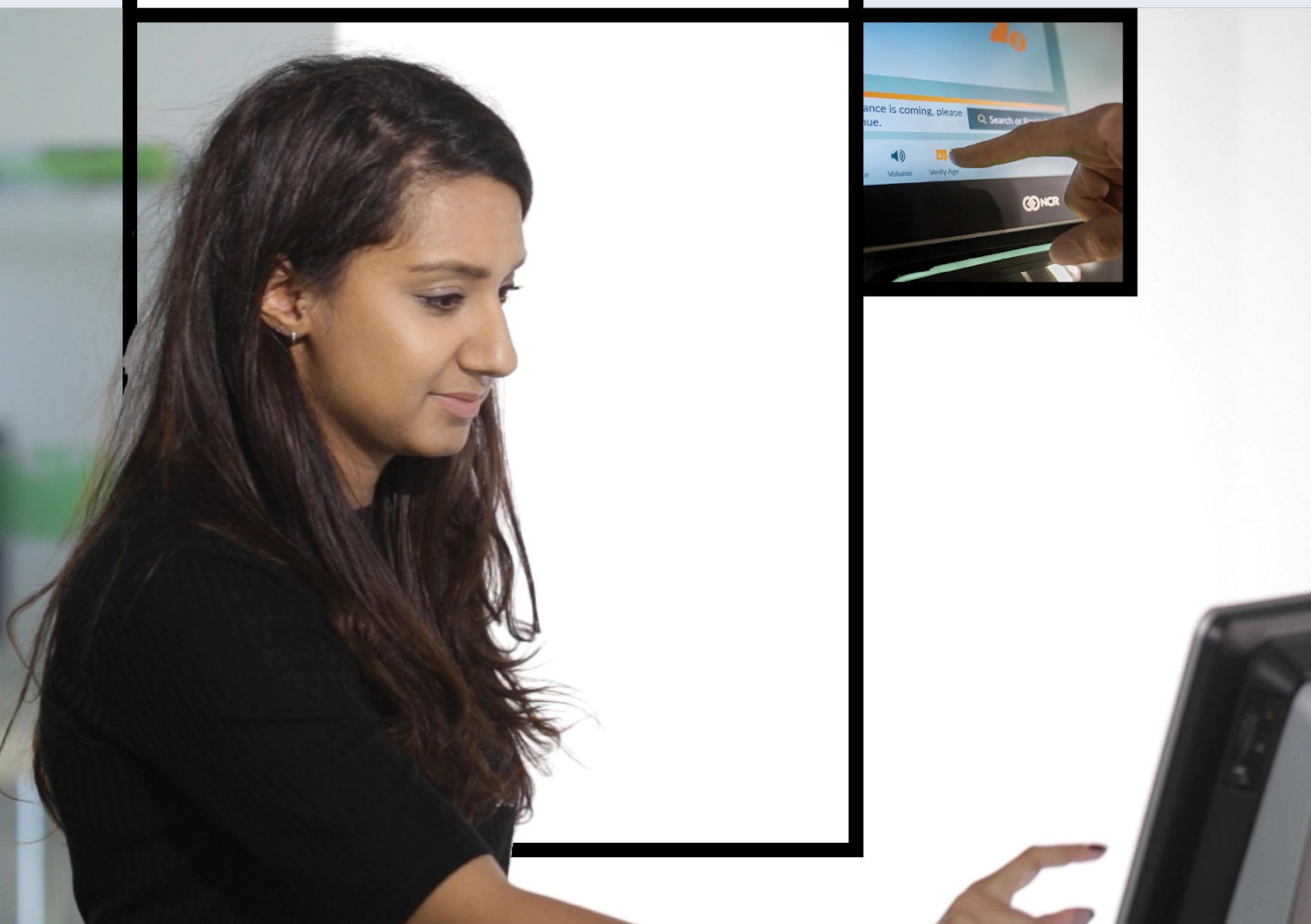
White paper

Yoti Age Scan - Public version

October 2020



Doing things differently

Age Scan ‘Powered by Yoti’ is just one innovative use of our digital identity technology.

We built it to give everyone a secure and private way of proving how old they are in different everyday cenarios: from age checking on social platforms and online stores, to supermarket self-checkouts, bars and clubs. In this white paper we’ll explain everything you need to know about the new way to prove your age.

What is Age Scan and what can it do?

Age Scan is a secure 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.

Age Scan 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, Age Scan 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, ideal for controlling access to age-restricted gaming, gambling and also adult content (pornography).

We believe Age Scan 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 Age Scan within its app to help identify user profiles where there is suspicion or doubt about the user's age, and flag these cases to its moderation team.

A further potential use is at the entrances to age-restricted premises such as bars, nightclubs and casinos. In this kind of application, Age Scan offers clear advantages – it does not get fatigued on a long shift[1], and it cannot show favour to personal friends, or bias against individual customers.

Age Scan is an emerging technology, and its age estimates are subject to a margin of error. To allow for this, the system is configurable to set whatever threshold a business or regulator requires, for example requiring those over 18 to be estimated as at least 21 or 23 – a buffer of three or five years. Where someone is over 18 but appears to be under chosen threshold, they can use either the Yoti app, where their account is anchored with a verified ID document, or undergo a manual photo ID check with a member of staff.

