Measurement of Age Assurance Technologies

A Research Report for the Information Commissioner's Office

AUTHORS

Allen, Tony - tony.allen@accscheme.com
McColl, Lynsey - lynsey@select-statistics.co.uk
Walters, Katharine - katharine.walters@accscheme.com
Evans, Harry - harry.evans@accscheme.com

Age Check Certification Services Ltd is an authorised conformity assessment body for the Age Check Certification Scheme carried out in accordance with BS EN ISO/IEC 17065:2012. It is a company registered in England at 3rd Floor, Broadstone Mill, Broadstone Road, Stockport, SK5 7DL (registered number 07357348).

Executive Summary

This research report sets out the approaches to the measurement of age assurance technologies.

Although the process of making age related eligibility decisions (such as when you purchase alcohol, tobacco or gambling products) is nothing new, in recent years there has been a significant growth in products and services that offer decision makers various levels of age assurance about their users. These products have emerged largely in a standards lacuna and, whilst the development of standards is rapidly catching up at national and international level, this research report was commissioned to help inform the Information Commissioner's Office (ICO) about how to measure and ensure confidence in the multiple approaches that have emerged in the meantime.

The report starts by defining age assurance and its various components (such as self-declaration, deployment of artificial intelligence, hard identifiers, digital identity services and other current or potentially emerging technical measures which could be deployed).

The report also touches 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, and individual efforts within different agencies, conformity assessment bodies and government/regulators to understand and define age assurance systems.

The emerging consensus is that a simple approach to describing the levels of confidence achieved by different assurance components would assist service providers, relying parties and those that regulate them.

So far, five specific levels of confidence have emerged in discussions at national and international standards fora:

The aim and intention of the standardisation process is to provide formulae, tolerances, descriptions and parameters to these five levels 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.

The report proceeds to explore four key pillars of the measurement of accuracy for age assurance technologies:

1. Efficacy - *does it work* - or the ability of the age assurance system to perform a task to a satisfactory degree. The report explores how to measure this, how to apply tolerances to it, how to report on those measures and which of the identified measures are most appropriate.

2. Equality - *does it treat different people fairly and equally* - or at least, are the outcomes of the age assurance process equally as good across different protected characteristics (such as skin tone and gender). The report also identifies, but does not explore in detail, issues associated with people who are 'identity challenged' - that is they struggle to have the appropriate means to identify themselves digitally. In this case, the broader range of components available for age assurance are potentially beneficial when compared against the challenges people face with proving who they are (instead of just how old they are).
3. Comparability - *can you compare 'apples' with 'pears'* - so can you take one type of age assurance component (say, a driving licence check) and compare that with the efficacy of another type of component (say, facial age analysis). Such comparability has the potential to enable a well-functioning competitive marketplace for age assurance technologies.
4. Repeatability - *can you repeat and reproduce the results of testing* - have you got sufficient samples, test protocols, consideration of environmental conditions (particularly ambient lighting and capture devices) and measurement rules to ensure confidence in the conformity assessment of different technologies.

The report examines 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).

In summary, the report concludes that the most appropriate measures are as follows:

For **continuous approaches** to age assurance, namely age estimation where the closer the estimate is to the true age of a person, the more accurate is the estimate:

> We recommend that Mean Absolute Error and Standard Deviation when taken together provide an effective means of measurement of age estimation systems. This is in contrast to the current common practice of just stating the Mean Absolute Error which is, in our view inadequate.

Mean Absolute Error (MAE)

- $MAE = \frac{\sum_{i=1}^{n}|(p_i - o_i)|}{n}$
- The central value of the absolute errors of the sample.

Standard Deviation (SD)

- $SD_{AE} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(AE_i - MAE)^2}$
- The amount of variation or spread over the distribution of absolute errors in the sample.

For **binary approaches** to age assurance, namely age verification where there is a positive declaration with only two possible states of 'yes' or 'no':

> We recommend that True Positive Rate, False Positive Rate and Positive Predictive Value when taken together provide an effective means of measurement of age verification systems. In our view, and borne out by the analysis set out in this report, the current common practice of just stating the false positive rate is inadequate.

True Positive Rate (TPR)	False Positive Rate (FPR)	Positive Predictive Value (PPV)
• $TPR = \frac{TP}{TP+FN}$ • Is the sensitivity of the technology's ability to correctly detect people who are over the age threshold.	• $FPR = \frac{FP}{FP+TN}$ • Is the technology's probability of false alarm (i.e., incorrectly identifying someone as being over the age threshold).	• $PPV = \frac{TP}{TP+FP}$ • The PPV is the proportion of the sample correctly identified as being over the age threshold given that they have been predicted as being over the age threshold.

To secure equality and fairness, all age assurance systems should be tested and be required to state their outcome error parity across, as a minimum, the protected characteristics set out in equalities legislation (such as skin tone and gender).

The report considers approaches to testing, analysis and certification. The ICO has existing powers in s.17 and Schedule 5 of the Data Protection Act 2018 to maintain overview and approval of certification criteria and to apply its tasks and powers under Article 42 of UK GDPR.

The report considers the key factors that need to be taken into consideration when assessing the approach to testing of age assurance systems.

These include 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 attack detection (spoofing) is undertaken in accordance with the relevant international standards.
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 efficacy).
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.

The report examines equality, efficacy and outcome fairness and proposes the best measures to quantifiably assess how a technology owner has implemented all four forms of fairness (data fairness, design fairness, outcome fairness and implementation fairness). One method identified in the report is to ensure that error rates are equitably distributed across different subgroups of the population.

The report concludes with eight recommendations, which are, in summary:

1. The ICO should continue to support international standards development for age assurance technologies.

2. The ICO should recognise different ways to measure accuracy and efficacy for age estimation vs age verification.

3. For age estimation, the comparable measure should be the **Mean Absolute Error (MAE)**, but only if it is published with information about the distribution of errors (the **Standard Deviation (SD)** and **outcome error parity** across protected characteristics (such as skin tone and gender).

4. For age verification, the comparable measure should be the **True Positive Rate (TPR) and** the **False Positive Rate (FPR) and the Positive Predictive Value (PPV)** published together with the **positive prediction value parity** across protected characteristics (such as skin tone and gender).

5. The ICO should consult on and publish applicable tolerances for these measures.

6. The ICO should consider further research into the implications of Trust Frameworks and interoperability between multiple systems and the potential to use multiple sources to elevate the level of confidence in age assurance outputs.

7. The ICO should explore how its tasks and powers under Article 42 of UK GDPR could be further extended to maintain oversight and approval of conformity assessment measures in the field of age assurance.

8. The ICO should publish supplementary guidance on the Children's Code on the measurement and reporting of age assurance technologies to ensure the upholding of information rights, whilst taking into account a need for an open, fair and comparable marketplace in the provision of such technologies to relying parties.

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 is a graduate of Edinburgh University and spent four years working in the European Parliament.

Harry Evans - Certification Officer at the Age Check Certification Scheme.
Harry is one of the lead auditors at ACCS specialising in data protection and privacy, the application of the Age Appropriate Design Code and the assessment and evaluation of age assurance measures from a data processing perspective. Harry is a Graduate of History from the University of Wales Trinity Saint David and a qualified ISO 27001 Lead Auditor.

Abstract

This research report was commissioned by the UK Information Commissioner's Office (ICO) to provide a technical study of measures to assess the accuracy of age assurance, including age estimation and age verification. The report explores approaches to measuring efficacy, equality, comparability and repeatability. Calculation of error in age estimation and age verification is studied with conclusions drawn about the appropriateness of them and a proposal to recommend 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.

The research investigates the measurement of inherent bias, equality and fairness in the context of handling personal data and potentially making decisions which may affect the rights and freedoms of individuals. The research also considers the main age assurance techniques, sensitivity and specificity, comparability across techniques and the impact of combinations and permutations of them. The report highlights approaches to measurements, testing, sampling and certification with a view to establishing repeatable and reproducible test protocols for analysis of age assurance systems.

Research Brief

The Information Commissioner's Office commissioned a technical study of measures to assess the accuracy of age assurance methods to develop a picture of the best and most appropriate indicators of accuracy and to understand the potential for greater consistency in approaches to measurement across the age assurance industry.

In response to the Research Brief, we have undertaken to provide:

1. A series of definitions that enables standardisation across the sector, including clarity around 'age assurance' (including both 'age estimation' and 'age verification'), 'levels of assurance' and the technical terms that will be applicable to the recommended measurement approach.
2. An assessment of the most appropriate measurements and/or indicators of accuracy, both alone and in combination, where these are designed to deliver a practical result that can be universally applied across all methods of age assurance but are as simple to understand and explain as possible.
3. An assessment of the methods available to test the full range of age assurance techniques, including any known limitations and recommendations for how to address those.
4. An assessment of measurement uncertainties or bias; including how these can be measured, appropriate tolerances to be applied and their impact upon the statements of efficacy of age assurance systems.

1. Introduction to Age Assurance

Although the concept of age restricted goods and services is not new in UK law or that of other legislatures across the world, age assurance technologies are still in their relative infancy and evolving rapidly. Age assurance has emerged in recent years as an umbrella term to include age estimation and age verification which have been more widely used, historically, to describe the multiple approaches which exist. It follows that definitions are not, necessarily, hard and fixed. The standards community, both within the UK and globally, is actively addressing current gaps and it is likely, too, that definitions will be subject to deliberation as draft legislation undergoes Parliamentary scrutiny.

This section explains what age assurance is, sets out how it is currently defined in draft national laws and standards, how that is evolving through the ICO's Children's Code and how that translates into different techniques for age assurance.

It is recommended, however, that existing definitions are reviewed periodically, especially given the nature of rapidly evolving technologies. We recommend that the ICO take an holistic approach, seeking to draw on the widest possible range of opportunities to achieve age assurance, within the definition. We particularly note that the ICO should ensure that potentially privacy preserving approaches (including those that do not involve establishing broader identity attributes) should be carefully retained within the range of activities that contribute to gaining age assurance about individuals.

1.1 Defining 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. The definition of age assurance is evolving, but at present there are several principal sources to consider.

