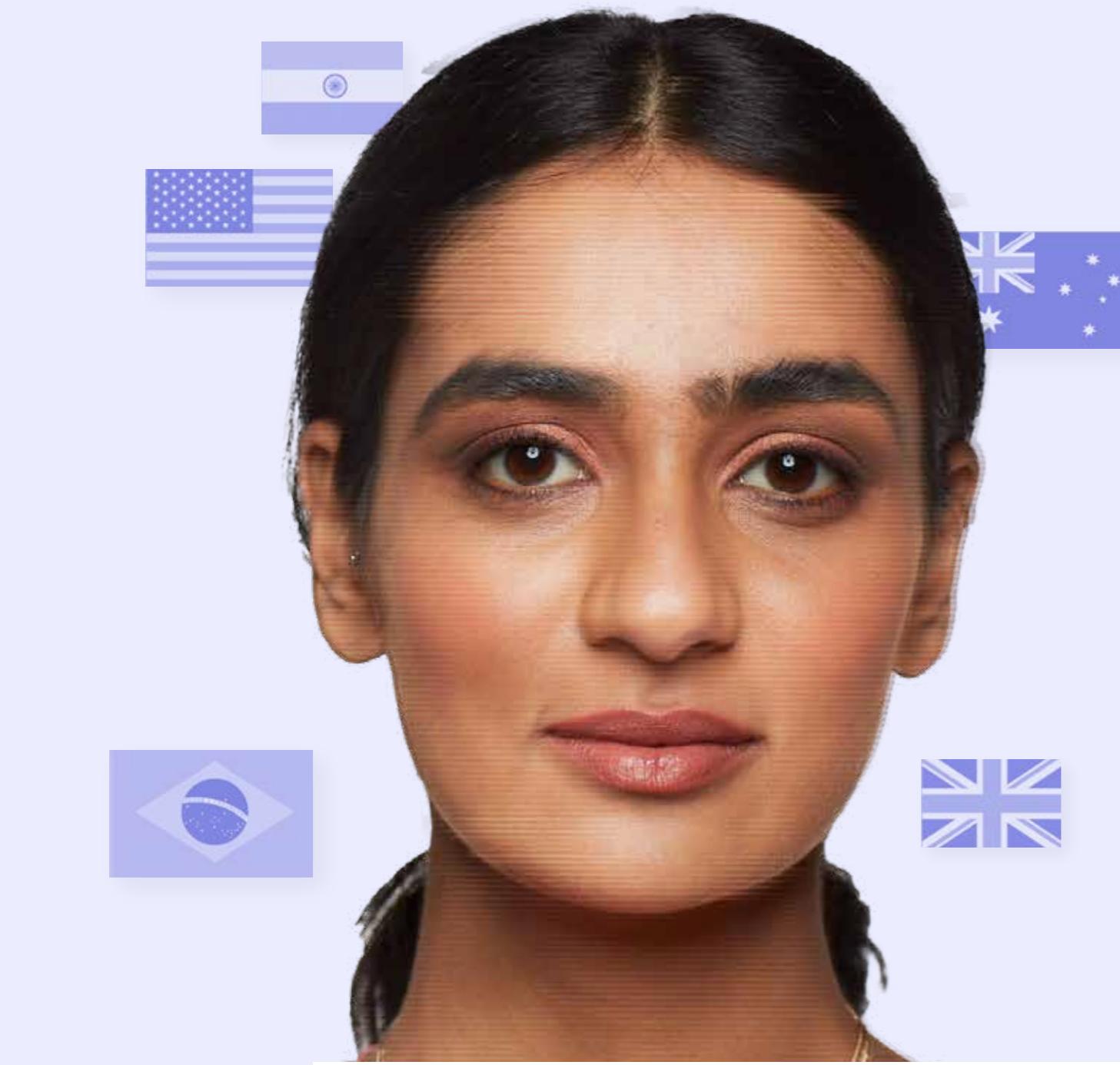
At Onfido, the three categories we monitor are gender, age, and geography.

When Onfido verifies an identity using biometrics, the user uploads a photo of their ID and a static or video selfie.

This means we can use data extracted from their identity documents to define categories for measuring bias. Gender and age are relatively simple, but geography is more complex.

5 Some countries are introducing “X” as a third gender category, e.g., from April 11, 2022, people can select “X” as their gender for [US passports](#). [New Zealand](#) has allowed this since 2012. Once we have collected a sufficient amount of data, we will expand our gender bias analysis to include this category.





Ethnicity and race are intimately tied with personal identity and so are predominantly self-reported.

The geography category is in line with NIST's use of country-of-birth to measure the demographic biases of face recognition algorithms⁶.

In 2019, we entered the UK Information Commissioner's Office Regulatory Sandbox with the aim of ensuring that our work on algorithmic bias mitigation respected the rights and freedoms of individuals when processing their personal data and that it was done in accordance with data protection law⁷. In line with the objectives outlined in our Sandbox Plan, we trialed and developed methodologies to group and label data, tested the performance of Onfido's face recognition models, retrained them, and measured the resulting performance improvement.

Ideally, we want to measure face recognition bias with respect to race and ethnicity, but defining these categories and acquiring consistent labels is very difficult. For example, the UK census defines

19 categories for ethnicity⁸, whereas the US Census Bureau defines five categories for race⁹. Ethnicity and race are intimately tied with personal identity and so are predominantly self-reported. Because self-reported data is difficult to acquire at scale, academic face recognition datasets usually automate the classification of images into broad racial categories without consulting the impacted people¹⁰.

Using geography as a category to evaluate racial bias is a practical compromise between what we would like to do and what we can do. While some countries, such as the UK and the US are racially diverse, we can infer from census and migration data that for many other countries the document issuing country is a reasonable proxy for race. Furthermore, for Onfido clients operating in particular countries, the geographic groups are of interest in their own right because they are a better representation of the population than the global dataset.

6 Patrick Grother, Mei Ngan, Kayee Hanaoka, [Face Recognition Vendor Test \(FRVT\), Part 3: Demographic Effects](#) (NIST, 2019)

7 ICO, [Regulatory Sandbox Final Report: Onfido](#) (2020)

8 Grouped into five larger categories these are: Asian (Indian; Pakistani; Bangladeshi; Chinese; any other Asian background), Black (Caribbean; African; any other Black, Black British, or Caribbean background), Mixed (White and Black Caribbean; White and Black African; White and Asian; any other mixed background), White (English, Welsh, Scottish, Northern Irish or British; Irish; Gypsy or Irish Traveller; Roma; any other White background); Other (Arab; Any other ethnic group). ([Ethnicity facts and figures, UK Government Website](#))

9 These are White, Black or African American, American Indian or Alaska Native, Asian, and Native Hawaiian or Other Pacific Islander ([United States Census Bureau](#)).

10 For example, the '[Racial Faces in the Wild: Reducing Racial Bias by Information Maximization Adaptation Network](#)' dataset uses four race categories (Caucasian, Indian, Asian and African) and obtains labels through a combination of a public database lookup and an automatic classifier.

Measuring performance

Having segmented the data, the next step is measuring performance. Algorithmic bias is an empirical concept. Once a dataset is split into groups, we can measure whether a particular algorithm exhibits bias. And while this measurement will tell us whether we have a problem with bias, it will not explain why bias exists.

For example, most face-matching algorithms are less accurate for females compared to males, but it's not fully understood why. Researchers have explored various hypotheses, such as dataset imbalance, prevalence of makeup, or the amount of face covered by hair, without finding a definite answer or a method to eliminate the bias despite repeatedly observing its presence¹¹.

At the same time, every measurement fits into a bigger historical context. Analogue film was historically engineered to best capture white skin tones. A similar imbalance has existed in datasets used to develop face matching algorithms; women and people of color are often under-represented¹². NIST's evaluation of face recognition algorithms¹³ and the Gender Shades' analysis of commercial gender classifiers¹⁴ showed the same biases appearing globally. This is a global problem – which is why it's so important for every company developing biometric algorithms to actively monitor and mitigate bias.



Analogue film was historically engineered to best capture white skin tones.

