Yoti Age Estimation

White Paper | Full version

MAY 2022

Executive summary

Yoti's facial age estimation is an effective, secure age-checking service that can estimate a person's age by looking at their face. There is no need for a physical check of documents or human intervention.

Yoti's facial age estimation is built in accordance with the 'privacy by design' principle in the UK GDPR. No individual can be identified by the model and it encourages data minimisation because it only needs a facial image. Yoti immediately deletes all images of users. The model cannot infer anything else about a person nor can it uniquely identify a person.

Yoti's facial age estimation is performed by a 'neural network', which we have trained to be able to estimate human age by analysing a person's face. Our technology is accurate for 6 to 12 year olds with a mean absolute error (MAE) of 1.36 years and of 1.52 years for 13 to 19 year olds. These are the two age ranges regulators are most focused upon in order to ensure that under 18s do not have access to age restricted goods and services.

Our True Positive Rate[1] (TPR) for 13-17 year olds being correctly estimated as under 23 is 99.65%. This gives regulators a very high level of confidence that nobody underage will be able to access adult content. Our TPR for 6-11 year olds being correctly estimated as under 13 is 98.91%. Our solution is configurable to meet any regulations that requires prior consent before age estimation is used.

At Yoti, we take our ethical responsibilities as a company developing new technology very seriously. All the data (face image and month and year of birth only) used to **train** the algorithm is obtained by Yoti in accordance with the UK GDPR during the onboarding process for the Yoti apps or using consented data collection exercises. See page 20, Appendix Data used to build the model, for details.

This May 2022 release is our first algorithm that estimates the age from 6–70, using anonymous images that have been given consent to be used for age estimation training purposes. We are pleased to report the algorithm continues to show improvements in accuracy on this iteration. Some small deviations in this trend are best explained by demographic changes in the underlying training and testing data (see page 30 for a detailed discussion).

We are delighted to announce that Yoti's Facial Age Estimation technology has now been approved by the German regulator, KJM, for the highest level of age assurance. This means it can now be used, alongside the Yoti digital ID app, to check the age of German located individuals accessing 18+ adult content.

1. True Positive Rate - the probability that an actual positive will test positive, such as an 18 year old is correctly estimated to be under 2

About 'Mean Absolute Error'

Yoti facial age estimation can make both positive and negative errors when estimating age (that is, it can estimate too high, or it can estimate too low). By taking 'absolute' values of each error we mean ignoring whether the error is positive or negative, simply taking the numerical size of the error. We then take the average (or 'arithmetic mean') of all those absolute error values, producing an overall 'MAE'.A table of MAE by year can be found in the appendix on pages 26-28.

3.

Expanding the data set & improving accuracy

Our first white paper, published in December 2018, contained accuracy across age ranges of 13-60. Since September 2021 we published our 6-12 data for the first time, and we now include data for age range 60-70 From the outset we have built the data for 6-12 year olds with a balanced ratio of data across skin tone and gender.

We are pleased to report the algorithm continues to show improvements in accuracy on this iteration. Some small deviations in this trend are best explained by demographic changes in the underlying training and testing data (see page 27 for a detailed discussion).

Key takeaways

- Mean Absolute Errors (in years) are 2.96 for 6-70, 1.52 for 13-19 & 1.36 for 6-12.
- Users are not individually identifiable
- Helps organisations to meet Children's Codes or Age Appropriate Design Codes
- Does not result in the processing of special category data
- Gender and skin tone bias minimised.
- TPR for 13-17 year olds correctly estimated as under 23 is 99.65%.
- TPR for 6-11 year olds correctly estimated as under 13 is 98.91%.
- Training data collected in accordance with the UK GDPR.
- Independently tested and certified.
- A secure, privacy respecting solution that protects individuals.
- Yoti liveness and age estimation is very hard to 'fool'.
- Over 500 million checks performed worldwide.
- Solution is fast and scales to tens of millions of checks per day.
- Deployments: 'Lite' model on device and full model on premise (law enforcement).

Skin tones

For skin tone, our research team tagged the images using a scheme based on the widely used Fitzpatrick dermatological scale. Fitzpatrick uses six bands, from Type I (lightest) to Type VI (darkest). For the present, we have presented our data in three bands (based on Fitzpatrick Types I & II, Types III & IV, and Types V & VI).

Skin tone scale

Mean Absolute Error by age band

Yoti facial age estimation accuracy | Mean estimation error in years split by gender, skin tone and age band

Gender	Female				Male				All
Skintone	Tone 1	Tone 2	Tone 3	All	Tone 1	Tone 2	Tone 3	All	
6-12	1.31	1.38	1.58	1.42	1.25	1.34	1.30	1.30	**1.36**
13-17	1.41	1.72	1.91	1.68	1.22	1.46	1.64	1.44	**1.56**
18-24	2.43	2.31	2.52	2.42	2.04	1.96	2.08	2.03	**2.22**
25-70	2.94	3.37	4.79	3.70	2.73	3.24	3.77	3.25	**3.47**
6-70	2.59	2.92	3.97	3.16	2.38	2.76	3.16	2.77	**2.96**

With age estimation, once you know you're dealing with a child you can…

Turn off excessive notifications.

Minimise the data you collect - don't store it.

Set geolocation to off but give the child the ability to turn it on if needed.

Shield their data. It shouldn't be used for things not in their interest.

Provide age-appropriate content.

Use child-friendly language to explain platforms.

Be certain the online community is within the same age threshold.

Always be sure to treat a child like a child.

Be certain the online community is within the same age threshold.

What is facial age estimation and what can it do?

Yoti facial age estimation is a secure, effective age-checking service that can estimate a person's age by looking at their face. We consider it to have wide application in the provision of any age-restricted goods and services, both online and in person. It is also a means to combat social exclusion for the significant numbers of individuals around the world who do not possess a state-issued photo ID document.

Yoti facial age estimation is designed with user privacy and data minimisation in mind. It does not require users to register with us, nor to provide any documentary evidence of their identity. It neither retains any information about users, nor any images of them. The images are not stored, not re-shared, not re-used and not sold on. It simply estimates their age.

In a retail setting, facial age estimation can be used at a point-of-sale terminal with a dedicated camera, letting a consumer use a self-checkout without the need for staff assistance. This is not only quicker and less of a nuisance for shoppers, but can greatly reduce friction between them and retail staff.

For general online use, it can be embedded into web pages or incorporated into apps, and receive an image of the user's face from a webcam connected to their computer or the camera in their mobile device. This is ideal for controlling access to age-restricted gaming, gambling and also adult content (pornography).

We believe facial age estimation can play an important role in safeguarding and child protection online, not only in preventing minors from accessing adult content, but also in preventing predatory adults from accessing social media spaces for children and teenagers. This is illustrated well by Yoti's partnership with the Yubo social networking platform. Yubo uses facial age estimation within its app to help identify user profiles where there is suspicion or doubt about the user's age, and flags these cases to its moderation team.

Deployment on premise and on device

Facial age estimation can also be deployed on premise by law enforcement to assess ages of victim and perpetrators in child abuse images. We have also developed a more efficient and lightweight age estimation model that can run on platforms with limited or low computational resources and mobile devices. This lightweight age estimation model provides much faster results, has no reliance on internet connectivity and is just 8.4% less accurate than our production model.

A further potential use is at the entrances to age-restricted premises such as bars, nightclubs and casinos. In this kind of application, facial age estimation offers clear advantages – it does not get fatigued on a long shift,[2] and it cannot show favour to personal friends, or bias against individual customers. It is very hard for under 18s to 'fool'. It also reduces the burden on retail staff to try and estimate customer ages and it can be used to reduce abuse to staff.

2. Studies have shown that the objectivity of human judgement of this kind can be significantly affected by hunger and fatigue – see for instance Danziger, Levav, Avnaim-Passo (2011) *Extraneous factors in judicial decisions*, PNAS April 26, 2011 108 (17) 6889-6892; *https://doi.org/10.1073/pnas.1018033108*

Data privacy and network security

Yoti's facial age estimation has been designed with data privacy and security as primary considerations.

The user does not have to register to use the service, and does not have to provide any information about themselves. They simply present their face in front of the camera. Their image is not stored locally on the point-of-sale terminal. It is securely transmitted to the Yoti backend server (currently hosted in the United Kingdom), secured by TLS 1.2 encryption. After the age estimate is performed, the captured facial image is deleted from Yoti's backend servers.

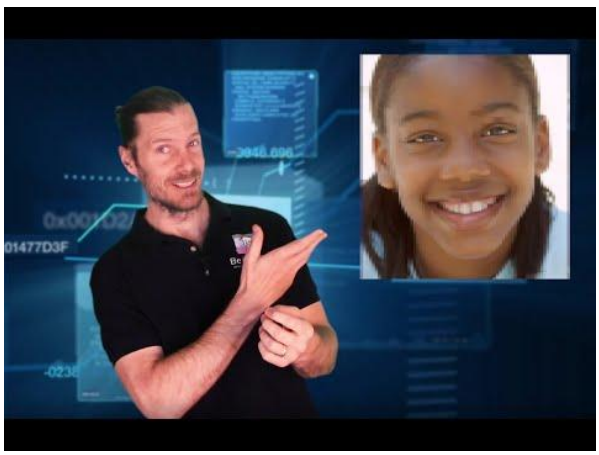
Although Yoti facial age estimation works by processing a facial photograph, under the GDPR definition of biometric data it is not a 'biometric' method of age checking, as our means of processing does not allow the "unique identification or authentication of a natural person". Instead it deletes the captured photograph and merely returns an age estimate.

The photograph is not viewed by any Yoti staff. In GDPR terms, Yoti is a data processor for the facial age estimation service. The relying party (Yoti's customer) is the data controller. As such, the relying party will decide the lawful basis for their use of facial age estimation (if required under EU / UK privacy law). In some jurisdictions, the individual will need to provide consent. The facial age estimation user interface is configurable so that relying parties can build in this request for consent. This feature is enabled by default for our US customers.

How does it actually work?

Facial age estimation is based on a computing technique known as a 'neural network', which we have trained to be able to estimate human age using a process of 'machine learning'. This is a form of artificial intelligence (AI), and is increasingly used in a wide variety of applications, from driverless cars to medical diagnosis, from tailoring online advertising to detecting credit card fraud. We discuss machine learning in more detail on the next page, but first some context on the problem we are using it to solve.

3 minute video explanation of Facial Age Estimation, delivered by Yoti partner Be in Touch



https://www.youtube.com/watch?v=6KCUO2vln3M

Tackling the challenge of age determination

Determining a person's exact age in the absence of documentary evidence of their date of birth is a difficult task. Indeed, the truism that 'age is just a number' could be said to have a sound scientific basis. By 'ageing' in a medical sense, we mean the physiological changes which occur when individuals develop and grow from juvenile to mature forms, and then the types of damage that progressively accumulate within the human body as time passes. The important point is that the rate at which human bodies 'age' in this way is influenced by numerous external factors other than simple passage of time. Factors that affect the ageing process, both in the long and short term, can include: quality of diet and nutrition, exposure to disease, adverse environmental conditions, use of narcotics, physical labour, stress and lack of sleep. Clearly, there are large variations throughout populations as to how different individuals are exposed to these ageing factors. The more extensively we look through different countries, ethnicities, and socio-economic groups, the wider these variations in exposure to ageing factors become.

