Age Check
Certification Scheme

Measurement of Age Assurance Technologies

Part 2 – Current and short-term capability of a range of Age Assurance measures

A Research Report for the Information Commissioner's Office (ICO) and the Office of communications (Ofcom)

AUTHORS

Allen, Tony – tony.allen@accscheme.com
McColl, Lynsey - lynsey@select-statistics.co.uk
Walters, Katharine – katharine.walters@accscheme.com
Lyon, Matthew – matthew.lyon@accscheme.com

www.accscheme.com | 0345 257 0018 | info@accscheme.com

AVID Certification Services Ltd (14865982)
Unit 321 Broadstone Mill, Broadstone Road, Stockport, SK5 7DL, UK

Executive Summary

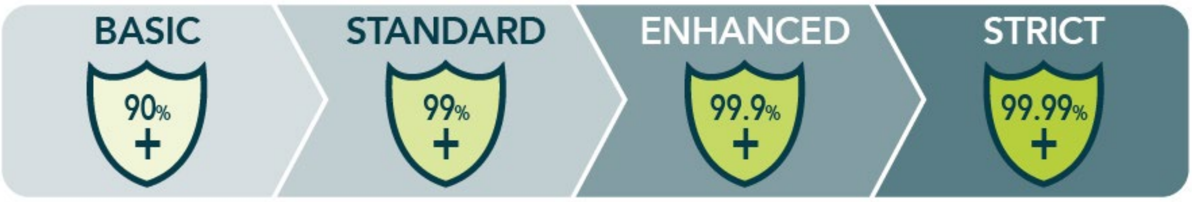
In 2022 the UK Information Commissioner's Office (ICO) commissioned ACCS to produce a technical study about the measurement of age assurance technologies. It was published in October 2022 and is referred to in this technical study as 'Part 1'. In conjunction with Ofcom, both regulators have commissioned this second technical study into understanding the current and short-term capability of a range of age assurance measures (referred to as 'Part 2'). It is intended to provide an understanding of the practicability and feasibility of developing a methodology for measuring the effectiveness and/or accuracy of age assurance systems across different services. The ICO and Ofcom have asked us to explore various age assurance methods across various industries and providers, including combined approaches, alongside an assessment of current effectiveness and anticipated effectiveness over the next five years.

Headline Measure of Accuracy

Part 1 identified statistical measures based on whether the age assurance measure delivered an estimation (continuous) output or a verification (binary) output. It set out how the use of measures such as mean absolute error, false positive rates, outcome error parity and other options could assess the accuracy of age assurance technologies. Part 1 pointed to the ongoing development of international standards setting out a framework for age assurance systems where proposed indicators of confidence are emerging through collaboration and building consensus. The content of Part 1 was also used by the authors of this part of the technical study (referred to as Part 2) in contributing to the development of those standards.

Part 2 maintains that the statistical measures identified in Part 1 are appropriate. It goes further, however, to hypothesise that a headline statement of overall accuracy of the age assurance measure could be provided, which could be more directly aligned to indicators of confidence. This could enable a quick, easy, and readily accessible indication of accuracy to be provided to an unfamiliar audience. When presented with other indicators, such as error rates, privacy and security controls, fairness measures and distribution of results, a holistic understanding of the effectiveness of the age assurance measure(s) overall could be established.

This short research project has focussed on one aspect – accuracy. We have tested our hypothesis, developed with scientific and technical advice, in the context of the current 'state-of-the-art' of age assurance measures. From this we have explored the outcomes that may be derived from aligning a headline measure of accuracy to the proposed indicators of confidence in international standards as follows:



ILLUSTRATIVE EXAMPLE - HOW OUR HYPOTHESIS COULD BE ALIGNED TO INDICATORS OF CONFIDENCE

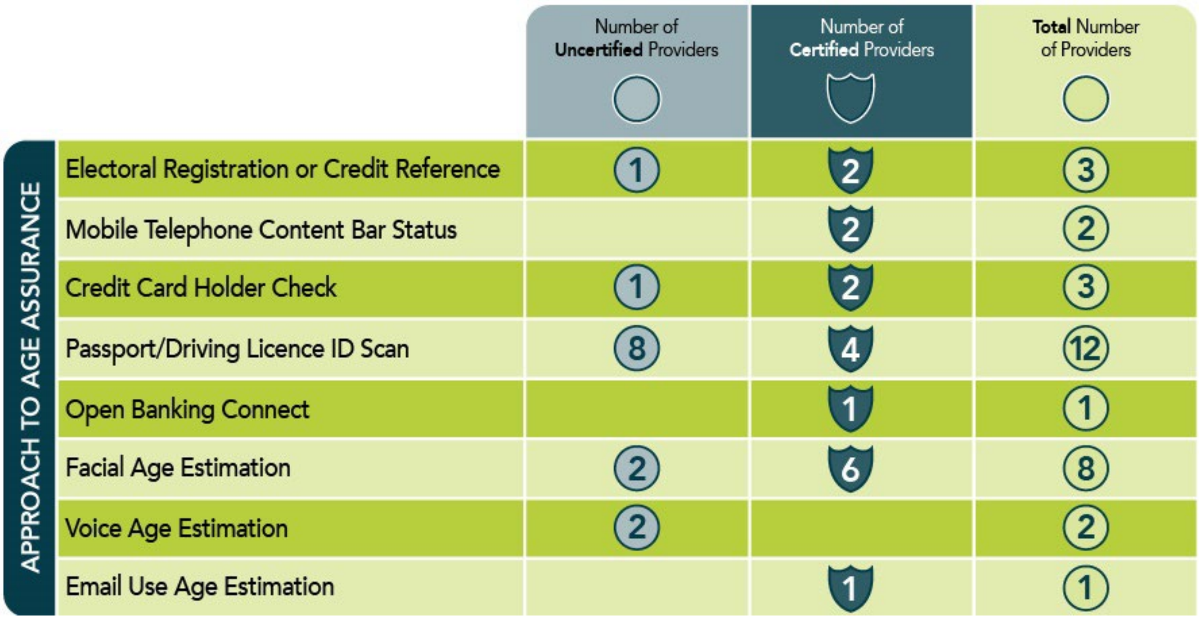
These numbers are an illustrative example of proposed indicators of confidence as are emerging through the consensus process of the development of ISO/IEC 27566 – Information security, cybersecurity, and privacy protection – Age assurance systems – Framework described in section 1.1 of Part 2.

Approach to Research

Our approach to Part 2 has been to open the findings of Part 1 to scrutiny and challenge. We have achieved this through establishing a Scientific and Technical Advisory Cell (STAC), through two workshops and through industry questionnaires. More detail about the approach taken is covered in Section 1.3 and Appendix 3. Part 2 focuses on applied research, by testing statistical theory and formulae against the current range of techniques capable of being deployed by age assurance providers.

In parallel to technical exploration, the project undertook a review of published academic articles about these innovative technologies, with a particular focus on applied statistical theory relating to binary and continuous approaches to measurement.

We identified age assurance service providers, a selection of age assurance methods (identified from those set out in Part 1 as being most likely to be prevalent in the UK market) and conducted primary research with the providers to identify which methods were commercially available and whether they were certified by independent third party testing.



		Number of **Uncertified** Providers	Number of **Certified** Providers	**Total** Number of Providers
APPROACH TO AGE ASSURANCE	Electoral Registration or Credit Reference	1	2	3
	Mobile Telephone Content Bar Status		2	2
	Credit Card Holder Check	1	2	3
	Passport/Driving Licence ID Scan	8	4	12
	Open Banking Connect		1	1
	Facial Age Estimation	2	6	8
	Voice Age Estimation	2		2
	Email Use Age Estimation		1	1

INDUSTRY SELF-DECLARED ANALYSIS OF AVAILABILITY OF AGE ASSURANCE MEASURES

Most of the approaches are currently employed by at least two providers, with half used by at least three. It is likely that more approaches to age assurance will emerge in future years.

In cooperation with the providers, we sought to examine the accuracy of these age assurance methods and explore these against the proposed hypothesis for indicators of confidence. This exercise was subject to our independent scrutiny and validation of the providers' data and claims, as we explain further in section 2.5. This was not an audit of any claims made by providers, nor verification or validation of their conformity against any standards, measures, tolerances, or schemes. In addition, we provide an objective assessment of the relative maturity of available methods in the context of the 'state-of-the-art'[1].

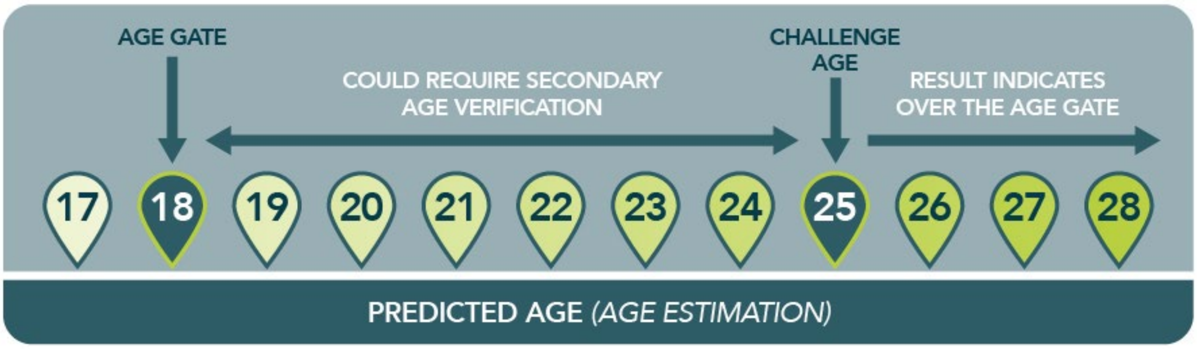
Conclusions and Observations

We conclude that the two separate approaches to measurement set out in Part 1 ('continuous' and 'binary' measures) could be expressed as a single binary headline range of accuracy. This initial indicator should be accompanied by transparency statements about the specific measures identified in Part 1 (such as Mean Absolute Error (MAE), Standard Deviation of AE (SD), True Positive Rate (TPR), False Positive Rate (FPR), Positive Predictive Value (PPV) and Outcome Error Parity (OEP)). This would assist with understanding of the accuracy of the component, while also maintaining the statistical detail for those that need deeper granularity for risk management decisions.

To achieve this, the proposed ranges of accuracy explored in our hypothesis would need to be established following further research and consultation, then age assurance technologies would need to be assessed according to a stated 'age gate,' which is the age or age range at which a technology is being tested. This could be any age gate selected by the provider submitting their system for test and will likely be driven by market demand. It could, for example, be 13, 16, 18, 21; or it could be 5 to 9, 10 to 12, 13 to 16; or any age range.

Some age assurance systems employ a workflow where initial age estimation processes are used to filter out individuals that are over a threshold (such as being over 25) before proceeding with secondary age assurance methods for those identified as under that threshold. We explored the implications for conversion of continuous (age estimation) measures to binary (Yes/No) outcomes and the risk that the measurement of accuracy of these would be disadvantaged by estimations close to the age gate (i.e., people just over or under the age gate are harder to accurately estimate than those much older or younger than the age gate).

For those systems, we suggest that the overall statement of accuracy would need to be stated together with the threshold age identified.


AGE GATE
CHALLENGE AGE
COULD REQUIRE SECONDARY AGE VERIFICATION
RESULT INDICATES OVER THE AGE GATE
17
18
19
20
21
22
23
24
25
26
27
28
PREDICTED AGE (AGE ESTIMATION)

ILLUSTRATIVE EXAMPLE OF HOW AGE ESTIMATION MEASURES COULD BE INCLUDED IN INDICATORS OF CONFIDENCE

[1] 'State-of-the-art' is a term used in the United Kingdom General Data Protection Regulation (UK GDPR) in respect of both security and data protection by design and default. It has significance in the context of deployment of appropriate organisational and technical measures and is explained in more depth in Section 2.2 of the report.

Part 2 of our technical study has been focussed on measurement of accuracy, but this is different from the overall effectiveness of age assurance systems. There are some important considerations that require further research, including bias, presentation attack vectors, fairness, and overall effectiveness. Part 2 also highlights some remaining challenges, not least of which being the availability of appropriate testing data sets, which may need to include biometric, demographically representative, and fairly distributed data. That series of practical, ethical, privacy and security concerns require further research and analysis.

Contents

List of Tables and Figures

Researchers

Tony Allen – Founder and Chief Executive of the Age Check Certification Scheme.

Tony is a Chartered Trading Standards Practitioner with over 25 years of experience in age restricted sales, law, and practice. He is Chair of the UK Government's Expert Panel on Age Restrictions. Tony holds a Master of Science (by Research), a Diploma in Trading Standards, a Diploma in Management Studies, and a BA (Hons) in Consumer Protection Law.

Lynsey McColl - Managing Director of Select Statistics

Lynsey is a Chartered Statistician with the Royal Statistical Society. She holds both a Master’s Degree and Ph.D. in Statistics. Lynsey is the Managing Director of Select Statistical Services and has experience of working in both public and private sector organisations applying her statistical knowledge to a wide range of Real-World problems.

Katharine Walters – Head of Policy and Regulation at the Age Check Certification Scheme.

Katharine was Head of Government Relations and Public Affairs at the Co-op for more than 20 years. She was a member of the former BEIS Retail Policy Forum, the British Retail Consortium’s Policy Board, and an advisor on many of the campaigns led by the Association of Convenience Stores. She spent four years in Brussels as a Researcher in the European Parliament. She holds a MA (Hons) in French with Politics from the University of Edinburgh.

Matthew Lyons – Associate Researcher

Matthew is a PhD student in the Machine Learning group at the University of Manchester, where he researches computer vision and deep learning models. In addition, Matthew has worked as a research software engineer in several research institutions, providing him with valuable experience in the field. He holds a Master of Medical Physics from the University of Sydney and a BSc (Hons) in Physics from the University of Warwick.

We provide more information about the Age Check Certification Scheme in Appendix 2.

Abstract

This is a technical study commissioned by the ICO and Ofcom into the current and short-term capability of a range of age assurance measures. The study builds on the approaches to measuring effectiveness, equality, comparability and repeatability set out in a study for the ICO published in October 2022 (referred to in this technical study as 'Part 1').

This study (referred to as Part 2) aligns continuous and binary measures for age assurance system to a headline measure of accuracy. We test the hypothesis that the indicators of confidence could be expressed as accuracy measures of 'Basic (90%+) – Standard (99%+) – Enhanced (99.9%+) – Strict (99.99%+)'. Our analysis of current performance of commercially available age assurance systems is then against these headline measures. We also undertake an objective assessment of the 'state-of-the-art' of current approaches.

We analyse the implications of this simplified approach and the potential loss of granularity of data available for technical specialists undertaking risk-based assessments when selecting appropriate measures for a particular use case. We suggest the retention and transparency of statistical measures identified in Part 1. Whilst also retaining other factors in Part 1(such as bias, liveness detection, fairness) covering the overall effectiveness of the system that require further research. We identify challenges with availability of data sets for testing.

Research Brief

The Research Brief issued jointly by the ICO and Ofcom stated that Part 2 should cover:

- Various age assurance methods deployed across various industries;
- Multi-methods of age assurance (e.g., a combined approach);
- A variety of providers;
- Current effectiveness, assessed using the measures outlined in Part 1 which was prepared for the ICO; and
- Anticipated effectiveness for the next 5 years, assessed using the measures outlined in Part 1 which was prepared for the ICO.

In response, ACCS put forward a structured 'sprint' research programme conducted over 10 weeks to deliver an appropriate analysis of the questions posed in Part 2.

This included:

Multiple Methods	Multiple Providers	Permutations & Combinations	Current Effectiveness
• Examining at least **eight** current techniques of establishing age assurance • Focus on those which are viable commercially and in practical usage • Direct applicability to the age appropriate design code and to video sharing platforms	• Working with at least **five** current, certified and active providers of age assurance solutions • Review of their own internal testing, analysis of results and independent validation • Anonymised results to ensure commercial confidence	• Examining the impact of the 'waterfall method' of age assurance • Measurement of the effectiveness of that approach, including statistical theorem • Modelling the different types of combinations (particularly the merging of continuous and binary approaches)	• Modelling the tolerances in different bands • Linking to the draft ISO/IEC and IEEE Standards, including the five bands of indicators of confidence proposed • Proposing regulatory risk characteristics and assessment of measures

Research Methodology

The overarching methodology for Part 2 was agreed with the ICO and Ofcom.

Part 2 contains one significant difference compared with Part 1. Whereas Part 1 was a largely theoretical study, deliberately undertaken with minimal interaction with age assurance providers, Part 2 focuses on applied research, by testing statistical theory and formulae against the current range of techniques capable of being deployed by age assurance providers.

There were several implications arising from this shift from the theoretical to the applied approach which guided the methodology. They include, but are not limited to, the ways in which the research team:

- Identified and assessed approaches currently undertaken by industry to understand and interrogate the current claims of age assurance providers about accuracy;
- Sought anonymised data from providers which underpin such claims;
- Validated[2] that such claims can be supported and explored options for achieving consistency and comparability across industry;
- Maintained rigorous independence of thinking in the findings and observations that form part of Part 2, and
- Ensured the rights and freedoms of data subjects and / or research participants, including anonymity and confidentiality.