11 Vitor Albiero, Kai Zhang, Michael C. King, Kevin W. Bowyer, [Gendered Differences in Face Recognition Accuracy Explained by Hairstyles, Makeup, and Facial Morphology](#) (2021)

12 Dr David Leslie, [Understanding bias in facial recognition technologies](#) (2020).

13 Patrick Grother, Mei Ngan, Kayee Hanaokam, [Face Recognition Vendor Test \(FRVT\), Part 3: Demographic Effects](#) (2019).

14 Joy Buolamwini, Timnit Gebru, [Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification](#) (2018).

Impact of bias on identity verification

At Onfido, our mission is to simplify digital identity for everyone – granting legitimate users seamless access, and stopping fraudulent users.

We do this with our **Real Identity Platform**, which allows businesses to orchestrate a mix of document verification, data verification, fraud signals and biometric verification. For biometric verification, the key performance metrics are:

FRR

False rejection rate (FRR) measures the probability that a genuine **user will be prevented** from accessing a service.

FAR

False acceptance rate (FAR) measures the probability that we incorrectly let a **fraudster gain access** to a service.

The impact of bias is different for each of these metrics: a large false rejection rate (FRR) bias, i.e., big differences in FRR across groups, means that users from some groups will be more likely to be rejected. The impact of automated rejection will depend on the application, but it will usually involve higher scrutiny of the user, longer approval times and perhaps require the user to submit additional documents. FRR bias is experienced directly by the affected users.

On the other hand, a large false acceptance rate (FAR) bias means that users from some groups will be more likely to incorrectly pass through the system: the user is not directly impacted by this bias. But there might be an indirect impact, since the company performing the identity verification knows that certain groups have higher FAR and so might impose higher scrutiny on these groups to limit their fraud exposure. Therefore legitimate individuals could still be indirectly penalized for algorithmic failings of the identity verification (IDV) provider.

The consequences of algorithmic bias are specific to each application. For document and biometric verification, FRR bias is felt directly by the user, while FAR bias primarily exposes the company to higher fraud risk.

Group-level bias vs. individual-level bias

Both FRR and FAR are group-level metrics; they measure aggregate performance of an algorithm over a group of people. This is a reasonable choice for applications such as IDV where each individual user interacts with the system only a few times. However, for applications such as authentication or face recognition-based phone unlocking, this is not enough. If a user is expected to interact with a system repeatedly, we need to look at bias at an individual level.

It's statistically possible for an algorithm to have an overall FRR of 1%, despite some individual users experiencing a FRR of 30% or more¹⁵. This does not mean that the group-level view is wrong, [but it does not show the full picture either](#).

It has to be supplemented with individual-level metrics to fully understand the biases of the algorithm.

Next, we will focus on our strategies to reduce the FRR bias for gender, age and geography. Bias is not binary; it exists on a scale. It is unrealistic to expect that we can completely eliminate bias, but it can be minimized. Our responsibility is to always strive for improvement and transparency.

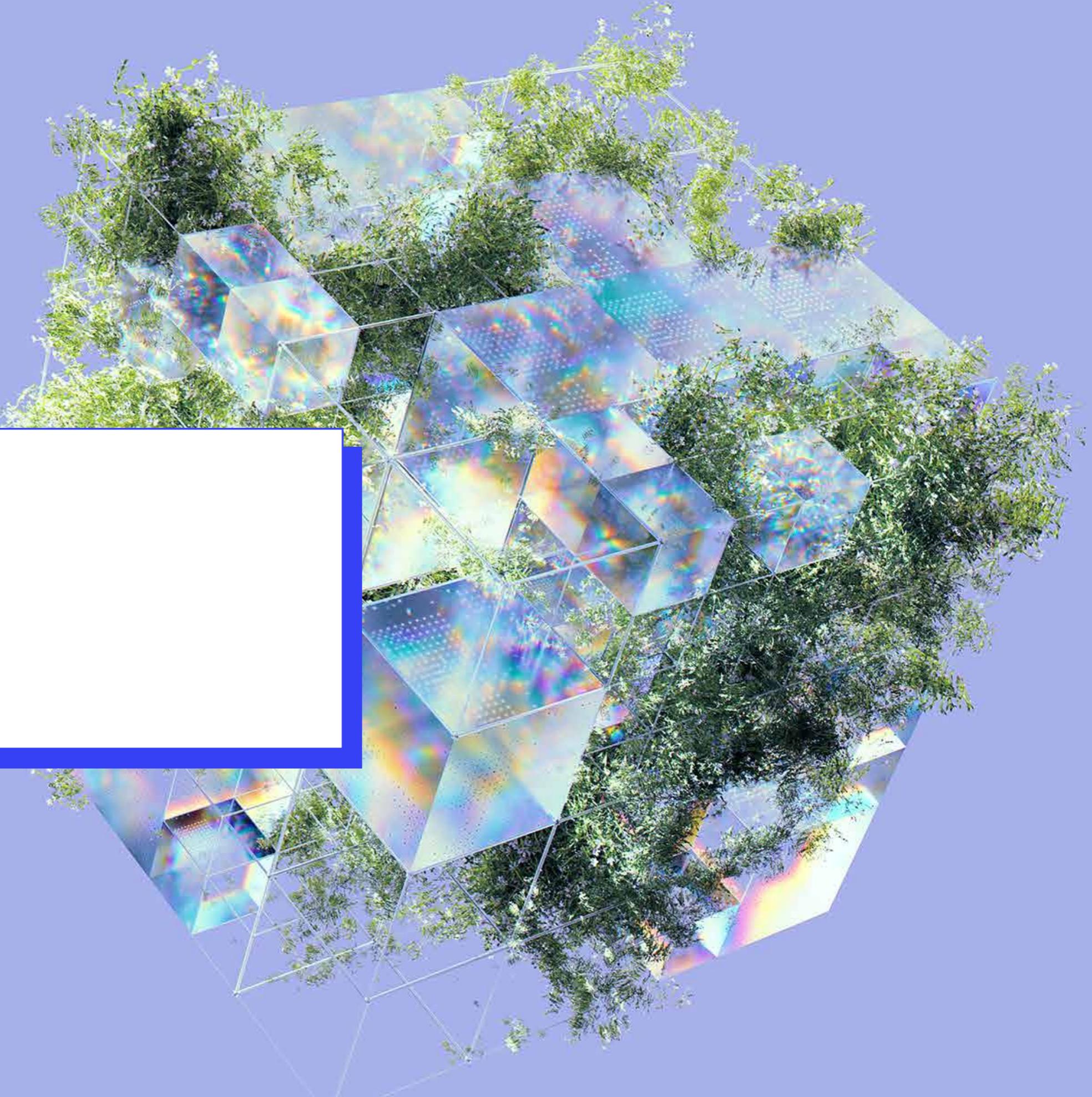
¹⁵ The personal experience of one the authors was about a 50% success rate at the automatic passport gates on the UK border, measured over multiple years and many flights. The overall success rate of all passengers in the arrival hall was well above 90%, but something about the author's passport made the individual experience significantly worse.



If you're purchasing biometric-based verification or authentication solutions, make sure you understand how your individual users will be impacted by potential biases — [do not just rely on overall metrics](#).

02

Strategies for bias mitigation



The developer's toolbox contains multiple strategies for making algorithms less biased.

These can be broadly classified into three categories:

1. Data preparation

Modify the dataset used for model training before training takes place.

2. Training

Change the training procedure itself by modifying the model architecture or the loss function.

3. Post-processing

Apply post-processing to the outputs of a trained model.

1. Data preparation

Data preparation strategies are critical to mitigating bias.

If a dataset is not sufficiently diverse, i.e., if some groups are not sufficiently represented, then we cannot expect a model trained on this dataset to perform well on these subgroups.

Diverse datasets

Diversity goes beyond simply counting the number of samples in each subgroup.

For example, a dataset may contain many male images, but if they are exclusively clean-shaven, we cannot expect the resulting model to perform well on bearded faces. Diversity is the foundation upon which all other strategies build and is why access to data is incredibly important for us to continue to mitigate bias.

Diversity is the foundation upon which all other strategies build.

Balanced datasets

We need diverse datasets, but we don't necessarily need balanced datasets.

In a diverse dataset, each subgroup is represented. In a balanced dataset each subgroup is the same size. Often we simply don't have an equal number of samples per subgroup. For example, we want to include people with face tattoos in the dataset, but because they are relatively rare it's tough to create a balanced dataset. If collecting

more samples from a group is not an option, then the only other way to create a balanced dataset would be to not use as many samples from over-represented groups – in other words, to artificially shrink the dataset. This can be done but will usually result in a lower-performing model compared to other approaches.