It may be surprising to learn that there are currently no entirely reliable medical or forensic methods to determine human age. Two of the more commonly attempted medical techniques focus on trying to ascertain whether the subject is above or below the legal age of maturity. These are X-ray or Magnetic Resonance Imaging of bone structure in the wrists (the degree to which the cartilage between the carpal bones has ossified) and dental X-rays (examining the maturity of wisdom teeth). However, both of these methods have a typical margin of error of at least two or three years, and for individuals with an atypical history to the general population, the error can be significantly worse. Due to this unreliability, their use has proved controversial – for instance, their use by immigration authorities to attempt to differentiate between child and adult refugees who have no documentation. It is also completely impractical to try and x-ray shoppers' teeth at the self checkout.

Other medical techniques examine 'biomarkers' taken from blood or tissue samples. Examples include measuring the degree of DNA methylation present, the length of the 'telomere' portion of chromosomes, or the serum levels of the metabolite C-glycosyl tryptophan. Whilst these biomarker techniques tend to provide good indicators of ageing processes in an individual, they do not correlate reliably with their chronological age from date of birth.

Ultimately, it could be argued that much of the difficulty in trying to measure 'age' (that is, a person's chronological age from their date of birth) arises because 'age' defined this way is a rather arbitrary quantity that does not mean anything definite in physiological terms. Science can accurately measure the extent to which a person's body has aged (that is, how to what extent it has developed, grown, matured and decayed), but cannot always reliably determine how many years it took for their body to arrive at that state.

Human ability to determine age

Notwithstanding the difficulty in devising an accurate forensic test for age, people still possess a reasonably good ability to guess someone's age simply by looking at them. And some people can come within a few years of the right answer. How do we manage it? In terms of facial features, what are the tell-tale signs we look for?

The most obvious visual cues include bone structure (bones grow and develop as we pass from child to adulthood), skin tone (wrinkles, elasticity) and hair colour (greyness), male baldness or facial hair after puberty. We could add many more cues to this list. However, whatever the detailed nature of the visual cues, the more general point is this: as humans, we simply learn "that's what people of a particular age look like". As we go through life, we encounter other people, we see what they look like and we learn how old they are, with varying degrees of precision (e.g. "a baby", "14", "mid-40s", "79" and so on). We accumulate this information and experience throughout our lives, and our brains can use it to make quick intuitive judgements. The extent of our previous experiences will be an important factor in how good our guesses are. We will be more accurate at guessing the age of someone from our own familiar peer group than from one we've not encountered.

However, whilst some people are good at estimating age, others are less good and this variability can frustrate teenagers and young adults who are often age estimated and asked to provide physical proof of age.

A study in this area[3] reported an MAE in human guesses of 4.7 years across an age range of 0 to 70; across an age range of 16–70, this rose to an MAE of 7.4 years.

It is worth emphasising that, although we might be able to retrospectively rationalise or refine our guess at someone's age, our initial judgement is more or less intuitive. We are not consciously following some step-by-step, rule-based method (for instance "add five years if there are wrinkles", or "add ten years for grey hair"). In effect, we don't 'know how we do it' – generally, our brains process the image and form an instinctive judgement, in line with what we've learnt from past experience, faster than any conscious deliberation or systematic evaluation of facial features. It turns out that this 'black box' approach to describing our cognitive process (that is, simply training our brain with data, without worrying too much about how it works) can actually be employed as a successful technique in machine learning too.

3. H. Han, C. Otto, X. Liu and A. K. Jain (2015) *Demographic Estimation from Face Images: Human vs. Machine Performance*, IEEE Trans. PAMI, Vol. 37, No. 6, 1148–1161 https://doi.org/10.1109/TPAMI.2014.2362759. See also Clifford CWG, Watson TL, White D. (2018) *Two sources of bias explain errors in facial age estimation.* R. Soc. open sci. **5**:180841. http://dx.doi.org/10.1098/rsos.180841 and Voelkle, Ebner, Lindenberger & Riediger (2012) *Let Me Guess How Old You Are: Effects of Age, Gender and Facial Expression on Perceptions of Age.* Psychology & Aging, **27** No.2 265–277. https://doi.org/10.1037/a0025065

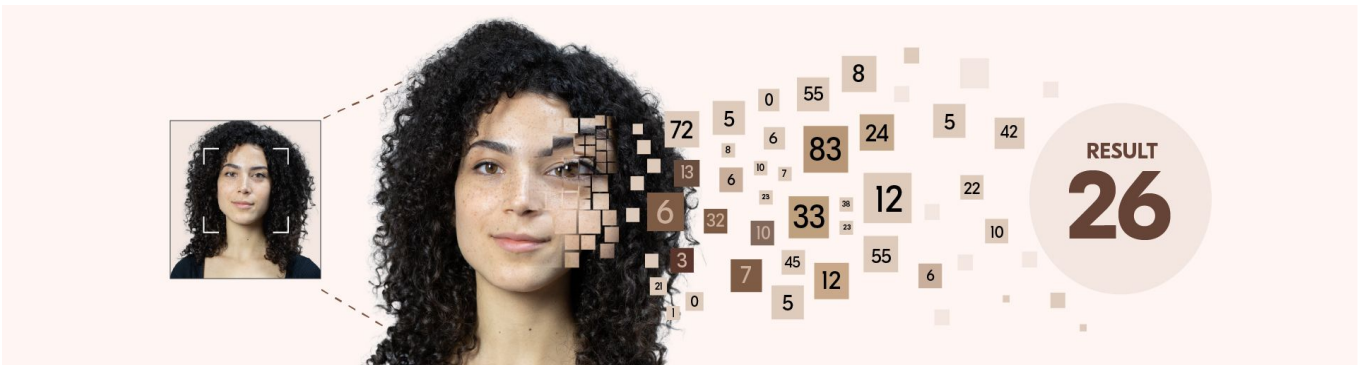
Supporting Children's Codes

Given the growing importance of age checking online for younger children and teenagers we have recently introduced additional training data to enhance our algorithm to estimate 6 to 12 year olds.

The Age Appropriate Design Code, originating in the UK, is driving a movement globally to design online interaction 'age appropriately' across the 4 C's - be that content, conduct, contact or contract[4]. The challenge for designers and platforms is to enable young people to be supported to thrive online whilst also enabling age appropriate interaction, protecting against detrimental content, grooming and supporting age appropriate content moderation. We can support platforms to recognise child users and so not employ nudge techniques or encourage children to provide unnecessary personal data, or make a child's real time location publicly available. Children should no longer be encouraged to stream to large groups of unknown adults. There are a growing number of countries around the world also reviewing legislation for a range of age restricted goods and services; in particular age assurance for access online. There are also adult content sites already using Yoti age estimation successfully to prevent children from accessing their websites.

Obtaining consented data to develop our software to accurately estimate 6 to 12 year olds has been a significant challenge. We have worked hard to ethically obtain parental consent to use anonymous images of children in our training data; that is facial images with month and year of birth. We can now correctly estimate 63% of images of 16 year olds to be between 15 and 17. For 6-12 year olds, our first MAE results are already within 1.36 years, so could be used effectively for triaging access to 13+ apps.

Over the coming months we will continue to invest more to improve our accuracy to make the internet safer for young people.



Detect face

A face is detected in an image and reduced to pixels. Each pixel is assigned a number that the AI can understand.

Compute numbers

The numbers are computed by a neural network that has been trained to recognise age by looking at millions of images of faces.

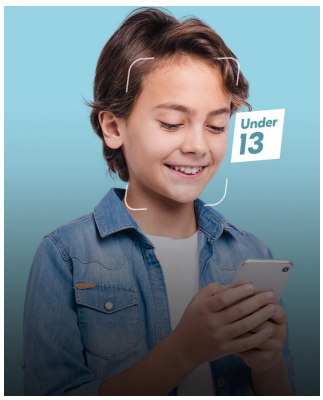
Determine age

The AI finds a pattern in the numbers and produces an age.

4. Livingstone, S. and Stoilova, M. "The 4 Cs, Classifying Online Risk to Children." SSOAR, 2021. https://doi.org/10.21241/ssoar.71817.

More on how it works

The first challenge for facial age estimation is ‘face detection’. It has to examine the image it gets from the camera, and work out which bit of it is an actual human face. Only this portion of the image is then fed into the neural network to get an age estimate. This stage also allows for basic error checking: if the system can’t find a face in the image (for example, because a customer didn’t position themselves properly in front of the camera, or some inappropriate object is put there) then the system can return an error message instead. This is also the stage when Yoti can check to be sure the face is a real face in front of the camera.



Note that this is not ‘facial recognition’ (where a computer system is trying to match a particular face against a database, to confirm that person’s identity). It is simply detecting whether or not there is anything in the captured image that looks like a human face.

We now come to the interesting bit. The facial image is made up of pixels. To the computer, each pixel is just a set of numbers. These numbers are fed into the artificial neural network. This is a network of mathematical processing nodes, arranged in layers, that is roughly analogous to the connections in the human brain. Whilst a typical brain has around 100 billion neurons, the artificial neural network has just hundreds of thousands of nodes. We feed numbers (pixel data) in, and they percolate through the neural net. Each node performs a mathematical function on the pixel data, and passes the result on to nodes in the next layer, until a number finally emerges out the other side. This number is an age estimate.

It’s an obvious question to ask ‘how is the neural network processing the data? What is it looking for – wrinkles? grey hairs?’ and so on. However, this is a rather human way of thinking about it, and it’s not really a very useful question to ask: to the computer, it is just being fed numbers. It doesn’t ‘know’ what the numbers represent or what they mean. We don’t try to tell it that. What we have told it, in the training phase when facial age estimation was being developed, was what the right answers were. In the training phase, we fed it millions of diverse facial images, for which we knew the subject’s age with confidence. The neural network keeps digesting the pixel data from each image, processing the numbers, and trying to get a result which matches the right answer. It keeps repeating the process, adjusting the processing, keeping the variations which bring it closer to the right answer, rejecting the variations which don’t help – in other words, it is ‘learning’.

After repeating the process a huge number of times, it arrives at sets of processing formulae which work best. To a human, these formulae would be bafflingly long and complex, and next to meaningless (and no, we’re not going to print them here…for one thing, they wouldn’t fit on the page!). However, it has effectively created a very complex model of age determination that is far superior to relying on a set of handcrafted instructions that a human programmer might supply.

The quality of the training data is crucial to any machine learning process. To train our facial age estimation algorithm, we use millions of images from Yoti users who have opted in to this use of their data. The process is explained to them at onboarding, and is discussed in more detail in the Appendix to this paper. They are free to opt out of this research at any time simply by selecting this in the Yoti app's settings. Most Yoti users want Yoti to make their lives safer and simpler, and they understand that using their data for internal research purposes is how we are able to improve and develop the products and technology to achieve this. We will publish white papers that demonstrate such applications.

For facial age estimation, these research images are tagged with only two attributes taken from a verified ID document that they have uploaded: their gender and their month and year of birth. Supported documents include passports, driving licences and national ID cards. We believe the size, diversity and verified age accuracy of this training data set gives Yoti's facial age estimation an advantage over competing solutions.