Scientific and Technical Advisory Cell (STAC)

To assist Part 2, a Scientific and Technical Advisory Cell (STAC) was formed which included members of the ACCS Project Team, technical specialists from the ICO and Ofcom, and representatives from age assurance service providers. These latter members were in the

[2] As set out further in section 2.5 of this report, this validation exercise was based on self-declaration by the age assurance service provider and, in the context of this research, was not an audit of any claims made by providers, nor verification or validation of their conformity against any standards, measures, tolerances or schemes.

minority but were able to provide the desired technical insight and expertise from industry. The STAC met four times during the project.

We deliberately ensured that STAC members did not review individual company performance results.

Technical Engagement Workshops

Two project workshops were held at the Digital Security Hub (DiSH), supported by Barclays Eagle Labs, in central Manchester.

The workshops explored the main findings of Part 1 to test the continuing relevance; sharpened focus on the enablers and barriers to data gathering; and sought to identify consequences of the proposed approach to measurement of age assurance technologies, using a simplified overall measure of accuracy. More detail about the workshops is included in Section 1.3 and Appendix 3.

Data Gathering

To identify and assess how age assurance providers are substantiating claims about accuracy, an initial research questionnaire was sent to 24 global age assurance service providers. Of those 24, 15 responses were received[3], and analysis with their Trade Association suggests that these represent most of the providers who are currently active in the UK market. These responses are analysed in Section 2.5 of Part 2.

Respondents to the initial questionnaire were then offered a one-to-one interview to discuss the data gathering exercise in more depth, which 14 of the 15 respondents agreed to. Discussion of the research context, methodology and initial hypothesis proved useful in setting appropriate expectations and providing reassurance about the confidentiality and security of any relevant data that providers may volunteer.

Participating providers were then invited to self-declare their own analysis of the performance of their systems (be that individual components or combinations of components) against the measures set out in Part 2. This self-declaration was accompanied by disclosure to the Project Team of analytical data (no personal data), reports, white papers, certification, or research undertaken by the provider or any third party to underpin that claim.

The data gathering exercise focussed on eight age assurance techniques (identified from those set out in Part 1 as being most likely to be prevalent in the UK market)[4]:

- Five considered to be 'age verification' measures (likely to be binary measures):
  - Electoral registration or credit reference;
  - Mobile telephone content control measures;
  - Credit card holder check;
  - Passport / driving licence ID scan; and

[3] The sprint nature of the project drove tight timescales and providers were given a week to respond to the questionnaire.
[4] These measures were explained in Section 3 of Part 1 including their overall descriptor, a simple explainer and a technical or legal definition of each.

    - Connections to bank account information.
- Three considered to be 'age estimation' measures (likely to be continuous measures):
    - Facial analysis age estimation;
    - Voice age estimation; and
    - Usage of an e-mail address over time as a method of age estimation.

Data Validation

The Project Team explored how to validate the claims of individual age assurance providers and explored options to achieve consistency and comparability across industry. Statistical and data science colleagues independently reviewed the analysis of data, reporting, and presentation of all or any claims about accuracy to assess whether they could be substantiated by us as an independent third-party. This exercise also sought to establish some early benchmarking analysis about the current 'state-of-the-art'.

The Project Team also tested initial hypotheses around tolerances to assess which tolerance spans may be most appropriate for the two age gates which currently drive most use cases, namely whether a person is over or under the age of thirteen, or over or under the age of eighteen.

It is important to note that this was not an audit of any claims made by providers, nor verification or validation of their conformity against any standards, measures, tolerances, or schemes. The purpose of this exercise was purely to support the objectives of Part 2 and this report, with guaranteed company anonymity, to the ICO and Ofcom.

Final Report

This report is intended to provide an independent technical study of the observed and gathered feedback, evidence, data, analysis, and validation. The report has been subject to an impartiality review in accordance with ACCS' process and policy for securing impartiality under ISO 17065:2012 – Requirements for bodies certifying products, processes, and services.

It is the understanding of the Project Team that Part 2 will be published by the ICO and Ofcom at some future point.

Glossary

Most terms used in Part 2 are defined in Part 1, but to aid understanding, we have provided a glossary of some terms at the back of this report together with a bibliography of referenced material.

Acknowledgements

The Project Team would like to thank all the participants at the two workshops and for the STAC for all the consideration of the topics under examination and the clarity of thought and comment on the issues arising. This has really helped to guide and direct Part 2 of the technical study to support the objectives set by the ICO and Ofcom.

1.Background

This introductory section summarises at a high level the initial findings of Part 1 of the technical study commissioned by the ICO and published in October 2022, referred to in this technical study as 'Part 1'. The analysis contained in Part 1 and its recommendations have formed the foundation of this Part 2 study. It includes a brief overview of the ICO and Ofcom and their responsibilities with regards to age assurance and a brief description about the work and status of the Age Check Certification Scheme (ACCS). Further detail about the organisations can be found in Appendices 1 and 2.

1.1 Findings in Part 1 of the research

The first technical study for the ICO was published in October 2022. It made eight recommendations, and this new Part 2 technical study, jointly commissioned by the ICO and Ofcom, is a step on from recommendation 5 in Part 1; namely that further work should be undertaken to:

> *"Identify, consult on and publish appropriate levels of tolerance for acceptable age assurance systems. These could be expressed as a risk-based approach depending on the level of confidence for the age assurance needed commensurate with the risk identified. To align this with the forthcoming international standard, the levels of confidence should be based on 'Asserted - Basic - Standard - Enhanced - Strict' approaches."*

Part 1 also touched on the efforts currently underway by the International Standards Organisation (ISO) to develop ISO/IEC 27566 – Information security, cybersecurity, and privacy protection – Age assurance systems – Framework[5], and individual efforts within different agencies, conformity assessment bodies and government/regulators to understand and define age assurance systems. The lead author for Part 2 is also acting as the Technical Editor for the standards development set out above, having been nominated by the British Standards Institution (BSI), the UK's National Standards Body for ISO.

It is suggested that a simple approach to describing the indicators of confidence achieved by different assurance components would assist service providers, relying parties and those that regulate them. Part 1 highlighted how international standards were developing around five indicators of confidence 'Asserted - Basic - Standard - Enhanced - Strict'.

The aim and intention of the standardisation process is to provide formulae, tolerances, descriptions, and parameters to these five indicators of confidence to enable policy or decision makers to apply their risk assessment considerations to the appropriate and proportionate level that is needed for the relevant age-related eligibility decision.

[5] See https://www.iso.org/standard/80399.html

For the purposes of Part 2, we do not look any further at 'Asserted' (the equivalent of self-declaration of age), but it does remain relevant to some use cases.

Part 1 examined multiple statistical methodologies for the assessment of these technologies – built around the core principles that the output of the process is either continuous (i.e. an estimation) or binary (i.e. a verification).

It also considered approaches to testing, analysis, and certification. Part 1 considered the key factors that need to be taken into consideration when assessing the approach to testing of age assurance systems. These included ensuring that:

a) The test protocols applied to secure repeatability and reproducibility of age assurance testing results are appropriate;
b) The identification and controls associated with the data capture subjects and data capture devices are considered and recorded;
c) The approach to both human and document presentation tack detection (spoofing) is undertaken in accordance with the relevant international standards[6];
d) Testing is undertaken in the appropriate ambient lighting for the use cases of the age assurance system (lighting has a significant impact on system effectiveness); and
e) The assessment considers the appropriate sample size and depth of evaluation, potentially applying different evaluation assurance levels commensurate with the level of confidence sought in the age assurance technology.

1.2 ICO and Ofcom interest in age assurance

The roles and responsibilities of both ICO and Ofcom in this area of emerging policy are set out in Appendix Two. In summary:

The ICO has issued the Children's code (known formally as the Age appropriate design code) which articulates how online services should safeguard children's personal data. The code states that organisations should either establish an appropriate level of certainty about the age of their users or apply the standards in the code to all their users[7]. This requires organisations to ensure the protections are appropriate for the age ranges of their users, by tailoring what they offer and putting the necessary safeguards in place for each age range. The Children's code is underpinned by the UK data protection legislation which organisations must comply with and notably Article 8 of the United Kingdom General Data Protection Regulation (UK GDPR) sets the age at which children can consent to the processing of their personal data in the context of an ISS at 13 years old.

In November 2020, Ofcom started regulating video-sharing platforms (VSPs established in the UK) and is required to ensure that such VSPs take 'appropriate measures' to protect minors from content which may impair their physical, mental, or moral development. Ofcom is now committed, as part of its VSP regulation, to driving forward the implementation of robust age

[6] Such as ISO/IEC 30107-1:2016 - Information technology — Biometric presentation attack detection — Part 1: Framework
[7] See particularly https://ico.org.uk/for-organisations/guide-to-data-protection/ico-codes-of-practice/age-appropriate-design-a-code-of-practice-for-online-services/3-age-appropriate-application/

assurance to protect children from the most harmful online content, including pornography.[8] Ofcom is also due to become the Online Safety regulator for regulated user-to-user services, regulated search services, and regulated online pornography services.

1.3 Technical Engagement Workshops

We established two technical engagement workshops drawing together specialists from the age assurance industry, statistical and measurement analysts, biometric scientists, and conformity assessment specialists. The purpose of the workshops was to open the findings of Part 1 to scrutiny and challenge and to provide an opportunity to explore our hypothesis for a headline measure of accuracy aligned with the proposed indicators of confidence.

Detail about the two workshops, including how they were structured, an overview of the types of participants, the questions posed and an overview of the discussion that followed is included in Appendix Three. Those who seek a deep understanding of how the Project Team have formed its observations are encouraged to delve into that accompanying detail.

In summary:

- Participants felt that the focus on binary vs continuous outcomes and the conversion of continuous methods into binary outputs was appropriate. There was agreement that continuous measures will, in reality, just be binary with logic applied to the output.
- They also welcomed new thinking about the overall confidence measure as being positive (albeit challenging from a testing and/or accreditation perspective). There was a clear recognition of the issues about good quality and accessible data sets and the specification of test data for use case/ real world deployment e.g., age restricted sales verification should focus the error analysis on the most relevant age ranges.
- Participants noted the potential impact of cultural bias and the critical need for a careful approach to data ethics noting that it is a tricky problem to solve, especially for under 18s.
- Participants were also aware of the potential liability of getting it wrong, proposing a need to consider the potential for age discrimination of 18 – 25-year-olds without formal ID documents.
- Some participants felt that it is possible that considerations in this work are too focussed on age estimation approaches, which is just one solution.
- It is important to limit the false negative rate to help protect users from excessive collection of data.

The issue that attracted the most feedback from participants in Workshop 2 remained the development, curation and maintenance of an independent data set for testing the overall effectiveness of age assurance systems. Participants noted that there should be more thought given to the different categories included in the testing data set e.g., ethnicity, skin tone, gender etc.

Participants also commented on the security of age assurance models e.g., susceptibility to model evasion attack.

[8] See Section 3.170 – 3.201 of Ofcom Video-sharing platform guidance: Guidance for providers on measures to protect users from harmful material (2021).

However, there was consensus that the proposed hypothesis was a good place to start and created some interesting outcomes. For instance, in some age assurance methods, particularly those that are sourcing hard identifier data, the results presented were either ‘an indication of age’ or ‘no data’; thus, leading to a situation where the actual accuracy measure was that of the provider of the hard identifier data and not the age assurance solution.

A particular example of this was the use of the electoral register established by local authorities in the United Kingdom. An age assurance provider enquiring into the presence or otherwise of a person and/or their date of birth on the electoral register would return a ‘yes – there is data, and this is what it says’ or ‘no – there is no data.’ The accuracy is not about the age assurance system, but the quality of the register of electors. In this case, we noted research by the Electoral Commission[9] that the error rate for date of birth on the register of electors[10] is less than 0.1% - indicating a reliability to an ‘enhanced’ level of confidence based on our hypothesis.

[9] https://www.electoralcommission.org.uk/who-we-are-and-what-we-do/our-views-and-research/our-research/accuracy-and-completeness-electoral-registers/2019-report-2018-electoral-registers-great-britain/accuracy-great-britain

[10] The date of birth held on local registers of electors may only contain entries for recent attainers of the right to vote (which differs for distinct types of election in the UK), and once established on the register, entries may simply be inferred that they are over 18.

2. Data Gathering, State-of-the-Art, Industry Metrics

In this section we explore what the 'state-of-the-art' means in the age assurance sector in the United Kingdom. This has been based on our own primary research and engagement with age assurance service providers to understand and interrogate their claims relating to accuracy of their systems.

We started by outlining our approach to data gathering and defining 'state-of-the-art' – particularly in the context of legislation, such as UK GDPR. We then assessed the data provided to us by age assurance service providers and sought to validate it. This was not an audit or certification process, and we did not require the participating providers to be certified to take part in this analysis.

2.1 Approach to Data Gathering

As set out in our Research Methodology (see page 10), we approached our data gathering exercise in stages. Firstly, we sought to identify the age assurance providers that we believed were or could be operating in the UK marketplace.

We then identified eight age assurance component types from the list in Part 1 of our research. These eight were considered to be at or near to market deployment and provided a good spread of different activities for our research. We discuss in more detail whether these eight methods can be considered 'state-of-the-art' in section 2.2. The eight types that we identified are:

- Electoral Registration or Credit Reference
- Mobile Telephone Content Bar Status
- Credit Card Holder Check
- Passport/Driving Licence ID Scan
- Open Banking Connect
- Facial Analysis Age Estimation
- Voice Analysis Age Estimation
- E-Mail Use Age Estimation[11]

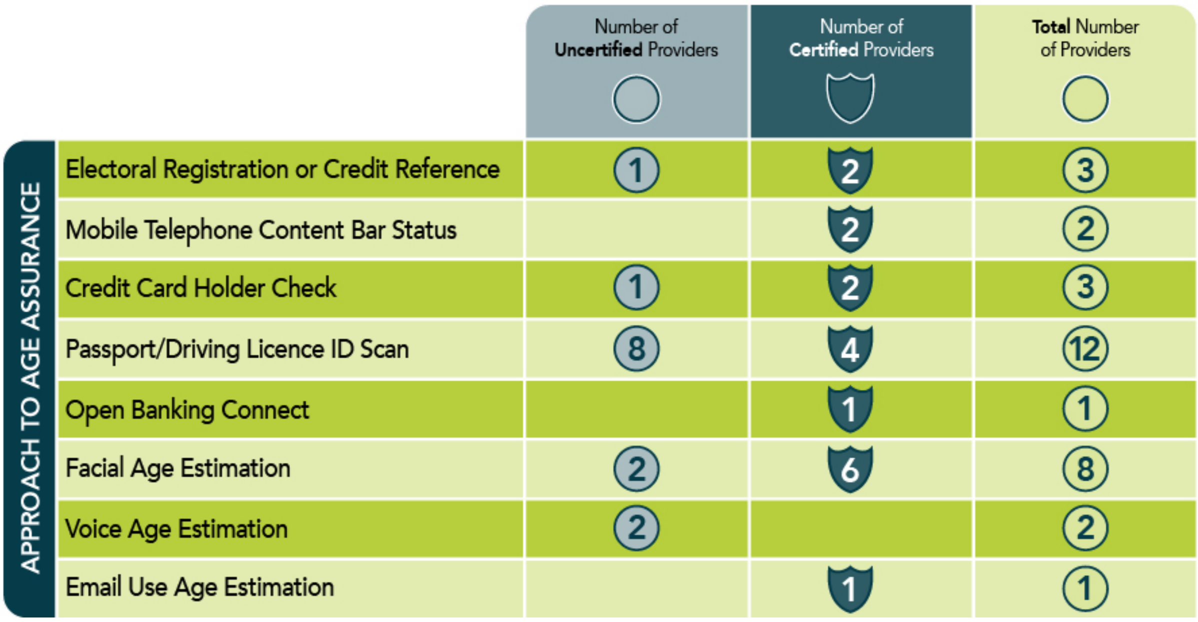
From open-source research, the providers' websites and our own records, where applicable for our clients, we identified the age assurance component types provided by each of the 24 identified providers – some provided just a single type, some provided multiple types. We also sought to identify if any of the methods had been independently validated or certified, either by the ACCS or another independent conformity assessment body.

We wrote to 24 global age assurance providers seeking confirmation of our initial segmentation and received replies from 15 of them confirming or updating our assumptions. We also asked if they would be willing to participate in further analysis of their approach to age assurance for the purposes of this research, which all but one agreed to.

[11] An e-mail address is formally known as a Fully Qualified Domain Address (FQDA)

In keeping with our undertaking to maintain their trust by making sure that Part 2 of this technical study was anonymous the report does not reference any company names of age assurance providers.

The responses indicate the number of providers of each age assurance component in the market is as follows:



APPROACH TO AGE ASSURANCE		Number of **Uncertified** Providers	Number of **Certified** Providers	**Total** Number of Providers
	Electoral Registration or Credit Reference	1	2	3
	Mobile Telephone Content Bar Status		2	2
	Credit Card Holder Check	1	2	3
	Passport/Driving Licence ID Scan	8	4	12
	Open Banking Connect		1	1
	Facial Age Estimation	2	6	8
	Voice Age Estimation	2		2
	Email Use Age Estimation		1	1

FIGURE 1 - TABLE OF AGE ASSURANCE MEASURE AVAILABILITY

In seeking responses, we asked age assurance providers if their solutions had been subject to independent third party testing and certification.

2.2 Approach to 'State-of-the-Art'

The 'state-of-the-art' is a relevant consideration in identifying the appropriate technical and organisational measures that may be available to an organisation, but there are others as stated in Article 25 of UK GDPR, such as cost of implementation and potential risks identified.