1. 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*

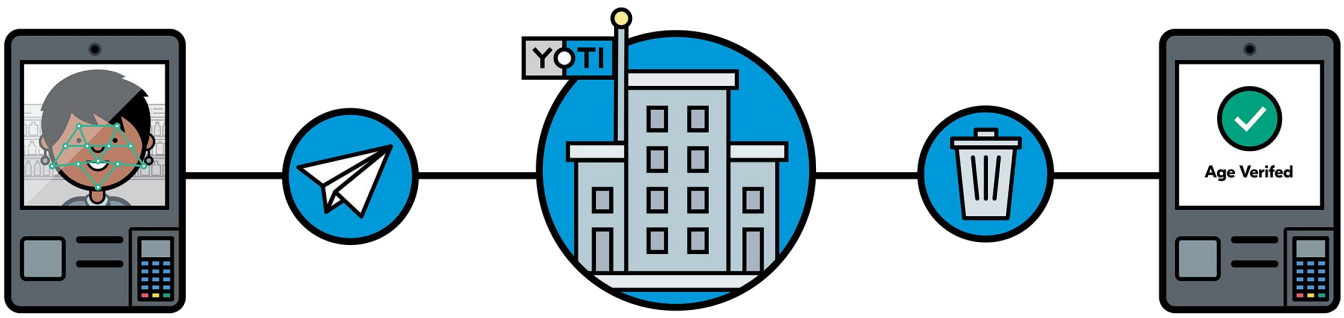
Data privacy and network security

Age Scan 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.

How does it actually work?

Age Scan 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 below, but first some context on the problem we are using it to solve.



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.

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

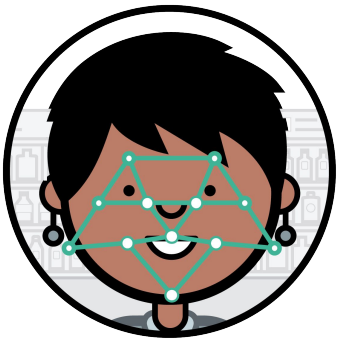
Notwithstanding the difficulty in devising an accurate forensic test for age, people still possess a reasonably good ability to guess someone's age simply by looking at them. We can all do it, usually coming within a few years of the right answer. How do we manage it? In terms of facial features, what 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 dozens 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.

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.

More on how it works

The first challenge for Age Scan 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.



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.

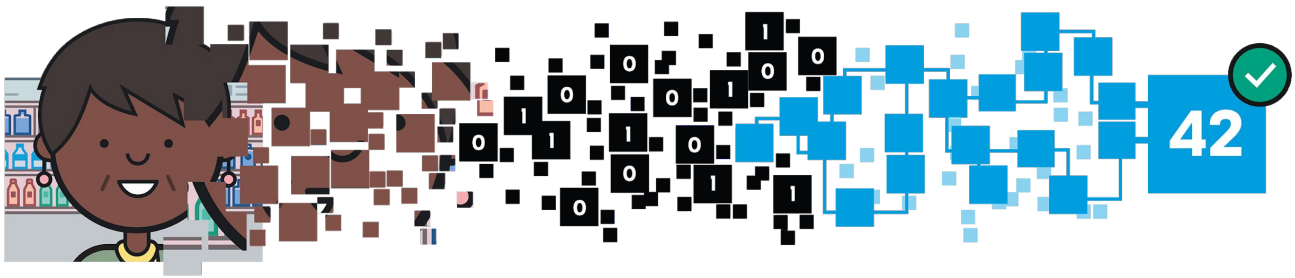
We now come to the interesting bit. The facial image is made up of pixels. To the computer, each pixel is just a set of numbers. These numbers are fed into the artificial neural network. This is a network of mathematical processing nodes, arranged in layers, that is roughly analogous to the connections 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.

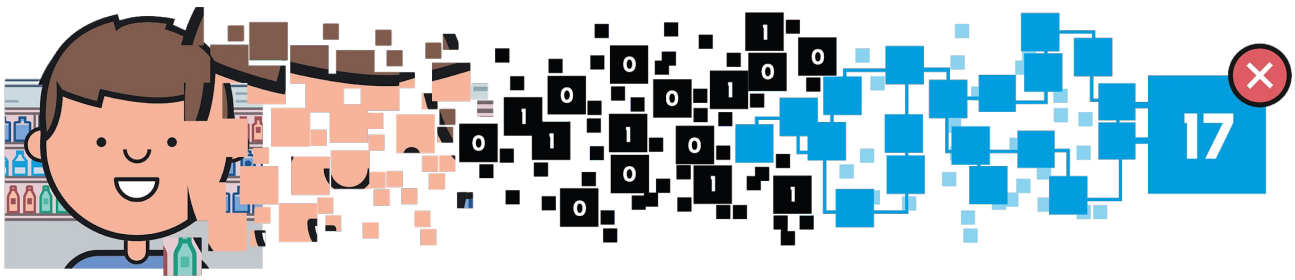
It’s an obvious question to ask ‘how is the neural network processing the data? What is it looking for – wrinkles? grey hairs?’ and so on. However, this is a rather human way of thinking about it, and it’s not really a very useful question to ask: to the computer, it is just being fed numbers. It doesn’t ‘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 Age Scan was being developed, was what the right answers were. In the training phase, we fed it hundreds of thousands 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 Age Scan algorithm, we use many thousands of images from Yoti users who have opted in to this use of their data. The process is explained to them at onboarding, and (as 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 Age Scan, these research images are tagged with only two attributes taken from a verified ID document that they have uploaded: their gender and their 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 Age Scan an advantage over competing solutions.