Deployment

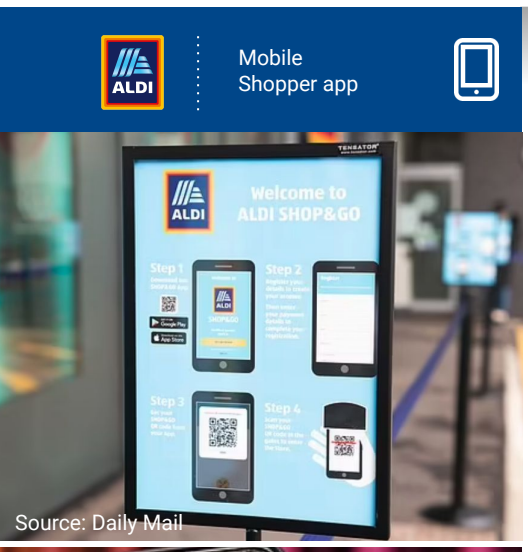
- **On a terminal** - retail EPOS, gambling terminals
- **SAAS** - assessing age or age band for person to access online ecommerce (tobacco/vaping/alcohol, pharma), social media/live streaming, gambling, gaming, adult content, dating
- **SAAS** - assessing age of adult to support parental consent
- **On premise** - law enforcement to assess age of victims & perpetrators in CSAM
- **On device** - eg prevent CSAM/nude sharing from phone / *VR headset*...
- **On device** - ensures solution effective even when connectivity is unavailable

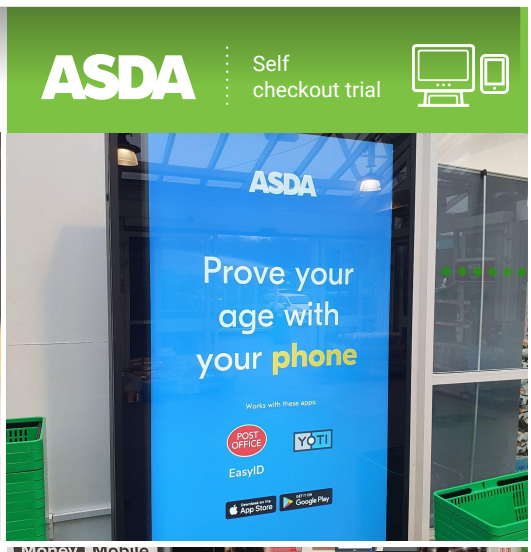
Practical use

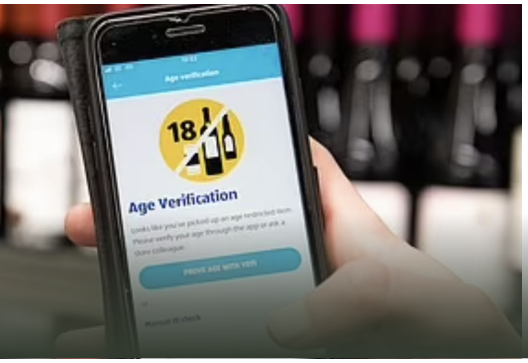
Facial age estimation works quickly, returning an age estimate in around 1 to 1½ seconds. The user needs to present their face to the camera, uncovered (although glasses do not usually present a problem). We recognise that in some areas internet speed can be challenging; we can cater for small image sizes of 50-100KB. We have scaled to handle tens of millions of checks per day and currently can handle up to 130 checks per second but can easily scale higher.

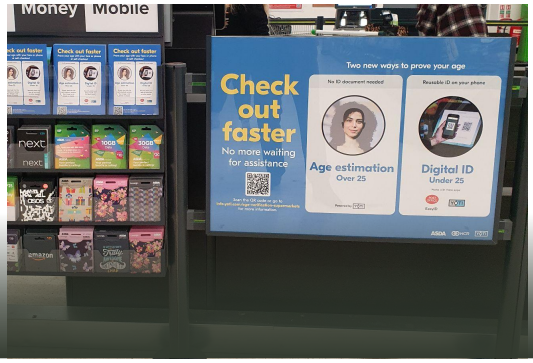
Dim lighting is not helpful; bright ambient light works best. Our research has found that the effect of beards and facial disfigurement can make a minor impact, however does not materially affect estimated ages. In response to the ongoing COVID-19 pandemic, we have been researching how facial age estimation copes when a person is wearing a mask covering the lower half of the face. Results suggest that whilst accuracy is reduced somewhat, acceptable performance can usually still be achieved as long as a larger safety buffer is used.

Live trials in the UK & European markets









Product developments

Our R&D and product teams are always striving to improve our service, not just in terms of the accuracy of the age estimation algorithm, but also solving practical problems of deploying the service in different environments. We work very closely with our partners to ensure their needs are met, within a vast global network of regulatory challenges and user environments.

On device - facial age estimation ‘Lite’

We have developed a much smaller and efficient ‘Lite’ version of the model that performs offline or directly on device without having to make calls to our servers. Our first on device “Lite” model benefits from being 87% smaller but is just 8.4% less accurate and provides much faster results with no reliance on connectivity.

Interoperable Age tokens

A problem for internet users has been having to prove their age on different sites, many times in a short period. Age tokens store the result of an age check as an anonymised attribute that can be re-used to gain access to integrated sites and services. They work a bit like a cookie and don't store any personal information, just that someone is a required age. Age tokens can be verified at the point of use by an integrated site, which defines their criteria for a token inline with business and regulatory requirements. The elegance of an age token ecosystem is enabling a wide set of age checking providers to provide the maximum utility for consumers; reducing the number of times they have to keep proving their age.

Inclusive facial age estimation

Our goals are twofold - to enable anyone in the world to prove their age for free in seconds and to provide an age estimation that any organisation can rely on. This can be both online and in person. To build trust, we strive to provide transparency as to the accuracy; bias levels, data sets. We also undergo independent industry certification and gain where available regulatory approval. We work with many trusted brands.

Anti-spoofing technology

We’ve developed proprietary anti-spoofing technology to prevent fake images being used for age estimation. Our passive liveness technology analyses the depth of an image to make sure it’s a real person and not a photograph, video or bot. As one would expect we measure bias on our liveness detection technique. We intend to seek independent review and or certification of the bias of liveness detection.

Legal compliance

Whilst legal compliance is a complex area, it is important to cover given there are understandable concerns about the potential unlawful use of personal and biometric data by governments and businesses.

Yoti's facial age estimation complies with the UK GDPR and the EU GDPR, and also our own ethical approach to user data and privacy. When clients use facial age estimation to age verify their users Yoti acts as the data processor, with clients as the data controllers. Our clients therefore need a legal basis to use facial age estimation. The technology processes personal data so the legal basis for processing will either be consent, performance of a contract between the client and the user or legitimate interests of the client that do not unfairly prejudice the user.

The Yoti Age Portal has a consent option built in so clients can easily collect consent if that is the lawful basis the client decides upon.

Yoti's facial age estimation does not involve the processing of special category data - this has been confirmed by the UK Information Commissioner's Office. This is because the age estimation model is unable to allow or confirm the unique identification of a person and it is not being used for the purpose of identification (and that is the key test for special category data). Put simply, if you put the same face into the model several times the model would have no idea it is the same face (and no way of working that out) and it would give slightly different age estimation results each time. The model was not trained to recognise any particular individual's face, but instead to categorise that face it is presented with into an age.

Definition of special category data in Article 9 of the UK GDPR:

> *Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data,* ***biometric data for the purpose of uniquely identifying a natural person****, data concerning health or data concerning a natural person's sex life or sexual orientation*

Recital 51 of the UK GDPR further says that:

> *The processing of photographs should not systematically be considered to be processing of special categories of personal data as they are covered by the definition of biometric data only when processed through a specific technical means allowing the* ***unique identification or authentication of a natural person.***

For more information about why Yoti's facial age estimation does not process biometric data, please see our blog here.

Fair, standardised measurement

Up until now in our white papers, we have published the accuracy levels (MAE) in age ranges: 13-15, 16-17, 18-24, 25-29, 30-39, 40-49, 50-60. However based on feedback from stakeholders, in this edition of the white paper we have decided to change the format. You will now find the accuracy levels table at the front of this paper for ease of access and we now include the accuracy levels (MAE) for each year of age, across gender and skin tone, from age 6 to age 70. So the previous overall MAE from the September white paper was the average of all the testing data (2.19). Now the overall MAE is weighted equally by year of age and is 2.96.

Why have we introduced this?

The reason for stating MAE for each year is so that customers and regulators can look at the age ranges that interest them; rather than only see an average from a group of years, where there may be imbalances in test set sample sizes, particularly within large age ranges. This avoids an imbalance in one year, or a small number of years, skewing the average results over a large age range. We continue to strive to get as diverse testing data as possible - across age, gender and skin tone. The ideal would be that our test data would be equal, 50:50 on gender and equal weighting for each of the six skin tone categories on the Fitzpatrick scale.

Do MAEs always improve?

Our test set changes over time due to regulatory data retention requirements. We currently only retain test data for three years, so we need to regularly introduce new data. This is perhaps an unintended consequence of privacy regulations which require a deletion period for all data. There is a strong case for regulators to allow a longer retention period for this type of AI performance testing data to ensure trustworthy measurement of performance improvement over time. As we add more diverse data and equalise testing data the MAEs of darker skin tones will continue to improve, but the average for certain years may temporarily deteriorate. Therefore, there may be instances when we have to accept marginal non improvement in a small number of years, until we have removed all evidence of material bias.

Recommendations for fair and accurate measurement of Facial Age Estimation:

- Test image data must be separate from training image data.
- Test subject identities must be separate from training identities – different images of the same individual should not be in both training and test data.
- Test data needs to be diverse in terms of gender and skin tone and the percentages should be published by year of age, gender and skin tone (categorised based on the Fitzpatrick Scale).
- Suppliers must be prepared to accept independent external review and publication of results.

MAE

- MAEs should be published for each year, gender and skin tone (six demographics).
- MAEs for these 6 demographics should be averaged for every year of age and average MAEs should be averaged for any published age range.

How accurate is facial age estimation?

We believe that when presented with a clear facial image, automated facial age estimation compares very favourably with human abilities.

Furthermore, when viewing a succession of faces, a person's judgement tends to be influenced by the preceding faces they have just seen, which is not a problem that affects facial age estimation. Humans tend to systematically underestimate the ages of older people, and overestimate the age of younger people, and our ability to estimate accurately tends to decrease as we ourselves get older. The latter problem clearly has particular implications for provision of age-restricted goods and services, where we need to check whether teenagers are above or below a required legal age.

Currently, the MAE across the entire data set, de-skewed to give equal weighting to male and female subjects, is 2.96 years and just 1.52 for 13-19 year olds. Further detail on our algorithm's accuracy, broken down by gender, skin tone and age range, is presented in this paper's appendix.

The vast majority of organisations who need to check age need to check whether individuals are over the age of 13, 18 or 21. We recognise that we still have further to go to reduce bias for older age groups, particularly individuals with skin tone V & VI. However, these older individuals are not materially disadvantaged when the age of interest is for example is 18-21 and the thresholds are usually between 25 and 30.

Yoti's facial age estimation has been certified by the Age Check Certification Scheme for use in a Challenge 25 policy area. The ACCS report is available at: https://www.accscheme.com/registry.

Age Check Certification Scheme

About 'Mean Absolute Error'

Yoti facial age estimation can make both positive and negative errors when estimating age (that is, it can estimate too high, or it can estimate too low). By taking 'absolute' values of each error we mean ignoring whether the error is positive or negative, simply taking the numerical size of the error. We then take the average (or 'arithmetic mean') of all those absolute error values, producing an overall 'MAE'.