For this project, we were asked to assess the relevant 'state-of-the-art' of current and potential age assurance technologies.

The relevant legislation

- Article 25[12] of UK GDPR requires data controllers to implement data protection by design and by default:

> *"Taking into account the **state of the art**, the cost of implementation and the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing"*

[12] Article 32 (security) also uses the term 'State-of-the-Art'.

- Section 368Z1 of the Communications Act 2003 requires UK-established video sharing platform services to take appropriate measures (as listed in Schedule 15A) for the purposes of protecting under 18s from videos and audiovisual commercial communications containing restricted material. VSP providers must determine whether it is appropriate to take a particular measure according to whether it is practical and proportionate to do so, taking into account a number of factors, including the size and nature of the VSP service, the nature of the material, and the harm that may be caused by it.

In the current Online Safety Bill[13], the development of Codes of Practice are proposed to be covered by some general principles, including at proposed Schedule 4, clause 2(c):

> *"the measures described in the code of practice must be proportionate and technically feasible: measures that are proportionate or technically feasible for providers of a certain size or capacity, or for services of a certain kind or size, may not be proportionate or technically feasible for providers of a different size or capacity or for services of a different kind or size"*

Although legislation uses different terminology, all of it relates to the implementation of technical measures that can be proportionate, feasible and reasonable.

State-of-the-art is not defined and is, by its nature, a nebulous and contemporary concept. The European Data Protection Board[14] have issued guidance that:

> *"In the context of Article 25 [GDPR], the reference to "state of the art" imposes an obligation on controllers, when determining the appropriate technical and organisational measures, to take account of the current progress in technology that is available in the market. The requirement is for controllers to have knowledge of and stay up to date on technological advances; how technology can present data protection risks or opportunities to the processing operation; and how to implement and update the measures and safeguards that secure effective implementation of the principles and rights of data subjects taking into account the evolving technological landscape.*
>
> *The "state of the art" is a dynamic concept that cannot be statically defined at a fixed point in time but should be assessed continuously in the context of technological progress. In the face of technological advancements, a controller could find that a measure that once provided an adequate level of protection no longer does. Neglecting to keep up to date with technological changes could therefore result in a lack of compliance with Article 25."*

[13] At the time of writing, the Bill was in Committee Stage in the House of Lords and this quoted clause was as the Bill stood at the start of that Stage in January 2023.

[14] Guidelines 4/2019 on Article 25 - Data Protection by Design and by Default - Version 2.0 - Adopted on 20 October 2020 - The EDPB Guidelines are used to enable consistency across data protection supervisory authorities of the European Union. They are no longer binding under the UK regime, but it is the view of ACCS that they remain helpful guidance when considering similar provisions in the UK data protection regime.

In our approach to this, we have used the guidance[15] produced by the EU's Agency for Cybersecurity (ENISA) on assessing the state-of-the-art of IT security components. This is referenced in the EDPB Guidance noted above and stems from the Kalkar case[16] in the German Constitutional Federal Courts in 1978. In the context of the state of technological advancement in IT products and services, it is broadly recognised as the correct approach.

It is important to note that 'state-of-the-art' in the context of UK GDPR involves an objective approach where multiple and diverse approaches to a particular technological challenge can be identified. In other contexts, such as marketing for instance, it could be referring to the 'best-of-the-best' or a single 'outstanding exemplar' in a market. Here we look at the legal context.

'State-of-the-art' is best described as the layer of technology that sits between those technologies that are within existing scientific knowledge and research but may not yet be commercially viable services; and those technologies that are so embedded in everyday life and usage, they have become generally accepted rules of technology. These features are amplified by the level of general industry recognition the technologies have and the extent to which they are proven to work in practice.

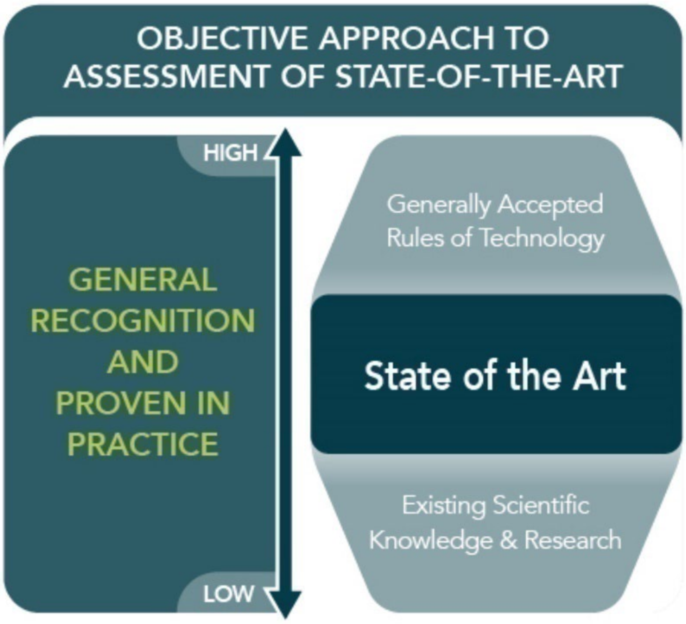

OBJECTIVE APPROACH TO ASSESSMENT OF STATE-OF-THE-ART
HIGH
GENERAL RECOGNITION AND PROVEN IN PRACTICE
LOW
Generally Accepted Rules of Technology
State of the Art
Existing Scientific Knowledge & Research

FIGURE 2 - OBJECTIVE APPROACH TO 'STATE-OF-THE-ART'

The classification of age assurance technologies requires a clear distinction between subjective and objective criteria. The 'state-of-the-art' criterion is purely objective. The subjective aspects consider the deployment or use of the technologies; however, they do not concern the definition of the 'state-of-the-art' itself.

As a result, the 'state-of-the-art' can be described as the procedures, equipment, or operating methods available in the trade in goods and services for which the application thereof is most effective in achieving the respective legal protection objectives.

[15] See What is "state of the art" in IT security? — ENISA (europa.eu)
[16] BVerfGE, 49, 89 (135 f)

Technical measures at the “existing scientific knowledge and research" stage are highly dynamic in their development and pass into the ‘state-of-the-art’ stage when they reach market maturity (or at least are launched on the market).

2.3 Application of State-of-the-Art to Current Age Assurance Technologies

The scope of our research has included what could be state-of-the-art in the domain of age assurance technologies. As set out in Section 2.2, this is an objective assessment made at a given point in time – as in the state-of-the-art at the date of this Part 2 of the technical study. We have taken the eight age assurance types and have assessed them based on the ENISA Guidelines[17].

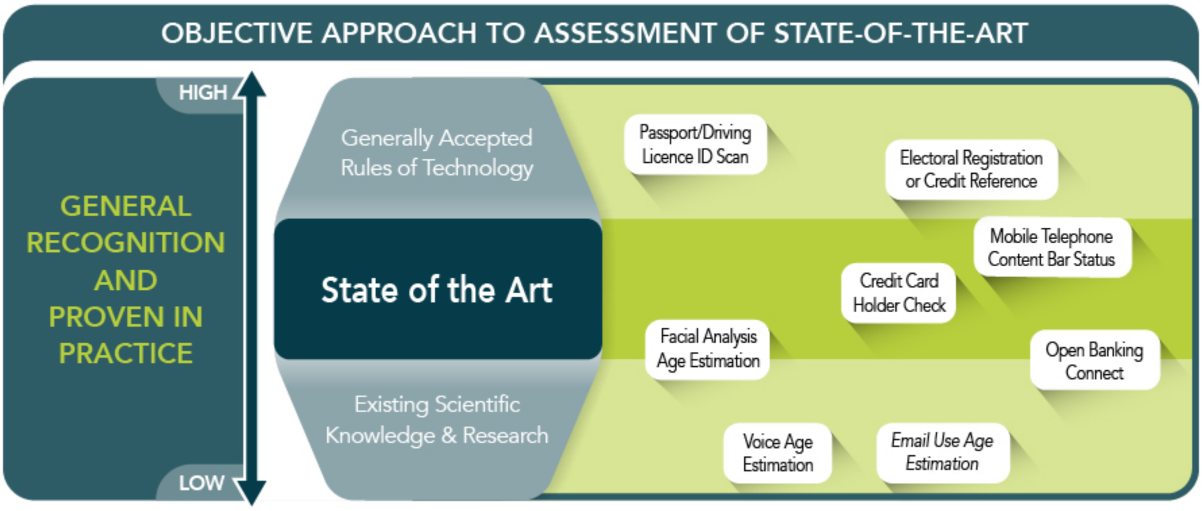

OBJECTIVE APPROACH TO ASSESSMENT OF STATE-OF-THE-ART
HIGH
LOW
GENERAL RECOGNITION AND PROVEN IN PRACTICE
Generally Accepted Rules of Technology
State of the Art
Existing Scientific Knowledge & Research
Passport/Driving Licence ID Scan
Electoral Registration or Credit Reference
Mobile Telephone Content Bar Status
Credit Card Holder Check
Facial Analysis Age Estimation
Open Banking Connect
Voice Age Estimation
Email Use Age Estimation

FIGURE 3 - APPLICATION OF STATE-OF-THE-ART TO AGE ASSURANCE TECHNOLOGIES

Our assessment of the current state-of-the-art, considers each component in isolation. However, it is worth noting that many age assurance providers use a combination of methods, such as through a ‘waterfall’ technique[18].

Age assurance components with high general recognition and that have been proven in practice may have existed for many years and be generally accepted rules of technology. Techniques such as Passport/Driving Licence scanning, electoral registration lookup or credit reference agency checks have formed the bedrock of long-standing age-related eligibility requirements such as gambling, credit, or access to licensed premises.

Certain less data intensive ‘quick check’ methods have also developed but may be less generally recognised or subject to less objective proof of effectiveness and accuracy in practice, such as through independent third-party conformity assessment. These include measures such as Mobile Telephone Content Controls or Credit Card Holder checks[19]. These are still in the ‘state-of-the-art’ level.

[17] See footnote 15.
[18] We describe ‘waterfall techniques’ in Part 1 of our Research on pages 36-37.
[19] In the UK, a minor (under 18) does not have the capacity to enter a legally binding consumer credit arrangement, such as a credit card (this does not apply to debit, pre-pay, or gift cards)

Other technologies exist which are transitioning from existing scientific knowledge and research into state-of-the-art. This includes Facial Analysis Age Estimation, where there is now proof in practice and growing general recognition of this technology as ‘state-of-the-art.’

With less general recognition, but nevertheless some proof in practice, are Open Banking Connect or E-Mail Use Age Estimation. These are not widely deployed (with single providers identified in this research). They could be regarded as close to the ‘state-of-the-art’ layer and may quickly emerge.

2.4 Future Development of Age Assurance Technologies

ICO and Ofcom have invited us to consider the technologies that may emerge over the course of the next five years. This requires analysis of the existing and potential scientific knowledge and research. It also requires a cost-benefit analysis for commercial age assurance providers to invest in further development of the technology. This requires an identifiable and applicable market for the technology.

There are a number of age assurance technologies currently in development. For example, voice age estimation technology is currently undergoing highly active development, research, and live testing. There are multiple providers seeking to prove its effectiveness but is not yet widely available in the market.

There is also recent scientific research[20] into hand modalities, including how to distinguish children and adults for fingerprint, palmprint, hand-geometry and digit print biometrics. There may be significant privacy protection advantages of, for instance, using the size and movement characteristics of a person’s hand to determine if they are likely to be an adult or fall into one of the suggested age and developmental stages included in Annex B of the Children's code. However, no solution has yet transferred to the commercial market.

More recently, research[21] into the potential for age classification using the electrocardiogram (ECG) trace of a person has been published. For age classification (adults and children), this research claimed classification accuracies up to 99% (a ‘basic’ to ‘standard’ indicator of confidence). The researchers highlighted how “such promising outcomes generated the feasibilities of further experimentation and possible practical implementation of ECG for anonymous age verification”.

There are many opportunities to explore new age assurance components. However, we assess that over the next five years, focus will concentrate on achieving interoperability between providers in the marketplace. This is likely to be through standards, protocols, taxonomy development and deployment through ‘hubs’ or ‘exchanges’. Interoperability is likely to be important, to achieve the deployment and creating a smooth user experience.

[20] A. Uhl and P. Wild. “Comparing Verification Performance of Kids and Adults for Fingerprint, Palmprint, Hand-geometry and Digit print Biometrics.” In Proceedings of the IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems (BTAS’09), 6 pages, Washington, DC, USA, September 2830, 2009.

[21] A. Adib, W. -P. Zhu and M. O. Ahmad, "Adult and Non-Adult Classification Using ECG," 2022 IEEE 7th Forum on Research and Technologies for Society and Industry Innovation (RTSI), Paris, France, 2022, pp. 174-179, doi: 10.1109/RTSI55261.2022.9905194

2.5 Analysis of Current Age Assurance Technologies

As part of the research, we asked providers to assess where they believe their age assurance components would fit on the Basic – Standard – Enhanced – Strict error classification rates presented in our hypothesis. We requested data to substantiate those claims and our data science team undertook a validation exercise to assess the reliability of these claims for the purpose of Part 2 of this technical study.

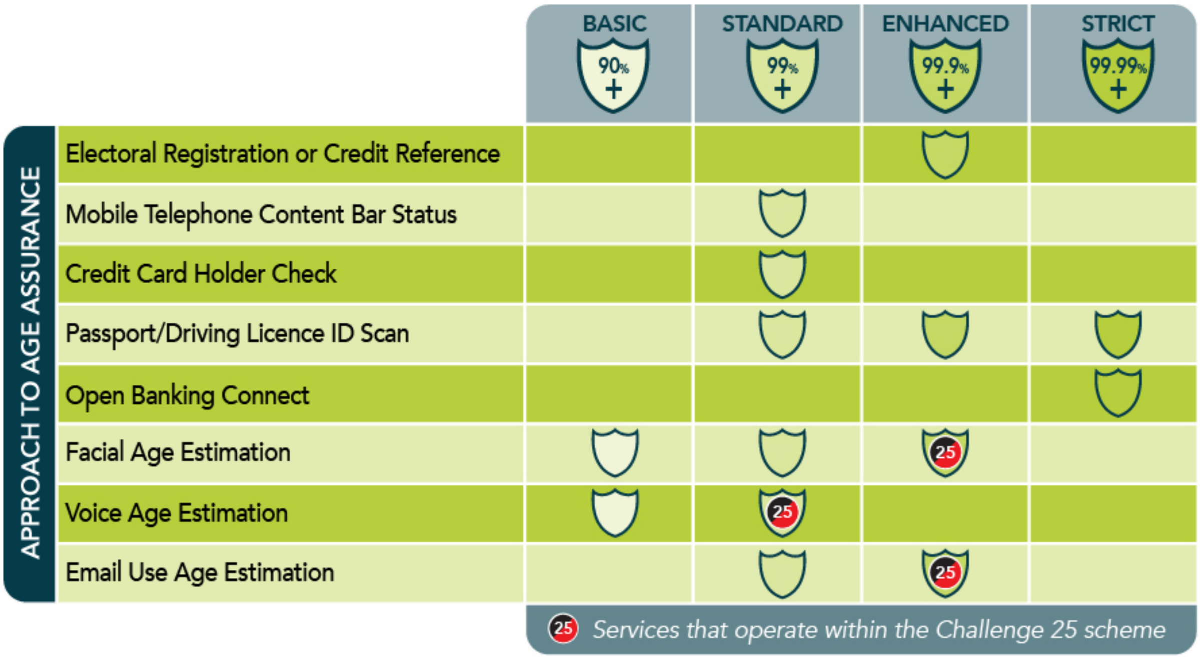
It is worth noting that this is self-declared data and although we have independently validated that data, this does not constitute an audit and certification process. This validation included a review by our own data scientists that the information provided was accurate, complete, and consistent, and that it was free from errors, anomalies, and inconsistencies. We have no reason to believe that the claims of conformity are inaccurate (and they accord with our understanding of the 'state-of-the-art'), but they should be regarded with some caution unless independently certified. For instance, we cannot confirm if the assessments are against comparable data sets or testing protocols; nor how they may have addressed statistical outliers.

Throughout this project we have undertaken not to publish individual performance or create any kind of 'league table' and we maintain that stance here.

We put the following scenario to age assurance providers:

> *"It is accepted that age assurance processes are not fool proof, and that 100% accuracy is difficult to achieve 100% of the time in 100% of all actual deployments. For instance, occasionally they will classify somebody as over 18 when they are not. This exercise is about understanding how infrequent that actually is.*
>
> *We have 2 questions:*
>
> 1. *Based on your own knowledge of your own age assurance methods (taken individually) how rarely do you believe your system may get the answer 'Is this person over 18?' wrong?*
>
> 2. *Have you got any data, analysis, white papers, internal reports, or external certifications that can provide any evidence to support your assessment?*
>
> *Here we are only considering your age assurance methods being presented with a user and being asked to answer the question – is this user over 18?*
>
> *For the purposes of this exercise, the answer that your system provides will be YES or NO. This is an assessment of all users of all ages. What we are seeking to discover is how often it might get that answer wrong.*
>
> *If 10,000 users are put through one of your age assurance components on its own (i.e., not multiple or waterfall techniques), how many times do you believe that process may result in an outcome that is wrong.*
>
> *We are just looking at 'normal everyday usage' by the range of users that are presented to your system, and so please ignore anything relating to mean absolute errors or spoofing."*

Notwithstanding this research, there remain several challenges to reliably assess the accuracy of age assurance technologies, and accuracy is not equivalent to effectiveness (as described in Part 1). Some issues include the lack of available independent test datasets and the discrepancy between real-world and test-lab results conditions. With these limitations in mind, providers shared data indicating how well their own solutions perform against the headline statement of accuracy and associated indicators of confidence. The indicative results are set out in the below table. It is important to note that comparability should be focused on outcomes, rather than process, ensuring that the context of the use case is of paramount importance.