The draft international standard on age assurance systems states:

> *"Age Assurance is the process of establishing, determining, and/or confirming either age or an age range of a natural person"*

[SOURCE: ISO/IEC 27566 Information security, cybersecurity and privacy protection – Age assurance systems – Framework – WORKING DRAFT]

The draft Online Safety Bill states:

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

[SOURCE: Draft Clause 189, Online Safety Bill]

The Information Commissioner's formal opinion on age assurance states:

> *"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 to their age.*
>
> *[SOURCE: Information Commissioner's opinion: Age Assurance for the Children's Code, 14 October 2021]*

The Age-Appropriate Design Code[1] does not define age assurance but refers to several techniques that could contribute to gaining a degree of certainty about the age of users of information society services. These include a non-exhaustive list of:

- *Self-declaration* - This is where a user simply states their age but does not provide any evidence to confirm it. The Code states that it may be suitable for low-risk processing or when used in conjunction with other techniques.
- *Artificial intelligence* - The Code identified that it may be possible to make an estimate of a user's age by using artificial intelligence to analyse the way in which the user interacts with the service.
- *Third party age verification services* - Such services typically work on an 'attribute' system where the information society service requests confirmation of a particular user attribute (in this case age or age range) and the age verification service provides a 'yes' or 'no' answer.
- *Account holder confirmation* - This is where a logged-in or subscription-based service allows the main (confirmed adult) account holder to set up child profiles, restrict further access with a password or PIN, or simply confirm the age range of additional account users.
- *Technical measures* - There are processes which discourage false declarations of age, or identify and close under-age accounts. Examples include neutral presentation of age declaration screens (rather than nudging towards the selection of certain ages) or preventing users from immediately resubmitting a new age if they are denied access.
- *Hard identifiers* - This involves confirming age using solutions which link back to formal identify documents or 'hard identifiers' such as a passport.

Age assurance techniques are, at present, a nebulous concept with multiple different methods, approaches, measurement challenges and propensity to define. Equally it is important to recognise that these technologies are constantly evolving and will continue to emerge as further use cases and age-related eligibility challenges are identified.

In this project, the brief asked for consideration of the measurement and definition challenges of:

- Use of hard identifiers
- Use of verified information (e.g., in centralised databases)
- Digital identity services

[1] The Age-Appropriate Design Code, known as the 'Children's Code' is issued by the Information Commissioner in accordance with s.123 of the Data Protection Act 2018

- Tokenised attribute exchange models
- Device-level intervention
- Credit/debit card confirmation
- Account confirmation processes
- Email verification processes
- Using facial images
- Using iris
- Using gait
- Natural language processing
- Analysing behavioural traits
- Profiling
- Parental control software
- Self-declaration

Age assurance covers a wide range of activities. Many of these activities are privacy preserving, particularly where they do not involve establishing broader identity attributes in order to provide age assurance. The definition will evolve and, in a world where age attributes are often merely seen as a single identity attribute, it will be important for the ICO to take an holistic approach, seeking to draw on the widest possible range of opportunities to achieve age assurance, within the definition. We particularly note that the ICO should ensure that potentially privacy preserving approaches should be carefully retained within the range of activities that contribute to gaining age assurance about individuals.

This report splits age assurance measurement into two parts - continuous and binary. These are explained in more detail in the body of the report. Many long-standing approaches to gaining age assurance have been based on hard identifiers providing a binary outcome (yes or no to the age question posed, i.e. is this person over 18?). Increasingly, technology is delivering age assurance processes that are continuous - such as age estimation techniques. These are converging in efficacy and may in the future provide for sufficient alignment to result in the measurement methodologies and levels of confidence in results to also converge.

1.2 Using the Term "Assurance"

The term "Assurance" is used in a wide variety of contexts in national and international laws, regulations, standards and measurement techniques. In this document, the term "Assurance" will appear in multiple contexts when referring to other standards and documents.

The Oxford English Dictionary defines assurance as a positive declaration intended to give confidence. In international standards, there are many references to the use of the term "assurance", such as:

> *"quality assurance is part of quality management focused on providing confidence that quality requirements will be fulfilled."*

[SOURCE: ISO 9000:2015 – Quality management systems — Fundamentals and vocabulary, 3.2.11]

"evaluation assurance is grounds for justified confidence that a target of evaluation meets the security functional requirements"

[SOURCE: ISO/IEC 15408-1:2009 - Information technology — Security techniques — Evaluation criteria for IT security — Part 1: Introduction and general model, 3.1.4]

"identity assurance is the process of establishing, determining, and/or confirming a subject identity."

[SOURCE: ISO/IEC 30108-1:2015 – Information technology — Biometric Identity Assurance Services — Part 1: BIAS services, 4.7]

There will continue to be multiple uses of the term "assurance" in different contexts and this project is unlikely to solve that problem.

1.3 Using "Levels"

Similarly, the use of the term "Levels" can give rise to confusion. Levels can be used to describe a position on a scale of amount, quantity, extent, or quality. Many aspects of standards describe levels as:

- a **score** 1, 2, 3, 4, 5, etc.;
- or as a **descripto**r 'Low'; 'Medium'; 'High', etc.;
- or as an **output** 'Basic'; 'Strict'; 'Enhanced' etc.

Care needs to be taken not to confuse different levels assigned to different measurement components for different things.

A 'level of confidence' can be used to describe the extent to which a positive declaration of assurance can be relied upon. However, in statistical terminology, it is also used to describe the confidence coefficient (the value (1−α) of the probability associated with a confidence interval or a statistical coverage interval). In this context, the confidence level is expressed as the likelihood that the true result falls within a range of results and is set (for present purposes) at 95%.

In addition, "Levels" are used to describe the depth of evaluation. So, in ISO/IEC 15408-1 - Evaluation criteria for IT security, referred to below in *section 9.9 Depth of Evaluation,* "Evaluation Assurance Levels" (EAL) are used to describe a set of assurance requirements, usually involving documentation, analysis and testing, representing a point on a predefined assurance scale, that forms an assurance package.

Evaluation Assurance Levels (which are numbered 1 - 7; with 1 being the lowest and 7 being the highest) describe the depth of testing and analysis undertaken for an assurance component to verify and validate that it is operating in the way it is meant to.

In the UK Government's Good Practice Guide: How to prove and verify someone's identity (GPG45[2]), there are 4 different levels of confidence assigned to the identify profile created by scoring different elements of a claimed identity. The four levels of confidence assigned are:

- low confidence
- medium confidence
- high confidence
- very high confidence

Although age assurance is not about identity (a person's age is an attribute of their identity, but it is not necessarily the case that you need to establish a person's identity to gain assurance about their age), there are parallels to the process of establishing an identity profile to establishing age assurance.

Care should be taken when referring to numerical or alphabetical 'levels' - if you have a declaration that something is "Level 5" that could indicate that it is 'very high', or it could indicate that it is 'very low' rather depending on the context of the scale 1 - 5 or 5 - 1. Therefore, descriptors such as 'low', 'medium' or 'high' are preferred.

Ultimately, this report recommends that the 'levels of confidence in age assurance' from providers should be categorised around outputs: asserted - basic - standard - enhanced - strict.

1.4 ISO/IEC 27566 – Age Assurance Systems (*currently at the ISO's working draft stage*)

The International Standards Organisation (ISO) are currently preparing a draft international standard of age assurance systems. This standard, known as ISO/IEC 27566 - Information security, cybersecurity and privacy protection - Age assurance systems - Framework, is seeking to:

- define the key terms, definitions and abbreviations applicable to the age assurance process
- specify the requirements for establishing the indicators of confidence associated with age assurance systems
- specify the roles, responsibilities and procedures of key actors in the age assurance process, including the requirement to establish age assurance policies
- give guidelines about attack vectors and countermeasures (i.e., anti-spoofing techniques), presentation attack detection, algorithms, or sensors.
- specify the data protection, privacy and security objectives specific to the age assurance process

In the draft international standard for age assurance, five indicators of confidence in age assurance are described in Table 1 below. For reusable age assurance products, Levels of Assurance (LoA) for authentication are referenced and these are as set out in the UK

[2] https://www.gov.uk/government/publications/identity-proofing-and-verification-of-an-individual

Government's Good Practice Guide: Using authenticators to protect an online service (GPG44[3]).

TABLE 1 – SCHEMATIC: INDICATORS OF CONFIDENCE IN AGE ASSURANCE

Asserted	Basic	Standard	Enhanced	Strict
• Based on self-asserted age attributes • No validation or trust elevation deployed • No attempt has been made to address contr-a indicators • Could be utilised in low risk or only where indicative age is required • Unlikely to be satisfactory for legally defined age-related eligibility	• Based on self-asserted age attributes with a single age assurance component that has low evaluation assurance level • Partial or simple validation or trust elevation; contra indicators may still be present • Could be used for unregulated age gateways	•Based on at least one age assurance component with standard evaluation assurance levels •For reusable age attributes, authenticated entity to LoA (level of assurance) 1 at least every 3 months •Validated and contra indicators addressed or accepted • Considered to be the minimum standard required for regulated age related eligibility unless a higher or lower level is specified by the policy maker	• Based on two or more age assurance components with higher levels of confidence and standard evaluation assurance levels •For reusable age attributes, authenticated entity to LoA (level of assurance) 2 at least every week • Validated and contra indicators addressed or accepted • Likely to be useful for enhanced risk goods, content or services age-related eligibility	•Based on two or more age assurance components with higher levels of confidence and higher evaluation assurance levels •For reusable age attributes, authenticated entity to LoA (level of assurance)3 at least every day •Validated and contra indicators addressed or accepted •Likely to be useful where age-related eligibility is critical to safeguarding or protecting the rights or freedoms of individuals

The draft standard describes how indicators of confidence can be applied within an Age Assurance System which consists of one or more assurance components indicating a person's age, and then how those indicators of confidence can be communicated to a relying party (the person or organisation that needs to make an age-related eligibility decision).