Balanced sampling

At Onfido, instead of aiming to create perfectly balanced datasets, which is often infeasible, we use diverse datasets together with balanced sampling.

Balanced sampling means that we take an equal amount of samples from each group. For example, if there are two groups, A and B, then balanced sampling takes one sample from A, followed by one sample from B regardless of their relative sizes. Consequently, the model will see an equal number of samples from each group. If we look at the dataset and if we assume that A is twice as large as B, then by the time we will have sampled all elements of A, we will have sampled each element of B twice.

Representative datasets

It's important to distinguish diverse and balanced datasets from representative datasets. In a representative dataset, each subgroup has the same relative size as it does in the overall population. Representative datasets are useful during the evaluation phase to estimate the expected overall performance of the algorithm. They are less useful for addressing algorithmic bias, since under-represented groups in the overall population will remain under-represented in the dataset.

Sampling strategy

At Onfido, we want to reduce bias for multiple overlapping groups. This makes it difficult to maintain data diversity.

For example, the two geographies ‘Western Europe’ and ‘North America’ are demographically more similar to each other than to ‘East Asia’. This means that we have to go beyond balanced sampling. Sampling weights can be chosen statically (kept constant throughout training) or dynamically (adjusted based on model performance during training). The groups used for sampling can be defined explicitly based on metadata such as gender or geography or derived automatically using clustering methods, i.e., a machine learning algorithm splits data into groups based on some measure of similarity. Which sampling strategy performs best will depend on the application, the dataset, and other training parameters.

Data augmentation is another strategy that can be used to increase the diversity of a dataset. Data augmentations can range from simple variations of brightness and contrast, to more sophisticated applications of style transfer¹⁷ to simulate the style of various

ID documents. Methods such as MixUp¹⁸, CutMix¹⁹ and related methods²⁰ also allow us to increase the diversity of the dataset by interpolating existing samples.

Advances in synthetic image generation using generative adversarial networks such as StyleGAN have enabled the generation of fake photorealistic portraits. Follow-up works, such as DiscoFaceGAN²¹, allow us to create images while controlling pose, lighting and identity²². These images are difficult to visually distinguish from real photos. If we could use synthetic images to train face recognition and anti-spoofing models, it would solve a lot of diversity and data privacy problems. However, research shows²² that models trained on synthetic images perform significantly worse when evaluating real images than models trained using real images. Even though synthetic images can look indistinguishable from real photos, there is a domain gap that prevents ML models trained on one type of data to successfully generalize to the other. The exact nature of this domain gap is not fully understood. **But considering the potential benefits of synthetic images, it is an area of active research.**

16 Leon A. Gatys, Alexander S. Ecker, Matthias Bethge, [A neural algorithm of artistic style](#) (2015).

17 Hongzi Zhang, Moustapha Cisse, Yann N. Dauphin, David Lopez-Paz, [Mixup: beyond empirical risk minimization](#) (2018).

18 Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, Youngjoon Yoo, [CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features](#) (2019).

19 Dominik Lewy, Jacek Mańdziuk, [An overview of mixing augmentation methods and augmentation strategies](#) (2022).

20 Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, Xin Tong, [Disentangled and Controllable Face Image Generation via 3D Imitative-Contrastive Learning](#) (2020).

21 It is a valid question, what ‘identity’ means for computer-generated images. In this context we mean that the GAN generates multiple images that a face recognition model considers to be the same ‘identity’.

22 See SynFace: [Face Recognition with Synthetic Data](#) and [SFace: Privacy-friendly and Accurate Face Recognition using Synthetic Data](#).

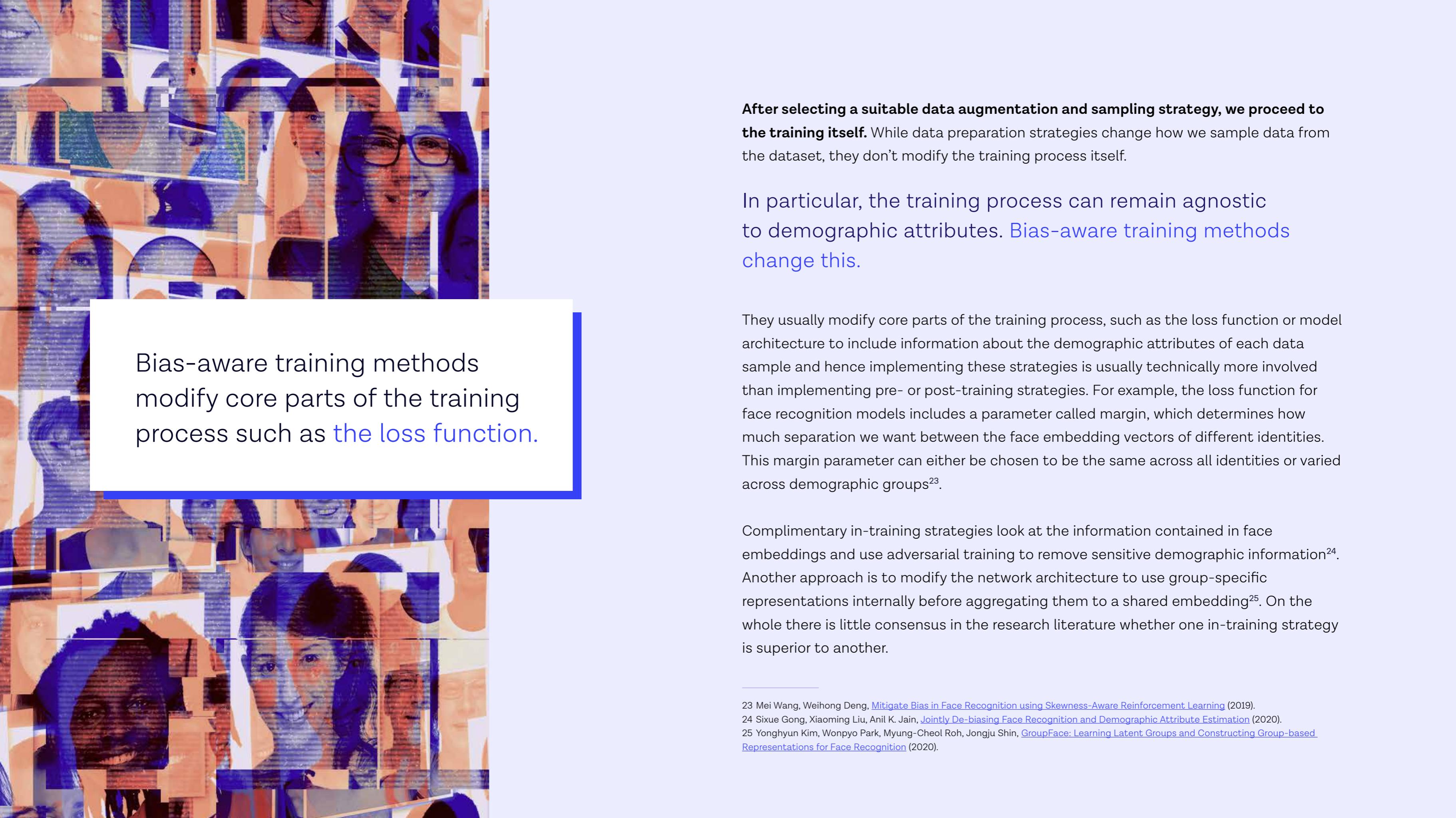


Recent advances in synthetic data generation open a promising area of research for dataset enrichment.



2. Training

Making bias reduction an explicit goal
of the training process.



Bias-aware training methods modify core parts of the training process such as [the loss function](#).

After selecting a suitable data augmentation and sampling strategy, we proceed to the training itself. While data preparation strategies change how we sample data from the dataset, they don't modify the training process itself.

In particular, the training process can remain agnostic to demographic attributes. [Bias-aware training methods](#) change this.