APPROACH TO AGE ASSURANCE	BASIC 90% +	STANDARD 99% +	ENHANCED 99.9% +	STRICT 99.99% +
Electoral Registration or Credit Reference			✓	
Mobile Telephone Content Bar Status		✓		
Credit Card Holder Check		✓		
Passport/Driving Licence ID Scan		✓	✓	✓
Open Banking Connect				✓
Facial Age Estimation	✓	✓	25	
Voice Age Estimation	✓	25		
Email Use Age Estimation		✓	25	

25 Services that operate within the Challenge 25 scheme

FIGURE 4 - ANALYSIS OF ACCURACY OF AGE ASSURANCE METHODS AGAINST OUR HYPOTHESIS

We noted that based on this self-declared data there appears to be a range of available approaches to age assurance in the marketplace for each of the hypothetical indicators of confidence against our hypothesis for the range of accuracy within each indicator. As we expected, the strict level tended to be applicable to those measures using ‘hard identifiers,’ such as uploads of passports and driving licences. We know that this may present data minimisation and security challenges, so that ought to be considered if specifying a ‘strict’ level of confidence on the parameters set out in this Part of the technical study.

2.5.1 Challenge 25

Some age assurance systems employ a workflow where initial age estimation processes are used to filter out individuals that are over an age threshold (such as being over 25) before proceeding with secondary age assurance methods for those identified as under that threshold. We explored the implications for conversion of continuous measures (i.e. age estimation) to binary (Yes/No) outcomes, and the risk that the measurement of accuracy of these would be disadvantaged by estimations close to the age gate. For example, people just over or under the age gate are harder to accurately estimate than those much older or younger than the age gate.

We also developed our exercise in relation to age estimation technologies where we applied a challenge age of 25 to the data that we gathered from age estimation providers. Effectively, this meant asking the same question – is this person over 18? – but ignoring any results that were returned between the age of 18 and 25. This replicates what may happen in the real world: a customer is initially age estimated by eye and if the seller thinks that they look under 25, they are then asked to produce ID to prove that they are over 18 (a process known as 'Challenge 25' or 'Think 25' in the retail, hospitality, and gambling sectors).

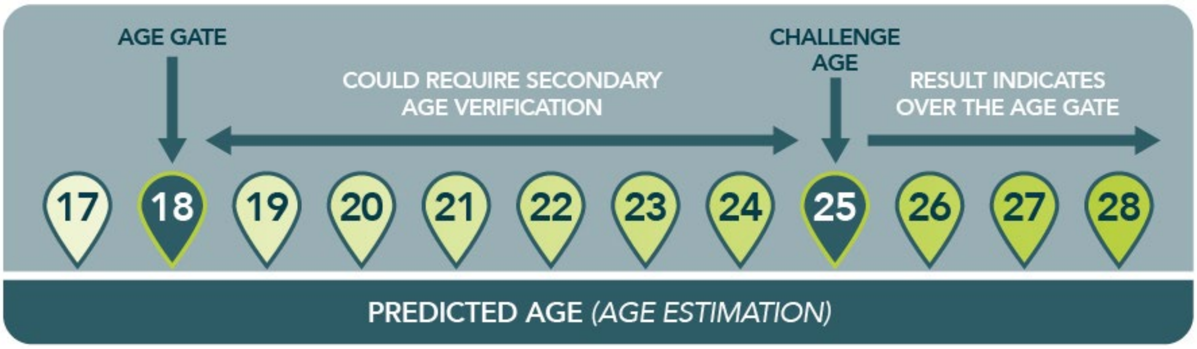

AGE GATE
CHALLENGE AGE
COULD REQUIRE SECONDARY AGE VERIFICATION
RESULT INDICATES OVER THE AGE GATE
17
18
19
20
21
22
23
24
25
26
27
28
PREDICTED AGE (AGE ESTIMATION)

FIGURE 5 - AGE GATES AND CHALLENGE AGE

If this is applied to the data, it upgrades the accuracy of the outputs, so a facial age estimation system that performs as 'standard accuracy' across all age groups, could perform at 'enhanced accuracy' under a challenge age of 25. It is worth noting that this approach requires some access to a secondary method of age assurance to address those people over 18, but not assessed by age estimation as being over 25 (effectively false negatives).

3.Refined Methodology for Measuring Accuracy

In this section, we explore the options to develop the approach set out in Part 1 into a simpler form, while providing an overall measure of performance that could be readily understood. The section highlights the opportunities and risks associated with any potential simplification. Overall, however, we conclude that it is possible to provide a simple measure of accuracy in response to a binary question (where the answer to a question such as 'Is this user over 18?' is 'Yes' or 'No'). This needs to be anchored in a statement of how the component(s) are used and the age gate(s) that are applicable.

We also explain the need to retain the measures set out in Part 1. They are and remain valid. We therefore suggest that in addition to the primary focus on overall accuracy, they should be referenced as secondary measures to provide more in-depth understanding and greater transparency.

We also explore the depth of testing that may be required and set out other aspects, first highlighted in Part 1, including effectiveness (where live detection and presentation attack are important) and equality (which includes bias). Further work and research is required on those aspects.

3.1 Introduction

When assessing the accuracy of an age assurance technology, it is important that the testing procedures and the measures of accuracy reflect how that technology is deployed. In Part 1, we set out various measures that can be used to assess the accuracy of both age estimation and age verification technologies. These measures considered the fact that the outcome from an age estimation technology is continuous, whilst the outcome from an age verification technology is binary.

Following the two technical engagement workshops, there was consensus that the deployment of an age assurance technology (whether it is age verification or age estimation) will, in fact, result in a binary outcome. For example, is the subject presented to the technology older or younger than 18? For an age estimation technology, an age threshold would simply be applied to the estimated age to identify whether a subject was older or younger than 18.

Considering this, we have identified further findings and observations for the measurement of accuracy to focus primarily on the binary metrics discussed in Part 1. Below, we set out how this might work. Throughout this section we refer to terms and measures that are defined and set out in Sections 5 and 6 of Part 1 of the technical study (we do not repeat the definitions here).

3.2 Age Gate

An age gate is the age or age range that is of interest to the relying party where an age-related eligibility decision is required. Age assurance technologies need to be assessed according to a stated age gate. This could be any age gate selected by the provider submitting their system for test and will likely be driven by market demand. It could, for instance, be 13, 16, 18, 21; or it could be 5 – 9, 10 – 12, 13 – 16; or any age range.

There are three different age gate scenarios to consider for testing, all of which can be treated as binary:

- Scenario 1: Over an age gate (e.g., 13 or 18) to stop access to age-inappropriate products/materials/services. In this case a person is identified as being over the age gate (positive) or under (negative).
- Scenario 2: Under an age gate to access safe places where no adults are allowed for safeguarding reasons (except for appointed safeguarding monitors). In this case a person is identified as under the age gate (positive) or over (negative).
- Scenario 3: Between one specified age and another (see, for example, age and developmental stages set out in Annex B of the Children's code). In this case a person is identified as within the specified range (positive) or outside (negative). Although this scenario involves two age thresholds, it can still be treated as binary since we simply convert the solution into between the two thresholds or outside of the two thresholds (there are two answers).

In all cases here, the testing scenarios have been set up to reflect that a false positive (i.e., incorrectly identifying someone as being positive) will cause harm. In scenario 1, a false positive would allow a minor to access age-inappropriate content. In scenario 2, a false positive would allow an adult into a child’s safe place. In scenario 3, a false positive would place an individual into an age group that does not reflect their true age.

3.3 Primary Measure: Overall Accuracy

To simplify the proposed approach, our hypothesis is that a headline statement of overall accuracy of the age assurance measure could be provided, which could be more directly aligned to indicators of confidence. This could enable a quick, easy, and readily accessible indication of accuracy to be provided to an unfamiliar audience. When presented with other indicators, such as error rates, privacy and security controls, fairness measures and distribution of results, a holistic understanding of the effectiveness of the age assurance measure(s) overall could be established (a detailed explanation of effectiveness is provided in Section 3.8).

Given the focus on binary metrics, we propose that overall accuracy could be used as this primary measure. This is the proportion of correctly classified subjects by the technology and is a useful overall measure of performance. This short research project has focussed on one aspect – accuracy - and in our hypothesis, which has been developed with industry and specialist input and advice, we have explored what outcomes may be derived from aligning a headline measure of accuracy to the proposed indicators of confidence as follows:

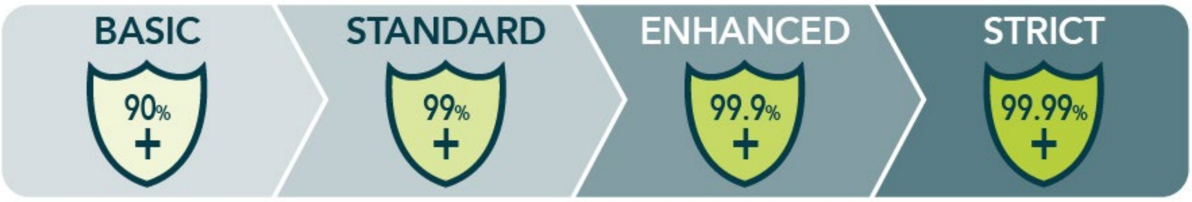


FIGURE 6 - ILLUSTRATIVE EXAMPLE - HOW OUR HYPOTHESIS COULD BE ALIGNED TO INDICATORS OF CONFIDENCE

The potential advantage of overall accuracy is that it is a simple metric that gives an overall measure of classification and is easy to understand for non-technical audiences.

There are two potential disadvantages that must be considered:

1. It does not provide any information on the type of errors that are present (false positives or false negatives).
2. It can be misleading if the test data are imbalanced. By this we mean that the number of test subjects below and above the age gate are not equal.

The first of these disadvantages is addressed by having a set of secondary measures that includes more detailed information on the types of errors (see more details below for the secondary measures).

The second is important to acknowledge and ensure that in any testing scenario the test data set is appropriately balanced. If, for a particular reason, this is not possible then we would suggest that the balanced accuracy is reported, which is calculated as follows:

$$Balanced\ Accuracy = \frac{1}{2}(TPR + TNR)$$

The balanced accuracy is the average of the true positive rate (TPR) and the true negative rate (TNR). It is the same as the accuracy if the test data set is balanced.

It is important to acknowledge that there are other metrics that can provide an overall measure of performance for a binary classifier, examples include:

- F-score
- Diagnostic Odds Ratio
- Phi Coefficient
- Mathew’s Correlation Coefficient
- Rand Score
- Cohen’s Kappa
- Youden’s J Statistics

Like all metrics, there are advantages and disadvantages to each. We have prioritised accuracy here given the importance of interpretation; many of the metrics above are more technical and therefore potentially harder to interpret and set tolerances against from a regulatory perspective.

3.4 Secondary Accuracy Measures

We propose that a set of secondary measures are reported in conjunction with the primary measure of accuracy. These secondary measures are presented to provide a more detailed picture of performance and are made available for those who wish to delve deeper and gain a better understanding of the risks and benefits of deploying the technology. They are important to include for transparency and because there may be deployment scenarios where, for instance, accuracy may be less important than the proportion of false positives, such as when age assurance technologies are being used to deliver challenge age decisions for the sale of alcohol.

The secondary measures are based on those that are recommended in Part 1[22]:

- True Positive Rate (TPR)
- False Positive Rate (FPR)
- Positive Predictive Value (PPV)

For age estimation technologies continuous measures of accuracy can also be included such as those recommended in Part 1:

- Mean Absolute Error (MAE)
- Standard Deviation of the AE (SD)

In Part 1, we identified outcome fairness or outcome error parity as a means of quantifiably assessing how a technology has implemented fairness throughout its design and implementation. Outcome fairness is assessed by ensuring the error rates are equitably distributed across different subgroups of the population. Consequently, outcome error parity should also be reported as a secondary measure and could include:

- Overall accuracy parity
- True positive parity
- False positive parity
- Positive prediction parity

It must be identified which protected characteristics are at risk of bias or discrimination and therefore error parity should be examined for these chosen characteristics. While it is relatively simple to examine protected characteristics individually, it is important to acknowledge the potential for intersectional biases where there are biases within combinations of protected characteristics (such as race and gender in combination).

3.5 Further Considerations of the Continuous Measures

Reflecting on the continuous measures for age estimation technologies and following feedback from the workshops, we acknowledge that these are complex and can be more challenging to understand for a non-technical audience. Focussing on the binary metrics not only reflects the reality of deployment, but also provides metrics that are often simpler to interpret.

[22] Other metrics associated with binary metrics are detailed in the first report which could also be included in this list (the list is not exhaustive).

The continuous metrics can, however, add an additional level of detail for a user who wishes to deploy an age estimation technology. Typically, when describing a continuous distribution, a measure of central tendency and dispersion are reported together. In Part 1, we recommended the mean absolute error (MAE) which is widely used throughout the industry and the standard deviation of the absolute errors (SD) but acknowledged that other measures are also available.

It is important to note that the distribution of the absolute errors is not symmetrical, and this can have implications on the metrics depending on how skewed that distribution is. As discussed in Part 1, alternative measures of central tendency and dispersion for when distributions are heavily skewed are:

- Median of the absolute errors
- Interquartile range of the absolute errors

Finally, a metric that is also widely used is the cumulative score (CS). The CS is the proportion of samples where the absolute error is less than a given number of years. It can be calculated for a range of different years; for example, 1 – 10 years to understand how the accuracy varies by size of error (and plotted as an error statistic curve with CS on the y-axis and size of error on the x-axis). Typically, the CS is used as a complementary measure to the MAE, for example.

The above highlights that there are many possible metrics that can be included and reported in the secondary measures. Which should be reported may depend on the deployment of the technology, the shape of the distribution of the absolute errors and what is of primary importance when testing.

3.6 Implications for Age Estimation Technologies

Following feedback from members of the STAC, it was identified that while this framework can be applied to age estimation technologies, it is unlikely that these types of technologies would currently be used in isolation to identify whether a person would pass or fail an age gate.

This becomes more apparent when considering subjects who are close in age to the age gate: for example, subjects who are between 17.5 and 18.5 years when the age gate is 18. It is unlikely that the accuracy of the technology to correctly classify these subjects is sufficient to pass with an enhanced or strict level of confidence. However, the confidence increases for those test subjects that are further away in age from the age gate.

Alternatively, the age estimation technology may be deployed within a challenge age setting. If a subject is identified as being under a given challenge age, they must go through a second age assurance step to verify that they are over the age gate. We refer to this multiple gateway approach as the waterfall technique and illustrate it in Figure 5 below.

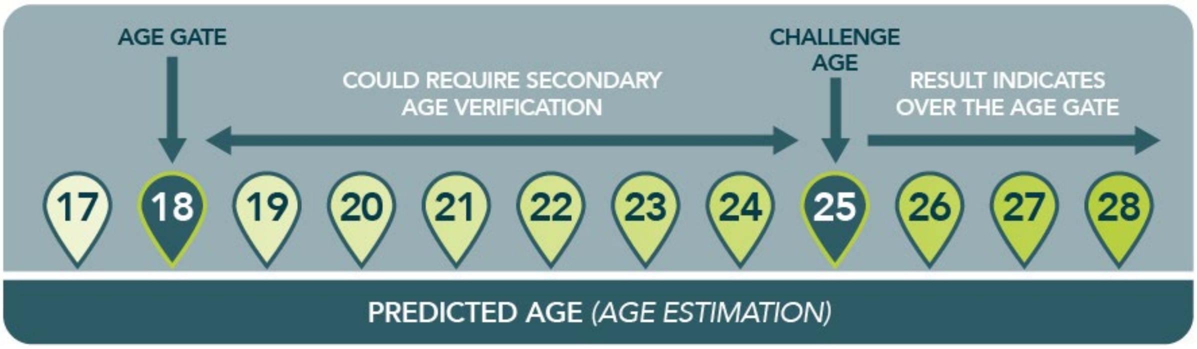

AGE GATE
COULD REQUIRE SECONDARY AGE VERIFICATION
CHALLENGE AGE
RESULT INDICATES OVER THE AGE GATE
17
18
19
20
21
22
23
24
25
26
27
28
PREDICTED AGE (AGE ESTIMATION)

FIGURE 5 - AGE GATES AND CHALLENGE AGE

In these instances, we suggest two approaches:

1. Evaluate the accuracy of the combined gateways rather than each individually. Where multiple gateways are relied upon to assess an age gate, the test could be set up such that it reflects the entire deployment including all gateways rather than each gateway individually. This would mean that the test would accurately reflect the true deployment and the overall accuracy of the combined gateways.
2. Provide an additional secondary measure based on a challenge age. If it is not possible to evaluate the accuracy of the full deployment when it relies on multiple gateways, an age estimation technology could also report an additional secondary measure based on a challenge age. This measure would declare the age at which an age estimation technology can accurately assess a person as being over a given age gate at a basic, standard, enhanced, and strict level. Additional measures highlighting, for example, the probability of being incorrectly classified as under the age gate (and therefore having to go through a second age assurance gateway despite being over the challenge age) can be included to assist in risk-based decisions. We note that this can have data privacy, intrusion and anonymity issues depending on the use it is deployed in.