Practical use

Age Scan 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). Dim lighting is not helpful; bright ambient light works best. The effect of beards and facial disfigurement are further areas of research. In response to the ongoing COVID-19 pandemic, we have been researching how Age Scan copes when a person is wearing a mask. Preliminary results suggest that whilst accuracy is reduced somewhat, acceptable performance can usually still be achieved so long as a larger safety buffer is used. We will provide more details on our research in this area in a future edition of this paper.





How accurate is Age Scan?

We believe that when presented with a clear facial image, Age Scan's ability to estimate age compares favourably with human abilities.

Research in this area[3] suggests that the root mean square error in human guesses across an age range of 7 to 70 approaches 8 years. 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 Age Scan. 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 mean absolute error across the entire data set, de-skewed to give equal weighting to male and female subjects, is 2.35 years. Further detail on our algorithm's accuracy, broken down by gender, skin tone and age range, is presented in this paper's appendix. We believe this accuracy will improve still further in years to come, as Age Scan is trained on an ever greater set of data from Yoti users. We intend to continue comparing Age Scan's accuracy against that of ordinary human estimators, and against people who believe they have a special aptitude at estimating age, to demonstrate that Age Scan is usually a more accurate approach (and cheaper and faster).

3. 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 Age Scan. To manage this potential for errors, we recommend using Age Scan as part of a strategy such as the British Beer & Pub Association's 'Challenge 21'[4] , 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.

Age Scan can be configured to work with legal age thresholds in a similar way. Furthermore, and unlike human staff, Age Scan's capacity for error is well quantified statistically. This makes it easier to choose a suitable buffer that is comfortably outside Age Scan'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 Age Scan might choose to set an initial threshold of 30. If Age Scan estimates that the customer is at least 30 years old, then no further age checking is required.

If Age Scan 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 0.07% of under 21 year olds would pass unchallenged by Age Scan, which compares 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 a considerable amount of 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 14–25 year old 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 Age Scan for this demographic, we now suggest a buffer of 3–5 years is most appropriate for the 14–25 age band. And 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 0.03% percent error rate. With a threshold set to 25 years, Age Scan's current error rate is 0.14%. For a threshold of 23 years, the error rate is 0.44%.

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

4. See https://beerandpub.com/campaigns/challenge-21/

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.

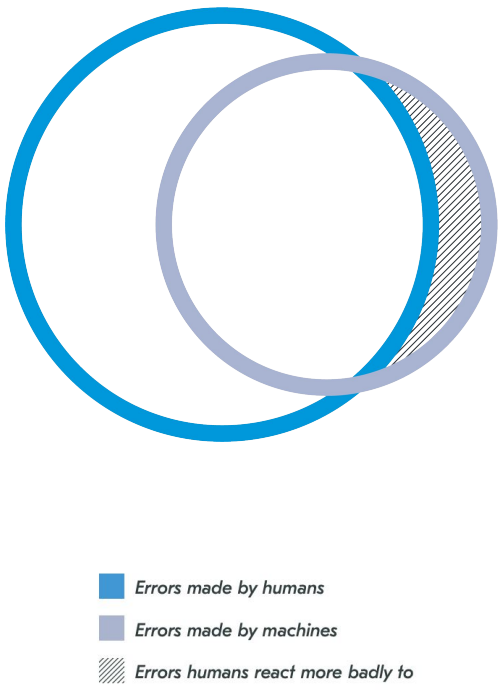
More statistical detail on Age Scan's 'false positive' rates for a selection of different thresholds and buffers is presented in the appendix of this paper. It is also worth considering 'false negatives' too (where Age Scan incorrectly estimates someone as being younger than the threshold age), as these 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 Age Scan, it is worth considering a general point about machine learning and the public's attitude to AI technologies of this kind: 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 Age Scan generally compares 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, and objectively, the machine learning system is no worse than the human judgement it is replacing (sometimes it may even be better overall!). Nevertheless, the general public may often unduly focus their attention on the machine failings, until they become comfortable with the new technology.



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 organisations seeking to assess these services. 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 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 PAS1296 code of practice[5]. 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 first company in the UK to achieve this certification[6].

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.

We have hosted two 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.

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 signed the Safe Face Pledge[7], which encourages companies using artificial intelligence to ensure that facial recognition technology is not misused.

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 Age Scan algorithm.

5. 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.
6. https://www.bbfc.co.uk/about-bbfc/media-centre/bbfc-statement-age-verification-under-digital-economy-act
7. https://www.safefacepledge.org

Appendix

This appendix provides further detail on the current accuracy of Age Scan's estimates. 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.