They usually modify core parts of the training process, such as the loss function or model architecture to include information about the demographic attributes of each data sample and hence implementing these strategies is usually technically more involved than implementing pre- or post-training strategies. For example, the loss function for face recognition models includes a parameter called margin, which determines how much separation we want between the face embedding vectors of different identities. This margin parameter can either be chosen to be the same across all identities or varied across demographic groups²³.

Complimentary in-training strategies look at the information contained in face embeddings and use adversarial training to remove sensitive demographic information²⁴. Another approach is to modify the network architecture to use group-specific representations internally before aggregating them to a shared embedding²⁵. On the whole there is little consensus in the research literature whether one in-training strategy is superior to another.

23 Mei Wang, Weihong Deng, [Mitigate Bias in Face Recognition using Skewness-Aware Reinforcement Learning](#) (2019).

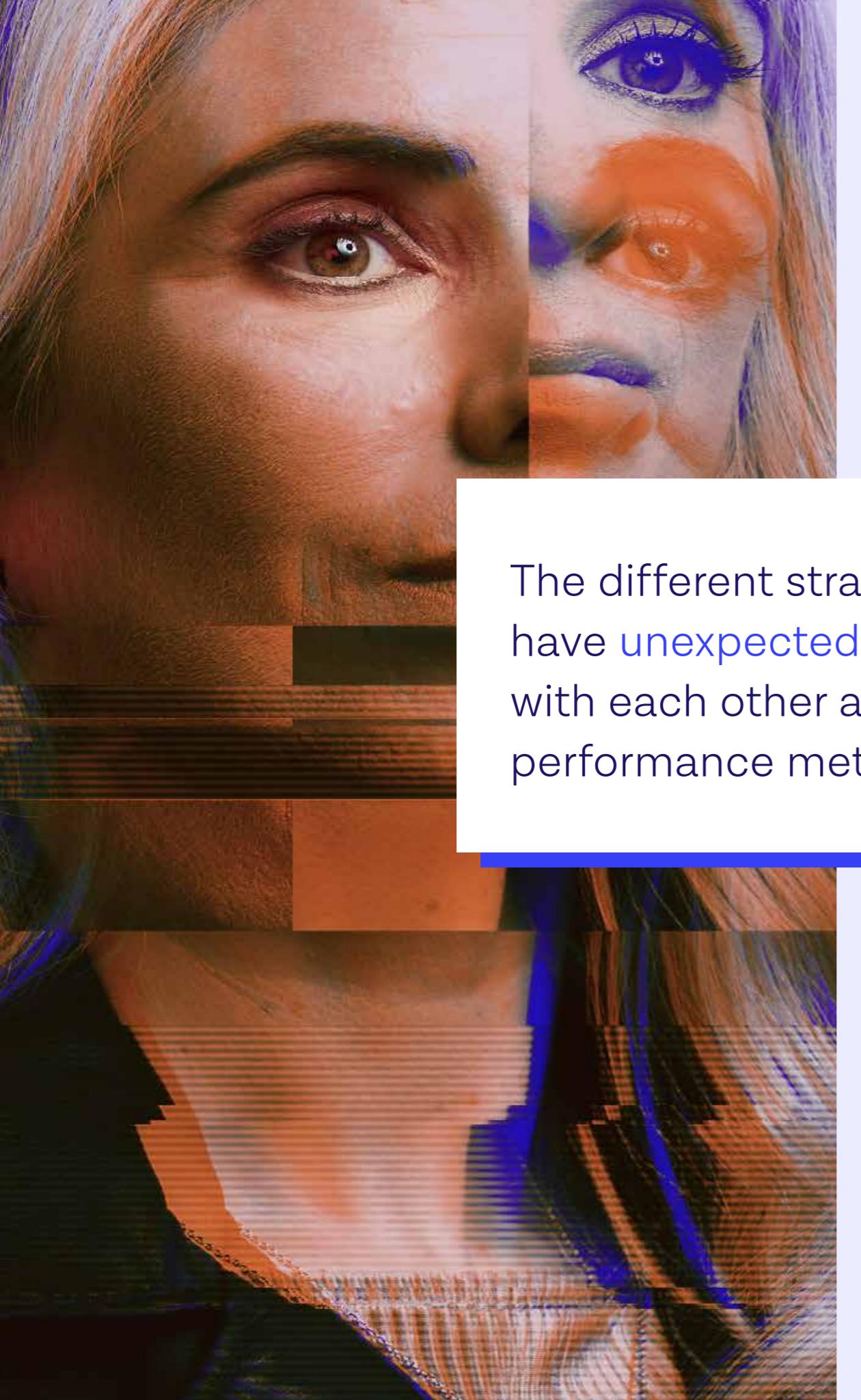
24 Sixue Gong, Xiaoming Liu, Anil K. Jain, [Jointly De-biasing Face Recognition and Demographic Attribute Estimation](#) (2020).

25 Yonghyun Kim, Wonpyo Park, Myung-Cheol Roh, Jongju Shin, [GroupFace: Learning Latent Groups and Constructing Group-based Representations for Face Recognition](#) (2020).

3. Post-processing

Applying post-processing strategies to the model outputs.





The different strategies can have unexpected interactions with each other as well as the performance metrics.

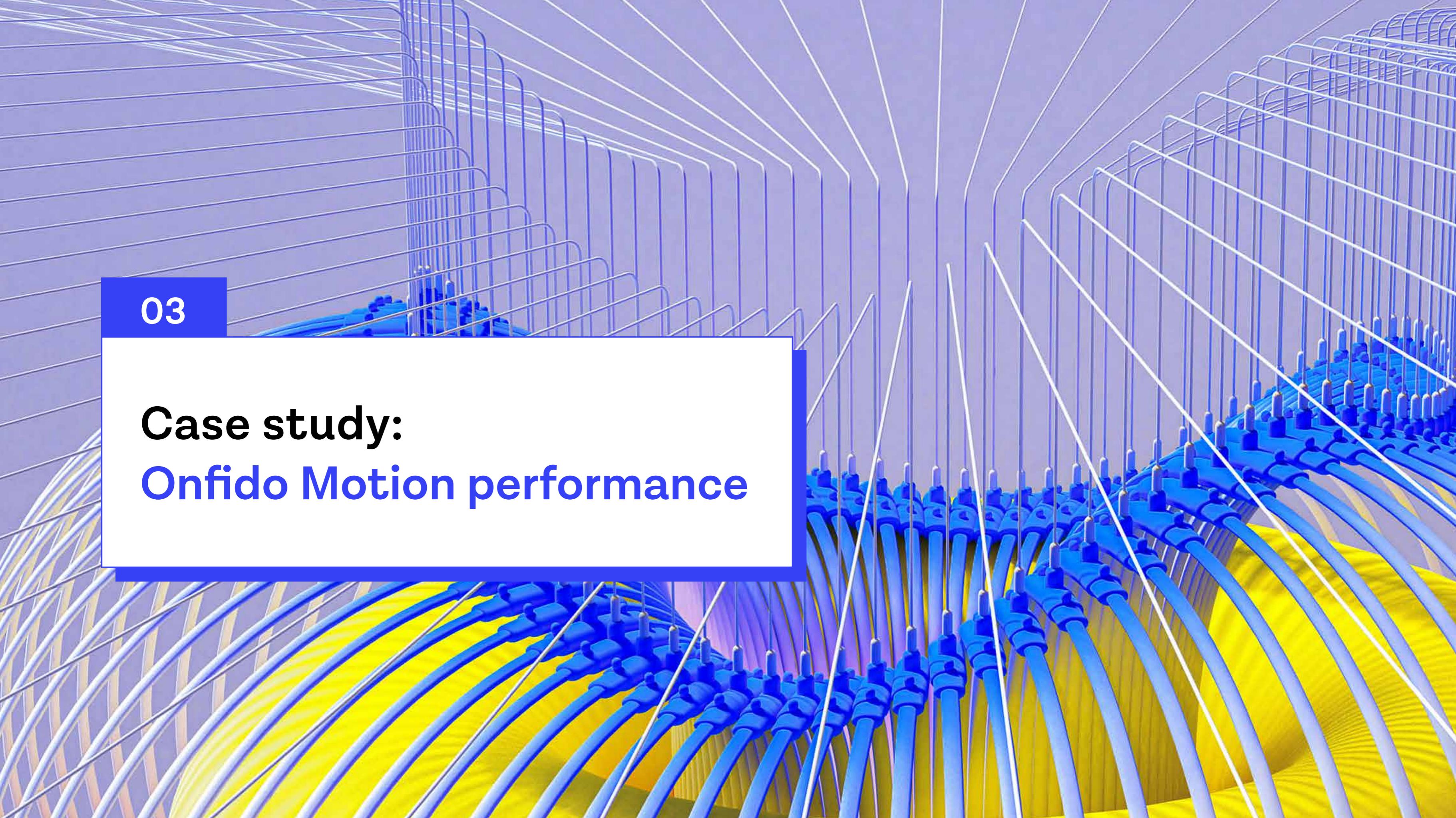
Inspired by the process of calibrating classifiers, calibration strategies apply group-specific transformations to the face matching similarity scores to equalize error rates across demographic groups²⁶.