The assurance components may include:

- A claimed age attribute by the person - known as a self-asserted age attribute
- A process or system deriving an age attribute from an identity document or record from an authoritative source - for example, an 18 plus attribute derived from the date of birth in a passport
- A process or system deriving an age attribute from primary or secondary credentials, a data set, an attribute attestation provider or identity service provider

[3] https://www.gov.uk/government/publications/authentication-credentials-for-online-government-services

- A process or system deploying artificial intelligence to ascertain age from one or more biometric identifiers, behaviours, characteristics or actions of individuals
- A process or system deploying social proofing to obtain or verify age attributes
- A process or system based on the attestation of trusted parties (such as parents or legal guardians) about the age of a person
- A process or system based on the profiling or tracking of the existence of consistent age attributes over time
- An assessment led by a trained human assessing elements that consider a person's appearance, demeanour, background and credibility in person or online
- A process or system that derives age attributes from any other method that can establish levels of confidence

An age assurance processing sub-system may include:

- A process or system for gathering assurance components from multiple sources
- A process or system for identifying attack vectors, protecting against presentation attack and assessing the liveness of individuals
- A process or system for identifying and addressing contra indicators
- A process or system for elevating the trust in an age attribute through multiple sources
- Facilities for individuals to exercise data rights
- A process or system for dissemination of age attributes, to a stated level of age assurance, to relying parties
- A process or system for monitoring, continuously improving and learning from age assurance activities
- A process or system for processing of entity authentication factors

2. About the Age Check Certification Scheme

This section provides background detail about the Age Check Certification Scheme, who have been commissioned to produce this report. It describes their work on age assurance standards to date, and also their accreditation by the United Kingdom Accreditation Scheme (UKAS) and by other international accreditation schemes. It also highlights that its certification criteria are the first to be formally approved by the Information Commissioner's Office under Article 42 of UK GDPR.

The Age Check Certification Scheme is an independent third-party conformity assessment service operated by Age Check Certification Services Ltd. 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"

2.1 UKAS Accreditation

Age Check 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[4] 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.

2.2 ICO Approval

The criteria that Age Check Certification Services utilise for the assessment of data protection and privacy of identity and age assurance services (ACCS 2:2021[5]); and for the assessment of the age-appropriate design of information society services (ACCS 3:2021[6]), have been approved by the Information Commissioner's Office.

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

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)
  - obligation to notify data breaches (Art 33)
  - obligation of DP by design and default (Art 25)
  - whether a DPIA has been completed where required (Art35(7)(d)
  - technical and organisational measures put in place to ensure security (Art 32).
- 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[7].

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

2.3 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 report, 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 (we discuss age buffers and challenge categories in more detail in *section 6.7 Age Buffer*s). 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

[7] 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

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.

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.

2.4 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*[8].

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.
- 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
    - Licensing action, conditions or restrictions imposed by Licensing Authorities

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

3.Defining Age Assurance Components

This section identifies the key age assurance techniques (referred to as components in the draft international standard). It seeks to provide both a simple explainer of what that component is, but also a more detailed technical or legal definition of what the component is - the type of definition that may appear on a certificate of conformity for instance. Not all the identified components are in existing commercial use and technological innovation is constantly identifying more components.

It is intended to provide a useful guide to current and potential future components and how they could be described and defined for regulatory and conformity assessment purposes. **It is recommended, however, that given the constantly evolving nature of techniques, or components, that it is kept under review to reflect future technological developments and also published as online guidance by the ICO.**

This report sets out a three-level approach to defining age assurance techniques:

- Descriptor
- Simple Explainer
- Technical or Legal Definition

This is a list that ought to be kept under review and could usefully be included in interactive online guidance. Any formalised testing or certification process will require a technical or legal definition of the technique or component under test (often referred to as the 'Target of Evaluation' (ToE)). These can be very specific, as shown in Table 2. For ease of understanding, these should be accompanied by a simple, plain-English, explainer and a short-hand descriptor.

TABLE 2 – DEFINITIONS OF AGE ASSURANCE TECHNIQUES

Descriptor	Simple Explainer	Example Technical or Legal Definition
Voter's Roll	Access to the Electoral Register	A list maintained by a local registration officer in accordance with the requirements of the Representation of the People Act 1983 and, by virtue of s.1(1)(d), 2(1)(d) and 9(2)(a) of that Act, that person appears to the local registration officer to be 18 years of age or over.
Credit Card	A payment card issued to users (cardholders) to enable the cardholder to pay a merchant for goods and services.	The presentation of a payment card with a number issued in accordance with ISO/IEC 7812-1:2017 Part 1 that relates to a type of account that requires the cardholder to be a person aged 18 years or over, such as by offering credit facilities under the terms of the Consumer Credit Act 1974.
Government Issued ID (these are sometimes	An official document issued by a government that contains a given person's identity.	The presentation and capture of an identity document referenced on the Public Register of Authentic Travel and Identity Documents Online (PRADO) maintained by the Council of

Descriptor	Simple Explainer	Example Technical or Legal Definition
called 'hard identifiers')		the European Union and the extraction of the age attribute of the presenter of the document to demonstrate that they of a certain age.
Smart Device Content Controls	Content Bar Status	The content bar applicable to a UK-issued Mobile Station International Subscriber Directory Number (MSISDN) based upon age verification carried out by a third-party OFCOM-licensed mobile network operator in accordance with the Mobile UK Code of practice for the self-regulation of content on mobiles (V3, 1/7/13).
Digital Identity	Identity Service Provider (or Attribute Service Provider)	An age attribute derived from an Identity Service Provider or Attribute Service Provider under the terms of the UK Digital Identity and Attributes Trust Framework[9].
Trusted Account Administrator	Open Banking, Regulated Utilities, Solicitors, Regulated Professions	An age attribute derived from an organisation charged with the administration of an account and who is authorized and regulated by a sectoral regulator that imposes requirements that the administrator of that account must undertake age assurance to a specified level of confidence.
Self-declaration	An asserted claim of age made by an individual.	Where a user states that their age or date of birth or confirms that they are over a certain age.
Cognitive Testing	Assessment of a person's cognitive ability relating to age	An assessment of the general mental capability of individuals involving reasoning, problem solving, planning, abstract thinking, complex idea comprehension, and learning from experience.
Automated Facial Analysis	Assessment of facial features to estimate age[10].	An assessment of anthropometric features of the face to carry out probabilistic analysis of the likely age of the person.
Automated Voice Analysis	Assessment of voice features to estimate age.	An assessment of a combination of both aural (listening) and spectrographic (instrumental) comparison of a voice to carry out probabilistic analysis of the likely age of the person.
Hand geometry	Assessment of the size of a hand, pinch gesture or span	An assessment of the span, measurement and touch points on an interactive screen of hand-based modalities to carry out probabilistic analysis of the likely age of the person.
Gait analysis	Assessment of the size and nature of a stride of a person.	An assessment of the size, measurement and muscular action of a stride of a person to carry out probabilistic analysis of the likely age of the person.
Behavioural profiling and inference	Social proofing	An assessment of the social connections and activity of a person (online or offline) including their likes, behaviours, social norms,

[9] UK digital identity & attributes trust framework: updated version - GOV.UK (www.gov.uk)
[10] We are careful to distinguish this from face recognition technology (FRT). In most systems that have been submitted for test, the age analysis function is not attempting to recognise the face or match the face to a pre-determined face database.

Descriptor	Simple Explainer	Example Technical or Legal Definition
		behaviours or activities or the actions or activities of people that they relate to carry out probabilistic analysis of the likely age of the person[11]. Other potential behavioural profiling sources include public posts, private chats, Natural Language Processing, mouse-stroke and key board analysis, browsing habits, purchasing habits, service activity e.g., likes, birthday posts, etc.
Parental Controls	**Account Confirmation, often in the form of Parental Controls**	A process where a user's age, or age range, is confirmed by another connected and trusted accountholder, e.g., a parent or legal guardian, using their account to confirm the ages of their children or provide permissions for their children to access different levels of content or services.
Reusable Age Assurance	Creating an age assurance account or reusing a previously created account	A process where use of an age-assured account with one service to establish an account or access another connected service.
Device-led Authentication	Storing an age assurance token or signal on the user's device	A process where the use of an age assurance process is to control access to either the device or functions on the device that itself enables online access to content or services (such as a laptop, phone, or games console). This differs from Smart Device Content Controls (often known as content bar status) as set out above.
Peer reporting and flagging	Potential identification of contra-indicators	A process for the provision of tools for users to report where they believe other users do not meet a certain age requirement.
Technical Measures	Software solutions	Service design features that give extra assurance to other assurance measures (e.g., not allowing a user to submit their age twice, or blocking circumvention of age assurance measures).
Longevity of identifying factors	Consistent existence of an identity factor over time	A process for identifying the first registration or transfer of a specific identity factor (such as an e-mail address for instance) and, from the date of that transfer, infer that the address (known as a fully qualified domain address (FQDA) by the Internet Engineering Task Force) and thus the holder has existed in time for a certain period.

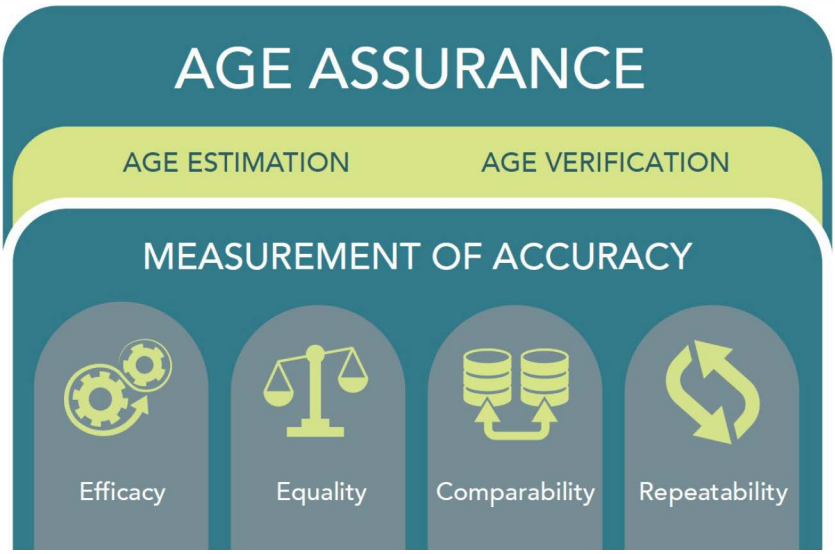
[11] In July 2021, Facebook (now Meta) announced that it intended to deploy artificial intelligence social proofing techniques to estimate the real age of its users.

4.Measurement of Accuracy

This section sets out the approach to the measurement of accuracy, including ensuring the efficacy of the age assurance components, approaches to measurement of equality and fairness (particularly in the context of fair processing of personal data and tackling inherent bias), how the results of testing can be compared across different assurance components and techniques and how to secure that testing results are repeatable and reproducible. It recommends that accuracy needs to be understood on a multi-dimensional level, with (at least) four elements used for assessment.

In the following chapter, we define a list of measures that can be used to assess accuracy which we split according to whether the technology objective is estimating a person's age (age estimation) or the pass/fail of an age-threshold (age verification).