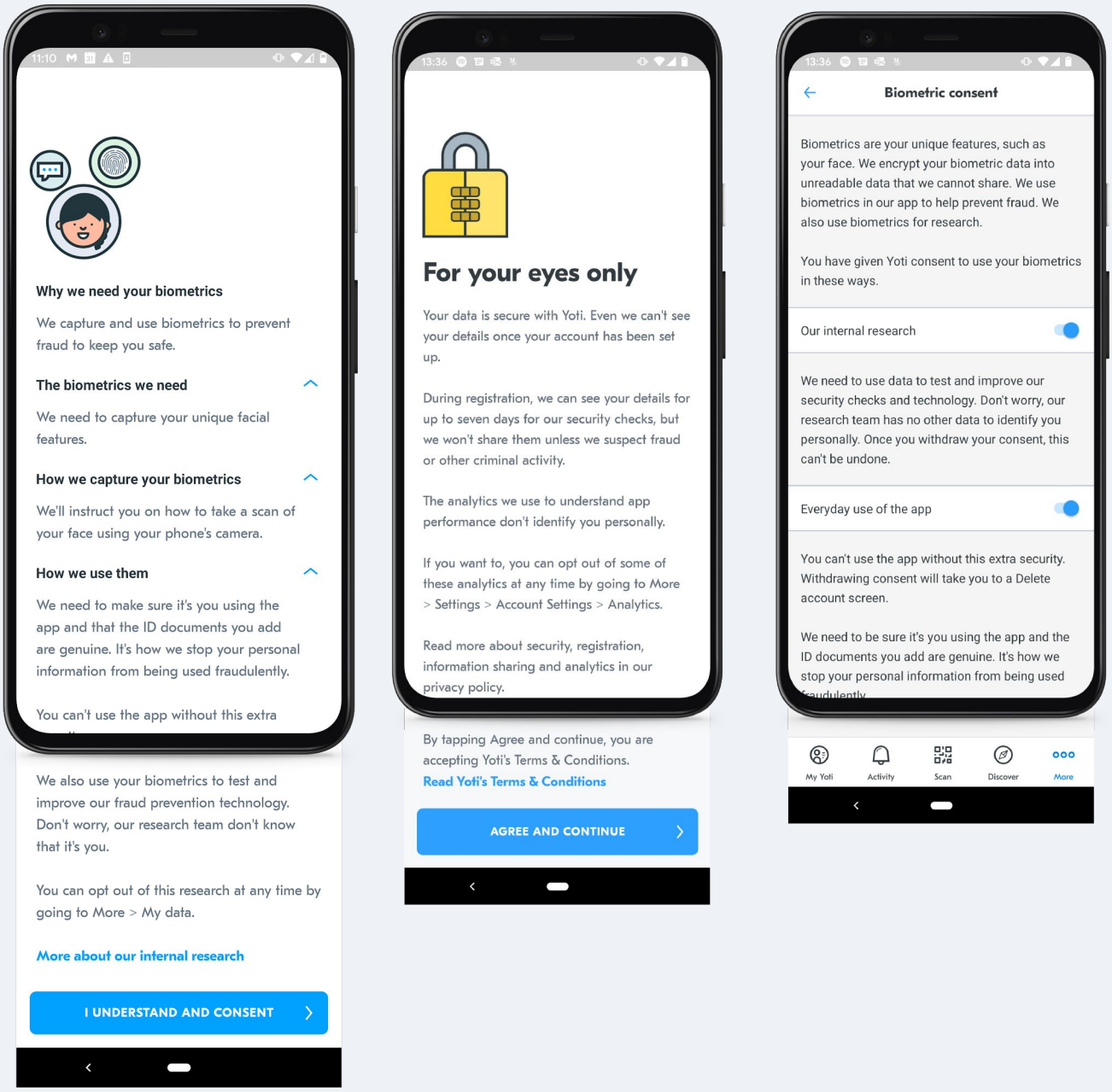
The current production model of Age Scan (October 2020) was built using a training data set taken mainly from Yoti users. We provide information to users at onboarding about our use of biometrics with links to more details, including the Privacy Notice[8] where the use of user data by our R&D team for 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)[9]. 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.

A small additional amount of data was obtained from a public domain source[10]. To enhance our coverage of particular demographics, further age-verified images were gathered by Yoti with consent in Nairobi, Kenya.

8. https://www.yoti.com/privacypolicy
9. 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
10. Images were taken from the Computer Vision Center and University of Barcelona's APPA-REAL Database, http://chalearnlap.cvc.uab.es/dataset/26/description/. These form only 0.03% of our production model training data.