The advantage of these methods is that they don't require a (usually expensive) retraining of the model. And if the group-specific transformations are determined using an unsupervised clustering method, then we also don't require knowledge of the sensitive attributes during inference to apply the calibration.

The different strategies can have unexpected interactions with each other as well as the performance metrics. For example, when tuning sampling weights for a data sampling strategy, one can observe a trade-off between FAR and FRR. Increasing the sampling rate for a particular group can reduce the FAR for the group, but too aggressive oversampling can lead to an increased FRR relative to other groups.

There are many different bias mitigation strategies available to the developer to choose from. Which combination of strategies works best for a given application has to be determined through experimentation.

²⁶ See [Post-comparison mitigation of demographic bias in face recognition using fair score normalization](#) and [FairCal: Fairness calibration for face verification](#).



03

Case study: Onfido Motion performance

At Onfido, we build better products that businesses and their customers will love

Demand for [fair, fast, and secure biometric verification](#) is higher than ever. Digital identity fraud is at an all time high, having increased 41% since 2020²⁷.

Businesses are contending with more sophisticated biometric fraud with fraudsters employing techniques like display attacks and 2D and 3D masks. User expectations of onboarding experiences have also never been higher, and businesses who fail to meet them are seeing those customers quickly take their business elsewhere, so there's a real need to move beyond traditional, manual verification methods.

Our customers are rightly challenging us to provide solutions with cutting edge fraud

mitigation performance without compromising on the experience of their users. **In response to these challenges we built our latest advancement in biometric technology, Onfido Motion.**

Motion delivers active liveness verification by asking customers to complete a simple head turn — no complex actions, camera maneuvering, or speaking out loud required. It only takes seconds to complete, 95% of verifications are returned in under 15 seconds, and it offers 10x better anti-spoofing performance compared to our previous video product. It doesn't compromise on security or experience, and we worked to ensure it doesn't compromise on fairness — which is why we included it as a key goal from the start of development to make Motion our fairest ever biometric product.

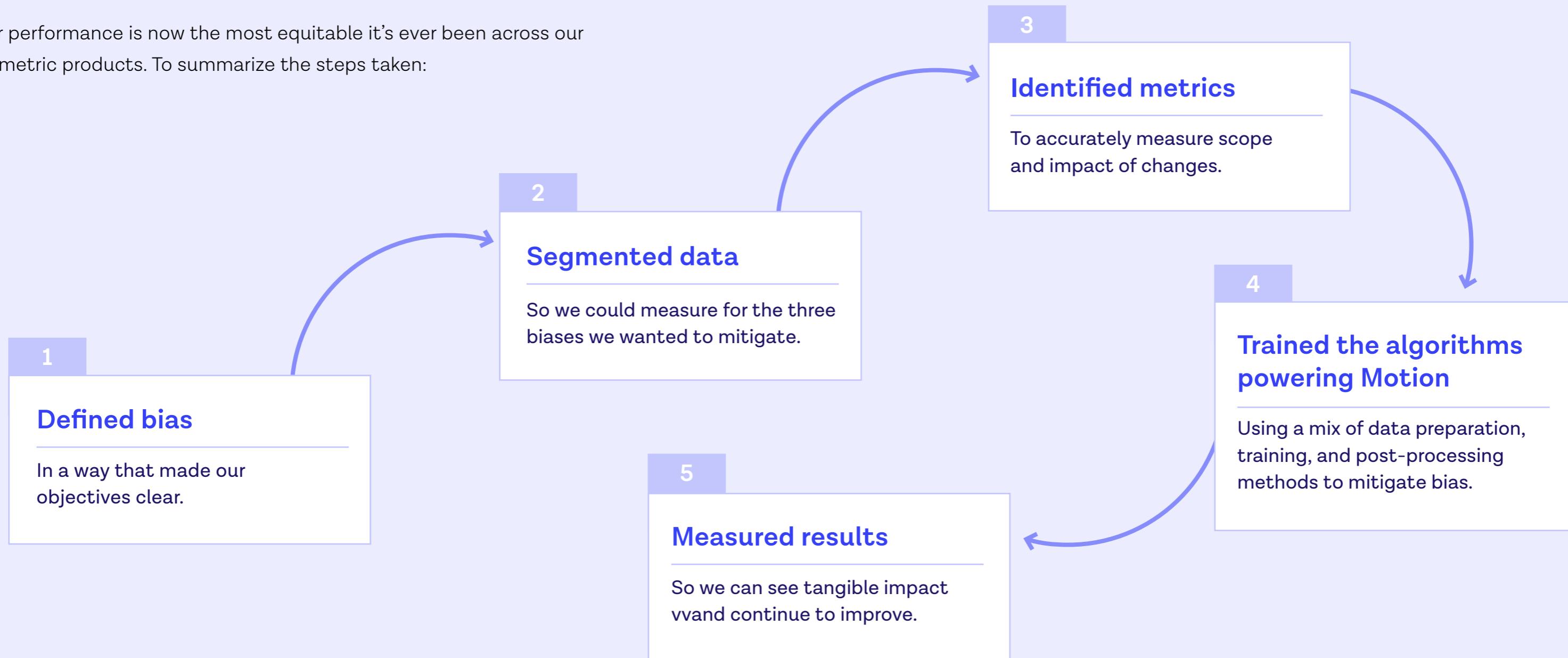


²⁷ Onfido, [Identity Fraud Report 2022](#) (2022).

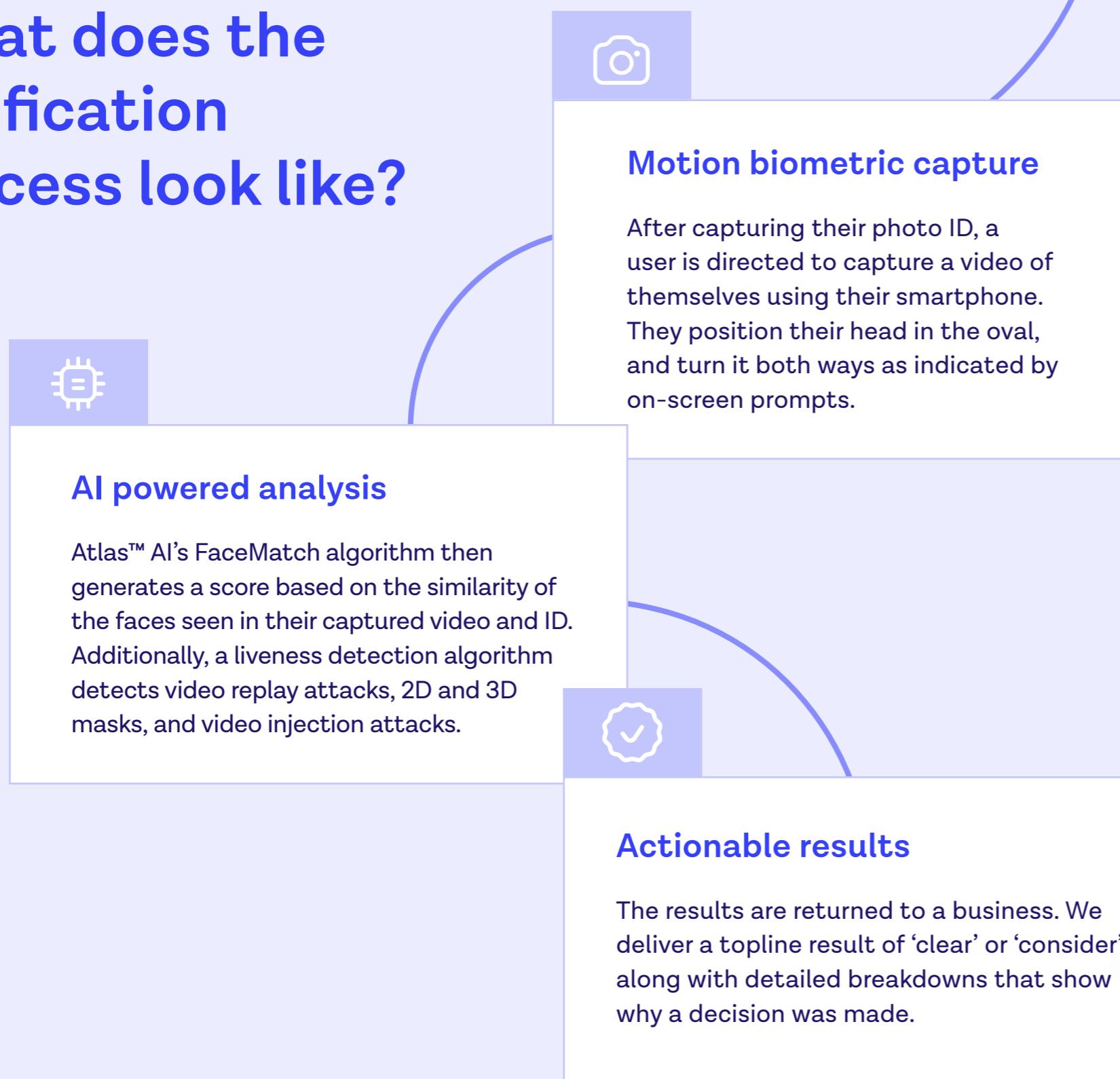
What steps did we take?