The starting premise is that "accuracy" is not a single concept and that there are (at least) four elements against which the methods and technique should be assessed.



4.1 Efficacy

Efficacy is the ability to perform a task (such as age estimation) to a satisfactory degree. In this context, efficacy will be examined via measures of accuracy. Measurement accuracy is often defined as the closeness of agreement between a measured quantity and a true quantity value of a measurand (i.e., the quantity intended to be measured) (ISO-JCGM 200, 2008)[12].

[12] ISO-JCGM 200, 2008 International Vocabulary of metrology- Basic and general concepts and associated terms (VIM)

However, the most appropriate measure of accuracy will necessarily depend on the outcome of the technology. For example, different measures are required for a technology that is estimating a person's age versus the pass/fail of an age-threshold (such as those set by Challenge Age policies).

It is important to note that the definition of what is satisfactory efficacy is one that must ultimately be set by regulators.

In this report, we explore two approaches to the measurement of efficacy:

- measures applicable to continuous age assurance outputs, where there is an estimation of age based on algorithms or assessments; and
- measures applicable to binary age assurance outputs, where there is a positive declaration with only two possible options: - 'yes' or 'no'

It is important to note that age assurance systems can contain multiple components (as set out in section *3.1 Using "Levels"* above. It is also possible that a measure could start as continuous (i.e. this person is likely to be between 55 and 65), but when applied to an age assurance threshold, it becomes binary (i.e. is that same person over 18: yes).

There will, undoubtedly, continue to be multiple uses of the term "assurance" in different contexts and this project is unlikely to solve that problem once and for all.

4.2 Equality

Equality involves ensuring that technologies treat different people fairly and equally with respect to protected characteristics such as gender and race. Whilst there is no single definition of fairness, potential assessment measures could include:

- **Anti-classification**: The model is fair if it does not use protected characteristics (except age itself in this context) or proxies from which protected characteristics can be inferred (i.e., a protected characteristic is not used to predict age).
- **Classification or outcome error parity**: The model is fair if protected groups receive an equal proportion of positive outcomes, or an equal proportion of errors.
- **Calibration**: The model is well-calibrated if the predicted ages reflect the actual ages in real life for the observations given those predictions. Equal calibration definitions of fairness say that a model should be equally calibrated between protected attribute groups.

In section 7 of this report, entitled *Issues of Equality, Parity and Fairness*, we explore in further detail approaches to determining equality, parity and fairness of age assurance techniques, particularly in the context of securing the protection of protected characteristics (such as race, gender, etc.) and the lawfulness of processing personal data in accordance with UK GDPR.

4.3 Comparability

Comparability is the extent to which differences between statistics from different age assurance technology testing, or over time, can be attributed to differences between the true values or the statistical analysis and testing.

Comparability could be more easily described as how to discuss the differences and similarities between 'apples' and 'pears'. This is an important aspect that underpins a well-functioning competitive marketplace. If economic decision makers (i.e., those procuring age assurance technologies for implementation) are not able to compare one product effectively and efficiently with another, the market for age assurance technology will be deficient.

Testing techniques should result in metrics that users are able to use in a comparable manner to either rank or distinguish their service from others that are operating in the marketplace. It is important that, in an open fair market, age assurance technology descriptions are not misleading. Section 4(d) of the Misleading Marketing Regulations 2008[13] states that *inter alia* comparative advertising must ensure that:

> "it objectively compares one or more material, relevant, verifiable and representative features of [the product]"

4.4 Repeatability

Repeatability is a measure of precision which quantifies the degree to which repeated measurements under the same operating conditions show the same results. This is in contrast to reproducibility which is where a test environment can reproduce the results found in-house, for example.

Knowledge of the uncertainty associated with measurement results is essential to the interpretation of the results. Without quantitative evaluations of uncertainty, it is impossible to decide whether observed differences between results reflect more than experimental variability, whether test items comply with specifications, or whether laws based on limits have been broken. Without information on uncertainty, there is a risk of misinterpretation of results.

Repeatability is, therefore, a fundamental principle of testing protocols. A test that is not repeatable will undermine confidence in the test laboratory and has implications for accreditation.

We shall assess these qualities against a range of age assurance techniques, including both estimation and verification. Prioritisation will be given to those techniques which are already operational for commercial age assurance purposes.

[13] SI 2008:1276 - The Business Protection from Misleading Marketing Regulations 2008 (legislation.gov.uk)

5. Approaches to Measurement of Continuous Age Assurance

This section identifies the measures applicable to continuous age assurance outputs, where there is an estimation of age based on algorithms or assessments. Continuous approaches do not result in a binary outcome (i.e., yes or no), but result in a range of outcomes within parameters, as such, the approach to measurement of them is very different to binary techniques.

This section looks at the statistical analysis of age estimation techniques. It explores only the efficacy of measurement. Issues of equality, comparability and repeatability are discussed later in the report.

5.1 Age Estimation

Continuous age assurance technologies provide an estimate of a person's age. The closer this estimate is to the true age of that person, the more accurate the estimate. In the following table a set of measures are defined that can be used to assess the accuracy of the age technology given a set of samples or testing data. Since the outcome, age, is a continuous outcome, the measures below can all be applied to continuous data. Each measure is defined using the following parameters:

- The true (or observed) age of sample i: o_i.
- The predicted age of sample i: p_i.
- The number of samples tested: n.

TABLE 3 – MEASURES APPLICABLE TO AGE ESTIMATION TECHNOLOGIES

Measure	Definition	Meaning/Notes
Error	$E_i = p_i - o_i$	The error is the difference between the predicted and true age of sample i. It is impacted by whether the prediction is an over or underestimate of the true age (it will be positive for the former and negative for the latter). The distribution of errors across all n samples can be visualised by a histogram, which will highlight the shape of the

Measure	Definition	Meaning/Notes
		distribution (is it symmetrical or skewed) and the range of errors across the full sample.
Absolute Error	$AE_i = \|p_i - o_i\|$	The absolute error is the absolute difference between the predicted and true age of sample i. The error is the magnitude of the size of the difference between the predicted and observed ages (i.e., it is positive irrespective of whether the prediction is an over or underestimate). The absolute error is a useful overall measure of accuracy, and we will focus on it below when defining measures of central tendency and spread over the sample distribution of absolute errors. Note that there are cases when understanding whether an age estimate is over or underestimating the true age is informative; particularly for model performance improvements and checking for differences between protected characteristics, for example. As above, the distribution of absolute errors across all n samples can be visualised by a histogram, which will highlight the shape of the distribution (is it symmetrical or skewed) and the range of absolute errors across the full sample.
Mean Absolute Error	$MAE = \frac{\sum_{i=1}^{n}\|(p_i - o_i)\|}{n}$	The mean absolute error is the central value of the absolute errors; it is the average value of the sample. There is another measure of central tendency that can be useful, particularly if the distribution of the errors suffers from outliers, known as the median (note the mean and median are identical in symmetric distributions).
Median Absolute Error	The median error ($MEDAE$) is the middle number in the sorted (ascending or descending) list of absolute errors.	The median is sometimes used rather than the mean when the distribution of absolute errors is heavily skewed. In this instance, the mean may be influenced by outliers (i.e., a small number of samples with particularly

Measure	Definition	Meaning/Notes
		large errors) and not be a reliable measure of central tendency. The mean is the most frequently used measure of central tendency and will continue to be the focus here, but the median is worth consideration in these specific circumstances.
AE Standard Deviation	$SD_{AE} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(AE_i - MAE)^2}$	The standard deviation is a measure of the amount of variation or spread over the distribution of absolute errors. A low standard deviation indicates that the values are close to the MAE and a higher value indicates a larger spread. Other measures of spread can be calculated for example, the range (the maximum minus the minimum absolute errors) and the interquartile range.
MAE 95% Confidence Interval	$CI_{MAE} = MAE \pm 1.96\frac{SD_{AE}}{\sqrt{n}}$ Note at a minimum the lower bound is 0 and should be truncated if needed	A confidence interval quantifies the uncertainty associated with an estimate, such as the MAE. The interval is calculated from the sample and is the range of values in which we estimate the MAE to lie with 95% confidence. A 95% confidence level is recommended as this is what is used most in ISO standards and the statistical community. All estimates such as the MAE should be reported with a confidence interval to understand and quantify their uncertainty. Without this additional measure, they are not very informative. In this example, the number 1.96 is the critical value of the Normal distribution based on a 95% confidence level. It is dependent on the data meeting the Central Limit Theorem which establishes that for a large enough sample, the sample average tends to a normal distribution. Typically, a sample size of more than 30 is deemed large enough.

Measure	Definition	Meaning/Notes
AE 95% Prediction Interval	$PI_{AE} = MAE \pm 1.96SD_{AE}$ Note at a minimum the lower bound is 0 and should be truncated if needed	A prediction interval or predictive confidence interval quantifies the uncertainty associated with the absolute error of a single individual. It is the range of values in which we estimate the absolute error of the individual to lie with 95% confidence.

5.2 Measures Discounted from Consideration

Several other measures were considered but found not to provide useful information over and above those discussed above. We provide them below for completeness together with an explanation of why they are not recommended (but we do not provide the formulae to calculate them).

- Mean Absolute Scaled Error (MASE): suitable for assessing the accuracy of forecasts through time. This is not appropriate here as the data recorded by an age estimation technology does not have a time element to it (time series data measure observations over time whereas age estimates are independent of time).
- Mean Squared Error (MSE): used as a measure of quality but can be heavily influenced by outliers (unlike the MAE). This means that if there are a small number of age estimates that differ greatly from the other estimates, these can influence the value of the MSE (the MAE does not suffer from this problem to the same extent). It is preferable to have a measure of central tendency that is not unduly influenced by outliers.
- Mean Absolute Percentage Error (MAPE): this measure is biased towards under rather than over predictions. This error places a heavier penalty on over rather than underestimates. An error that places equal weight on the direction of the error is preferable.
- Average Absolute Deviation: an alternative measure of dispersion or spread compared to calculating the AE standard deviation, but in this case, we favour the standard deviation since it is to be used in the confidence interval calculation.