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



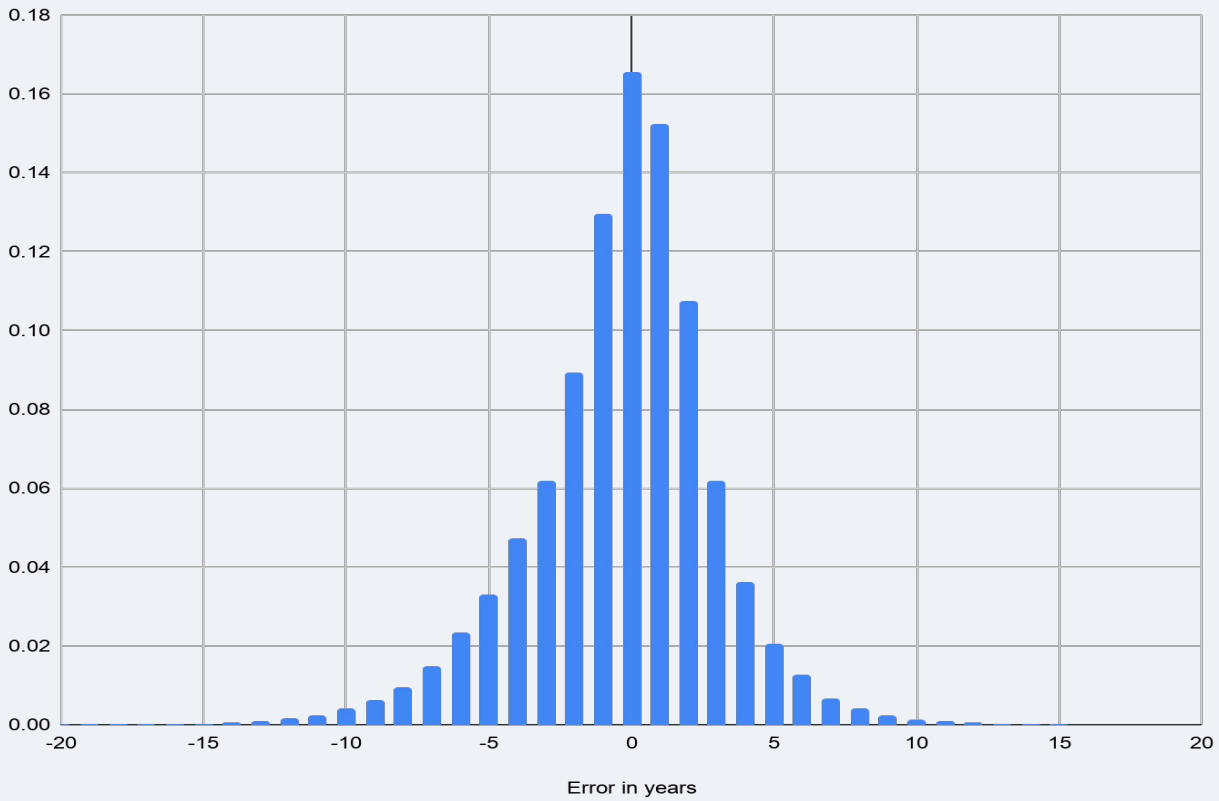
We provide information to users at onboarding about our use of biometrics with links to more details, including the full privacy notice, where the use of user data for R&D is extensively detailed. Users can opt out of their data being used for R&D activity at any time, via the settings on the app.

Data used for testing

Our testing data is also taken from Yoti users worldwide, 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. Nevertheless, we recommend that Age Scan is tested for parity of outcomes with different demographics (including ethnicities) in any particular use case where it is to be deployed.

Accuracy across the entire dataset

Accuracy across the entire dataset In our most recent testing of the model, (performed October 2020), we used test data comprising over 100 thousand facial images of verified age. The mean absolute error (MAE) in age estimates (across entire data set) was 2.23 years. This is illustrated in the scatter plot on page 8. However, as we discuss below, this MAE figure reflects that the test data currently has a greater proportion of male subjects. The gender-weighted average MAE, calculated as (MAE for males + MAE for females) ÷ 2, is 2.35 years. The range of errors tends towards a normal distribution, with a standard deviation of 3.91. This is illustrated in the chart below.



Accuracy by age, gender and skin tone

We have explored how the accuracy (mean absolute error) of Age Scan varies with age, gender and skin tone. Over 100 thousand 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 manually tagged the images using a scheme based on the widely used Fitzpatrick[11] 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). 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 Age Scan's age estimates in six classes: female (for three different skin tones), and male (for three different skin tones). There were at least 230 test subjects in every class.

For each age band, the table also displays:

- the weighted 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 weighted 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 weighted average MAE, calculated as (weighted average MAE for females + weighted average MAE for males) ÷ 2

These weightings attempt to deskew the test data set, so as to present equal contributions from the three skin tone groupings and both genders