We've produced excellent results by developing Motion with bias mitigation in mind from the start.

Our performance is now the most equitable it's ever been across our biometric products. To summarize the steps taken:



What does the verification process look like?



Onfido Motion performance

The key metric we chose to measure bias in Onfido Motion is FRR (false rejection rate), for a number of reasons:

1. FRR is the metric that most directly represents the experience of real users who would be rejected should the system be biased.
2. Measuring FAR bias requires a dataset of spoofs (impersonation attempts) with demographic attributes. We obtain demographic information from document images, which are either missing or inaccurate for most spoofs in our dataset. Thus our ability to measure FAR bias for Motion is limited.

In the tables below we compare the group FRR against the overall FRR with respect to geography, gender and age. The number shown is the ratio between group FRR and overall FRR together with a 95% confidence interval.

Value = 1 Indicates no difference between the group's FRR and the overall FRR.

Value > 1 Indicates the model is biased against that group — more people are falsely rejected in this group compared to overall.

Value < 1 Indicates that the model is preferential towards that group — fewer people are falsely rejected in this group compared to overall.

Our aim is for the values to be **as close to 1 as possible** — with values above 2 being unacceptable.

A note on confidence intervals

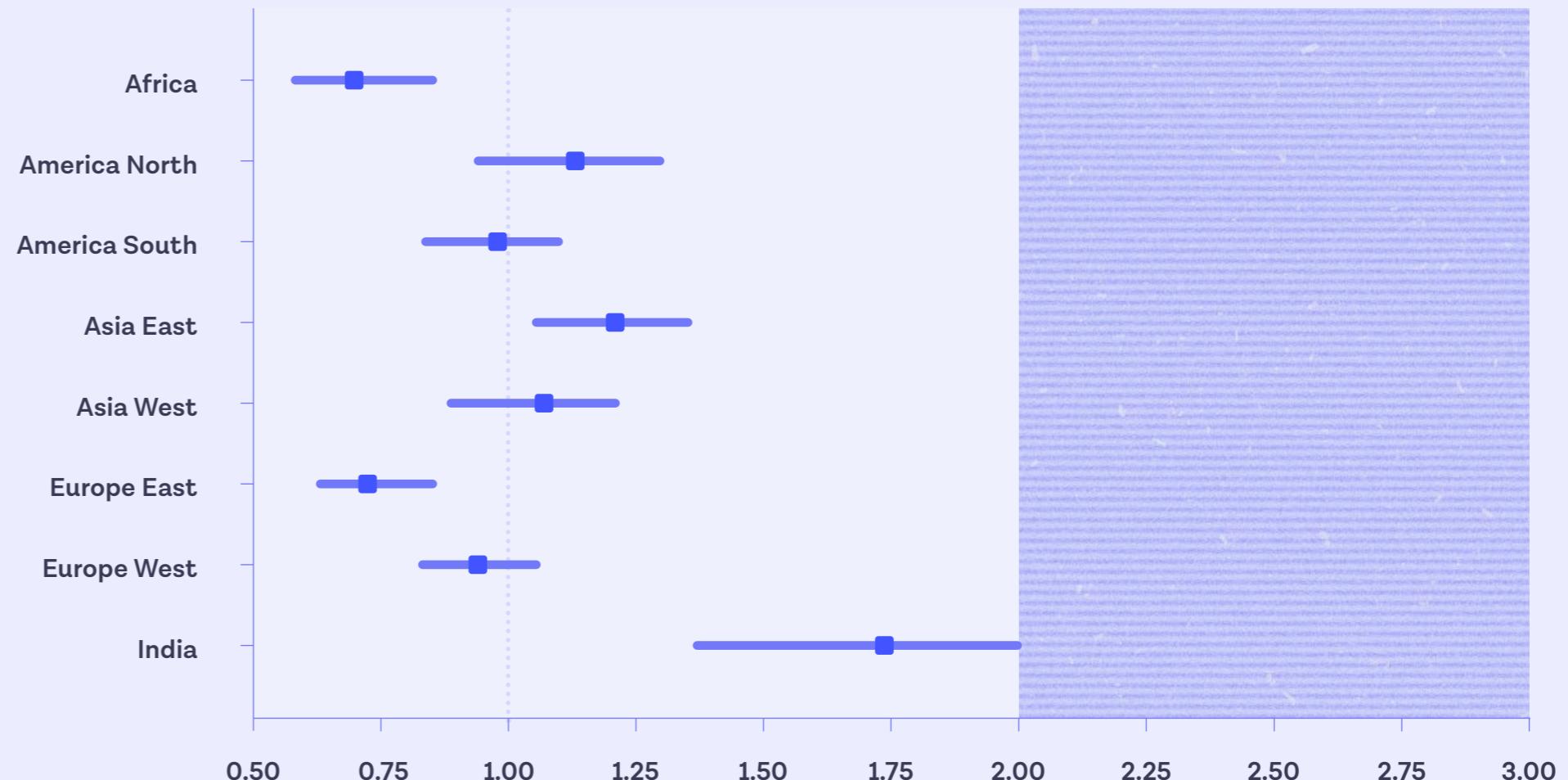
We report all our results along with confidence intervals. A confidence interval represents the uncertainty in our measurement of the model's performance. This uncertainty is present because we can only estimate model performance on a limited number of samples (the test set). If we had an infinite amount of data, the uncertainty (and thus width of the confidence interval) would decrease to 0. Because we only have a limited test set on which to measure performance, the confidence intervals cover a range of values. All confidence intervals are reported for the 95% confidence level.

Our models are very accurate, making very few mistakes on our test set, and producing a very low FRR. At such a low FRR, just a few additional mistakes can cause a large relative change in FRR. It is possible that our test set includes slightly more difficult samples than is representative — or slightly fewer. This is what leads to the uncertainty in our FRR estimates. This uncertainty is reflected in the width of the confidence intervals we report.

Geography bias

We are pleased to report that in 7 of 8 geographies, the ratios are between 0.7 and 1.2 – meaning of these groups none are more than 1.2 times more likely to be falsely rejected.