5.3 Observations on Age Estimation Measurement

There are several key points to bear in mind:

1. The MAE is a useful overall measure to summarise the accuracy of an age estimation technology on average. The MAE is a measure of central tendency of the sample. An age technology with low MAE tells you that you have a good “average” performance over the sample or training data set.
2. However, the MAE should not be looked at in isolation and, on its own, is not sufficiently informative. Reporting the absolute error standard deviation quantifies the spread of the

distribution. If the standard deviation is low as well as the MAE, then the performance is not only good on average, but also across the entire dataset. For example, if two different technologies have both been assessed with a MAE of 2.5 years they could be assessed as having the same level of accuracy, but this is not the case if one has an AE standard deviation of 0.25 years and the other has an AE standard deviation of 1.5 years. Looking at the MAE on its own would not have highlighted that the performance of the technology with the lower standard deviation is better overall.

3. The spread of the distribution can be quantified further by producing a 95% absolute error prediction interval. For example, for a standard deviation of 2, an individual is predicted to lie within +/- 3.92 years of the MAE with 95% confidence, compared to +/- 1.96 years with a standard deviation of 1.

4. The distribution of absolute errors should be visualised using a histogram to understand its shape and the spread or range of absolute errors. If the distribution contains outliers, the MEDAE should also be reported.

5. To understand whether an age technology is over or under predicting ages, the distribution of the errors (rather than absolute errors) will help. This can be useful for identifying areas to improve performance and for investigating differences between protected characteristics, for example).

6. The MAE should not be reported without its associated 95% CI to quantify its uncertainty. The smaller the 95% CI the more precise the estimate.

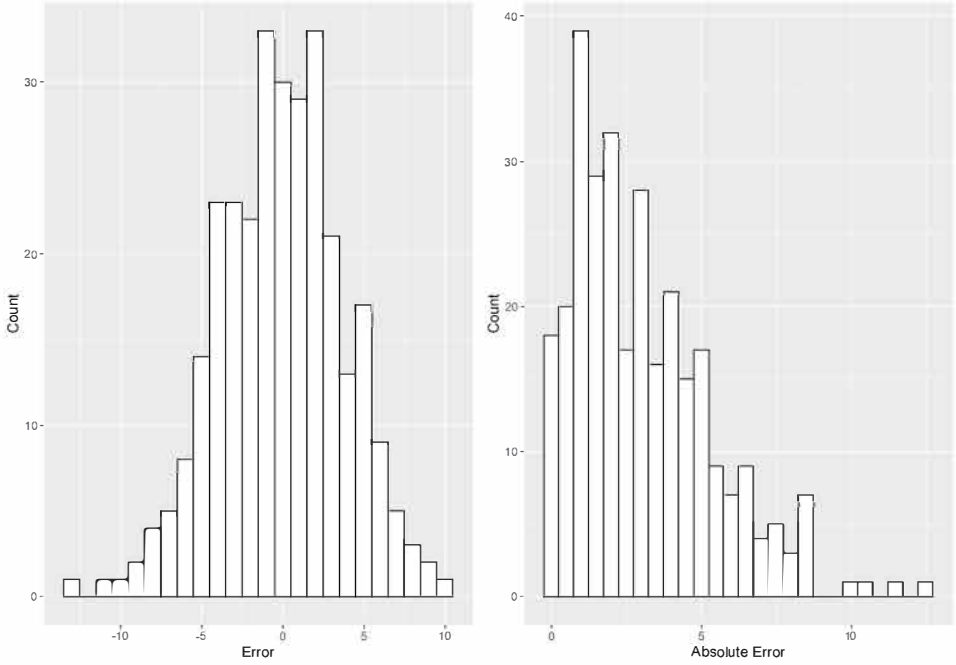
Identifying what is an acceptable level of MAE and AE standard deviation (i.e., how low does the MAE need to be for the accuracy of the age estimation technology to be deemed acceptable) is a decision for regulators. We discuss options for tolerances in *section 9.10 Regulatory Options and Tolerance Levels* below.

5.4 Worked Example for Age Estimation Measurement

A worked example of these measures is given below based on a pseudo data set made up of 300 samples aged between 14 and 18.

Firstly, the errors and absolute errors are calculated for each of the three samples and plotted using histograms below (with the errors on the left and absolute errors on the right).

FIGURE 1 - HISTOGRAM OF ERRORS FOR AGE ESTIMATION MEASUREMENT



The histogram of the errors illustrates that they are reasonably symmetrical (indicating that there is unlikely to be a bias towards over or under prediction) and the errors range between -12.7 and 10.34.

The histogram of the absolute errors illustrates that these range between 0 and 12.7 with the peak of the distribution greater than 0, but less than 5.

The mean absolute error (MAE) is calculated to be 3.0 years and the median absolute error (MEDAE) 2.6. These two measures are similar as there are no large outliers within the data set.

To illustrate the impact of outliers on the mean if we added another three observations with errors greater than 25 the MAE changes to 3.3 but the MEDAE remains 2.6. The impact of the outlier is therefore not too large, and the data set would have to suffer from very large outliers to suggest that the MAE was not reliable.

The 95% confidence interval of the MAE is [2.8, 3.3] indicating that the MAE estimate is reasonably precise with a margin of error of 0.35 years. The standard deviation of the absolute errors is 2.3 years and 95% of the absolute errors lie between 1.2 and 8.5 years. The 95% prediction interval for the absolute error of an individual is [0, 7.5].

6. Approaches to Measurement of Binary Age Assurance

This section identifies the measures applicable to binary age assurance outputs, where there is a positive declaration with only two possible states - 'yes' or 'no'.

Binary age assurance techniques are the output of posing a question to which there are only two possible answers to a question- e.g. Is this person aged over 18? Yes, or No. These approaches are more generally associated with age verification methods using access to information, data, documents or records to gain a level of confidence in the truthfulness of the binary outcome.

It includes statistical analysis of age verification techniques. It explores only the efficacy of measurement. Issues of equality, comparability and repeatability are discussed later in the report.

6.1 Age Verification

The objective of age verification is to identify whether a person is:

- Scenario 1: Over an age threshold (e.g., 13 or 18) to stop access to age-inappropriate products/materials/services.
- Scenario 2: Under an age threshold to access safe places where no adults are allowed for safeguarding issues (except for appointed safeguarding monitors).
- Scenario 3: Between one specified age and another. In the ICO's Children's Code, these are pre and early-literacy (0-5), core primary school years (6-9), pre-teen years (10-13) and transition to adulthood years (14-17), and adults (18+) to access services in each age group, but they could be any categorisation of age.

A technology may simply provide verification alone or could be an age estimation technology that applies the age threshold to the estimated age. In either case, the outcome is binary as follows:

- Scenario 1: A person is identified as over the age threshold (positive) or under (negative).
- Scenario 2: A person is identified as under the age threshold (positive) or over (negative).
- Scenario 3: A person is identified as within the specified age range (positive) or outside (negative).

As such, the measures to assess accuracy must be tailored to a binary outcome. For those technologies that produce a continuous outcome, the accuracy measures defined in the age estimation section can be applied.

Measures that can be used to assess the accuracy of the age verification technology are defined below. Scenario 1 is used to define and illustrate these measures, but they are applicable to all three scenarios described above. The measures are all based around the confusion matrix[14] below that gives the frequency of the results according to the observed and predicted age thresholds of a sample or training data set.

TABLE 4 – CONFUSION MATRIX DESCRIBING THE PERFORMANCE OF THE AGE VERIFICATION TECHNOLOGY

		Predicted	
		Positive: Over Threshold	Negative: Under Threshold
Observed	Positive: Over Threshold	True Positives (TP)	False Negatives (FN)
	Negative: Under Threshold	False Positives (FP)	True Negatives (TN)

- True Positives: the number of samples that are both observed and predicted to be over the threshold (i.e., the number of samples correctly classified as over the threshold).
- True Negatives: the number of samples that are both observed and predicted to be under the threshold (i.e., the number of samples correctly classified as under the threshold).
- False Positives: the number of samples that are observed to be under the threshold but predicted to be over it (i.e., the number of samples incorrectly classified as being over the threshold).
- False Negatives: the number of samples that are observed to be over the threshold but predicted to be under it (i.e., the number of samples incorrectly classified as being under the threshold).

In an ideal scenario, all samples would either be true positives or true negatives, which means that no sample had been incorrectly classified.

Possible measures to assess accuracy are defined below.

[14] In the field of machine learning and specifically the problem of statistical classification, a confusion matrix, also known as an error matrix, is a specific table layout that allows visualisation of the performance of an algorithm, typically a supervised learning one.

TABLE 5 - MEASURES APPLICABLE TO AGE VERIFICATION TECHNOLOGIES

Measure	Definition	Meaning/Notes
True Positive Rate (TPR) **Also known as: Sensitivity, Recall, or Probability of Detection**	$TPR = \frac{TP}{TP + FN}$	The sensitivity is the technology's ability to correctly detect people who are over the age threshold. It is the proportion of the sample who have been predicted as being over the age threshold among those who are over the age threshold.
True Negative Rate (TNR) **Also known as: Specificity or Selectivity**	$TNR = \frac{TN}{FP + TN}$	The specificity is the technology's ability to correctly detect people who are not over the age threshold. It is the proportion of the sample who have been predicted as being under the threshold among those who are under the age threshold.
False Positive Rate (FPR) **Also known as: Fall-Out or Probability of False Alarm**	$FPR = \frac{FP}{FP + TN}$	The false positive rate is the technology's probability of false alarm (i.e., incorrectly identifying someone as being over the age threshold). It is the proportion of the sample who have been predicted as being over the threshold among those who are not over the age threshold.
False Negative Rate (FNR) **Also known as: Miss Rate**	$FNR = \frac{FN}{TP + FN}$	The false negative rate is the technology's miss rate (i.e., incorrectly identifying someone as being under the age threshold). It is the proportion of the sample who have been predicted as being under the threshold among those who are over the age threshold.
Accuracy	$ACC = \frac{TP + TN}{TP + TN + FP + FN}$	The accuracy is the proportion of the sample that have been

Measure	Definition	Meaning/Notes
		correctly classified as being over or under the age threshold. Note assumes that the balance between samples of over and under the age threshold is reasonable.
Positive Predictive Value (PPV) **Also known as: Precision**	$PPV = \frac{TP}{TP + FP}$	The PPV is the proportion of the sample correctly identified as being over the age threshold given that they have been predicted as being over the age threshold.
Negative Predictive Value (NPV)	$NPV = \frac{TN}{TN + FN}$	The NPV is the proportion of the sample correctly identified as under the age threshold given that they have been predicted as being under the age threshold.
False Discover Rate (FDR)	$FDR = \frac{FP}{FP + TP}$	The FDR is the proportion of the sample incorrectly identified as over the age threshold given that they have been predicted as being over the age threshold.
False Omission Rate (FOR)	$FOR = \frac{FN}{FN + TN}$	The FOR is the proportion of the sample incorrectly identified as under the age threshold given that they have been predicted as being under the age threshold.
Positive Likelihood Ratio (LR+)	$LR+= \frac{TPR}{FPR}$	The positive likelihood ratio is the value in performing the test. It is the ratio of the true positive and false positive rates. The greater the value over 1 indicates the greater the probability that a positive test

Measure	Definition	Meaning/Notes
		result is evidence that the person is over the age threshold.
Negative Likelihood Ratio (LR-)	$LR- = \frac{FNR}{TNR}$	The negative likelihood ratio test is the value in performing the test. It is the ratio of false negative and true negative rates. The closer the value to 0 the greater the probability that a negative test result is evidence that the person is under the age threshold.