The average MAE can be measured as;

i) the average of each year's MAE - eg. there are 65 year MAEs in the 6-70 age range,
ii) the average of each summary age range MAE - there are 10 age ranges shown in the table on page 24,
iii) the average of all the images in the training data (but this data may be skewed towards certain ages with more training data).

4. H. Han, C. Otto, X. Liu and A. K. Jain (2015) *Demographic Estimation from Face Images: Human vs. Machine Performance*, IEEE Trans. PAMI, Vol. 37, No. 6, 1148–1161 https://doi.org/10.1109/TPAMI.2014.2362759. See also Clifford CWG, Watson TL, White D. (2018) *Two sources of bias explain errors in facial age estimation.* R. Soc. open sci. **5**:180841. http://dx.doi.org/10.1098/rsos.180841 and Voelkle, Ebner, Lindenberger & Riediger (2012) *Let Me Guess How Old You Are: Effects of Age, Gender and Facial Expression on Perceptions of Age.* Psychology & Aging, **27** No.2 265–277. https://doi.org/10.1037/a0025065

Safety buffers

As discussed above, just as human estimators have a capacity for error, so does facial age estimation. To manage this potential for errors, we recommend using facial age estimation as part of a strategy such as the British Beer & Pub Association's 'Challenge 21'[5], which is already widely adopted by publicans and their bar staff in England and Wales. This type of strategy works as follows: Certain goods and services can only be sold to customers over a particular age (e.g. 18 years old). However it is difficult for human staff to be sure whether someone is over 18 just by looking at them. Conversely though, it is fairly easy to tell if someone is significantly older than 18, and customers in this age range would find it an unjustifiable inconvenience to have to show ID to prove their age. Therefore, the store's policy is to only require customers to prove their age if they appear to be under 21. Most supermarkets in England use a Challenge 25 policy.

Facial age estimation can be configured to work with legal age thresholds in a similar way. Furthermore, and unlike human staff, facial age estimation capacity for error is well quantified statistically. This makes it easier to choose a suitable buffer that is comfortably outside facial age estimation's margin of error, and configure the system to estimate whether customers are above or below that threshold.

As an example, consider the situation in the USA, where the selling of alcohol is restricted to over 21s, and common practice today is for retailers to challenge people who appear to be under 40. In this case, a retailer using facial age estimation might choose to set an initial threshold of 30. If facial age estimation estimates that the customer is at least 30 years old, then no further age checking is required.

If facial age estimation estimates that the customer is below 30, then they will be directed into a user flow where they need to present documentary proof of their age (for example, using their Yoti app that is anchored to their passport, driving licence or national ID card). Testing on our current model shows that with a threshold set to 30, only 0.1% of under 21 year olds would incorrectly pass unchallenged by facial age estimation[6], which compares very favourably with the accuracy of human staff. This is great news for the 30 plus population – they will not need to provide ID document evidence of their age and they will be able to happily leave their documents at home.

Since early 2019, we have spent much time reviewing the appropriate size of buffer for a number of use cases. We have come to the conclusion that this depends on a number of variables. The primary one is the demographic of users. The under 18 age group is the chief area of concern for regulators globally in terms of age restricted goods and services. Given the improvements in accuracy of facial age estimation for this demographic, we now suggest a buffer of 3–5 years for highly regulated sectors (e.g. adult content, gambling, alcohol, tobacco) is most appropriate for the 13–25 age band, whereas no buffer may be deemed fine for social media or gaming use case. In some countries, more cautious regulators may initially look for a higher buffer. For a jurisdiction with legal age restriction of 18, and a threshold set to 28 (a 10 year buffer) we would currently have a 0.06% error rate (that is, only 1 in 1,666 14–17 year olds would be incorrectly let through). With a threshold set to 25 years, facial age estimation's current error rate is 0.2%. For a threshold of 21 years, the error rate is 1.04% assuming equal numbers of 14-17 year olds in the test sample[7].

For a demographic of senior citizens, such as for a travel entitlement use case, a regulator may consider a buffer of five to seven years would be more appropriate.

5. See https://beerandpub.com/campaigns/challenge-21/
6. For more information see page 31
7. For more information see page 30

However, there is not currently a commercial demand from relying parties or regulators for age estimation of this demographic. This will always be discussed with the relying party and with the relevant sector and jurisdiction regulator. Over time, as the accuracy of age estimation technology increases, regulators will be able to set lower buffers with confidence.

More statistical detail on facial age estimation 'false positive' rates for a selection of different thresholds and buffers is presented in the appendix of this paper. It is also worth considering 'false negatives' too (where facial age estimation incorrectly estimates someone as being younger than the threshold age), as these can be a source of unwanted friction. False negative rates are also discussed in the appendix.

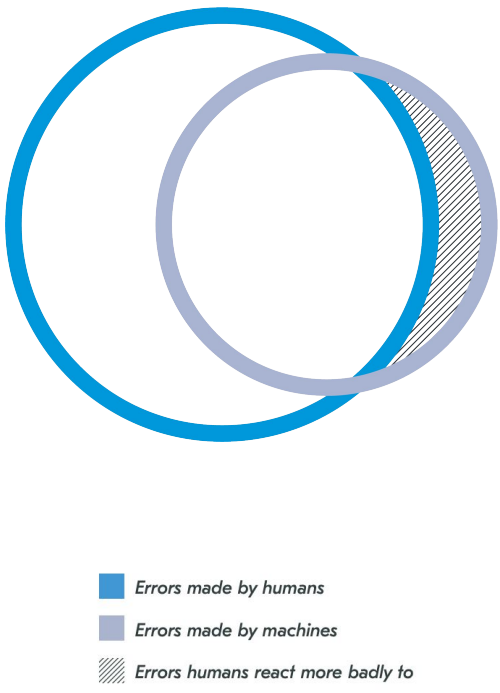
Public acceptance of AI technologies

When discussing the accuracy of facial age estimation, it is worth considering a general point about machine learning and the public's attitude to AI technologies of all kinds: namely, how unforgiving humans tend to be in regard to mistakes made by AI.

Whilst we feel it is fair to claim that the accuracy of facial age estimation generally compares very favourably with human judgement in the broad majority of cases, there will inevitably be rare occasions where it 'makes mistakes'. Of course, humans make mistakes too. However, sometimes machine learning systems make mistakes that no human would have made. This is illustrated in the Venn diagram below.

As can be seen, typically, humans make errors, just as a well-trained machine learning system does. Furthermore, in most of the cases where the machine system gets it wrong, a human would make the same mistake. However, humans tend to be much more bothered by the small percentage of cases on the right of the diagram – these are cases where the machine learning system makes a mistake, but a human would not have been fooled. It can be argued that this is an irrational reaction.

Nevertheless, the general public may often unduly focus their attention on the machine failings, until they become comfortable with the new technology. We believe that digital approaches can be harnessed to support age appropriate design of services, enabling data minimisation, improving online safety and countering certain online harms. One of the objectives of this white paper is to support the education of the public.



Yoti's commitment to ethical use of AI technologies

At Yoti, we take our ethical responsibilities as a company developing new technology very seriously.

Our Data Protection Officer has completed a formal Privacy and Ethics Impact Assessment for Yoti age-checking solutions, which is available on request to potential clients. It covers Yoti both as a data controller for our own use of agechecking solutions with our own users, and as a data processor when offering age-checking solutions to corporate customers.

We have set up an internal Ethics Committee with members from several different areas of our business, to consider ethical issues related to our technology and its use. We used frameworks such as 'Responsible 100' and 'Digital Catapult' as starting points for the scope of these considerations. Findings of the committee are shared with Yoti's senior management teams, Board of Directors and our Guardian Council.

External scrutiny

We have also obtained an ISAE 3000 assurance report from one of the top four global auditing firms, validating our age checking services as being in accordance with the British Standards Institution's PAS 1296 code of practice.[8]

In July 2019 our age checking solutions were assessed under the Age-verification Certificate Standard, a scheme run by the UK government's then Age-verification Regulator (the British Board of Film Certification).

The assessment considered whether a solution was effective and followed an approach of data protection by design and by default. Yoti were the the only company in the UK to achieve this certification[9].

The German Association for Voluntary Self-Regulation of Digital Media Service Providers (FSM) awarded us its Seal of Approval for our age verification solutions[10].

We have hosted three roundtable sessions to get feedback from a range of industry practitioners on unintended consequences of our approach. Participants from the UK included the University of Warwick, the University of Keele, the Home Office Biometrics Ethics Committee, the Children's Commissioner for England, the NSPCC, the ICO, GCHQ, and groups such as Women Leading in AI, and techUK[11].

We have also been actively reaching out to organisations representing various minority groups to seek their views and input, including the UK transgender charity, Sparkle.

We have asked the US Centre for Democracy & Technology to perform a deep dive with full access to our CTO and tech team. We have sought comment from World Privacy Forum and Future of Privacy Forum.

In addition, we commissioned a report from a leading academic which reviews the accuracy and bias mitigation of the facial age estimation algorithm.

Certified B Corporation

FSM

8. PAS 1296: 2018 *Online age checking—Provision and use of online age check services—Code of Practice*. Available from the British Standards Institute shop.bsigroup.com.
9. https://www.bbfc.co.uk/about-bbfc/media-centre/bbfc-statement-age-verification-under-digital-economy-act
10. https://www.fsm.de/de/fsm.de/yoti
11. https://www.yoti.com/blog/age-estimation-technology-tackles-grooming-online/

Appendix

This appendix provides further detail on the current accuracy of facial age estimation. Taking confidence from the trends we've seen in past months (illustrated below), we expect these figures to continue to improve as the volume and diversity of our dataset increases.

Data used to build the model ('training data')

We have invested significantly in building a leading R&D team since early 2015, working on a variety of AI initiatives.

The current production model of facial age estimation (May 2022) was built using a training data set taken mainly from Yoti apps' users (though not US users). We provide information to users at onboarding about our use of biometrics with links to more details, including the Privacy Notice[12] where the use of user data by our R&D team for internal research is extensively detailed. The screenshots overleaf show the current onboarding screen and the screen where users can opt out of their data being used for R&D activity.