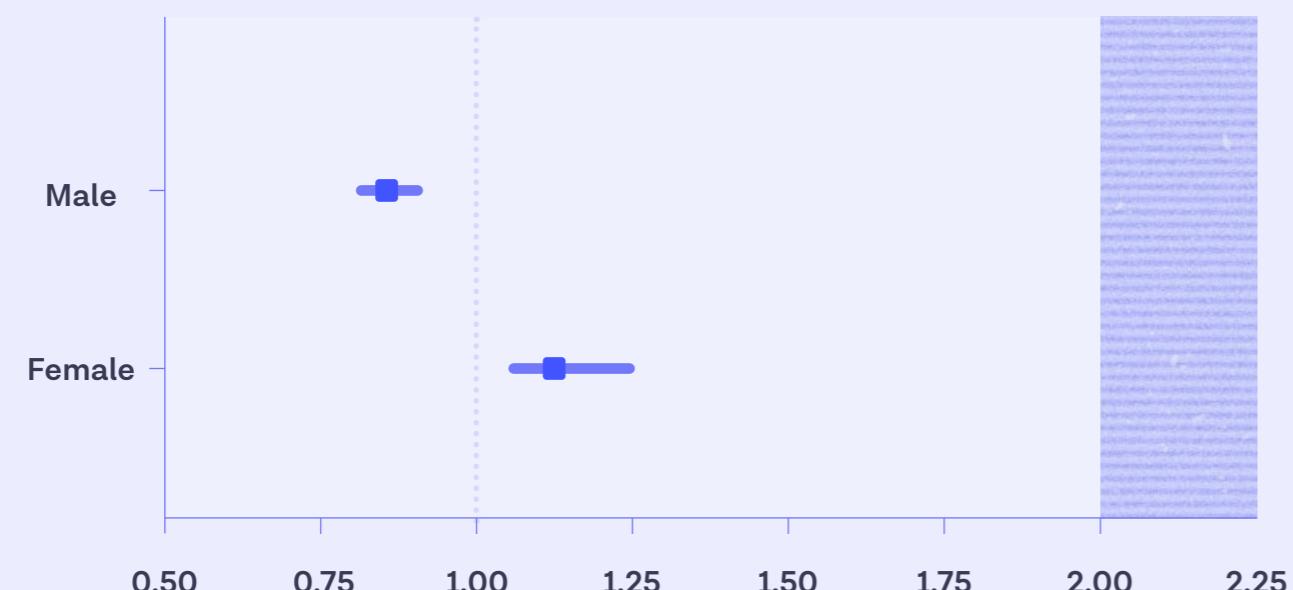
The outlier is India, where the FRR ratio is higher at 1.72. One of the contributing factors to this is varying image quality. The quality of identity documents varies across countries and in India in particular we observe a larger than average number of dark and blurry images printed in IDs. This is an area we want to continue to improve.



Group FRR / Overall FRR	Europe West	Europe East	America North	America South	Asia East	Asia West	India	Africa
(95% confidence interval)	0.94 (0.81 - 1.05)	0.72 (0.62 - 0.82)	1.12 (0.93 - 1.33)	0.97 (0.85 - 1.14)	1.20 (1.08 - 1.32)	1.06 (0.91 - 1.21)	1.72 (1.35 - 2.10)	0.70 (0.55 - 0.85)

Gender bias

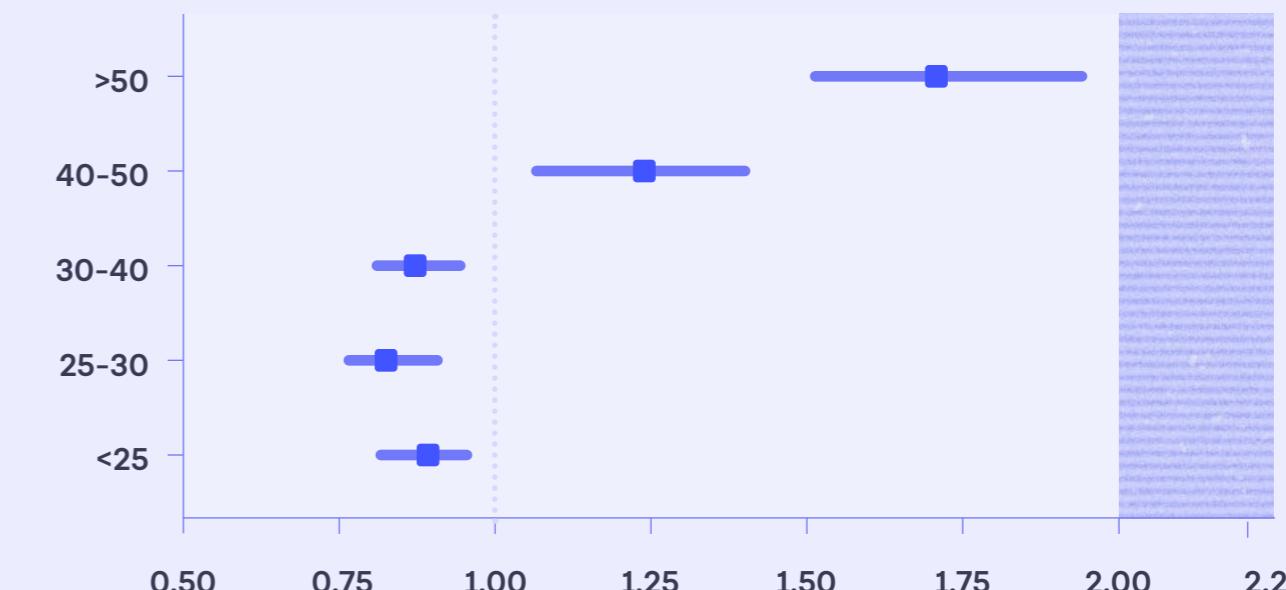
We observe some bias between male and female, with a ratio of **0.87** for male and **1.18** for female. This is in line with general observations that face matching models are slightly worse on females than males as previously explored in this paper.



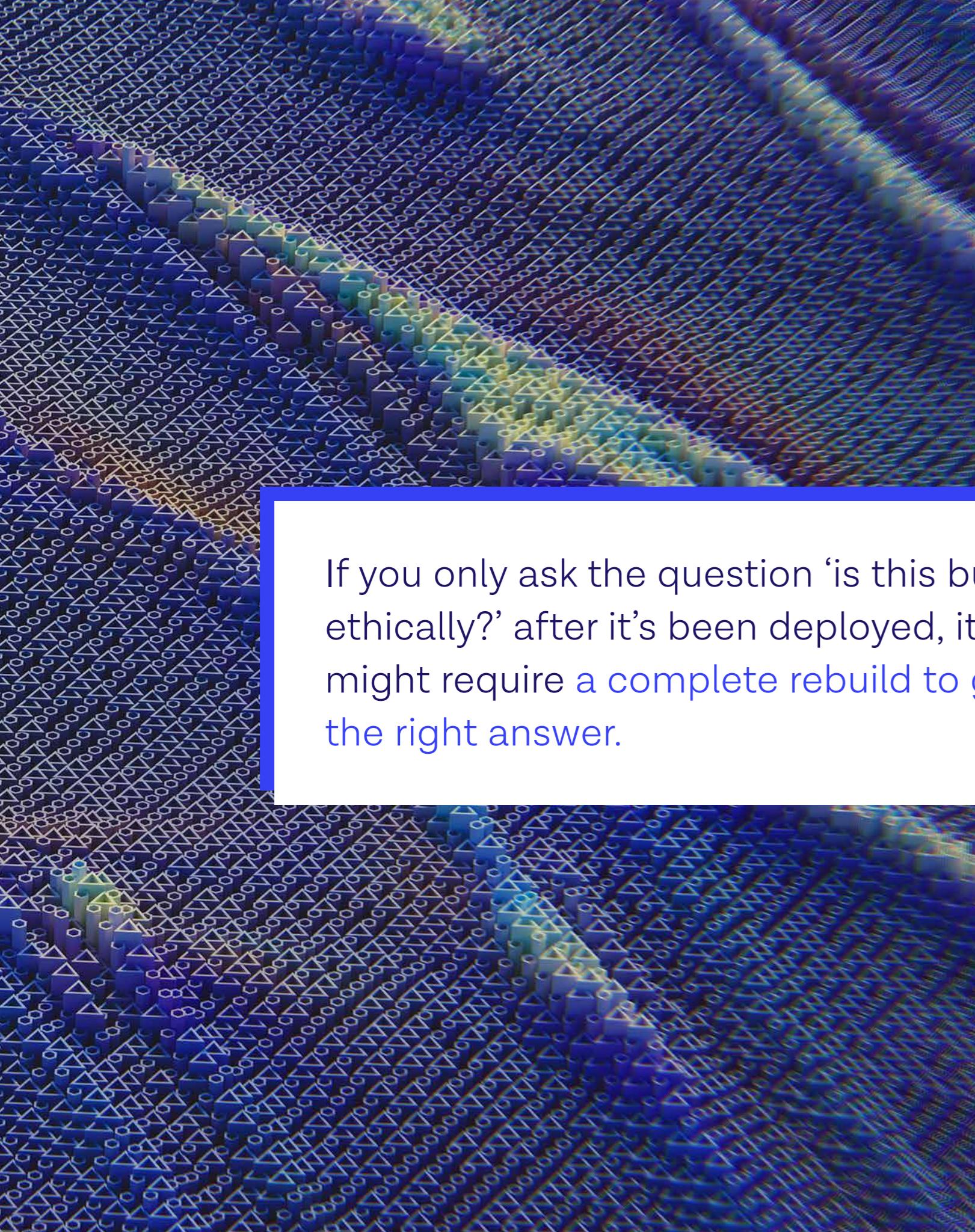
Group FRR / Overall FRR	Male	Female
(95% confidence interval)	0.87 (0.82 - 0.92)	1.18 (1.11 - 1.26)