3.7 Regulation and Tolerance Levels

A significant challenge for those seeking to understand and compare different age assurance systems is to have a simple indicator to derive a feel for 'what good looks like.' During the workshops and research, we explored a hypothesis of how this could be approached, although we note that it is not for this study to set the tolerance levels. The proposed hypothesis should not be seen as a pre-determined outcome and other bands of tolerance may emerge after further research, consultation, and engagement with wider stakeholder groups.

3.7.2 Secondary Measures of Accuracy

We acknowledge that setting tolerances for all secondary measures may not be practical. In discussions with STAC members and workshop attendees, two options were considered:

1. Reporting the actual figures of each measure; or
2. Defining a set of traffic light bands and reporting the secondary measures as red, amber, or green.

These options are not mutually exclusive, and both the actual figures could be reported alongside a traffic light banding. The benefit of a traffic light system is for ease of interpretation in identifying where metrics are good versus bad, but the limitation is that the banding would need to be updated over time to reflect the continual improvement of age assurance technologies otherwise they could become obsolete quickly.

We note that for the second option, defining appropriate bands for the standard deviation of the absolute errors (or any other measure of dispersion) is likely to be challenging as it is not well known what the range of absolute errors is across different technologies nor is it clear what an acceptable level of range may be.

3.8 Overall Effectiveness of the System

In Part 1 of our research, we highlighted a series of effectiveness measures, including outcome fairness, liveness detection and presentation attack. Although we have focussed on accuracy to explore further in Part 2, these other measures should not be overlooked when considering the overall effectiveness.

3.8.1 Outcome Fairness

Emphasis should be placed on minimising or eliminating the outcome error differences. Where that is not currently possible given the 'state-of-the-art' of technology, emphasis should be on determining what level of difference (if any) is acceptable.[23] We suggest that further discussions are needed to identify appropriate technical and organisational measures that seek to minimise or eliminate the outcome error differences.

3.8.2 Liveness Detection

It is important in the workflow of any online age assurance system to incorporate a process for seeking to determine that the user presenting the information is a genuine, live human being. This is a process known as liveness detection and ranges from simple processes, like CAPTCHA[24] whereby a human is prompted to solve a puzzle or identify features in photographs; through to advanced liveness detection against video injection attack or deepfake attacks. Approaches to testing liveness detection are set out in ISO/IEC 30107-1:2016 — Information technology — Biometric presentation attack detection — Part 1: Framework.

3.8.3 Age-Related Presentation Attack

An age-related presentation attack differs from broader presentation attacks and refers specifically to attempts by a user to spoof or fool systems about the age of the user. This could be through falsifying or altering the date of birth shown on an official document (to various levels of sophistication). It could be trying to use the identity of an older sibling to evade systems. It could be trying to work around or intercept the response token provided by an age assurance service provider and cause this to be falsified.

3.8.4 Privacy & Security Objectives

[23] The UK GDPR includes requirements on fairness. This ties into the 'state-of-the-art' provisions of Article 25 to allow for some degree of difference, but providers and relying parties should take into account the potential for bias and discrimination and aim for parity across sex and ethnicity. See https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/guidance-on-ai-and-data-protection/how-do-we-ensure-fairness-in-ai/what-about-fairness-bias-and-discrimination/

[24] Completely Automated Public Turing test to tell Computers and Humans Apart

The draft ISO/IEC 27566 — Age Assurance Systems — Framework[25], currently in development, lists a series of privacy and security objectives for the systems. These relate specifically to threats to an effective age assurance system and require further development and research, but include:

- Privacy Objectives
  - Unlinkability
  - Untraceability
  - Attributes minimisation
  - User consent
  - Transparency
- Security Objectives
  - linkage of the attributes to the legitimate individual,
  - detection of collusion attacks between individuals
  - prevention of an endless usage of evidence and
  - forwarding of a security token by an age assurance provider to another provider only if allowed.

3.9 Testing Data Sets

To evaluate accurately the deployment of an age assurance technology, a test data set must be used that is independent from the data set used to train it. During the workshops, it was highlighted that there were significant challenges with gathering, curating, and maintaining test data sets, particularly for the under-18 age group for which there are ethical and lawfulness considerations.

Some of the issues that need to be considered when compiling a test data set are set out briefly below.

Practical considerations include (but are not limited to):

- The ethics and lawfulness of collecting data for under 18s.
- Different test data sets are required for different technologies (e.g., facial vs. voice age estimation).
- Data sets should not just contain “perfect” quality images; they should be representative of lighting, camera quality and environmental considerations.
- Live vs. static data sets.

Statistical considerations include:

- Age range: The test data set must include a range of ages that will adequately evaluate the technology that is based on the Age Gate that is being assessed (and Challenge Age if relevant). For example, if an Age Gate of 18 is being evaluated, what is the lowest and highest age that a test data set needs to include? Test subjects in their 60s are unlikely to be relevant.
- Breakdown of test subjects: The test data set needs to be representative of age, gender, and skin tone. In many studies, probability sampling such as simple random sampling is used, which selects samples from a population based on the principle of randomisation. Since the sample are randomly selected then we can be confident that it is nationally representative. If this is not possible then non-probability sampling can

[25] https://www.iso.org/standard/80399.html

be used such as quote sampling. Care must be taken to ensure that subjects are chosen in such a way that the test data set is unbiased and representative of the population.

- Sample size of test data set: There are sample size formulae readily available that calculate the number of sample sizes required to accurately estimate an overall accuracy based on the estimated accuracy of the technology, the required margin of error and a confidence level. What is clear from some initial calculations is that the sample size required to estimate the accuracy of a technology at a strict or enhanced level (for example, accuracy of 99.9% or 99.99%) with any level of precision is very large (tens of thousands of subjects). This may necessitate a more in-depth review of evaluation assurance (discussed in section 3.10 below).

3.10 Depth of Testing and Analysis

The draft ISO/IEC PWI 7732 – Age Assurance Systems – Part 2: Measurement and Testing suggests adopting and adapting the Common Criteria Evaluation Methodology for the purposes of evaluating the effectiveness, security, and reliability of Age Assurance Systems. This is an existing, widely adopted, methodology.

The Common Criteria are set out in ISO/IEC 15408-1, ISO/IEC 15408-2 and ISO/IEC 15408-3 and the Common Evaluation Methodology is set out in ISO/IEC 18045.

3.10.1 Evaluation Assurance Levels

There are seven predefined Evaluation Assurance Levels (EAL1 to EAL7)[26] which correspond to increasing efforts for design verification and testing. The draft Age Assurance Systems standard suggests that they should be aligned as shown in Figure 7 below.

Evaluation Assurance Level (EAL)	Applicability to Age Assurance Systems
EAL1 Functionally tested	Equivalent for the testing of an age assurance component to a basic level of confidence
EAL2 Structurally tested	Equivalent for the testing of an age assurance component to a standard level of confidence
EAL3 Methodically tested and checked	Equivalent for the testing of an age assurance component to an enhanced level of confidence
EAL4 Methodically designed, tested and reviewed	Equivalent for the testing of an age assurance component to a strict level of confidence
EAL5 Semi-formally designed and tested	Not used in this standard
EAL6 Semi-formally verified, design and tested	Not used in this standard

[26] These are set out in ISO/IEC 15408-3:2020 - Information technology - Security techniques - Evaluation criteria for IT security.

EAL7 Formally verified, designed, and tested	Not used in this standard

FIGURE 7 - EXISTING PROPOSAL FOR EALS AND THEIR DESCRIPTION

We would suggest that the depth of testing and analysis be further reviewed. If we explore the detail of the assurance classes at the different evaluation levels, whilst level 1 may be sufficient for basic, we believe that level 3 should be aligned to standard, level 5 aligned to enhanced and level 7 aligned to strict. Thus levels 2, 4 and 6 would not be needed in this standard. This would amend the above table as follows:

Evaluation Assurance Level (EAL)	Applicability to Age Assurance Systems
EAL1 Functionally tested	Equivalent for the testing of an age assurance component to a basic level of confidence
EAL2 Structurally tested	Not used in this standard
EAL3 Methodically tested and checked	Equivalent for the testing of an age assurance component to a standard level of confidence
EAL4 Methodically designed, tested and reviewed	Not used in this standard
EAL5 Semi-formally designed and tested	Equivalent for the testing of an age assurance component to an enhanced level of confidence
EAL6 Semi-formally verified, design and tested	Not used in this standard
EAL7 Formally verified, designed, and tested	Equivalent for the testing of an age assurance component to a strict level of confidence

FIGURE 8 - AMENDED PROPOSAL FOR EALS AND THEIR DESCRIPTION

Our reason for this relates to the level and amount of data testing that would be required to undertake third party validation of data at levels 1 – 4. This will not be sufficient to assess the accuracy of an enhanced or strict indicator of confidence. Typically, testing at level 1 utilises around 30 presentations to the system; whereas, testing at levels 2 and 3 utilise 300 presentations to the system; and testing beyond those 3000 to 12,000 presentations. If there is a requirement to demonstrate accuracy to the hypothetical 'enhanced' and 'strict' indicators of confidence, then a suitably high number of presentations would be required to be able to demonstrate that.

3.10.2 Approach to Testing

The test deployment is adapted to the specific claimed capabilities of the age assurance service provider – including specifying the component(s); workflow and age gate(s). This is known as the Target of Evaluation – or ToE. The test process starts with gaining an understanding of the ToE and the platforms or services upon which the ToE operates (android, iOS, windows, etc). This can involve working through demonstration sites, test environments or deployed production environments depending on the maturity of the client's service deployment and state of readiness. It is not a pre-requisite that the client's product be available for end use by their clients or users.

The test level undertaken is selected by reference to the indicator of confidence sought (the higher the indicator of confidence, the more in-depth the testing that is required. This is set out in ISO/IEC 15408-3:2020 – Information technology – Security techniques – Evaluation criteria for IT security; in particular the Assurance Class 'Tests' – requiring the completion of the family test family ATE_IND[27].

The test deals with the degree to which there is independent functional testing of the ToE. This will be particularly necessary where the headline measure of accuracy claimed (for enhanced and strict indicators of confidence) is more than 99.9% - otherwise a test data set would need to contain millions of records and be expensive to administer and use. An appropriate mix of testing must be planned for each ToE, which considers the availability and coverage of test results and the functional complexity of the age assurance service provider.

ISO/IEC 15408-3:2020 (clause 14.4.4.2) requires the evaluator to devise and conduct tests with the objective that the ToE operates in accordance with its design representations including, but not limited to, the functional specification. The approach is to gain confidence in the correct operation through representative testing, rather than to conduct every possible test. The extent of testing to be planned for this purpose is a methodology issue and needs to be considered in the context of the particular ToE and the component(s) or workflows used.

[27] see ISO/IEC 15408-3:2020: clause 14.4.

4. Outcomes of the Study

This Part 2 of the technical study builds on the analysis of the approach to the technical measurement of age assurance technologies set out in Part 1 of the research. Testing those theories has both validated that they are appropriate measures, but also opened the opportunity to propose simplified approaches that may be more understandable and digestible to a non-technical audience.

We conclude that the two separate approaches to measurement set out in Part 1 of our research (measurement of binary and continuous outcomes) could be combined to provide a single ‘headline’ statement of accuracy, supported by measures identified in Part 1 being made available by providers for transparency. This would assist with understanding of the overall accuracy of the age assurance method, whilst also maintaining the statistical detail for those that need to know this for risk management decisions.

There is more work to be done on this. As the Online Safety Bill progresses through Parliament, the statutory definition of age assurance may evolve. Similarly, the work on ISO/IEC 27566 which contains a definition of age assurance may also evolve as the document heads through consultation and ballots to become an adopted international standard.

In our view, the output of age estimation approaches can be expressed as a binary ‘YES/NO this person is over/under the age of 13, 18, 21 or whatever the age of interest is’. Reducing age estimation outputs to binary is a truer reflection of how the technologies are deployed. However, this runs the risk of reducing the ability of age estimation as a tool for higher indicators of confidence.

Age estimation is more accurate the further away it gets from the age of interest - i.e., someone who is in their 50’s could be very accurately assessed as being over 18, whereas someone who is 19 may be less accurate. At or about the age of interest, it is inevitably the least accurate. We explored the introduction of a buffer or challenge age, whereby the accuracy can be stated if the results of those estimated to be younger than the buffer age are discarded (or diverted to an alternate age assurance process). In that case, age estimation could be deployed for use cases involving higher indicators of confidence.

We have built on this suggestion by looking at how the five indicators of confidence could provide a simple ‘accuracy rate’ and have these supported by secondary measures, which should be available for transparency.

In our view, if transparency measures were to be proposed, they should include (where relevant to the age assurance component in question):

- True Positive Rate (TPR)
- False Positive Rate (FPR)
- Positive Predictive Value (PPV)
- Mean Absolute Error (MAE)
- Standard Deviation of the AE (SD)
- Outcome Error Parity (OEP)

We have also explored the complexity of the continuous measures recommended in Part 1 and highlighted other measures that could also be reported in conjunction with the above where appropriate (such as cumulative score or the interquartile range of the absolute errors as an alternative measure of dispersion).

In practice, we understand that many age assurance service providers use multiple technologies to provide age checks. This was highlighted in Part 1 of our research when discussing waterfall techniques and various permutations and combinations of technologies. Where technologies use multiple gateways, we conclude that the test should be set up in such a way that it reflects the entire deployment or workflow rather than each gateway individually. This means that the test would reflect the overall accuracy of the complete workflow.

We have also explored issues around effectiveness of systems, or better described as, how prone they are to system-level attack, bias, or spoofing. We examined this in Part 1 of our technical study, and we conclude that there remains more work to be done on this issue.

The issue of appropriate testing data sets remain which includes both practical and statistical constraints.

5.Findings and Observations

The research brief asked for findings and observations identified from our research in the context of expectations for the measurement and testing of the accuracy of age assurance technologies.

Key Findings

1. Providing primary focus on measuring accuracy based on binary metrics (where the answer to a question such as ‘Is this user over 18?’ is ‘YES’ or ‘NO’) would more accurately reflect how the technologies will be deployed.

2. We consider that the overall accuracy of the age assurance component and/or series of components may be a suitable primary indicator of confidence because it is simple to interpret and provides an overall measure that can be readily understood.

3. Our research suggests that age assurance systems could effectively be assessed according to a stated age gate (i.e., ‘13’, ‘16’, ‘18’) representing the principal age of interest to the relying parties for a particular use case. In other words, it is important that the focus is on the age, or indeed the age range, at which a technology is being evaluated.

4. Some age assurance systems employ a workflow where initial age estimation processes are used to filter out individuals that are over a threshold (such as being over 25) before proceeding with secondary age assurance methods for those identified as under that threshold. Current observed practice is that these processes currently tend to move to an age verification approach for these cases, but that may change in the future.

5. Existing regulations and guidance require an objective assessment of the state-of-the-art of technical measures. In Part 2, we observe that based on (1) an approach to this objective assessment set out in guidelines published by the EU’s Agency for Cybersecurity (ENISA)[28], (2) our data gathering from participants in this research, and (3) our independent analysis and validation of that data, there are now a range of technologies in the market which could be described as ‘state-of-the-art’ available at each indictor of confidence.

Challenges and Complexities

6. Age assurance components that provide continuous outcomes (such as age estimation) are complex and as a result, we have identified some additional measures that could be used as alternatives to, or in addition to, the MAE and SD.

7. Where providers use multiple age assurance components for their system, the test might benefit from being set up to reflect the entire workflow in addition to, or instead of, each component individually. This means that the test would reflect the

[28] See Section 2.2

overall accuracy of the complete workflow, better reflecting the way that age assurance systems are used.

8. The approach to testing of age assurance systems would benefit from structured, ideally accredited and subject to availability of appropriate testing data sets, which may need to include biometric, demographically representative, and fairly distributed data. When considering age gates under 18, this could involve the processing of test data relating to children which creates a series of potential concerns, including practical, ethical, privacy and security concerns that may require further consideration.

Further Considerations

9. In our research, we used factors of 10 as hypothetical bands of tolerance for simple indicators of accuracy. These bands of tolerance could be aligned to the existing indicators of confidence in the draft ISO/IEC 27566 – Age Assurance Systems – Framework (Basic 90%+, Standard 99%+, Enhanced 99.9%+ and Strict 99.99%+ balanced accuracy). Identifying, consulting on, and adopting recognised tolerances for simple indicators of confidence could assist understanding of age assurance technologies.

10. Our findings suggest that, in addition to the statement of overall accuracy, a set of secondary measures to provide a holistic understanding of the performance of a technology for those needing to make risk-based judgements about the performance of the system should be provided. These should include the six metrics recommended in Part 1; namely Mean Absolute Error (MAE), Standard Deviation (SD), False Positive Rate (FPR), False Negative Rate (FNR), Positive Prediction Value (PPV) and Outcome Error Parity (OEP).

11. The project team consider that the measures of effectiveness described in Part 1 including measures that assess bias, liveness detection and presentation attack are important when making risk-based judgements for deployment of age assurance systems and this could be a desirable area for further research.

12. In Part 2, we have examined the issue of accuracy of age assurance systems. A further question arises, however, as to how often the age check should be deployed (i.e., every time a user visits, or periodically or just once) and how often a prior age assurance check of a user should be re-authenticated. This should be based on an analysis of risks and could usefully be subject to further research. This should not be confused with the overall measure of accuracy of the system – they are two distinct factors for consideration.