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

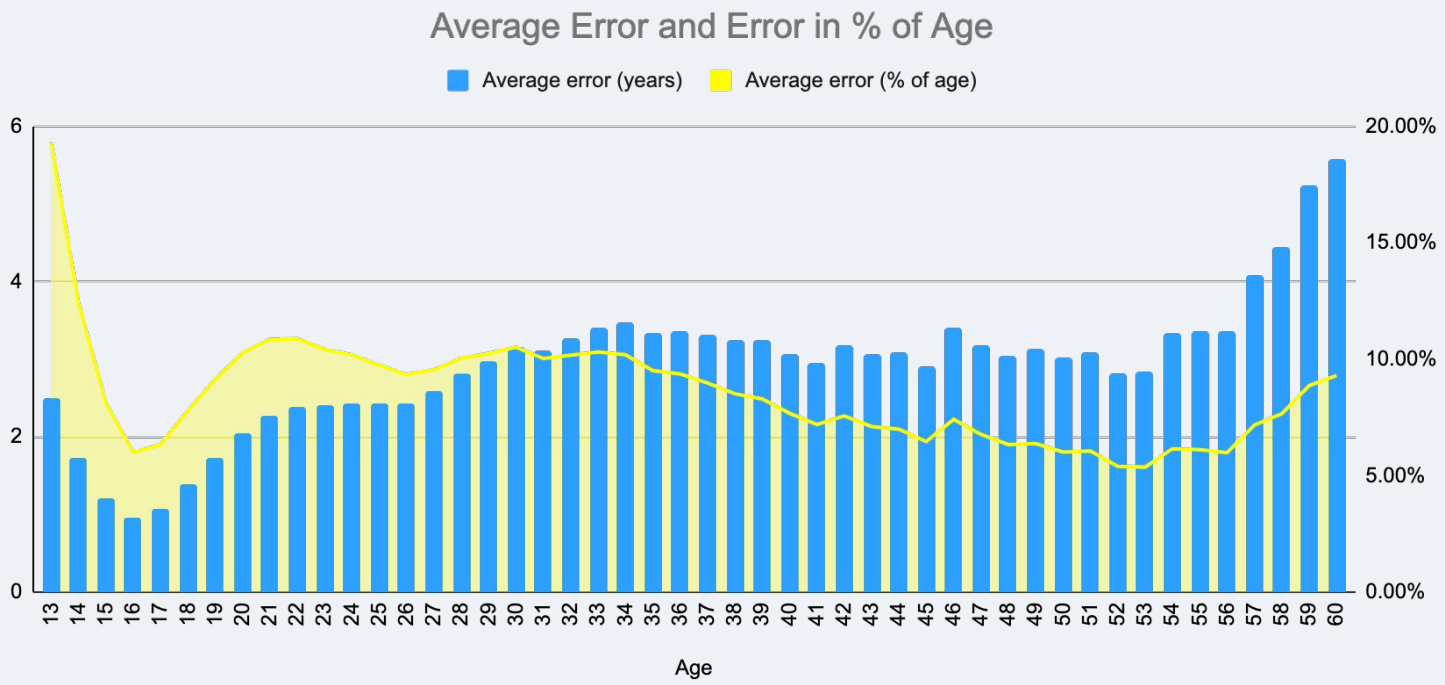
Age Band	Gender								
	Female				Male				All
	Skin Tone (based on the Fitzpatrick Scale)								
	Type I & II	Type III & IV	Type V & VI	All	Type I & II	Type III & IV	Type V & VI	All	
	MAE	MAE	MAE	Average MAE	MAE	MAE	MAE	Average MAE	Average MAE
13-15	1.39	1.74	2.03	1.72	1.15	1.46	1.57	1.39	1.56
16-17	1.09	1.13	1.21	1.14	0.88	1.11	1.04	1.01	1.08
18-24	2.14	2.09	2.04	2.09	1.75	1.92	2.00	1.89	1.99
25-29	2.84	3.34	4.22	3.47	2.37	2.53	2.53	2.48	2.97
30-39	3.48	4.36	4.60	4.15	2.87	3.18	3.34	3.13	3.64
40-49	3.15	3.91	4.49	3.85	2.81	3.19	3.24	3.08	3.46
50-60	3.39	4.68	5.64	4.57	3.18	3.86	3.91	3.65	4.11
All	2.16	2.71	2.41	2.43	1.95	2.42	2.46	2.28	2.35

Mean absolute error (MAE) of Age Scan 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 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.