Age bias

We see a tight grouping of ratios in all but the over 50 group. We believe this discrepancy to be down to the relatively low amount of data we have access to for that age group – we see fewer over 50s using our service. This is something we want to address in the future.



Group FRR / Overall FRR	<25	25-30	30-40	40-50	>50
(95% confidence interval)	0.89 (0.81 - 0.96)	0.83 (0.76 - 0.93)	0.87 (0.80 - 0.95)	1.24 (1.07 - 1.42)	1.71 (1.51 - 1.95)



Reflections and next steps

Creating AI ethically is a [company-wide initiative](#).

It cannot be solved by single teams working in isolation. It requires engineering to build infrastructure with bias-mitigation in mind. Legal and privacy teams need to define rules for how data is defined, labeled, and used during training. Security teams need to ensure this data is held securely and is managed throughout its lifecycle from capture to deletion. Operations teams need to monitor model performance in production and provide the right labeling for engineers and scientists. Procurement teams need to validate new and existing vendors to assess their processes. And crucially, leadership needs to make this a priority.

This is a process that takes time to embed into company culture, and if you delay investment, it's going to take a larger one later. A typical AI use case can take months or even years to take from concept to production. If you only ask the question 'is this built ethically?' after it's been deployed, it might require a complete rebuild to get the right answer.

At Onfido we're proud to have embedded AI bias measurement and mitigation into every step of the research and production processes. Our work is by no means done — and we'll continue to measure and mitigate for as long as there's any discrepancy in performance from one person to another.

If you only ask the question 'is this built ethically?' after it's been deployed, it might require a complete rebuild to get the right answer.

About Onfido

Onfido makes digital identity simple

We make it easy for people to access services by digitally verifying them using our [Real Identity Platform](#).

The Real Identity Platform allows businesses to tailor verification methods to individual user and market needs in a no-code, orchestration layer – combining the right mix of document and biometric verifications, trusted data sources, and passive fraud signals to meet their risk, friction and regulatory requirements.

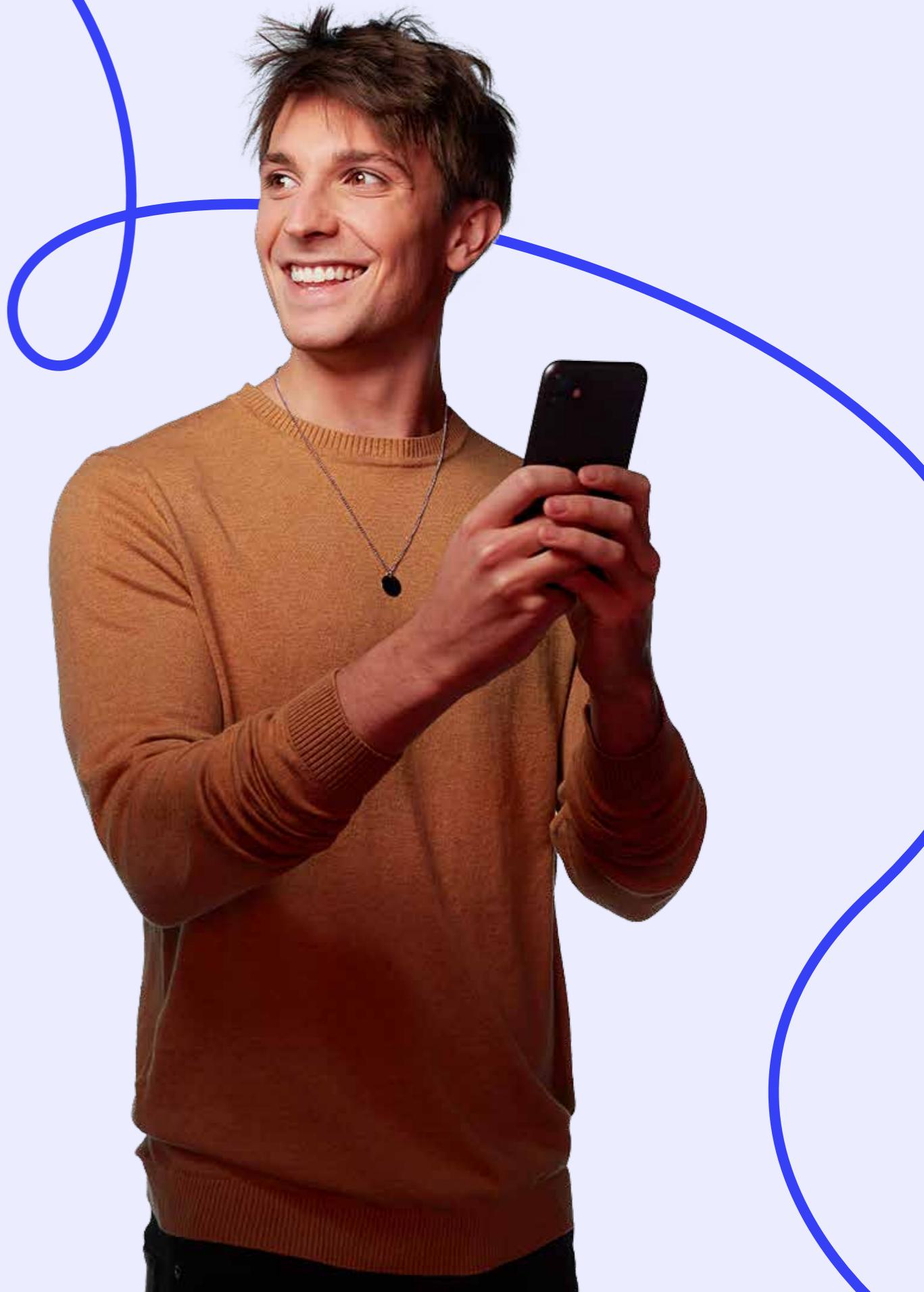
Onfido Atlas™ AI powers the platform's fully-automated, end-to-end identity verification.

Developed in-house for over 10 years, it's how we ensure our analysis is fair, fast and accurate.

Recognized as a global leader in AI for identity verification and authentication, We are backed by

TPG Growth, Idinvest Partners, Crane Venture Partners, Salesforce Ventures, M12 (Microsoft) and others. In 2021, we were awarded 'Artificial Intelligence and Machine Learning Hot Company' by CyberDefense Global Infosec Awards, 'Fraud Prevention Innovation of the Year' at the CyberSecurity Breakthrough Awards, and named to the CB Insights Fintech 250 for the fourth year running.

We partner with over 800 businesses globally to help millions access services every week – from billion dollar institutions to hypergrowth start-ups.
We support checks in 195 countries, and 2,500+ document types.



[Contact us](#)

Authors



Martins Bruveris

Senior Applied Scientist



Richard Tomsett

Senior Applied Scientist



Olivier Koch

Director of Applied Science

Contributors



Romain Sabathe

Applied Science Lead



Ananya Lahiri

Applied Scientist



Sarah Munro

Senior Director,
Biometrics



Giulia Di Nola

Product Manager,
Biometrics