Ideally, a technology would correctly classify all persons (i.e., 100% accuracy).But this is unrealistic. It is important that, based on the implications of an incorrect classification, technology minimises false positives which are defined as follows for each scenario:

- Scenario 1 False Positives: those under the age threshold are incorrectly classified as over it allowing access to age-inappropriate content.
- Scenario 2 False Positives: those over the age threshold are incorrectly classified as under it allowing adult access to safe spaces causing safeguarding issues.
- Scenario 3 False Positives: those outside of the age range incorrectly classified as within it allowing access to content tailored to a different age group.

In all cases false positives have the potential to cause harm (particularly in scenarios 1 and 2). False negatives should be minimised where possible (e.g., in scenario 1 where someone over the age threshold has been identified as under), but these are less critical since they result in inconvenience (and potential economic consequences if it results in users abandoning the technology) rather than harm. Therefore, when assessing the above measures, it is important to note that false positives are more critical to minimise.

It is worth noting that in all the metrics above, 95% confidence intervals could be calculated to quantify their uncertainty. However, if the sample size has been correctly calculated (with inputs that are aligned to the deployment and expected outcomes of the technology) then the confidence intervals of the metrics should be close to the margin of error defined in the sample size calculation (for more details, see *section 9.6 Sample size and breakdown* below).

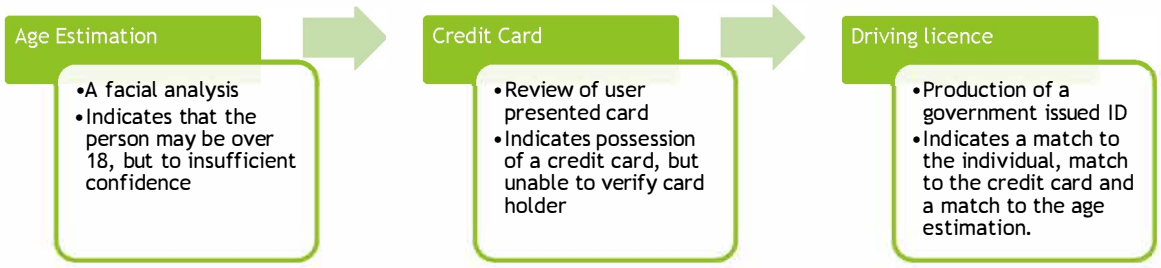
6.2 Age Verification: Waterfall Technique

The waterfall technique for age verification is a breakdown of age assurance activities into linear sequential phases, where each phase depends on the output of the previous one and

corresponds to a series of decisions providing greater or lesser levels of confidence in the age assurance gained from the process.

Some technologies rely on multiple gateways to assess whether a person is, for example, over an age threshold such as 18. At each gateway a new source of information or database is added (for example, electoral register, credit card reference data, mobile phone data etc.). At each gateway a person is assigned as over 18 (positive) or insufficient evidence to identify as over 18 (negative). The technology passes a certain number of gateways until they are confident that those have not been assigned as being over 18 are under 18.

The approach to a 'waterfall technique' is that the cumulative results of the age assurance components are greater than the individual results of each component on its own. The whole is greater than the sum of its parts. This presents a statistical difficulty, which needs to be explored further when considering Trust Frameworks and interoperability. Whilst, in theory, the whole is greater than the sum of its parts, in statistical theory, this propagation of uncertainty results in the errors associated with each part being multiplied together. This fails to recognise the cumulative knowledge gained by the multiple components, so a method of statistically recognising this is required.



Like any other age verification technique, the same binary measures can be applied to the final classifications after a person has reached the last gateway. To assess the accuracy of each individual gateway, the technology can also calculate the overall accuracy at each (assuming those who have not been assigned as over 18 are under 18 as there is no evidence to the contrary) with the expectation that this overall accuracy will improve with each additional gateway added.

A well-designed waterfall technique is privacy protecting, as the sequence of data gathering is directly tailored to the level of confidence sought before the process is completed. However, a poorly designed sequencing can lead to the collection of unnecessary data. It could also potentially be more intrusive and could breach the data minimisation principle of UK GDPR.

6.3 Permutations and Combinations

Whilst the waterfall technique describes the sequential building of sufficient levels of confidence to reach an age assurance decision, there is also the possibility of building multiple sources of age assurance. This could occur in one instance, perhaps through a single age assurance service provider, or over a period of time. This can also result from the use of pass-through rooms in information society services.

Imagine the concept that a user entering a lobby area of an online space controlled by a service provider (such as a social media platform) is permitted to enter that lobby area and to access some of the rooms unrestricted. However, if the user selects to enter rooms in that online space that require age assurance, a process is followed to gain the appropriate level of confidence in the age of that user, which is recorded. If, at a later date, the user then selects to enter another room which requires a higher level of confidence in their age, the original level is recalled (perhaps from a record on the user account or token on a user device) and then elevated by adding additional age assurance to gain the requisite higher level of confidence. In this way, the end-to-end user journey is a series of permutations and combinations of age assurance that evolve over time commensurate with the risk profile of the spaces visited by the user.

One way to describe this could be through a table, as outlined below, showing the options available for the higher levels of confidence discussed in *section 1*.3 *Using "Levels"* and detailed in Table 1 in *section 1.4*. Permutations and combinations are not relevant to the lower levels of confidence (self-asserted and basic).

TABLE 6 - PERMUTATIONS AND COMBINATIONS OF AGE ASSURANCE OUTPUTS

To achieve:	Option 1	Option 2	Option 3	Option 4	Option 5
Standard Level of Confidence	1 x Standard Age Assurance Component	2 x Basic Age Assurance Components	-	-	-
Enhanced Level of Confidence	1 x Enhanced Age Assurance Component	2 x Standard Age Assurance Components	1 x Standard Age Assurance Component **PLUS** 2 x Basic Age Assurance Components	4 x Basic Age Assurance Components	-
Strict Level of Confidence	1 x Strict Age Assurance Component	2 x Enhanced Age Assurance Components	1 x Enhanced Age Assurance Component **PLUS** 2 x Standard Age Assurance Components	1 x Enhanced Age Assurance Component **PLUS** 1 x Standard Age Assurance Component **PLUS** 2 x Basic Age Assurance Components	1 x Standard Age Assurance Component **PLUS** 4 x Basic Age Assurance Components

These permutations and combinations have the same statistical challenge as the waterfall technique. In theory, the whole is greater than the sum of its parts. However, in statistical theory, this propagation of uncertainty results in the errors associated with each part being multiplied together. This fails to recognise the cumulative knowledge gained by the multiple components, so a method of statistically recognising this is required.

This also raises the query of how to address authentication (i.e. binding the age attribute to the user) and whether or not the reliability of the age assurance output diminishes with time. These are questions addressed in *section 8 Approaches to Authentication* below.

6.3 Observations on Age Verification Measurement

The accuracy measure gives a good indication of the overall accuracy of the technology. However, on its own, it does not provide additional information on whether the technology's misclassifications are because of false positives or false negatives (and we know that here false positives are more problematic).

Key points to note are:

1. The results of an age verification assessment can be summarised by a confusion table, which details the four different combinations of possible results (true positives, true negatives, false positives and false negatives).
2. The overall accuracy (proportion of correctly classified samples) is a useful overall measure and should be reported. But in isolation, it does not provide any information on the type of errors that are present (false positives or false negatives).
3. Reporting both the sensitivity (TPR) and specificity (TNR) informs the user about the prevalence of different errors. The greater the sensitivity, the fewer the false negative errors and the greater the specificity, the fewer the false positive errors.
4. Maximising the TNR/minimising the FPR may be more of a priority than the TPR/FNR since false positive errors have the potential to cause harm.
5. Ultimately, the system will be judged on its False Positive Rate (FPR), but this should not be considered on its own without also considering the sensitivity and specificity of the age assurance system.
6. Predictive values are helpful to users of technology. Given that the technology has predicted a result, what is the probability that it is right? In this case maximising the PPV, minimises the FDR (the more critical errors).

6.4 Sensitivity & Specificity

The sensitivity and specificity of the age assurance component is a crucial element of understanding the overall efficacy of the system

- A high sensitivity (TPR) means that the technology will rarely misclassify those who are over the age threshold. The false negative rate is 1-TPR.

- A high specificity (TNR) means that the technology will rarely misdiagnose those who are under the age threshold. The false positive rate is 1-TNR.

Based on the above, the primary aim of the technology is to maximise TNR (and therefore minimise FPR). Of course, one way to have a 100% TNR and 0% FPR, is to assign everyone as under the age threshold (using scenario 1 as an example), but of course this is not practical. Therefore, there must be a trade-off between sensitivity and specificity, but the weighting to specificity is higher. Further options for tolerances are provided in *section 9.10 Regulatory Options and Tolerance Levels* below.

6.5 Predictive Values

The predictive values are likely to be helpful to users of the technology. Sensitivity and specificity condition on the true outcome, e.g., given the true outcome, what is the probability that the technology got the classification right? However, when the technology is being used, the true age of the person is unknown and therefore we need to ask: given that the technology says the person is over the age threshold, what is the probability that is correct? Both PPV and NPV are important, but maximising PPV is imperative; the probability that someone who is classified as being over the age threshold is over the age threshold. Maximising PPV by default minimises the False Discover Rate (FDR) since PPV = 1 - FDR and we want to reduce the chance of a false discovery (the probability that someone who is identified as being over the age threshold but is in fact under it).