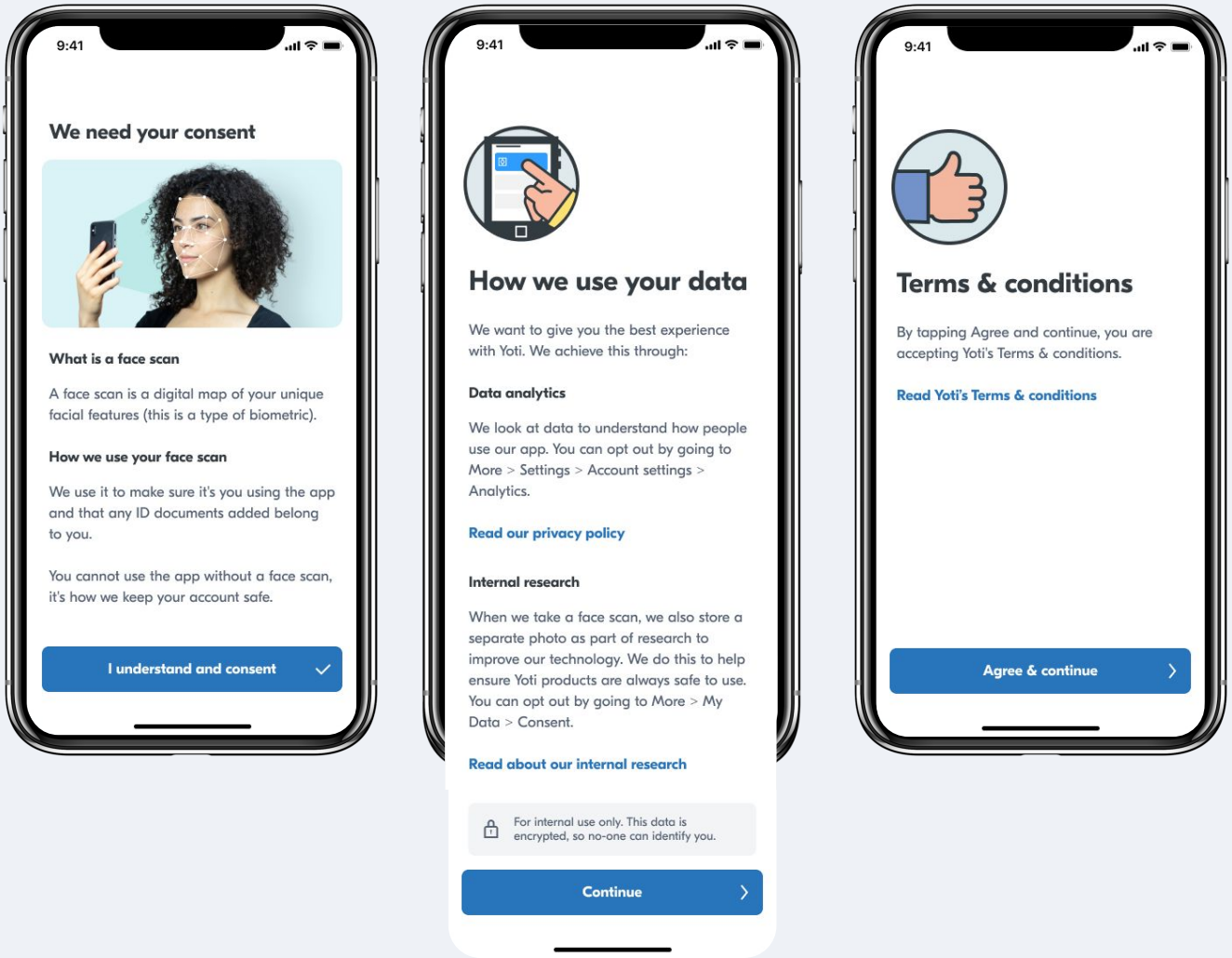
Any user can go to the app settings at any time and opt out of R&D use of their data. This prevents further data from that user being sent to R&D, and it deletes all the data associated with that user that is on the R&D server and available for R&D to use. We have chosen to automatically delete the existing data when a user opts out or deletes their account, even though we do not legally have to under the research provision in GDPR article 17(3)(d).[13] We employ a privacy-by-design approach (hashed numbering) so that although we can find data of a specific user to action the data deletion, there is no way to recreate a specific user's identity from that R&D data.

To enhance our coverage of particular demographics, further age-verified images were gathered by Yoti with consent in Nairobi, Kenya. Through the Share2Protect campaign, we have enabled parents and children to support the extension of the facial age estimation to extend to 6-13 year olds.[14] We have also purchased further parent consented child facial images, with month and year of birth. We undertook thorough due diligence on all our data sources.

In 2021 Yoti has been part of the ICO Sandbox to extend Yoti facial age estimation AI programme to under 13 year olds without ID documents.[15]

12. https://www.yoti.com/privacypolicy
13. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) https://eur-lex.europa.eu/eli/reg/2016/679/oj
14. https://www.yoti.com/blog/protecting-kids-safer-internet-day-2021/
15. https://ico-newsroom.prgloo.com/news/ico-supports-projects-to-strengthen-childrens-privacy-rights

On-boarding and R&D opt-out screens in the Yoti app



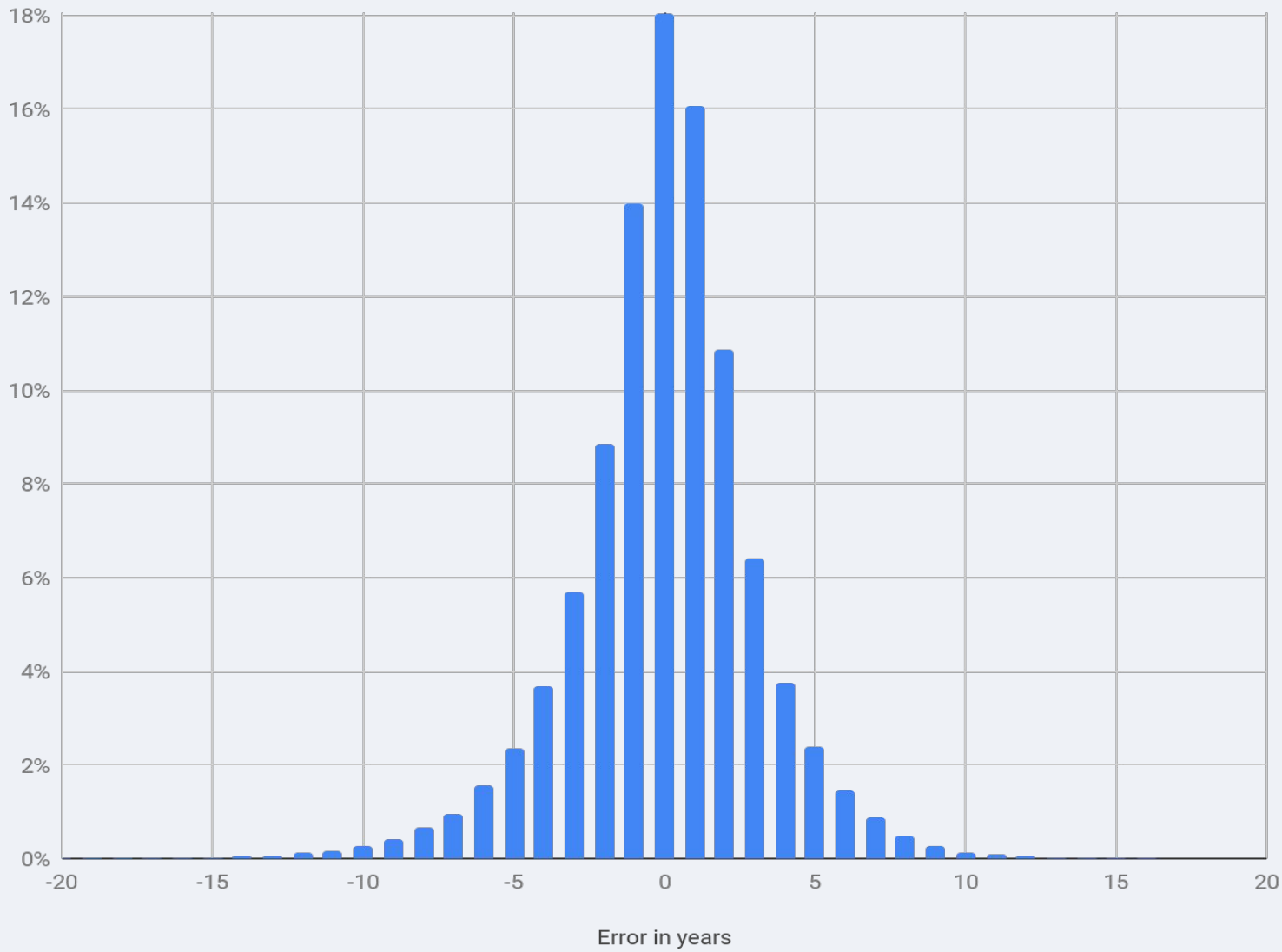
We provide information to users at onboarding about our use of biometrics with links to more details, including the full privacy notice, where the use of user data for R&D is extensively detailed. Users can opt out of their data being used for R&D activity at any time, via the settings on the app.

Data used for testing

Our testing data is also taken from Yoti users worldwide (not US users), in the same manner as the training data. We strive to ensure that it represents as broad a demographic as possible, considering age, gender and skin tone, giving us confidence that the results presented in this White Paper will be reproducible in a wide variety real world situations.

Accuracy across the entire dataset

In our most recent testing of the model, (performed May 2022), we used test data comprising over 125,000 facial images of verified age. The MAE across all years is now 2.96 years; for females it is 3.16, for males it is 2.77. This reflects a higher number of males in the training data across most years. The range of errors tends towards a normal distribution, with a standard deviation of 2.94. This is illustrated in the chart below. The standard deviation is a measure of the variance of the data around the mean.



Accuracy by age, gender and skin tone

We have explored how the accuracy (mean absolute error) of facial age estimation varies with age, gender and skin tone. Over 125,000 facial images of verified age in our test set were tagged with the subject's gender and skin tone. Gender was taken from the subject's uploaded identity document. For skin tone, our research team tagged the images using a scheme based on the widely used Fitzpatrick[16] dermatological scale. Fitzpatrick uses six bands, from Type I (lightest) to Type VI (darkest). For the present, we have presented our data in three bands (based on Fitzpatrick Types I & II, Types III & IV, and Types V & VI). The majority of the tagging was performed using a manual process, with some data tagged automatically. We have put quality procedures in place to help ensure our manual tagging is reliable and free from bias.

In presenting the data, we have grouped it into age bands, focusing particularly on bands which are of particular concern to regulators as regards the safeguarding of minors and access to age-restricted goods, services, websites and premises.

For each age band, we present the mean absolute error (MAE) in facial age estimation's age estimates in six classes: female (for three different skin tones), and male (for three different skin tones).

For each age band, the table also displays:

- the average MAE for females (of all skin tones), calculated as (MAE for Type I & II) + (MAE for Type III & IV) + (MAE for Type V & VI) ÷ 3
- the average MAE for males (of all skin tones), calculated as (MAE for Type I & II) + (MAE for Type III & IV) + (MAE for Type V & VI) ÷ 3
- the overall average MAE, calculated as (weighted average MAE for females + weighted average MAE for males) ÷ 2

The average attempts to deskew the test data set, so as to present equal contributions from the three skin tone groupings and both genders

Skin tone scale

16. Fitzpatrick, T, (1988) *The Validity and Practicality of Sun-Reactive Skin Types I Through VI*. Archives of Dermatology 1988; 124 (6): 869–871

Transition of MAE table from age bands to MAE for each year band

This table below attempts to provide a bridge with the previous white paper chart which was grouped into age bands.

Going forwards we will now publish each year MAE as per pages 3 and 4 in the Executive Summary. Below, we include the 9 age ranges that we have published in prior white papers and the additional age group for 60-70 year olds.

The row 'All' shows the average MAE for all test data, which equals 2.37.

The row Yearly Average shows the average MAE calculated by taking the average of each of the 65 MAEs for each year of age between 6 and 70, as shown on page 4, the Yearly Average was 2.96.