Appendix One – Regulator's Rationale for Research

A1.1 Ofcom

Ofcom is the regulator for the communications services that people use and rely on each day.

They regulate the TV, radio and video-on-demand sectors, video-sharing platforms, fixed line telecoms, mobiles, postal services, plus the airwaves over which devices operate. They make sure:

- People are able to use communications services, including broadband;
- A range of companies provide quality television and radio programmes that appeal to diverse audiences;
- Viewers and listeners are protected from harmful or offensive material on TV, radio and on-demand;
- People are protected from unfair treatment in programmes, and do not have their privacy invaded;
- The universal postal service covers all UK addresses six days a week, with standard pricing; and
- The radio spectrum is used in the most effective way.

In November 2020, Ofcom started regulating video-sharing platforms (VSPs) established in the UK. A service, or a dissociable section of a service, is a VSP if it is provided on a commercial basis, using an electronic communications network and the principal purpose of the service, or an essential functionality of it, is the provision of videos uploaded by users.

Ofcom is required to ensure that VSPs within the UK's jurisdiction take 'appropriate measures' in respect of videos that are available on their service to protect minors from content which may impair their physical, mental, or moral development.

One of their aims for Year 2 of their VSP regulation is to drive forward the implementation of robust age assurance to protect children from the most harmful online content, including pornography.

Ofcom is also due to become the Online Safety regulator for user-to-user, search, and online regulated pornography services.

In March 2022, the Government introduced the Online Safety Bill. The Bill, as initially presented to Parliament, included the following provision at clause 11(3), as an example:

"A duty to operate a service using proportionate systems and processes designed to—

(a) prevent children of any age from encountering, by means of the service, primary priority content that is harmful to children (for example, by using age verification, or another means of age assurance);

(b) protect children in age groups judged to be at risk of harm from other content that is harmful to children (or from a particular kind of such content) from encountering it by means of the service (for example, by using age assurance)."

The initial Bill also included a provision at clause 72(2), as follows, related to online services that host their own pornographic content (i.e., non-user-generated content):

"A duty to ensure that children are not normally able to encounter content that is regulated provider pornographic content in relation to the service (for example, by using age verification)."

A1.2 Information Commissioner's Office

The ICO is the UK's independent authority set up to uphold information rights in the public interest, promoting openness by public bodies and data privacy for individuals. The ICO has issued the Children's code (known formally as the Age appropriate design code), which articulates how online services should safeguard children's personal data.

The code sets out 15 interlinked standards that relevant organisations should conform with to ensure they comply with the UK's data protection laws. In the instance of an infringement of UK GDPR, the ICO can use its full range of enforcement powers against the organisations concerned, including fines of up to 4% of global turnover or £17.5 million, whichever is higher.

Standard three of the code focuses on age-appropriate application. It states that online services must have a level of certainty about the age of their child users, that is appropriate to the level of risks posed by the service to these users. "Risks" in this context refers to the potential negative impact on children's rights (as defined by the UN Convention on the Rights of the Child) that can arise from gathering and using their data. This includes children's rights to privacy, safety and wellbeing, physical and emotional development, access to information, and play.

The code states that organisations should either establish an appropriate level of certainty about the age of their users or apply the standards in the code to all their users.

As part of a wider package of external support to accompany the code, the ICO has developed an Opinion on the use of age assurance that sets out the Commissioner's current view of how age assurance methods can be used by organisations to conform with standard three of the Children's code. The ICO's Regulatory Policy Projects team is undertaking a project focused on age assurance. This will enable them to keep up with technological developments and deepen their understanding of how industry is responding to the code and the requirement on age assurance. It will also ensure that the guidance and support provided is relevant and helps the ICO to regulate effectively and fairly.

A1.3 The Digital Regulation Cooperation Forum (DRCF)

The ICO and Ofcom have worked together effectively for many years. Their collaboration has deepened since 2020, when the ICO launched the Children's code and Ofcom took on powers to regulate UK VSPs. In the same year, they co-founded the DRCF alongside the Competition and Markets Authority, with the Financial Conduct Authority joining subsequently.

In their joint statement (published 25 November 2022), they committed to ensure that their policies are consistent with each other's regulatory requirements and guidance and take into account each other's perspectives[29].

Developing an aligned approach to age assurance has been a priority for their joint work to protect children online. They have recently published joint research into families' attitudes towards age assurance[30]. They intend for this work to build on their commitment to achieve a better understanding of age assurance using their shared resources.

[29] Online safety and data protection: a joint statement by the ICO and Ofcom (publishing.service.gov.uk).

[30] https://www.ofcom.org.uk/about-ofcom/how-ofcom-is-run/organisations-we-work-with/drcf

Appendix Two - About the Age Check Certification Scheme

A2.1 Our Role

The Age Check Certification Scheme (ACCS) is an independent third-party conformity assessment service operated by AVID Certification Services Ltd and accredited by UKAS. The scheme is established to undertake standards-based assessments of age assurance services, digital identity services and age-appropriate design of information society services.

"We check that ID and age check systems work"

A2.2 UKAS Accreditation

AVID Certification Services is an accredited conformity assessment body under ISO/IEC 17065:2012 – Conformity assessment — Requirements for bodies certifying products, processes, and services. This is carried out in accordance with the Accreditation Regulations 2009[31] by the United Kingdom Accreditation Service (UKAS).

UKAS is recognised by Government to assess, against nationally and internationally agreed standards, organisations that provide conformity assessment services such as certification, testing, inspection, calibration, and verification.

Accreditation by UKAS demonstrates the competence, impartiality, and performance capability of these evaluators. In short, UKAS 'checks the checkers.'

The Schedule of Accreditation for our ACCS services is available on the UKAS website.

A2.3 ICO Approval

The criteria that AVID Certification Services use for the assessment of data protection and privacy of identity and age assurance services (ACCS 2:2021[32]); and for the assessment of the age appropriate design of information society services (ACCS 3:2021[33]), have been approved by the ICO.

To be approved, the certification criteria must be:

- Derived from UK GDPR principles and rules, as relevant to the scope of certification, i.e.:
    - Lawfulness of processing (Art 6-10)
    - Principles of data processing (Art 5)
    - Data subjects' rights (Art 12-23)
    - General obligations of controllers and processors (Chapter IV)
    - Obligation to notify data breaches (Art 33)

[31] SI 2009:3155 - https://www.legislation.gov.uk/uksi/2009/3155/contents/made
[32] ACCS 2: 2021 - Technical Requirements for Data Protection and Privacy
[33] ACCS 3: 2021 - Technical Requirements for Age appropriate Design for Information Society Services

    - Obligation of DP by design and default (Art 25)
    - Whether a DPIA has been completed where required (Art35 – 36)
    - Technical and organisational measures put in place to ensure security (Art 32)
    - International transfers (Chapter V);
- Formulated in such a way that they are clear and allow practical application;
- Auditable (i.e., specify objectives and how they can be achieved to demonstrate compliance);
- Relevant to the target audience;
- Inter-operable with other standards, for example ISO standards; and
- Scalable for application to different size or type of organisations.

The approval process is a formal function of the Commissioner exercising their tasks and powers under Articles 57 (1)(n) and 58 (3)(f) pursuant to Article 42(5) of the UK General Data Protection Regulation[34].

The Record of Approval of our ACCS certification criteria is available on the ICO website.

A2.4 ACCS 1:2020 – Technical Requirements for Age Estimation Technologies

The Age Check Certification Scheme has established a set of technical requirements for the assessment of age estimation technologies. This was a global first and without precedent and so we have not taken ACCS 1 as the underlying basis of our approach in this Part 2 of the technical study, although we have referred to some of the techniques in ACCS 1 as part of this research. We have challenged the original thinking that lay behind ACCS 1 with a wider overview of approaches to measurement and analysis of age assurance technologies.

ACCS 1 is based on testing the hypothesis of whether the age estimation technology is fit for deployment for a given challenge age category. For example, a Challenge 25 category means that anyone younger than 25 should be challenged for proof of age to ensure that they are over 18.

The technical requirements envisage that age estimation technology is rapidly advancing, and accuracy levels are always improving. In setting requirements around accuracy levels, these are assessed on the basis that technology is fit and safe to be deployed for the minimum ‘challenge age’ which has been identified. So, for instance, a particular age estimation technology may ‘pass’ and be certified as fit for use at ‘Challenge 25’ or ‘Challenge 28’ or indeed any other age.

It is worth noting that the applicable tolerance levels are much wider for the older the challenge age, so it is intended that users, seeking to commission this type of technology as a part of their age verification processes, can have greater confidence in those certified with a lower challenge age category.

[34] UK GDPR is implemented in the United Kingdom by the Data Protection Act 2018 as amended by various provisions to implement the European Union (Withdrawal) Act 2018

The methodology used to assess the accuracy of the technology has been developed in conjunction with Chartered Statisticians and considered by regulators, trade bodies and interested parties as an appropriate methodology.

A2.5 ACCS 4:2020 – Technical Requirements for Age Check Systems

ACCS 4 relates to the technical implementation of what is, at present, the only actually adopted and published standard for age check systems: *PAS 1296:2018 - Online age checking - Provision and use of online age check services - Code of Practice*[35].

PAS 1296:2018 provides a code of practice for age check providers, age exchanges or relying parties who undertake age check processes. As a code of practice, it does not set requirements, but does provide for organisations to make claims of conformity including through independent 3rd party validation of age check systems. To do that, it is necessary for the conformity assessment body to set out the technical requirements that it will apply, using PAS 1296:2018 as a framework, to assess whether, or not, to issue a certificate of conformity.

ACCS 4 aims to achieve the following:

- To validate and certify tools to help prevent harm to children and nuisance caused by young people from access to age-restricted content, goods, and services;
- To improve the quality, consistency and performance of age verification systems and procedures both online and offline;
- To provide consumers, purchasers, specifiers, regulators, law enforcement authorities, content providers, service providers and goods retailers with the assurance for them to identify suitable companies for conducting age verification;
- To help companies and individuals to demonstrate that their services or products meet an appropriate standard;
- To enable companies to demonstrate compliance with UK GDPR of processing operations by controllers and processors; and
- To mitigate the risks of non-compliance with age-restricted content, goods or services legislation including mitigating the risks of:
    - Criminal or disciplinary sanctions;
    - Civil or criminal action against the business and individual staff;
    - Damage to reputation leading to a loss of business; and
    - Licensing action, conditions or restrictions imposed by Licensing Authorities.

[35] https://shop.bsigroup.com/products/online-age-checking-provision-and-use-of-online-age-check-services-code-of-practice/standard

Appendix Three – Research Workshops

A3.1 Workshop 1 – Exploring the Key Issues

Our research for Part 2 included hosting two workshops. Workshop 1 focussed on testing the thinking and recommendations contained in Part 1. It explored eight particular elements, guided by the STAC[36], but also sought general views and feedback:

- Measures of accuracy
- Testing data sets
- Assessing bias
- Comparability
- Presentation attack
- Permutations and combinations
- Tolerances
- Continuous vs Binary Age Assurance Approaches

Other than staff from the ICO, Ofcom and ACCS, participants of the workshops were selected through registering their interest via a webpage we established to promote the workshops. A few individuals also expressed interest in joining the workshops after an event for the age assurance industry in January 2023 which focused on Part 1 of our technical study.

In Workshop 1, 21 individuals participated, including five representatives from the age assurance industry, an individual from academia and a representative from the cyber security sector. Workshop 2 involved largely the same people, with a few new additions. 22 people took part, with three individuals from organisations not present at Workshop 1. These were organisations who offered biometrics and identity technology services, a digital identity company and an anti-fraud identity service. The workshops also included STAC members.

It is important to note that the following record in this Appendix reflects the opinions, data and information collated from participants in the workshops. Where relevant and within scope, it has been captured in the body of this report.

A3.1.1 Measures of Accuracy

In Part 1 we set out various measures that can be used to assess the accuracy of both age estimation and age verification technologies. Part 1 recommended the use of mean absolute error (MAE) and standard deviation (SD) for age estimation; and false positive rate (FPR), true positive rate (TPR) and positive predictive value (PPV) for age verification to measure the effectiveness of age assurance systems.

In the workshop, we asked participants:

- Are the measures sufficient?
- Are there other measures that should be considered?
- Are there practical considerations for producing these estimates?

[36] Scientific and Technical Advisory Cell, see page 11.

The participants indicated that the measures were appropriate, but complex and difficult for non-statistical professionals to understand. They highlighted a need to build an overall picture of the error distribution, to include:

- The approximate centre of the distribution[37];
- The spread of results;
- The symmetry of results; and
- Addressing outliers.

Participants indicated a need to compare with real world scenarios as far as possible, such as the analogous performance of human beings in gaining age assurance against the performance of technology.

Participants identified a need to assess overall confidence (including spoof-ability known formally as 'presentation attack,' proxy ID use, etc.).

They felt these measures needed to deliver confidence in the technology, particularly in advancing the performance of either humans or technologies to better protect children.

A3.1.2 Testing Data Sets

To assess the accuracy of age assurance technologies, a test data set must be used that is independent from the data set used to train the technology (otherwise the accuracy may be artificially inflated) and verified to ensure that it is robust.

In the workshop, we asked participants:

- What are the main challenges for obtaining a test data set?
- How do we ensure a test data set is representative of what we see in the wild?
- Should there be one independent test data set that can be applied across technologies?
- How do we gather data now to understand current levels of accuracy given there is currently no one single test data set to use?

Participants indicated that there were significant challenges with gathering, curating, maintaining, and providing test data sets. There is a lack of suitably age-labelled and ethically obtained data sets of sufficient quality, particularly for under-18 age groups. The ethics and lawfulness of gathering such biometric and personal data about children for testing purposes needs to be addressed.

Concerns were raised about some current data sets involving 'data scraping' the internet, being automatically labelled, or using self-reported age mugshots. Participants highlighted the importance of data sets being appropriately 'ground-truthed'[38].

Having neat, 'square-on'[39], perfectly lit images with plain neutral backgrounds and no eye wear or factors obscuring the image is not representative of what we see "in the wild" (or in

[37] Also known as the central tendency.
[38] 'Ground-truthed' is a term used to describe the level of knowledge of the true identity, age, characteristics, and other meta data associated with the data set.
[39] 'Square-on' refers to a forward-facing image with 0% pitch, yaw and roll of the image.

real usage). The quality of images is usually more varied, and a test data set would need to reflect this.

The data set conditions should be aligned and should not be comprised of “perfect” quality images. They need to be representative of not only population characteristics, but also of various lighting, camera quality and environmental considerations.

Participants felt that the use of an independent test data set (or sets) could prevent leakage from the original data set. However, different age assurance technologies have different requirements e.g., video, 3D solution, static image, which presents another challenge for obtaining data sets.

Ideally, there should be a data set, held independently (by a regulator or conformity assessment body) that can be set aside for comparison. Currently there appears to be just one independent certification laboratory globally (the Age Check Certification Scheme).

The issues surrounding a reliable, available, suitably curated range of data sets was a recurring theme of the workshops and STAC discussions. This is something that requires further consideration, research, and development.

A3.1.3 Assessing Bias

In Part 1 we set out how we approached the issue of assessing biases in terms of equality of errors between protected characteristics[40].

In the workshops, we asked participants:

- What are the main potential causes for biases in age assurance technology?
- Are there other established measures of biases that we should consider?
- How do we assess the potential for other biases that are not defined by a person’s protected characteristics?
- Should a provider be analysing results of outliers with particularly large errors?

Participants highlighted the lack of sufficient data of sufficient quantity and quality to provide suitable analysis of bias. This is due to the lack of representativeness of data and unequal access to hard identifiers associated with the data sets that are available. Certain socio-economic demographics can be underrepresented, and people may feel uncomfortable interacting with certain technologies for testing and analysing bias. There was a considerable need to build trust and confidence in the technologies and unintended or undisclosed uses of data.

In addition to the protected characteristics, participants felt that bias could be introduced by:

- Neurodiversity – how might someone answer a question in different ways?
- Education – how might people struggle to interact with technology?

[40] Protected characteristics in the UK are set out in S.4 of the Equality Act 2010 and include: age, disability, gender reassignment, marriage and civil partnership, pregnancy and maternity, race, religion or belief, sex, sexual orientation. In the context of age assurance, the most relevant characteristics are gender and skin tone (race).

Participants indicated that there should be expectations for audit of how technology is trained and how often it is reviewed. There should be an approach to continuously validate and assess the technology.

Although some bias is inherited from the process (such as the camera quality or properties of light) there can be bias due to application set up e.g., model trained for indoor and tested on outdoor. In addition, the balance between recall and sensitivity will differ between products based on the use case/range they are trained to look for.

A3.1.4 Comparability

Part 2 is seeking to explore issues of comparability between age assurance components and systems. Being able to compare the accuracy results between age assurance systems is critical for regulators and service providers. The ability to compare the results will allow:

- Regulators to understand the baseline or current levels of accuracy;
- Consumers, users, and campaigners to understand the relative performance of different systems; and
- Relying parties and service providers to compare the accuracy between systems.

In the workshop, we asked participants:

- What are the key considerations for ensuring the measures of accuracy are comparable?
- How can we ensure that the testing protocols are set to ensure comparability?
- How do we ensure consistent statistical analysis of the results between age assurance providers?

Participants indicated that performance of systems needed to be considered at different age groups. The accuracies can be computed and averaged, providing a simplified indicator of confidence.