6.6 Information Retrieval

Information retrieval/AI often focus on precision (sensitivity) and recall (PPV), but these measures do not consider true negatives and therefore could bias predictions if they are the only focus. In information retrieval, the number of true negatives is unknown and much larger than the true positives; this does not hold in this application.

6.7 Age Buffer

It is likely that the misclassification rate will be higher for those persons who are closest to any age thresholds. For example, if the technology is estimating whether a person is over 13 or not, it is likely to be more accurate at classifying people who are 10 years or younger or 16 years and over, compared to someone who is 12. Therefore, it is not uncommon for users to apply an age buffer to a threshold. For example, if the age at which a person has access to services is 13, the application of an age threshold of 16 will increase confidence that those who are identified as above 16, are indeed over 13. This is illustrated in the Challenge Age scheme that asks individuals to prove they are over 18 if they look under 21 or 25.

In some circumstances, it may be appropriate to implement an age buffer in relation to electronic age assurance, particularly when applying the tolerance levels as set out in *section 9.10 Regulatory Options and Tolerance Levels*. In our view, however, the setting of an age buffer is only really relevant where there is a statutory penalty for non-compliance, such as for the sale of alcohol, weapons, tobacco, etc to under 18s. In these circumstances, the law requires retailers to take all reasonable precautions and exercise all due diligence to avoid the commission of the offence.

In the *London Borough of Enfield v Argos Ltd*[15], Moses LJ said:

> *"I would regard a policy that required members of staff to aim considerably higher than the age prohibited by statute as a reasonable and sensible one."*

[15] [2008] EWHC 2598

This case was centred on the actions of human beings in the assessment of age, whereas the actions of machines ought to be assessed around the continuous and/or binary outcomes of the age assurance process as set out elsewhere in this report.

6.8 Extension to Binary accuracy measures

For those cases where the technology estimates a person's age and age threshold is applied, it is possible to further explore how close to the age threshold an incorrect classification is. For example, in scenario 1, one possible incorrect classification is classifying someone who is under the age threshold as over. If the age threshold is 18 and the person who is incorrectly assigned as over 18 is 17, this may be deemed as less of a failure than someone who is 13 and incorrectly misclassified as over 18.

To measure the size of failures in these instances, the measures defined to assess age estimation are also appropriate here, but rather than comparing the predicted age with the true age, we compare the predicted age with the missed age threshold to better understand how close to this threshold the technology was.

For example, given the following parameters:

- The true (or observed) age of sample i :.
- The predicted age of sample i :.
- The age threshold: T_A.

The false positive absolute error for sample i can be calculated as:

$$FPAE_i = \begin{cases} |p_i - T_A| & if\ p_i > T_A\ and\ o_i < T_A \\ 0 & otherwise \end{cases}$$

It is possible to then calculate, for example, the mean false positive absolute error over all false positive results.

6.9 Worked Example for Age Verification Measurement

A worked example of measures is given below based on a pseudo data set made up of 300 samples aged between 14 and 22 each assessed as either over or under an age threshold of 18.

TABLE 7 – WORKED EXAMPLE FOR AGE VERIFICATION MEASUREMENT

		Predicted		
		Positive: Over 18	Negative: Under 18	
Observed	Positive: Over 18	118	27	145
	Negative: Under 18	34	121	155
		152	145	300

The overall accuracy of the data set is 79.7% since 61 of the 300 samples are incorrectly classified. Approximately 8 in 10 samples are correctly classified.

The sensitivity (or TPR) is 81.4% and the specificity (or TNR) is 78.1%, which suggests that the proportion of misclassifications over and above the threshold are similar (34 vs. 27 of the sample), but slightly higher for positive classifications (I.e., correctly classifying those who are over the age threshold). The FPR is 21.9% which indicates that 21.9% of the samples who are under the age threshold have been incorrectly predicted to be over the threshold.

The predictive values assists in understanding the probability of a classification being correct, given a sample has been predicted to be either over or under the threshold. The PPV (the proportion of the sample correctly identified as being over the age threshold given that they have been predicted as being over the age threshold) is 77.6% and the NPV (the proportion of the sample correctly identified as under the age threshold given that they have been predicted as being under the age threshold) is 81.7%. This means that if a sample is predicted to be over the age threshold, they have a 77.6% chance of being over the age threshold compared to an 81.7% chance of being under the age threshold if the sample has been predicted as being under. Ideally, the PPV would be higher as it means that 22.4% of samples who are predicted to be over the age threshold will not be (the false discovery rate).

7. Issues of Equality, Parity and Fairness

This section covers the risks involved in the collection of personal data, including too much data, and what considerations age assurance service providers should keep in mind when balancing accuracy versus compliance with legislative requirements.

As explained earlier in this report one of the pillars of accuracy is efficacy, which is the ability to perform a task to a satisfactory degree. For age assurance providers to provide their services to any degree, whether that be age estimation or verification, they require personal data. The greater the need for accuracy, the greater the quantity or sensitivity of personal data is needed.

The more personal data that is held increases the importance of ensuring that an organisation has appropriate security measures in place to protect personal data. This is especially the case with an increase in the amount of special category data[16].

Age assurance service providers and their relying parties need to be mindful of how much personal data is processed as holding too much data risks breaching the UK GDPR data minimisation requirements. So, in UK GDPR terms, when an age assurance service provider has gathered enough data to provide an age assurance output to the requisite level of confidence, gathering further data would be considered unnecessary (and therefore, unlawful) data processing. Taking a 'belt and braces' approach to data gathering for age assurance purposes to gain a higher level of confidence than that identified as necessary is, of itself, unlawful.

An example of excess data collection for age assurance could be scanning of a government issued identity document:

A UK driving licence contains the date of birth in two locations - at line 3 and within the machine-readable code at line 5.

However, it also contains other personal information not necessarily relevant to the age assurance process - such as home address, vehicle licence categories, etc.

[16] Article 9 of UK GDPR imposes additional conditions on processing special category data, which includes data revealing racial or ethnic origin, genetic data, biometric data (where used for identification purposes) and data revealing a person's sex life or sexual orientation (among other things).

7.1 Legislative framework

First, there is a need to understand the legislative framework to underpin age assurance. The UK GDPR comprises of six principles that help to set out the framework for the handling of personal data. In relation to age assurance the key reference is set out below:

Article 5(1) of UK GDPR states:

> *"1. Personal data shall be:*
>
> *(a) processed lawfully, fairly and in a transparent manner in relation to the data subject.*
>
> *(b) collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes; further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes.*
>
> *(c) adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed;"*

Processing personal data in a fair, lawful and transparent manner in the context of age assurance means that providers of age verification or estimation services must have identified their lawful basis for processing and that they do not process personal data outside of that lawful basis(es). Even if the provider is a data processor, they must ensure that personal data is not processed outside of the lawful basis(es) as stipulated by the data controller. They must ensure that the processing of personal data is fair and done in a way that individuals would expect. Providers should communicate the processing of personal data with individuals in a clear and transparent manner.

Collected for specified, explicit and legitimate purposes in this context means age assurance providers must only use personal data for the purposes for which it was originally gathered and not used for any other purpose without legitimate reason for doing so. Legitimate purposes could include sharing personal data with, for example, law enforcement agencies.

Adequate, relevant and limited to what is necessary, otherwise known as 'the minimisation principle' requires that the amount of personal data being processed is enough to fulfil their purpose and nothing extra. Data minimisation should be a key consideration for age assurance as a balance needs to be struck between having enough personal data to confidently predict age but not holding or processing (even if instantly deleted) too much personal data.

It is also important to highlight at this stage what the processing of personal data means.

Article 4 of the UK GDPR sets out the definitions and Article 4(2) states:

> *"processing means any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration,*

> *retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction."*

In other words, 'processing' when used in the context of personal data means any action that is taken with that data either by human intervention or via a virtual technical process.

7.2 Security requirements

The UK GDPR also sets out the security requirements for the processing of personal data. This includes the appropriate technical and organisational measures needed to secure the personal data being utilised for the age assurance process. This becomes even more important when processing data between a relying party and an age assurance service provider or in a Trust Framework, between age assurance service providers.

Article 32 states:

> *"1. Taking into account the state of the art, the costs of implementation and the nature, scope, context and purposes of processing as well as the risk of varying likelihood and severity for the rights and freedoms of natural persons, the controller and the processor shall implement appropriate technical and organisational measures to ensure a level of security appropriate to the risk, including inter alia as appropriate"*

Technical security measures relate to cyber security. It is the steps an organisation has put in place to protect its cyber environment, such as regular penetration testing, regular back-up process, anti-virus and malware protection, effective wiping and disposal of hardware etc.

Organisational security measures relate to personnel and physical security. People are often the biggest risk when it comes to protecting personal data and it is the organisation's responsibility to ensure that all staff receive regular data protection training as is appropriate for their role. Physical security is about making sure that all entry and exit points, especially where members of the public could be, are locked down using appropriate security measures.

It is important to consider what measures are appropriate for the size of the organisation and amount and sensitivity of the personal data being processed. It is unreasonable, for instance, for a start-up business that provides basic age estimation services as a processor to have the same security measures in place as a global organisation processing millions of age verifications.

7.3 Equalities

The Equality Act 2010 protects people against discrimination, harassment or victimisation in employment, and as users of private and public services based on nine protected characteristics:

- age,
- disability,
- gender reassignment,

- marriage and civil partnership,
- pregnancy and maternity,
- race,
- religion or belief,
- sex, and
- sexual orientation.

Although measures taken in relation to age related eligibility to receive goods and services through gaining age assurance is specifically permitted by the Act, actions taken during that process that treat people with different protected characteristics in different ways or producing different outcomes, would still be prohibited by the Act.

Age assurance techniques can result in bias and discrimination towards a sub-group of the population unintentionally and unknowingly. Some causes of potential bias include:

- Behavioural bias: Data-driven technologies can reproduce, reinforce, and amplify inequality and discrimination present in society.
- Generative models: the use of adversarial networks, variational autoencoders and automatically regressive models to stimulate and predict future outputs of the neural network.
- Representational bias: A lack of representation in the data from the population of interest.
- Sample bias: insufficient sample size for the population of interest.
- Corrective bias: where a deliberate action is taken to correct an inherent bias, but that itself leads to an over-correction or a correction that is not monitored or reviewed over time.
- Natural phenomenon: are things that occur in nature, such as biological processes, aging, genetics, physical processes, wave propagation, as examples.