Age Band	Gender								All
	Female				Male				All
	Skin Tone (Fitzpatrick Scale)								
	Type I & II	Type III & IV	Type V & VI	All	Type I & II	Type III & IV	Type V & VI	All	
	MAE	MAE	MAE	Average MAE	MAE	MAE	MAE	Average MAE	Average MAE
6-9	1.39	1.38	1.46	1.41	1.40	1.52	1.53	1.48	1.44
10-12	1.31	1.32	1.76	1.46	1.07	1.02	1.05	1.05	1.26
13-15	1.45	1.89	2.27	1.87	1.23	1.52	1.88	1.55	1.71
16-17	0.99	1.02	1.08	1.03	0.87	1.14	1.02	1.01	1.02
18-24	2.15	1.96	1.80	1.97	1.73	1.89	1.93	1.85	1.91
25-29	2.84	3.42	4.91	3.72	2.24	2.49	2.96	2.56	3.14
30-39	3.05	3.59	4.82	3.82	2.49	2.89	2.91	2.76	3.29
40-49	2.87	3.04	3.91	3.27	2.69	2.99	3.29	2.99	3.13
50-60	2.98	2.97	6.09	4.02	2.96	3.34	4.45	3.59	3.80
60-70	2.79	3.75	4.05	3.53	3.16	3.50	4.56	3.74	3.63
All	**2.59**	**2.92**	**3.97**	**3.16**	**2.38**	**2.76**	**3.16**	**2.77**	**2.96**

Mean absolute error (MAE) of facial age estimation for different genders and skin tones, across age bands of interest. The weighted columns give equal weight to each of the three skin tone groups, and equal weight to both genders

We believe the differing mean absolute error shown for different groups (age, gender, skin tone) correlates strongly with how well-represented those groups are in the training data set. Additionally it seems reasonable to hypothesise that any absolute error will tend to be higher for older people than younger people, because older people will have been exposed to various unpredictable environmental factors for longer. It should also be remembered that 8% inaccuracy is 4 years for 50 year olds but only 1.6 years for 20 year olds.

Mean Absolute Error by year

Age	Gender								
	Female				Male				All
	Skin Tone (Fitzpatrick Scale)								
	Type I & II	Type III & IV	Type V & VI	All	Type I & II	Type III & IV	Type V & VI	All	
	MAE	MAE	MAE	Average MAE	MAE	MAE	MAE	Average MAE	Average MAE
6	1.13	1.73	1.55	1.47	1.47	2.06	1.68	1.74	1.60
7	1.40	1.41	1.32	1.38	1.05	1.33	1.85	1.41	1.39
8	1.58	1.14	1.58	1.44	1.51	1.75	1.21	1.49	1.46
9	1.43	1.32	1.40	1.38	1.62	1.10	1.25	1.33	1.35
10	1.16	1.11	1.63	1.30	0.99	1.00	1.00	1.00	1.15
11	0.79	1.16	2.12	1.36	0.87	0.84	1.06	0.92	1.14
12	1.71	1.76	1.50	1.65	1.26	1.30	1.09	1.22	1.44
13	2.20	2.85	2.65	2.57	1.83	1.81	2.16	1.93	2.25
14	1.62	2.14	2.69	2.15	1.46	1.84	2.23	1.84	2.00
15	1.20	1.56	1.97	1.58	1.06	1.34	1.71	1.37	1.47
16	0.93	1.14	1.34	1.13	0.82	1.17	1.20	1.07	1.10
17	1.08	0.91	0.88	0.96	0.92	1.12	0.91	0.98	0.97
18	1.44	1.18	0.88	1.17	1.15	1.42	1.26	1.28	1.22
19	1.91	1.68	1.49	1.70	1.51	1.64	1.67	1.61	1.65
20	2.33	2.19	2.05	2.19	2.03	1.91	2.02	1.99	2.09
21	2.72	2.67	2.49	2.63	2.17	2.12	2.01	2.10	2.36
22	2.89	2.74	3.22	2.95	2.38	2.21	2.36	2.32	2.63
23	2.80	2.90	3.76	3.16	2.42	2.14	2.41	2.32	2.74
24	2.91	2.83	3.74	3.16	2.62	2.28	2.87	2.59	2.87
25	2.89	3.14	4.43	3.49	2.13	2.37	2.67	2.39	2.94
26	2.60	3.34	4.35	3.43	2.15	2.34	2.94	2.48	2.95
27	2.93	3.26	5.31	3.83	2.24	2.47	2.79	2.50	3.17
28	2.90	3.75	4.93	3.86	2.23	2.54	3.35	2.71	3.28
29	2.95	3.63	5.89	4.16	2.49	2.77	3.10	2.79	3.47
30	3.16	3.49	4.78	3.81	2.31	2.79	3.05	2.71	3.26

Mean Absolute Error by year

Age	Gender								
	Female				Male				All
	Skin Tone (Fitzpatrick Scale)								
	Type I & II	Type III & IV	Type V & VI	All	Type I & II	Type III & IV	Type V & VI	All	
	MAE	MAE	MAE	Average MAE	MAE	MAE	MAE	Average MAE	Average MAE
31	2.81	4.38	5.05	4.08	2.61	2.42	2.84	2.62	3.35
32	3.19	4.21	5.20	4.20	2.64	2.83	2.95	2.81	3.50
33	3.44	3.88	4.40	3.91	2.80	3.23	3.31	3.11	3.51
34	3.54	4.12	5.28	4.31	2.86	3.14	2.99	2.99	3.65
35	3.22	3.48	5.64	4.11	2.75	2.97	2.95	2.89	3.50
36	3.14	3.57	2.38	3.03	2.99	2.69	3.19	2.96	2.99
37	3.14	3.80	3.70	3.55	2.63	2.96	3.41	3.00	3.27
38	3.48	2.42	4.73	3.54	2.67	3.06	3.34	3.02	3.28
39	3.52	4.11	3.75	3.80	2.82	2.56	3.59	2.99	3.39
40	3.04	3.21	4.90	3.72	2.61	2.66	2.91	2.73	3.22
41	2.79	3.23	3.21	3.07	2.87	2.72	3.34	2.98	3.03
42	2.95	3.98	4.49	3.80	2.74	3.19	3.38	3.10	3.45
43	3.12	3.22	4.09	3.48	2.67	2.60	3.15	2.81	3.14
44	2.82	3.15	6.70	4.22	2.66	2.91	3.52	3.03	3.63
45	2.64	3.95	3.99	3.53	2.41	3.16	3.23	2.93	3.23
46	2.94	4.24	2.22	3.13	2.90	3.79	3.17	3.29	3.21
47	3.35	2.74	4.74	3.61	2.97	3.42	2.75	3.05	3.33
48	3.58	4.47	2.86	3.64	2.56	2.97	2.90	2.81	3.23
49	2.98	2.91	4.52	3.47	2.77	2.76	2.98	2.84	3.15
50	3.06	3.70	3.50	3.42	2.80	3.00	3.05	2.95	3.18
51	2.85	4.37	2.96	3.39	2.85	2.90	5.56	3.77	3.58
52	2.35	2.67	3.34	2.79	2.90	2.71	3.44	3.02	2.90
53	2.35	4.07	3.09	3.17	2.78	2.31	2.61	2.57	2.87
54	2.26	2.07	9.68	4.67	2.65	3.44	2.96	3.02	3.84
55	2.22	2.20	6.45	3.62	2.13	3.18	3.83	3.05	3.34

Mean Absolute Error by year

Age	Gender								
	Female				Male				All
	Skin Tone (Fitzpatrick Scale)								
	Type I & II	Type III & IV	Type V & VI	All	Type I & II	Type III & IV	Type V & VI	All	
	MAE	MAE	MAE	Average MAE	MAE	MAE	MAE	Average MAE	Average MAE
56	2.77	3.49	5.15	3.80	2.61	3.89	4.17	3.56	3.68
57	3.53	3.08	7.06	4.56	2.89	3.20	3.99	3.36	3.96
58	2.98	2.71	5.35	3.68	3.12	3.50	4.06	3.56	3.62
59	2.97	3.21	4.79	3.65	2.91	3.20	4.32	3.47	3.56
60	2.52	3.36	4.11	3.33	2.66	3.11	4.47	3.41	3.37
61	2.59	4.60	5.06	4.08	2.78	3.39	4.45	3.54	3.81
62	2.17	4.08	3.15	3.14	2.46	2.75	3.75	2.99	3.06
63	2.07	3.23	4.22	3.17	2.32	2.88	6.20	3.80	3.49
64	2.10	3.46	3.26	2.94	2.43	2.79	3.14	2.79	2.86
65	2.23	2.45	2.48	2.39	2.02	3.83	3.84	3.23	2.81
66	2.41	3.16	4.45	3.34	2.39	4.17	4.58	3.71	3.53
67	2.67	2.66	4.85	3.39	3.07	4.00	5.60	4.22	3.81
68	3.01	3.68	2.48	3.06	3.32	5.67	5.22	4.74	3.90
69	4.10	5.07		4.59	4.35	4.21	3.13	3.90	4.24
70	4.73	6.05	5.11	5.30	4.48	6.32	8.04	6.28	5.79
Avg	**2.59**	**2.92**	**3.97**	**3.16**	**2.38**	**2.76**	**3.16**	**2.77**	**2.96**

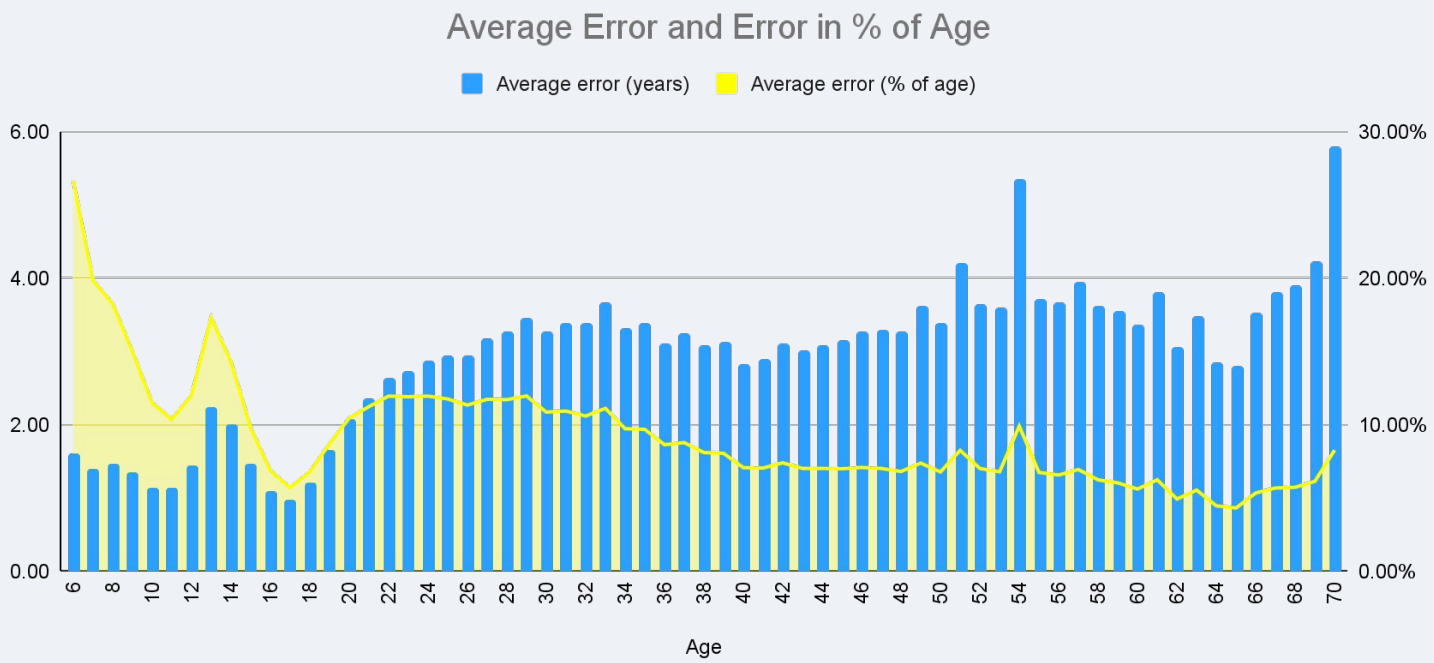
Absolute versus percentage errors

When we started publishing mean absolute error values for teenagers, a key age of regulatory interest, our MAE in April 2019 was 2.93 years and some stakeholders felt it unlikely that it would improve sufficiently to become an efficient age assurance technique. However our May 2022 MAE for teenagers is now 1.52 years. This is a 9.4% average error across the seven years.