Models that predict as 18+, 18- etc should be assumed to be categorical (i.e., a binary output) even if the method utilised to reach that conclusion used estimation (i.e., a continuous measure input). Comparability should be focused on outcomes, rather than process, ensuring that the context of the use case is of paramount importance. This is important because a more nuanced approach (rather than a binary approach) would be needed if a platform were trying to determine age ranges of users to ensure their service considers the needs of different users.

Participants suggested that since age assurance providers have different approaches, it will be ideal to work out how to do an “apple for apple” comparison. If provider A has a binary model and the other has continuous, critically it is too hard to compare the two unless they are converted to one format.

To achieve this, we need to consider:

- Using the same (good) dataset for each approach;
- Defining a specific testing procedure for all providers;
- Examining if there could be a certification approach for both continuous and categorical outcomes; and
- Whether to convert all outputs to categorical (binary) to compare the overall accuracy.

Participants agreed it was of fundamental importance to establish a clear achievable baseline and noted that comparability of technology in a laboratory setting could be used as a proxy to comparability in real life situations.

A3.1.5 Presentation Attack

Presentation attack detection is the process of determining if an age assurance technology is susceptible to being "spoofed". It will be necessary to continuously review and address threats which will become prevalent and easily accessible to young people seeking to circumvent age assurance systems.

In the workshops, we asked participants:

- What are the main types of attacks that age assurance systems are likely to be presented with?
- How does an age assurance system assess its vulnerability to these types of attacks?

Participants highlighted that, at present, international standards only covered presentation attack through impersonation – such as with photos, masks, videos. This needed to be extended to cover the use of documentary presentation; synthetic identities; and the emerging availability of deep fake and video injection attack.

In considering the vulnerability of age assurance systems to attack, the testing and certification needs to consider and report on:

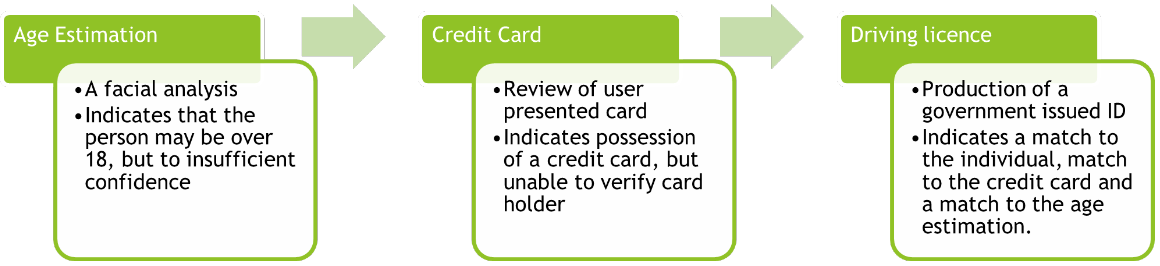
- The motivation or purpose of the attack;
- Incidents vs trends in attacks, particularly age assurance specific attacks;
- Threat modelling linked to risk analysis;
- Presenting the solution differently to users (A/B), to see if users are more willing to engage if age assurance is presented differently; and
- How many potential points of failure there are.

Participants felt that presentation attack detection performance was an important secondary measure of the performance and accuracy of the age assurance system.

A3.1.6 Permutations and Combinations

Some age assurance systems rely on building multiple sources of age assurance, sometimes with different steps before an age is verified. In Part 1, this was described in Section 6.2 and 6.3. An approach identified was to use a 'waterfall method' where the cumulative results of the age assurance components are greater than the individual results of each component on its own.

An example of this method is shown in the following chart (taken from Part 1):



Age Estimation

- A facial analysis
- Indicates that the person may be over 18, but to insufficient confidence

Credit Card

- Review of user presented card
- Indicates possession of a credit card, but unable to verify card holder

Driving licence

- Production of a government issued ID
- Indicates a match to the individual, match to the credit card and a match to the age estimation.

In the workshop, participants were asked:

- How common is it for users to go through permutations and combinations of age assurance?
- What are the challenges for assessing the accuracy where multiple sources of age assurance are used?

Participants said that it is common to have a waterfall approach, which reflects real life human age assessment, i.e., checking if someone looks over 25 and if not requiring the presentation of ID to prove they are over 18[41].

Participants noted that different indicators of confidence are attached to each age assurance method and different scenarios require different confidence levels. So, when considering deployment of age assurance, aligning these can be a challenge depending on the use case.

This would require quantifying the error, determining a ‘sweet spot’ of how many sources are being used, and ensuring the criteria for assessing the accuracy of each method does not exclude certain groups e.g., elderly people or minority groups. There are two aspects to this:

- Inclusivity – considering whether the verification bar is set at a level that would prevent some people from hitting it due to lack of official documents (i.e., children).
- Privacy intrusion - more generally, considering whether the requirement for multiple methods of verification/authentication be justified by the use case. This may be more relevant to a scenario looking at differing forms of age estimation.

To explore a simplified approach or headline measure, it is important to consider how errors may scale when moving away from the target age. Having too many age gates or permutations could result in an unfriendly user experience. Providers should also guard against the collection of unnecessary data that could end up being harvested for other purposes.

A3.1.7 Tolerances

The comparison of different age assurance systems to derive a feel for ‘what good looks like’ may be challenging. Although the outputs of the system can be measured, the outcomes are not realised unless set in the context of appropriate tolerances. Whilst the level of acceptable tolerance is a policy question that would be set by the regulators, we explored in our research how to define tolerances.

[41] This is commonly known as ‘Challenge’ or ‘Think’ 25 and is adopted as widespread best practice in physical face-to-face transactions where age assurance is necessary, such as the sale of alcohol, tobacco, weapons, or gambling.

In the workshop, participants were asked:

- Should there be different levels of tolerances applicable to age assurance?
- How should tolerances be set such that they can be reviewed and updated over time as technologies improve?
- What are the important aspects to be included in tolerances?

There was strong consensus between participants that there needed to be different levels of tolerances applicable to different use cases for age assurances. Participants felt that there should be at least three and not more than four distinct categories to provide an appropriate level of choice, although no granular detail of those was provided before we set forward our hypothesis relating to four indicators.

Participants felt that in some use cases, very hard boundaries are set, with criminal sanctions, and that they may be challenging for looser age assurance measures to achieve unless as a part of a waterfall methodology.

The tolerances need to be based on clear principles with clear communication and setting the right expectations. Whilst there is a clear attraction by a simple 'Basic – Standard – Enhanced – Strict' approach as suggested in the first technical study, it was also important that more granular detail was retained for those that needed to understand it.

The tolerances should be risk/harm related – so tolerance levels could be more stringent around significant age thresholds (i.e., tighter around 18 than 60). The point was made that a headline figure can only maintain confidence if the underlying performance were known/transparent and available to scrutiny. As expressed before also, any threshold is a proxy for accuracy which might change should the 'State-of-the-Art' change over time.

There is a need to understand that there may be a trade-off between inclusivity and tighter tolerance levels e.g., if tolerance levels are strict, over 18s from certain groups may be discriminated against (on basis of access to identity documentation, etc).

Participants indicated that tolerances could be linked to the level of access granted/ risk of service/ aspect of service being accessed.

It was important that tolerances were explained, not just arbitrarily imposed, but participants felt that there was a need to start somewhere, and they should be kept under continuous review to reflect technological advancement.

A3.1.8 Continuous vs Binary Age Assurance Technologies

In Part 1 we set out two types of measures:

- Measures that assessed the accuracy of age assurance technologies that provide **continuous estimates** of a person's age (age estimation); and
- Those measures that assessed the age assurance technologies that provide **binary estimates** (age verification).

In the workshop, participants were asked:

- For the purposes of regulation, do we require both continuous and binary measures?
- Will all technologies result in a simple pass/fail (even if the underlying input is initially an estimate of a person's age)?

- If we only assessed the binary measures, could we be missing important information/insights about the age estimation technology?

Participants indicated that continuous measures are confusing as there are many different measures and the results require some knowledge of statistical analysis. They felt that there were no practical use cases in the marketplace for continuous measures other than as a base threshold – clients only really care about the challenge age (i.e., the higher age than the age restriction that may be used to challenge for hard identifiers).

The use of mean absolute error does allow an assessment of confidence for a given age threshold and does aid the ability to explain the technology to clients.

If we were to convert the output of a continuous measure into a binary answer, then that could further aid clarity. There is a risk that this oversimplifies measures relating to bias or deviation but is likely to be more instructive. There is also a risk that age assurance service providers with certification at a particular age gate (such as 18) may be tempted to try to sell their product at other age gates without the necessary certification for them.

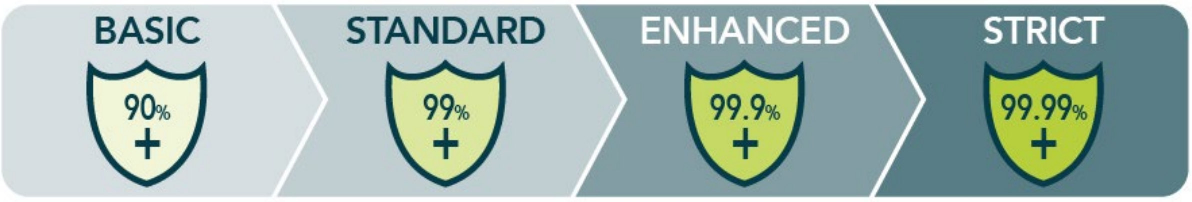
A3.2 Workshop 2 – Testing the Hypothesis of a Simplified Approach

In Workshop 1, we explored the key issues related to the conclusions and recommendations of Part 1. This tended to indicate that, whilst the measures identified were appropriate, important and ought to be transparently available for each solution, they were not easily understandable.

Based on guidance from the STAC, we took the opportunity in Workshop 2 to explore the possibility and consequences of a simplified approach by presenting a hypothesis.

A3.2.1 Classification Error Rates as a Hypothesis

If there is a stated age gate (e.g., 18), it is important to consider the implications of assessing and describing the accuracy of the age assurance method simply by the classification error rate.



To put this in other terms, out of every one million checks done by an age assurance system, how often would they be likely to get the answer right:

- At least 900,000 in a million – would provide a basic indicator of confidence;
- At least 990,000 in a million – would provide a standard indicator of confidence;
- At least 999,000 in a million – would provide an enhanced indicator of confidence; and
- At least 999,900 in a million – would provide a strict indicator of confidence.

It is worth noting that it was accepted that a zero level of error is neither a practical, nor a statistically valid outcome to be sought.

Participants were led through a series of exercises to explore if:

- The simplification made sense;
- Whether it resulted in a loss of granularity or important measures;
- Would result in comparable age assurance products across the market;
- Was standardisable and, therefore, certifiable; and
- Could be easily understood by regulators, campaigners and interested parties.

There was broad support for the hypothesis put forward, but participants also felt that we may need to retain some of the granular detail for those that may need to access it.

Having reviewed proposals in draft international standards, participants were concerned that the approach was too complex, with too many variables and being too open to abuse and gaming the standards.

Participants indicated that overall, the ‘headline measure’ would be a useful tool to help simplify and communicate the relative accuracy of different age assurance systems. However, they also felt that the measures identified in Part 1 ought to still be identified and reported on test certificates for complete transparency.

Appendix Four - Literature Review

A4.1 Introduction

Our research included identification and analysis of scholarly articles from open-source research. This literature review was intended to explore the applied statistical theory relating to binary and continuous approaches to measurement.

To quantitatively characterise the performance of age assurance methods, several evaluation metrics have been used within the literature. Broadly speaking these can be split into two categories: metrics that quantify a discrete variable, and metrics that quantify a continuous variable. A discrete variable is one where there is a finite number of values that it can take on. For example, in the case of binary classification the discrete variable can either be assigned to be 0 or 1. With regards to the problem of age estimation, a discrete variable will be used when classifying a datapoint as belonging to a predetermined age group, such as belonging to under 18. This classification task is sometimes referred to as classification age encoding (CAE) [1]. Continuous variables on the other hand can take a range of values. For example, in this context a continuous variable will be used to quantify the estimated age in years, also known as real-value age encoding (RVE).

A4.2 Scholarly Sources on Age Verification

A 2022 study analysed[42] age verification techniques employed by various popular online applications, such as WhatsApp, Instagram, and TikTok. Of the ten applications covered within the study, none employed any age verification outside of asking the user to input their age. Whilst the sample size of the study is small, it highlights the current lack of deployment of any meaningful age verification systems in social media applications.

Another study conducted by the Oxford Internet Institute[43] set out to explore lessons learnt by the online gambling industry with regards the successful application of age verification processes. The research sought to, amongst other things, understand the rationale for use or non-use of age verification across three case studies: online gambling, the online sale of age-restricted goods, and social gaming. The study came to several conclusions, including the proposal that children should not be age-gated at every step, as the recommendations are intended to strengthen existing regulatory frameworks limiting access to age-restricted goods, rather than to create new barriers, firm in the view that there is great value in free exploration of the Internet. Further to this, businesses across these different sectors were found to apply a range of age verification methods that afford various levels of assurance and are subject to differing levels of enforcement.

[42] L. Pasquale, P. Zippo, C. Curley, B. O'Neill, and M. Mongiello, "Digital Age of Consent and Age Verification: Can They Protect Children?" in IEEE Software, vol. 39, no. 3, pp. 50-57, May-June 2022, doi: 10.1109/MS.2020.3044872.

[43] V. Nash, R. O'Connell, B. Zevenbergen, and A. Mishkin, "Effective age verification techniques: Lessons to be learnt from the online gambling industry", University of Oxford, 2012.

A4.3 Scholarly Sources on Age Estimation

Mean absolute error (MAE) is the most common metric used for quantifying representative volume element (RVE)[44]. Given an age estimation x_i on the ith sample of a dataset of size N, and the corresponding true age y_i of that sample, MAE represents the average absolute error over the dataset.

$$\mathrm{MAE} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i=1}}^{\mathrm{N}} |x_i - y_i|$$

As this measure averages over an entire dataset, it gives an overall picture of the level of error present with a particular method. This averaging poses a potential drawback when applied to age assurance technologies, however. The error of an age estimation technology may be dependent on the age of the participant, which is to say: a technology may be more likely to under or over-estimate the age of an individual if they are very young or very old. An example of this is human-to-human age estimation, where several studies found that the human prediction error in both facial and voice prompts was positively correlated with the true age of the participant[45].

MAE can be extended to highlight the variations within age ranges by averaging over an age range basis as in the equation below.

$$\mathrm{MAE_k} = \frac{1}{n_k} \sum_{i=1}^{N_k} |x_i - y_i|$$

Here n_k represents the number of participants within a range in a given dataset. This can be a broad range, or even individual ages, which would give a series of MAE per age.

Cumulative score (CS)[46], as described by Yan Fu and Huang is another means of quantifying the accuracy of a continuous age assurance measure. Here $n_{e_i \leq \theta}$ is the number of predictions for which the age estimation has an absolute error no higher than θ years. n is therefore the total number of participants within the dataset.

$$\mathrm{CS}(\theta) = \frac{n_{e_i \leq \theta}}{n}$$

[44] Vincenzo Carletti, Antonio Greco, Gennaro Percannella, and Mario Vento. “Age from faces in the deep learning revolution.” In: IEEE transactions on pattern analysis and machine intelligence 42.9 (2019), pp. 2113-2132.

[45] Evelyne Moyse. “Age estimation from faces and voices: A review.” In: Psychologica Belgica 54.3 (2014).

[46] Yun Fu and Thomas S Huang. “Human age estimation with regression on discriminative aging manifold.” In: IEEE Transactions on Multimedia 10.4 (2008), pp. 578-584.

Quantifying age estimation error when there is no known ground truth is also a critical area of research[47] [48]. For this scenario, an apparent error ϵ_i can be used.

$$\epsilon_i = 1 - e^{-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}}$$

Here the age labels are provided by asking humans to estimate the age of participants within the dataset. The mean μ_i and variance σ^2, calculated from the distribution of guesses for participant i is then used. The absolute value of ϵ_i can be averaged across the participants within the dataset to form an apparent MAE.

For discrete variables, commonly used evaluation metrics within age group classification are accuracy, top-1 accuracy, and top-5 accuracy.

$$\text{accuracy} = \frac{\text{correct classifications}}{\text{all classifications}}$$

Classification models will often output a list of classes, e.g., age ranges, and a corresponding list of calculated probabilities that a given input will belong to a given class. Top-1 accuracy refers to the scenario where the class estimated as having the highest probability is the correct class. Top-5 refers to the scenario where the correct class is contained within the top 5 most probable classes that an input belongs to. Other noteworthy metrics are precision, true positive rate, false positive rate, and combinations thereof[49].

A4.4 Facial Age Estimation

In the advent of recent advances in deep learning research, facial image modalities are the most popular area of research for age estimation. Carletti et al.[50] presents the most comprehensive review on this topic at the time of writing. This review compiles results from 31 different studies, in 7 different publicly available datasets, evaluated on MAE between 2013 and 2019. MAE across the different studies ranged from 7.41 to 2.81 years, however the results were dependent on the composition of the datasets they were tested against. For example, in MORPH-II[51], a dataset designed for minimal facial pose variations, the best performing model[52] had an MAE of 2.81 years. Conversely in FACES[53], a dataset designed for high facial pose

[47] Sergio Escalera, Junior Fabian, Pablo Pardo, Xavier Baró, Jordi Gonzalez, Hugo J Escalante, Dusan Misevic, Ulrich Steiner, and Isabelle Guyon. “Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results”. In: Proceedings of the IEEE international conference on computer vision workshops. 2015, pp. 1-9.