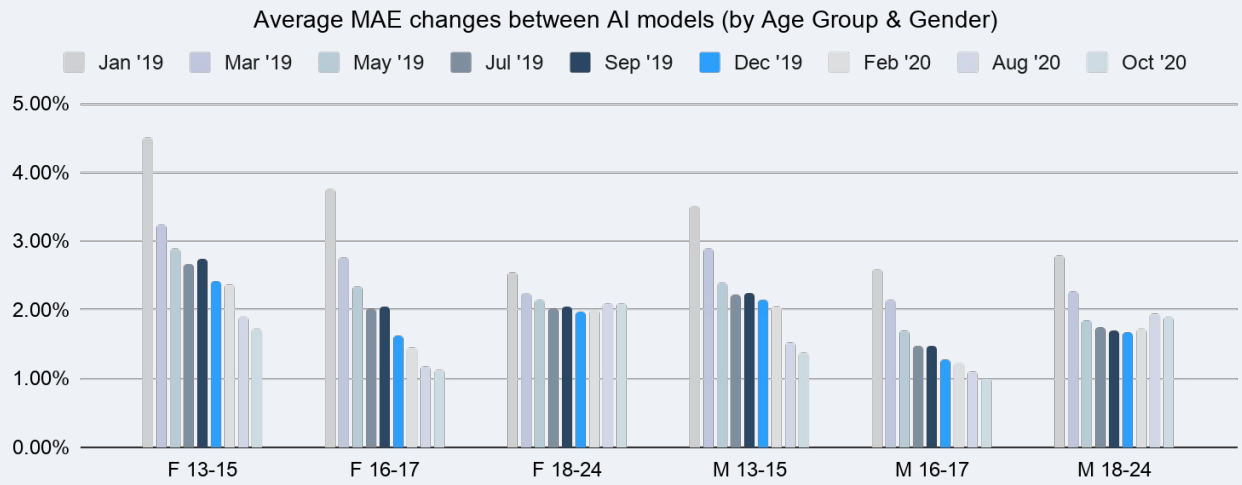
Absolute versus percentage errors

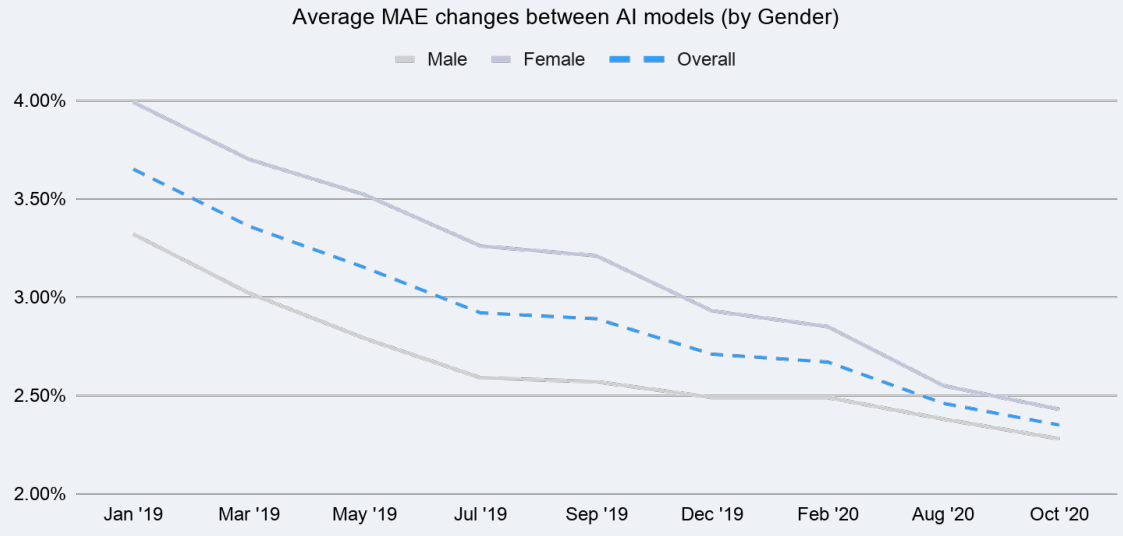
Additionally, it is worth noting that although the magnitude of error may appear larger for older age bands, that when considered as a percentage of the subject's age, it may be more accurate in relative terms. For instance, an error of 2 years for a 15 year old is a 11% error, whereas an error of 2 years for a 50 year old is an error of 7%. This is illustrated in the chart below.



Improvement in accuracy as the training data set grows

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 in the past two years. 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. We are optimistic that these trends will continue as we further broaden the diversity of our training data. Where appropriate we will endeavour to undertake further targeted fieldwork in this regard.





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 Age Scan 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 Age Scan 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 Age Scan to simply return ‘yes, they’re 18+’ or ‘no they’re not’.
However, as described earlier in this paper, Age Scan 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 Age Scan 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 Age Scan 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 Age Scan, 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 Age Scan 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, weighted the value equally for each age’s contribution (regardless of the number of test subjects for that age).

False Positive for a selection of thresholds, for an age of interest of 18 (October 2020)

		Actual Age				Average False Positive Rate (weighted equally for each age)
		14	15	16	17	
Test Sample Size		*2,626*	*6,929*	*10,357*	*10,367*	
Thresholds (years)	20	0.80%	1.00%	2.20%	4.95%	2.24%
	21	0.27%	0.49%	1.08%	2.70%	1.13%
	22	0.15%	0.22%	0.66%	1.50%	0.63%
	23	0.04%	0.14%	0.31%	0.87%	0.34%
	24	0.04%	0.09%	0.18%	0.56%	0.22%
	25	0.00%	0.07%	0.10%	0.27%	0.11%

False positive rates for a selection of thresholds, for an age of interest of 21 (October 2020)