For our first set of children aged 6-12, the average error is 1.36 years; after a much shorter period of research and smaller training data set. We believe it is very likely our MAE for 6-12 year olds will improve as our training set increases. The MAE of 1.36 years means Yoti can already offer a highly effective age estimation solution for businesses wishing to, or being required to, support with age appropriate design.

We have sufficiently high volumes of training data for males of all skin tones across the ages 6 to 29. Our average MAE for 6-29 year old males is 1.80. It is 1.68 for I&II (lightest) skin tone males, 1.95 for V&VI (darkest) skin tone males and 1.79 for III&IV skin tone males so there is less than 5% difference between the highest and lowest skin tone accuracy across these 24 years of age. However, we do not have as high volumes of training data for females compared to males, and in particular, V&VI (darkest) skin tone females. Our corresponding MAE for 6-29 year old females is 2.25. It is 1.98 for I&II (lightest) skin tone females, 2.15 for III&IV skin tone females and is highest at 2.63 or V&VI skin tone (darkest) females. This is a 22% difference between the highest and lowest skin tone accuracy across these 24 years of age. We will work hard to close this gap over coming months.

It is also worth noting that although the magnitude of error may appear larger for older age bands, when considered as a percentage of the subject's age, it often is more accurate in relative terms. For instance, an error of 2 years for a 15 year old is a 13% error, whereas an error of 2 years for a 50 year old is an error of 4%. This is illustrated in the chart below.



Improvement in accuracy as the training data set grows and changes

As mentioned above, we believe the differing mean absolute error shown for different groups (age, gender, skin tone) correlates strongly with how well-represented those groups are in the training data set. We have periodically retrained our age estimation model on an ever-expanding data set, as we continue to add further age-verified images taken from Yoti users at onboarding. The charts below illustrate the significant improvements in accuracy that we have observed over time. The size and composition of our test data has itself changed (diversified) over this period too, so the comparisons from one model's results to the next are not absolute, however the overall trend is clear and encouraging. Where appropriate we will endeavour to undertake further targeted fieldwork in this regard.

N.B. From September 2021 we have revised our approach to concentrate on achieving a reduction on bias, even where this may have a detrimental effect on accuracy.

As of this update (May 2022) we have removed some older images from both our training and testing data sets. This complies with our privacy policy on customer data retention, where if a user has been inactive for over 3 years we delete their data. This will have two implications to note:

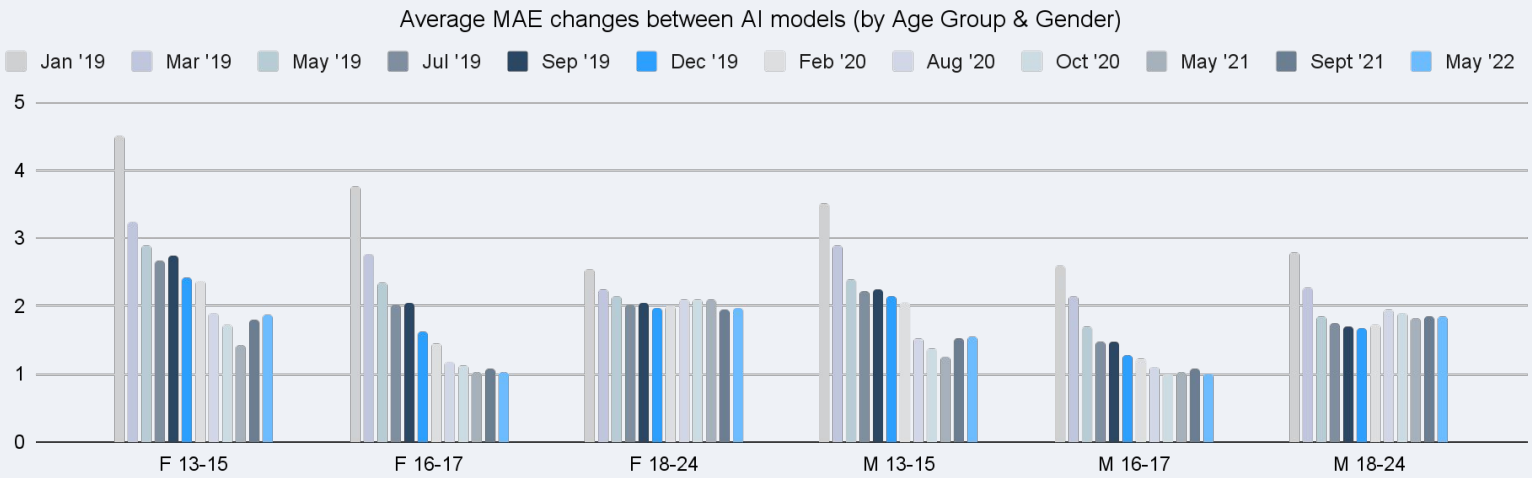
- **Training data** - where deleted data may have a skewed number of images in a certain subcategory, this may affect accuracy in that data range.
- **Testing data** - changes in this data set will mean results over time are not strictly 100% comparable as each model is not being tested against exactly the same set of test data.

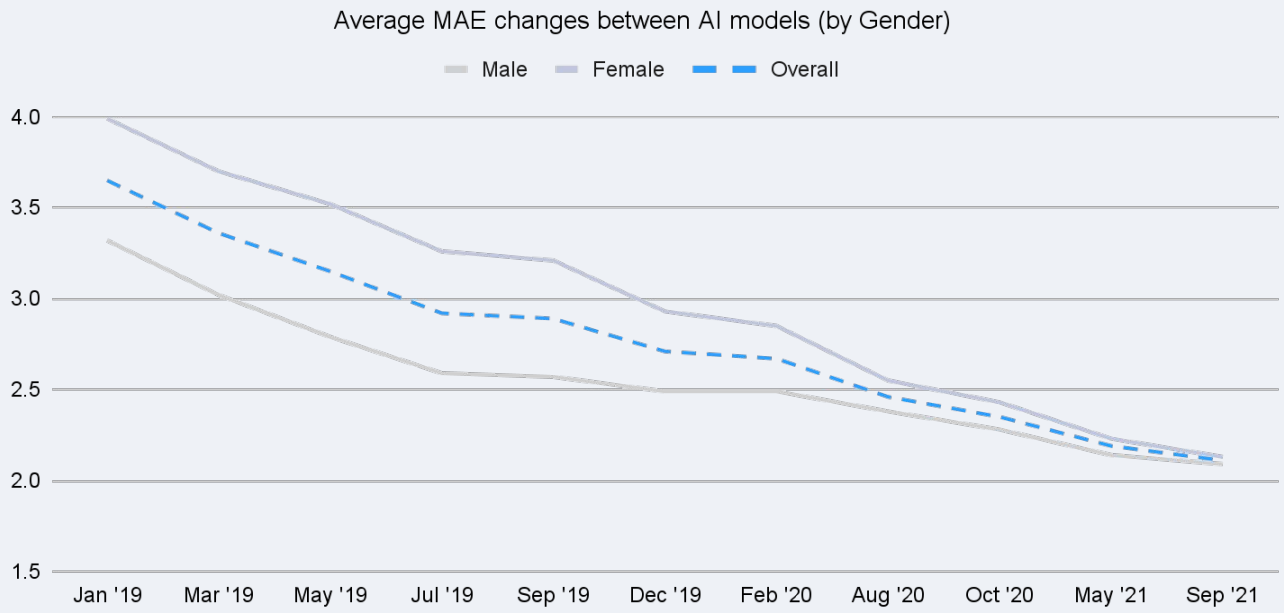
We do not believe the change in data will be statistically significant overall, both to the accuracy and testing results. We will also monitor churn of our data sets to ensure we replace data with the corresponding demographic that may have any significant effect on our accuracy or testing.

Summary of the performance of the new algorithm:

1. We have made improvements in MAE between the ages of 6-10 and 10-19. This was one of two key objectives.
2. The algorithm has reduced bias overall. The algorithm shows improvements in both (a) gender equality by improved performance for females and (b) skin tone equality by improved performance for skin tones 2 and 3 (though it has a very small deterioration for skin tone 1 male). This was the second of our key objectives.
3. There is a very small deterioration in accuracy for 20-31 year olds.
4. There is a large improvement in accuracy for the age ranges 31-46 and 58-70.

Improvement in accuracy as the training data set grows and changes





False positives

‘False positives’ are when we ask a question with a yes/no answer, and the answer comes back as ‘yes’ when it should have been ‘no’. So for example, when dealing with age-restricted goods or services, if we ask ‘Is this person old enough to buy alcohol?’ and facial age estimation tells us ‘Yes they are’, but actually they are not, then we have a ‘false positive’. In this kind of use case, we can regard false positives as a measure of facial age estimation being too lenient.

Let’s define some terms to help quantify things. When dealing with age-restricted goods and services, the **age of interest** is what we call the age stipulated in the relevant law or regulation. So for example, in many jurisdictions, the age of interest for buying alcohol is 18. In many use cases, we will ask ‘is this person above the age of interest?’ (e.g. ‘are they over 18?’), and configure facial age estimation to simply return ‘yes, they’re 18+’ or ‘no they’re not’.

However, as described earlier in this paper, facial age estimation has a margin of error, and we would expect some false positive replies when asking if a person was above the age of interest (particularly if their true age is close to it). For this reason, to try and avoid false positives, we recommend configuring a **threshold age** above the age of interest, to create a safety buffer. Instead of asking facial age estimation if the person is above the age of interest, we actually ask if they are above the threshold age instead. So for example, for an age of interest of 18, we might chose a threshold age of 23. We ask facial age estimation whether or not people are over 23. If the answer is ‘yes, they are’, we accept with confidence that they are over 18.

The challenge, therefore, is to pick an appropriate threshold for the given use case, which delivers an acceptably low false positive rate. The two tables below provide detailed statistics from our testing of facial age estimation, showing false positive rates for different ages of young people, for a succession of threshold ages. The first table considers a scenario where the age of interest is 18, the second table considers an age of interest of 21.

As is to be expected, the results show that it is much easier for facial age estimation to correctly estimate that young teenagers are below a threshold age than people who are only one year away from it. However when considering the acceptability of false positive rates for any given use case, the risk involved should be considered too: for example, the potential harm in a 14 year old purchasing alcohol is likely to be greater than for a 20 year old.