[48] Sergio Escalera, Mercedes Torres Torres, Brais Martinez, Xavier Baró, Hugo Jair Escalante, Isabelle Guyon, Georgios Tzimiropoulos, Ciprian Corneou, Marc Oliu, Mohammad Ali Bagheri, et al. “Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016”. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2016, pp. 1-8.

[49] Marina Sokolova and Guy Lapalme. “A systematic analysis of performance measures for classification tasks.” In: Information processing & management 45.4 (2009), pp. 427–437.

[50] See footnote 1 above [Carletti et al.]

[51] Karl Ricanek and Tamirat Tesafaye. “Morph: A longitudinal image database of normal adult age-progression.” In: 7th international conference on automatic face and gesture recognition (FGR06). IEEE. 2006, pp. 341-345.

[52] Shahram Taheri and Önsen Toygar. “On the use of DAG-CNN architecture for age estimation with multi-stage features fusion.” In: Neurocomputing 329 (2019), pp. 300–310.

[53] Natalie C Ebner, Michaela Riediger, and Ulman Lindenberger. “FACES—A database of facial

variation, the best performing model in this dataset[54] had an MAE of 3.82 years. It is worth noting that there is a significant lack of published data for 'in the wild' analyses, where challenging and heterogeneous variations of pose, illumination, and image quality are present. Additionally, current state of the art facial age estimation technologies will likely have lower MAE ranges compared to the studies evaluated in Carletti et al. as improvements have been made in the years following its publication.

Carletti et al. also looked at 14 studies evaluated using a top-1 group classification accuracy. Here the models were evaluated on a dataset[55] divided into seven age categories: 0-2, 3-7, 8-12, 13-19, 20-36, 37-65, and 66+. The best performing model[56] had an accuracy of 67.3% across the groups, whilst the worst yielded 45.1%. Whilst misclassification was at best above 30%, the dataset was constructed with a high degree of variation in lighting conditions, pose, and image quality.

A less comprehensive but more recent review published in 2020[57] compiled results from 18 different models. Both 'handcrafted' and deep learning models were included, and on average the deep learning models well outperformed the 'handcrafted' models. While there was significant overlap with the studies found within Carletti et al., the age estimation MAE ranged from 7.04 to 2.16 years within the MORPH-II dataset.

Another noteworthy review published in 2018[58] predominantly focused on more classical machine learning approaches to perform facial age estimation. This review included methods from 51 different age estimation studies, spanning from the years 1999 to 2016. Performance across the studies varies greatly, in part due to the inclusion of studies that only tested on private databases. The models that were validated against public datasets, such as FG-NET[59], had relatively moderate MAE's, averaging around 5 years.

expressions in young, middle-aged, and older women and men: Development and validation." In: Behavior research methods 42 (2010), pp. 351–362.

[54] Hao Liu, Jiwen Lu, Jianjiang Feng, and Jie Zhou. "Label-sensitive deep metric learning for facial age estimation." In: IEEE Transactions on Information Forensics and Security 13.2 (2017), pp. 292-305.

[55] Eran Eidinger, Roee Enbar, and Tal Hassner. "Age and gender estimation of unfiltered faces." In: IEEE Transactions on information forensics and security 9.12 (2014), pp. 2170–2179.

[56] Ke Zhang, Ce Gao, Liru Guo, Miao Sun, Xingfang Yuan, Tony X Han, Zhenbing Zhao, and Baogang Li. "Age group and gender estimation in the wild with deep RoR architecture." In: IEEE Access 5 (2017), pp. 22492–22503.

[57] Alice Othmani, Abdul Rahman Taleb, Hazem Abdelkawy, and Abdenour Hadid. "Age estimation from faces using deep learning: A comparative analysis." In: Computer Vision and Image Understanding 196 (2020), p. 102961

[58] Raphael Angulu, Jules R Tapamo, and Aderemi O Adewumi. "Age estimation via face images: a survey." In: EURASIP Journal on Image and Video Processing 2018.1 (2018), pp. 1–35.

[59] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Toward automatic simulation of aging effects on face images," IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 4, pp. 442–455, Apr. 2002.

A4.5 Bias

Several types of biases are known to be present when humans estimate the age of a person based on pictures or videos of their face[60]. Examples include the true age[61], whether the participant is smiling[62], ethnicity[63], gender, and facial occlusions. Bias within the literature is quantified by the difference in the average error of participants with a certain characteristic versus those without.

Within machine learning models, biases can often occur due to imbalances within the datasets that the models are trained on. For example, models often perform worse on data with characteristics that it has not seen many examples of[64]. Additionally, biases present in humans are often also present in trained models. It should be no surprise then that age estimation is no exception. One study published in 2022 found that across a selection of commercially available age estimation apps, that they were less accurate and more susceptible to both age and facial expression biases[65]. In particular, the study found that the models overestimated the age of smiling faces, and that the drop in accuracy due to facial expression also correlated with the true age of the participant. Overall, this study highlights the need for a diversity of facial expressions to be present within training and validation datasets.

Another study published in 2021[66] conducted an analysis using three facial recognition models to evaluate the observed bias in gender and ethnicity. The authors found that in the models tested, male MAE was lower, whilst there was no consistent trend in racial bias across the three models. The issue of bias is typically present in the training datasets as well, where a recent study[67] found that most existing large scale face databases were biased towards ‘lighter skin’ faces.

A4.6 Scholarly Sources on Measurement Tolerance

Measurement tolerance within the facial age estimation literature is relatively scarce. As demonstrated previously, demographic statistics play a significant role in model performance, and therefore introduce additional uncertainty that should be quantified. When deploying an

[60] Manuel C Voelkle, Natalie C Ebner, Ulman Lindenberger, and Michaela Riediger. “Let me guess how old you are: effects of age, gender, and facial expression on perceptions of age.” In: psychology and aging 27.2 (2012), p. 265.

[61] Tzvi Ganel and Melvyn A Goodale. “The effect of smiling on the perceived age of male and female faces across the lifespan.” In: Scientific reports 11.1 (2021), p. 23020.

[62] Naoto Yoshimura, Fumiya Yonemitsu, Kyoshiro Sasaki, and Yuki Yamada. “Robustness of the aging effect of smiling against vertical facial orientation.” In: F1000Research 11 (2022).

[63] Albert Clapes, Ozan Bilici, Dariia Temirova, Egils Avots, Gholamreza Anbarjafari, and Sergio Escalera. “From Apparent to Real Age: Gender, Age, Ethnic, Makeup, and Expression Bias Analysis in Real Age Estimation.” In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. (2018)

[64] Shen, Zheyan, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu and Peng Cui. “Towards Out-Of-Distribution Generalization: A Survey.” *ArXiv* abs/2108.13624 (2021)

[65] Tzvi Ganel, Carmel Sofer, and Melvyn A Goodale. “Biases in human perception of facial age are present and more exaggerated in current AI technology.” In: Scientific Reports 12.1 (2022), p. 22519.

[66] Andraž Puc, Vitomir Štruc, and Klemen Grm. “Analysis of race and gender bias in deep age estimation models.” In: 2020 28th European Signal Processing Conference (EUSIPCO). IEEE. 2021, pp. 830–834.

[67] Michele Merler, Nalini Ratha, Rogerio S Feris, and John R Smith. “Diversity in faces.” In: arXiv preprint arXiv:1901.10436 (2019).

age assurance system, the performance distribution statistics help quantify the level of risk involved given the deployment scenario. Despite this, performance statistics such as variance and standard deviation are not commonly quoted within the literature. For example, none of the three review papers covered within this work quote standard deviation or variance within their analyses. Whilst Cumulative Score goes some way to quantifying the amount of variability within an age estimation system, there is a general lack of published research that provide distribution statistics to a degree that would be suitable for an age assurance solution.

A4.7 Observations on Literature Review

The project team have examined the analysis of literature available on this topic and make the following observations:

- The performance of facial age estimation technology in academic study appears to be worse than the levels of performance identified in Part 2 of this technical study, either as claimed by age assurance service providers, or as validated in this research.
- Accuracy of age estimation technologies is dependent on the age of the participants. Therefore, to effectively verify the performance of these systems, error rates should be provided for individual ages or age groups.
- There is a need for public validation datasets with high variation in pose, lighting condition, and image quality.
- Age estimation models have been shown to present bias for facial expression and gender, and therefore these should be accounted for when validating age assurance solutions.
- There is a need for more quantification of performance distribution statistics within the literature. These statistics play a significant role in quantifying the risk involved in deploying an age assurance system.

Glossary

Part 1 includes many definitions about age assurance measures that are referred to in this Part 2 of the technical study, but we have also included here a brief glossary of terms referred to in Part 2.

Accuracy

Accuracy refers to the closeness of a measured or observed age to the true or actual age. Accuracy is not a single concept when it comes to measurement of age assurance technologies. Part 1 set out four elements against which methods and techniques should be assessed, including 1) efficacy, 2) equality, 3) comparability, and 4) repeatability. The focus of Part 2 is on measurement of accuracy through the lenses of what might be practical and feasible expectations.

Age assurance

Age assurance is a collective term used to describe the range of techniques used to provide age estimation, age verification or age assessment. Its definition is not yet universally agreed or accepted; however, the phrase is included in a few official publications and normative references including:

- ICO opinion: Age Assurance for the Children's code, 14 October 2021

  "Age assurance" refers collectively to approaches used to provide assurance that children are unable to access adult, harmful or otherwise inappropriate content when using ISS [Information Society Services]; and estimate or establish the age of a user so that ISS can be tailored to their needs and protections appropriate for their age".

- ISO/IEC AWI 27566 - Information security, cybersecurity and privacy protection – Age assurance systems – Framework (In Development), 29 April 2023

  "Age assurance is a declaration that provides a level of confidence in the length of time that a person has lived".

- Online Safety Bill (as presented to Parliament), 17 March 2022

  "Age assurance means measures designed to estimate or verify the age or the age-range of users of a service".

Age assurance providers

The providers are those organisations, primarily commercial businesses, which have developed age assurance systems which can be deployed to provide the assurance as described above.

Age assurance components

Components (or measures) are the techniques by which age assurance can be delivered. Section 3 of Part 1 of the technical study included a range of current and potential future techniques, with a simple explainer and technical/legal definition for each. That list, however, should be kept under review given the constant evolution of technologies.

Age assurance workflow

Some age assurance systems employ a workflow where initial age estimation processes are used to filter out individuals that are over a certain age threshold (such as being over the age of 25) before proceeding with secondary age assurance techniques for those identified as under that threshold. A workflow is a process which reflects the entire deployment of an age assurance system, rather than an individual component (or technique)

Age estimation

Age estimation is a technique (developed through a technology or technologies) which provide an estimation of age based on machine learning through algorithm or alternative form of machine learning. It provides a continuous output. Of the various age assurance techniques considered in Part 2 of this technical study, three are age estimation: facial analysis age estimation, voice age estimation and email usage estimation. They rely either on biometric or behavioural data.

Age gate

In Part 2, the concept of an age gate is described in section 3.2. It is used to describe the age or age range at which a technology is being tested. Most use cases are currently driven by two age gates, namely whether a person is over or under the age of 13, or over or under the age of 18.

Age verification

Age verification is the process of confirming a person's age to determine whether they are legally allowed to access goods, content or services. Age verification can be accomplished through various methods, including presenting identification documents, verifying birth dates through online databases, or methods that aim to establish the confidence in an age attribute through binary measurement approaches.

Tolerances

In statistics, tolerances refer to the acceptable range of variation or deviation from a specified value or standard. It is the amount of difference that is allowed between a measured or observed value and a target or nominal value without affecting the quality or performance of a product or process. In the context of Part 2, tolerances describe the stated range of measurable outcomes which may ultimately be defined and agreed by regulatory bodies for age assurance.

Secondary measures

Part 2 explores the concept of secondary age assurance measures to provide more in-depth understanding and greater transparency than a primary focus on overall accuracy, used alone. Part 2 sets out that there are many and various possible statistical metrics that could be used as secondary measures. In the context of age assurance, recommended secondary measures include (at least) True Positive Rate, False Positive Rate, Positive Predictive Value, Mean Absolute Error, Standard Deviation of Absolute Error and Outcome Error Parity.

“State-of-the-art”

The term "state-of-the-art" refers to the current level of development, advancement, or innovation in age assurance technology and scientific research. It describes the highest level of knowledge, expertise, and technology that is currently available or in use in a particular domain.

The “state-of-the-art”, as described in Part 2, has been used as an objective and temporal assessment, and has been grounded in its legal context under UK GDPR rather than marketing vernacular. The guidelines published by the EU’s Agency for Cybersecurity (ENISA) have been used to inform the relevant assessments.

Bibliography

This bibliography refers to articles and works both in the body of this report and in the Appendices.

Journals, Articles and Learned Works

Abdelkawy H, et al (2020). “Age estimation from faces using deep learning: A comparative analysis”. *Computer Vision and Image Understanding 196.*

Adewumi O, et al (2018). “Age estimation via face images: a survey”. *EURASIP Journal on Image and Video Processing*

Adib A, et al (2022). “Adult and Non-Adult Classification Using ECG”. *IEEE 7th Forum on Research and Technologies for Society and Industry Innovation (RTSI).*

Anbarjafari G, et al (2018). “From apparent to Real Age: Gender, Age, Ethnic, Makeup, and Expression Bias Analysis in Real Age Estimation”. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.*

Carletti, V et al, (2019). “Age from faces in the deep learning revolution”. *IEEE transactions on pattern analysis and machine intelligence.*

Cootes T. F, et al (2002). “Toward automatic simulation of aging effects on face images”. *IEEE Transactions on Pattern Analysis Machine Intelligence, vol 24, no.4.*

Cui P, et al (2021). “Towards Out-Of-Distribution Generalization: A Survey”. *ArXiv abs/ 2108. 13624*

Ebner C. N et al (2010). “FACES – A database of facial expressions in young, middle-aged, and older women and men: Development and validation.” *Behavior research methods 42*

Ebner C N, et al (2012). “Let me guess how old you are: effects of age, gender, and facial expression on perceptions of age”. *Psychology and aging 27.2*

Eidinger E, et al (2014) “Age and gender estimation of unfiltered faces”. *IEEE Transactions on information forensics and security 9.12.*

Escalera S, et al (2015). “Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results”. *Proceedings of the IEEE international conference on computer vision workshops.*

Escalera S, et al (2016). “Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016”. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*

Feng J et al (2017). “Label-sensitive deep metric learning for facial age estimation”. *IEEE Transactions on Information Forensics and Security 13.2*

Feris S R, et al (2019). “Diversity in faces”. *arXiv preprint arXiv:1901.10436.*

Fu, Y and Huang, T (2008). “Human age estimation with regression on discriminative aging manifold”. *IEEE Transactions on Multimedia 10.4.*

Ganel T and Goodale A M, (2021). “The effect of smiling on the perceived age of male and female faces across the lifespan”. *Scientific reports 11.1.*

Ganel T, et al (2022). “Biases in human perception of facial age are present and more exaggerated in current AI technology”. *Scientific Reports 12.1*

Gao, C, (2017). “Age group and gender estimation in the wild with deep RoR architecture”. *IEEE Access 5.*

Grm K, et al (2021). “Analysis of race and gender bias in deep age estimation models”. *2020 28th European Signal Processing Conference (EUSIPCO). IEEE.*

Lapalme, G., and Sokolova M (2009). “A systematic analysis of performance measures for classification tasks”. *Information processing & management 45.4.*

Moyse, E (2014). “Age estimation from faces and voices: A review”. *Psychologica Belgica 54.3.*

Nash V, et al (2013). “Effective age verification techniques: Lessons to be learnt from the online gambling industry”. University of Oxford.

Pasquale, L et al, (2022). “Digital Age of Consent and Age Verification: Can they Protect Children?”. *IEEE Software, vol 39.*

Ricanek K and Tesafaye T (2006). “Morph: A longitudinal image database of normal adult age-progression”. *7th international conference on automatic face and gesture recognition (FGR06).*

Sasaki K, et al (2022). “Robustness of the aging effect of smiling against vertical facial orientation”. *F1000Research 11.*

Taheri S and Toygar Ö (2019). “On the use of DAG-CNN architecture for age estimation with multi-stage features fusion.” *Neurocomputing 329*

Uhl A and Wild P. (2009). “Comparing Verification Performance of Kids and Adults for Fingerprint, Palmprint, Hand-geometry and Digit print Biometrics.” *IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems.*

Official Publications

European Data Protection Board (EDPB) Guidelines 4/2019 on Article 25 - Data Protection by Design and by Default - Version 2.0 - Adopted on 20 October 2020

Standards and Normative References

ISO/IEC 27566 (In development) Information security, cybersecurity, and privacy protection – Age assurance systems – Framework

ISO/IEC 30107-3:2017 - Information technology — Biometric presentation attack detection - Part 3: Testing and Reporting

ISO/IEC 17065:2012 – Conformity assessment — Requirements for bodies certifying products

PAS 1296:2018 - Online age checking. Provision and use of online age check services. Code of Practice

ACCS 1:2020 – Technical Requirements for Age Estimation Technologies

ACCS 2:2021 - Technical Requirements for Data Protection and Privacy

ACCS 3:2021 - Technical Requirements for Age appropriate Design for Information Society Services

ACCS 4:2020 – Technical Requirements for Age Check Systems