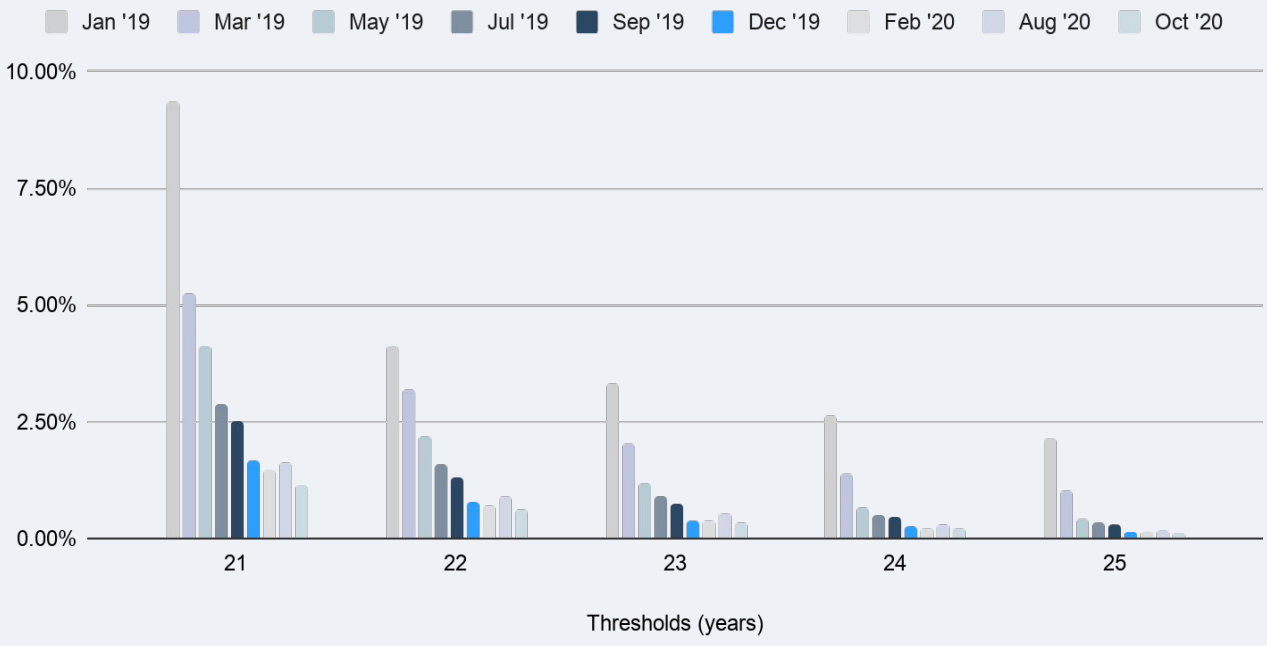
		Actual Age					Average False Positive Rate*
		16	**17**	**18**	**19**	**20**	
Test Sample Size		10,375	10,367	7,510	5,164	4,124	
Thresholds (years)	**24**	0.31%	0.78%	1.62%	3.37%	8.27%	**2.00%**
	25	0.17%	0.43%	1.01%	1.72%	4.53%	**1.11%**
	26	0.10%	0.22%	0.59%	0.77%	2.01%	**0.54%**
	27	0.05%	0.16%	0.33%	0.37%	1.02%	**0.29%**
	28	0.05%	0.12%	0.25%	0.23%	0.65%	**0.20%**
	29	0.05%	0.11%	0.23%	0.19%	0.34%	**0.15%**
	30	0.05%	0.10%	0.19%	0.14%	0.24%	**0.12%**
	31	0.04%	0.06%	0.15%	0.10%	0.17%	**0.09%**
	32	0.03%	0.05%	0.08%	0.08%	0.15%	**0.06%**
	33	0.03%	0.04%	0.08%	0.06%	0.12%	**0.06%**
	34	0.03%	0.02%	0.07%	0.04%	0.05%	**0.04%**
	35	0.03%	0.01%	0.07%	0.04%	0.05%	**0.03%**
	36	0.03%	0.01%	0.05%	0.04%	0.05%	**0.03%**
	37	0.03%	0.01%	0.04%	0.04%	0.05%	**0.03%**
	38	0.03%	0.01%	0.03%	0.02%	0.05%	**0.02%**
	39	0.02%	0.01%	0.03%	0.02%	0.05%	**0.02%**
	40	0.02%	0.01%	0.03%	0.00%	0.00%	**0.01%**

Improvements over time

Our false positive rates have shown steady improvement over the past year, and 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
21	9.34%	5.23%	4.12%	2.89%	2.50%	1.65%	1.46%	1.62%	1.13%
22	4.11%	3.20%	2.21%	1.58%	1.32%	0.78%	0.72%	0.91%	0.63%
23	3.31%	2.05%	1.19%	0.90%	0.75%	0.40%	0.38%	0.55%	0.34%
24	2.65%	1.39%	0.66%	0.49%	0.47%	0.24%	0.20%	0.31%	0.22%
25	2.14%	1.04%	0.44%	0.33%	0.31%	0.15%	0.14%	0.19%	0.11%



Trade-off between false negatives and false positives

'False negatives' are when we ask a question with a yes/no answer, and the answer comes back as 'no' when it should have been 'yes'. So for example, when dealing with age-restricted goods or services, if we ask 'Is this person old enough to buy alcohol?' and Age Scan tells us 'No, they're not', but actually they are, then we have a 'false negative'. In this kind of use case, we can regard false negatives as a measure of Age Scan being too cautious.
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[12, 13, 14]. 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 Age Scan, we have generally framed this in terms of trying to minimise false positives (effectively, where Age Scan 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 (Age Scan 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 Age Scan, 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 Age Scan would be likely let through. The next column shows the percentage of young people from 18–25 that Age Scan 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 Age Scan incorrectly estimated they were under it), but also 'genuine negatives' (where Age Scan has correctly estimated that the young person is over the legal age, but they are still below the chosen threshold age).

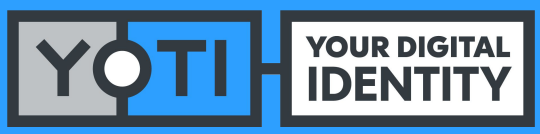
12. *An analysis of abuse and violence towards retail staff when challenging customers for ID* (Allen & Rudkin, 2017) https://www.underagesales.co.uk/user/Abuse%20and%20Violence%20Report%202.pdf.
13. *'It's not part of the job': Violence and verbal abuse towards shop workers–A review of evidence and policy* (Taylor, 2019) https://assets.ctfassets.net/5ywmq66472jr/22QfMejeWYbimJ9ykX9W9h/0e99f15c0ed24c16ab74d38b42d5129a/It_s_not_part_of_the_job_report.pdf.
14. *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 Age Scan 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	1.13%	**36.33%** (genuine negatives for 18-20 year olds ÷ false negatives for 21-25 year olds)
22	0.63%	**47.29%** (genuine negatives for 18-21 year olds ÷ false negatives for 22-25 year olds)
23	0.55%	**58.70%** (genuine negatives for 18-22 year olds ÷ false negatives for 23-25 year olds)
24	0.22%	**69.14%** (genuine negatives for 18-23 year olds ÷ false negatives for 24-25 year olds)
25	0.11%	**80.71%** (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.*



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