In the tables below we also present an average false positive rate for each threshold, weighting the value equally for each age’s contribution (regardless of the number of test subjects for that age).

False Positive rates for a selection of thresholds, for an age of interest of 18 (May 2022)

						Average False Positive Rate (weighted equally for each age)
		14	15	16	17	
Test Sample Size		*3,413*	*8,032*	*10,412*	*11,574*	
Thresholds (years)	20	0.44%	0.81%	1.61%	4.36%	1.81%
	21	0.18%	0.46%	0.96%	2.57%	1.04%
	22	0.12%	0.31%	0.58%	1.50%	0.63%
	23	0.06%	0.20%	0.41%	0.85%	0.38%
	24	0.03%	0.19%	0.21%	0.53%	0.24%
	25	0.03%	0.15%	0.17%	0.31%	0.17%
	26	0.03%	0.14%	0.12%	0.19%	0.12%
	27	0.03%	0.11%	0.08%	0.07%	0.07%
	28	0.00%	0.09%	0.06%	0.05%	0.05%
	29	0.00%	0.09%	0.05%	0.04%	0.04%
	30	0.00%	0.07%	0.03%	0.03%	0.03%

False positive rates for a selection of thresholds, for an age of interest of 21 (May 2022)

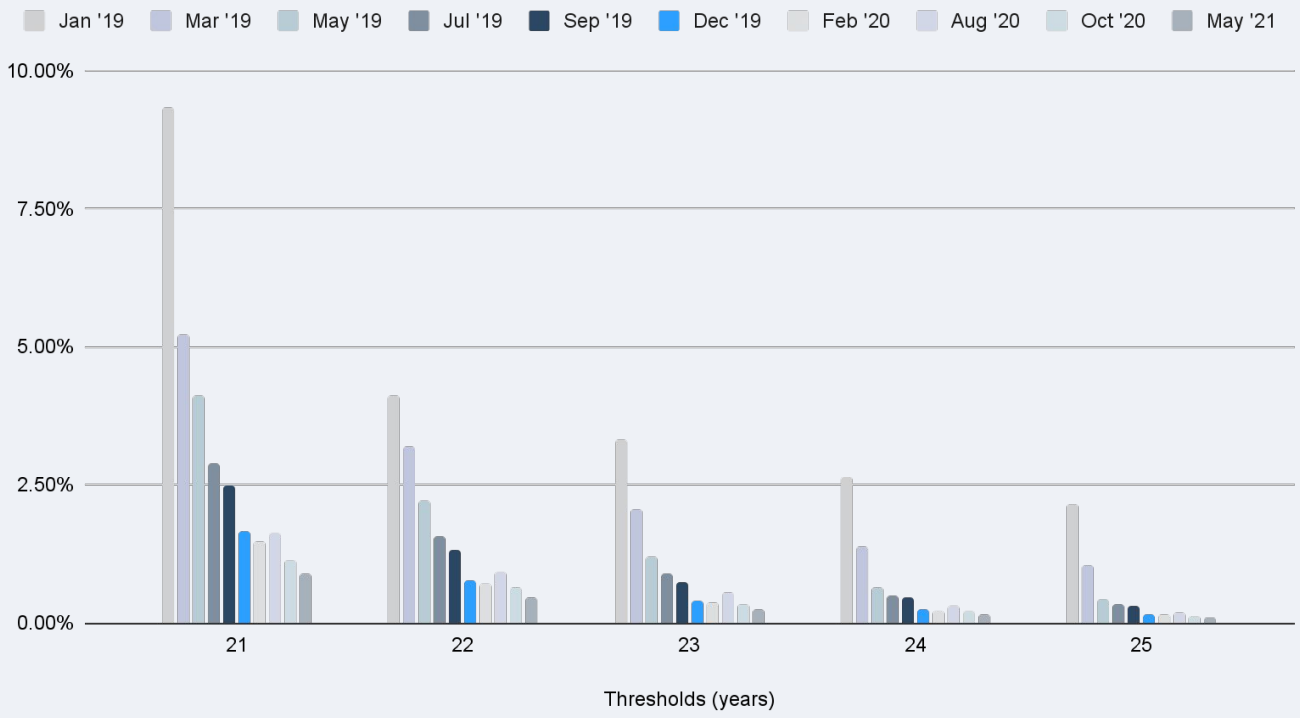
		Actual Age					Average False Positive Rate*
		16	**17**	**18**	**19**	**20**	
Test Sample Size		10,412	11,574	8,845	5,531	4,454	
Thresholds (years)	**24**	0.21%	0.53%	1.05%	2.48%	7.61%	**2.38%**
	25	0.17%	0.19%	0.61%	1.21%	4.18%	**1.27%**
	26	0.12%	0.19%	0.31%	0.71%	2.00%	**0.67%**
	27	0.08%	0.07%	0.15%	0.27%	0.94%	**0.30%**
	28	0.06%	0.05%	0.09%	0.14%	0.58%	**0.18%**
	29	0.05%	0.04%	0.08%	0.11%	0.27%	**0.11%**
	30	0.03%	0.03%	0.05%	0.07%	0.20%	**0.08%**
	31	0.03%	0.03%	0.03%	0.05%	0.09%	**0.05%**
	32	0.03%	0.03%	0.03%	0.00%	0.07%	**0.03%**
	33	0.02%	0.03%	0.02%	0.00%	0.07%	**0.03%**
	34	0.02%	0.03%	0.01%	0.00%	0.04%	**0.02%**
	35	0.02%	0.03%	0.01%	0.00%	0.00%	**0.01%**
	36	0.01%	0.03%	0.01%	0.00%	0.00%	**0.01%**
	37	0.01%	0.03%	0.01%	0.00%	0.00%	**0.01%**
	38	0.01%	0.03%	0.01%	0.00%	0.00%	**0.01%**
	39	0.01%	0.02%	0.01%	0.00%	0.00%	**0.01%**
	40	0.01%	0.02%	0.00%	0.00%	0.00%	**0.01%**

Improvements over time

Our false positive rates have shown steady improvement over the period between January 2019 and May 2021. We are confident this trend will continue as our training data set grows in volume and diversity. This is illustrated for a selection of thresholds in the table and chart below.

Average false positives for 14-17 year old (by threshold) - improvements over time

Thresholds (years)	Jan '19	Mar '19	May '19	Jul '19	Sep '19	Dec '19	Feb '20	Aug '20	Oct '20	May '21
21	9.34%	5.23%	4.12%	2.89%	2.50%	1.65%	1.46%	1.62%	1.13%	0.89%
22	4.11%	3.20%	2.21%	1.58%	1.32%	0.78%	0.72%	0.91%	0.63%	0.45%
23	3.31%	2.05%	1.19%	0.90%	0.75%	0.40%	0.38%	0.55%	0.34%	0.25%
24	2.65%	1.39%	0.66%	0.49%	0.47%	0.24%	0.20%	0.31%	0.22%	0.15%
25	2.14%	1.04%	0.44%	0.33%	0.31%	0.15%	0.14%	0.19%	0.11%	0.10%



We have not included our Sep 21 data because all of the other historic data was calculated on an average MAE for all testing data.

Trade-off between false negatives and false positives

False negatives are an annoyance to those trying to access an age-restricted service or purchase age-restricted goods. They can cause friction and conflict between customers and retail staff, with assaults and abuse being a growing problem[17, 18, 19]. It also means that customers have to revert to carrying physical ID documents with them. These documents (such as passports and driving licences) can be expensive to apply for and obtain, and a significant proportion of young people do not possess them. Large numbers of physical ID documents are also lost every year, increasing the risk of identity fraud as well as incurring a replacement cost.

Earlier in this paper, when discussing choice of a threshold age and safety buffer for use with facial age estimation, we have generally framed this in terms of trying to minimise false positives (effectively, where facial age estimation is too lenient), as these carry a greater risk of harm to young people. However it is also sensible to consider false negative rates too (facial age estimation being too cautious). Choosing higher thresholds will tend to decrease false positives at the expense of causing more false negatives. It is important for regulators (or businesses in unregulated sectors) to consider their risk tolerance for any given deployment of facial age estimation, and choose a threshold which is likely to deliver an acceptable balance between false positive and false negative rates.

The table overleaf illustrates this for comparison against a typical ‘Challenge 25’ retail scenario, where the ‘age of interest’ (the legal age for buying age-restricted goods) is 18.

For each threshold, the ‘false positives’ column shows the small percentage of under-age teenagers that facial age estimation would be likely let through. The next column shows the percentage of young people from 18–25 that facial age estimation would be likely to reject, meaning they would have to present physical ID to prove their age instead. Note that this not only includes ‘false negatives’ (young people who were actually older than the threshold, but facial age estimation incorrectly estimated they were under it), but also ‘genuine negatives’ (where facial age estimation has correctly estimated that the young person is over the legal age, but they are still below the chosen threshold age).

17. *An analysis of abuse and violence towards retail staff when challenging customers for ID* (Allen & Rudkin, 2017) https://nfrnonline.com/wp-content/uploads/Abuse-and-Violence-Report-2.pdf.
18. *‘It’s not part of the job’: Violence and verbal abuse towards shop workers–A review of evidence and policy* (Taylor, 2019) https://assets.ctfassets.net/5ywmq66472jr/22QfMejeWYbimJ9ykX9W9h/0e99f15c0ed24c16ab74d38b42d5129a/It_s_not_part_of_the_job_report.pdf.
19. *Freedom from Fear: Survey of violence and abuse against shop staff in 2018* (Union of Shop, Distributive & Allied Workers, 2018) https://www.usdaw.org.uk/2018FFFReport.

We feel these rates compare favourably with the current 'Challenge 25' scheme, where shopkeepers have to estimate young people's ages, and require all those they think are under 25 to produce physical ID. Depending on risk tolerance, we believe facial age estimation offers clear potential to maintain robust protection for under-18s whilst substantially reducing the numbers of young people over 18 who have to bring physical ID with them when they go shopping.

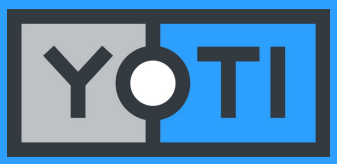
Comparison of false positives for underage teenagers versus rejection rates for young people over the legal age of interest (18), for a selection of safety buffer thresholds

Choice of Threshold (years)	**Average* False Positive Rate** (for ages 14-17)	**Combined average* rejection rate (false negatives & genuine negatives)** (for ages 18-25)
21	1.04%	**50.72%** (genuine negatives for 18-20 year olds ÷ false negatives for 21-25 year olds)
22	0.63%	**58.14%** (genuine negatives for 18-21 year olds ÷ false negatives for 22-25 year olds)
23	0.38%	**65.56%** (genuine negatives for 18-22 year olds ÷ false negatives for 23-25 year olds)
24	0.24%	**73.20%** (genuine negatives for 18-23 year olds ÷ false negatives for 24-25 year olds)
25	0.17%	**80.92%** (genuine negatives for 18-24 year olds ÷ false negatives for 25 year olds)

**Note that the numbers of subjects of each age in the test data set was not equal. Therefore to avoid skewing the results, the false positive and negatives figures in this table are averages, weighted equally for the contribution of each age.*

Memberships, associations and accreditations

Reviewed by



To find out more visit **yoti.com**