As an example, a face analysis system relies upon an image of a face taken through a camera lens and then the processing of that data (albeit instantly) to make an estimation of age. If the data received by the camera (the capture device) is deficient, this will affect the efficacy of the age estimation. There are many reasons why this may be the case. It may be a poor-quality camera, a poor connection, poor ambient lighting (see more in ***section 9.4 Ambient Lighting*** below) or inappropriate positioning of the camera.

However, there are at least two other potential factors at play that could influence how that camera sees and, consequently, how the age assurance system processes, data about people with darker skin.

- The data that is used to train artificial intelligence may itself be deficient, or under-representing data subjects with darker skin, so the neural network and algorithm has less available data to conduct its analysis against - this is an example of representational bias.
- The properties of light indicate that less light is reflected from a darker surface than a lighter surface, so the camera has less data to examine from a darker surface (a problem fixed by increasing levels of ambient light) - this is an example of natural phenomenon.

What this means in the real world is that, say an age estimation system is deployed at a kiosk to determine if a customer looks over 25 and, if not, to prompt for the customer to provide a form of ID. If, all other things being equal, the age estimation system is deficient in analysing people with darker skin, that will mean that it could be more likely to default prompting for that customer to produce ID. As a result of that inherent (and possibly natural phenomenon caused by the properties of light), that customer faces a direct discrimination based on the colour of their skin (or race, as the legally protected characteristic).

As such, the development and deployment of age assurance systems must consider the potential direct and indirect discrimination potentially caused by the system itself.

It is worth noting that there are other potential harms of an AI technology as set out by the Alan Turing Institute's guide to understanding artificial intelligence ethics and safety[17] such as denial of individual autonomy, recourse and rights or invasion of privacy. In this report, however, the focus is on the potential for bias and discrimination.

Following the Alan Turing Institute's guidelines, an assessment of fairness is recommended, which is based on the principal of discriminatory non-harm "No harm to others through the biased of discriminatory outcomes that may result from practices of AI innovation." The guidelines identify four forms of fairness:

- Data Fairness
- Design Fairness
- Outcome Fairness
- Implementation Fairness

Outcome fairness is the best measure to quantifiably assess how a technology owner has implemented all four forms of fairness and one method to do so is to ensure that error rates are equitably distributed across different subgroups of the population.

For continuous (age estimation) techniques that produce a continuous outcome, error parity is similarly the focus but in this case the measures include:

- Mean Absolute Error Parity: ensuring that the overall accuracy of the technology is equivalent between different population subgroups.

- Mean Error Parity: ensuring that the technology is not biased towards over or under prediction for different population subgroups.

For binary (age verification) techniques that produce a binary outcome, measures include:

- True Positive Parity: ensuring that the accuracy of the technology is equivalent between different population subgroups. Also known as 'equal opportunity' fairness.

- False Positive Parity: ensuring that the error rate of the technology is equivalent between different population subgroups.

[17] https://www.turing.ac.uk/research/publications/understanding-artificial-intelligence-ethics-and-safety

- Positive Predictive Value Parity: ensuring that the precision of the technology is equivalent between different population subgroups.

- In practice the accuracy or error rates for a technology will never be the same across different population subgroups due to the inherent variability of the technologies. Defining what an acceptable difference between these measures for subgroups to accept parity between the subgroups is one that must be defined by regulators. We discuss this further when we propose tolerances in *section 9.10 Regulatory Options and Tolerance Levels* below.

It must be identified which protected characteristics are at risk of bias or discrimination and therefore error parity 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). Investigating intersectionality is more difficult since there are likely to be many combinations to consider and the sample size within each combination will be small.

To investigate error parity fully, ideally there would be the equivalent sample size in each population subgroup as the size recommended for the full subgroup so that the estimate for each subgroup is estimated with the same level of confidence and margin of error. This is unlikely to be possible, but it is important to ensure that each subgroup has a reasonable sample size. To understand the impact of different sample sizes on the margin of error and level of confidence, see the sample size illustrations in *section 9.6 Sample size and breakdown*.

In the preparation of ACCS 1:2020 - Technical Requirements for Age Estimation Technologies, ACCS consulted with the Equalities and Human Rights Commission (EHRC). In that discussion, the EHRC highlighted that the error parity had to show a negligible difference between protected characteristics, but stated that there was no definition of 'negligible' and ultimately, it was for the Courts to determine. In ACCS 1, a difference of up to 0.25 years was established as being acceptable.

In 2020, the UK Government, through the Centre for Data Ethics and Innovation (CDEI) undertook a review into bias in algorithmic decision making[18]. That review concluded that:

- Regulation can help to address algorithmic bias by setting minimum standards, providing clear guidance that supports organisations to meet their obligations, and enforcement to ensure minimum standards are met.

- AI presents genuinely new challenges for regulation, and brings into question whether existing legislation and regulatory approaches can address these challenges sufficiently well. There is currently little case law or statutory guidance directly addressing discrimination in algorithmic decision-making.

- The current regulatory landscape for algorithmic decision-making consists of the EHRC, ICO, and sector regulators and non-government industry bodies. At this stage, we do

[18] Review into bias in algorithmic decision-making - GOV.UK (www.gov.uk)

not believe that there is a need for a new specialised regulator or primary legislation to address algorithmic bias.

- However, algorithmic bias means that the overlap between discrimination law, data protection law and sector regulations is becoming increasingly important. This is particularly relevant for the use of protected characteristics data to measure and mitigate algorithmic bias, the lawful use of bias mitigation techniques, identifying new forms of bias beyond existing protected characteristics, and for sector-specific measures of algorithmic fairness beyond discrimination.

- Existing regulators need to adapt their enforcement to algorithmic decision-making, and provide guidance on how regulated bodies can maintain and demonstrate compliance in an algorithmic age. Some regulators require new capabilities to enable them to respond effectively to the challenges of algorithmic decision-making. While larger regulators with a greater digital remit may be able to grow these capabilities in-house, others will need external support.

In *section 9.10 Regulatory Options and Tolerance Levels* below, we propose some tolerances for outcome error parity, but ultimately it is for Regulators to propose, consult on and then issue guidance on appropriate tolerances. There is no right or wrong answer, save that having no set or expected tolerance is both unachievable (in data and statistical terms - distribution of results may 'tend to zero' but never actually quite get to zero) and unwelcome (in public policy terms, having no upper expectation of tolerance may result in undesirable outcomes for those with protected characteristics).

8. Approaches to Authentication

This section addresses the issue of how you bind the claimed age attribute to the person that age attribute is about.

Authentication is not, in itself, the process of gaining age assurance. However, it is a critical part of the overall journey of a user when undertaking age assurance for the first time or seeking to re-use a previously verified age for a future or different process.

A relying party may need to know if someone has already proven their age to the requisite level of confidence before they are granted access to the service again or before they are granted access to a new service. This is called 'authentication' and can be useful if users need to sign in to the service more than once.

An authenticator could be some information (like a password), a piece of software or a device.

There are different types of authenticators. An authenticator will usually be one of the following:

- something the user knows
- something the user has
- something the user is

Sometimes an authenticator can fit into more than one of these categories.

8.1 Something the user knows

The most common way for users to sign in to a service is by entering a piece of information that only they know. This is called a 'secret'.

A secret could be something like:

- a PIN
- a password
- an answer to a question that only the user knows the answer to - also called knowledge-based verification (KBV)

8.2 Something the user has

A user might be able to sign in to a service using something called a 'token'. A token can be something physical, like a chip and PIN card or a mobile phone. A token can also be something digital, like a single use authentication code or a digital certificate. This could include an age attribute to be stored on the user's device.

Using a token by itself might not be appropriate if the service needs a high level of protection. This is because tokens can be easily lost, stolen or shared.

Some tokens can contain information about:

- the person that is using it to sign in to the service
- the organisation that issued the token (for example, the age assurance service provider that issued the initial token)

8.3 Something the user is

A user might be able to sign in to a service using their biometric information. Biometric information is a measurement of someone's:

- biological characteristics, such as their fingerprint or face
- behavioural characteristics, such as their signature

The app or device may use facial recognition software to check the user looks like the person who created the account or registered the phone. If there is a match, the user can access the service.

There is a chance someone could try to impersonate another user by recreating their biometric information. For example, they could:

- hold up a photo of the user
- wear prosthetics or a mask to make themselves look like the user
- play a recording of the user's voice
- use a copy of the user's fingerprint

Some types of biometric information will be easier to recreate than others. These are called 'presentation' or 'spoofing' attacks. Although attacks can be detected by the system that is used to capture biometric information, there is always a risk that a fraudster could successfully sign in to a service this way. We discuss this again *in section 9.2 Presentation Attack Detection* below.

It is also possible that the system can make a mistake when it is matching someone's biometric information. It could either:

- wrongly match a user to another person (called a 'false match')
- not be able to match a user to anyone, even though a record of their biometric information exists (called a 'false non-match')

This all gives rise to the question of whether, or not, in the measurement of age assurance technologies, it is also necessary and relevant to measure the efficacy of authentication used. If it is, we suggest that reference should be drawn from ISO/IEC 29115:2013, Information technology - Security techniques - Entity authentication assurance framework for an appropriate approach to securing levels of authentication (LoA).

9. Approaches to Testing, Analysis and Certification

This section describes the approach to testing of age assurance systems, including how to identify the appropriate test protocols, test subjects, environmental considerations (particularly ambient lighting) and capture devices. The section explores spoofing (more formally known as presentation attack detection) for both the data subject and for documentation presented to age assurance systems. An approach to sample size calculation is recommended together with exploring the appropriate depth of evaluation to be conducted by independent 3rd party conformity assessment bodies. Finally, the section underlines the tasks and powers of the ICO to maintain oversight and approval of certification criteria under Article 42 of UK GDPR.

9.1 Test Protocols

Any test laboratory or conformity assessment body should have appropriate test protocols in place to secure effective, repeatable testing of the target of evaluation - the system that is under test. Test protocols should describe the capture methodology setting out the subjects, devices and environmental circumstances that will be used to present the test to the target of evaluation.

The capture subject[19] describes the individual who is going to be subject to the age verification, categorisation or estimation process. A conformity assessment body may use members of a test crew, who are real people with true identities - called bona fide identity subjects - or they could use a series of simulated identities which have existed over time (i.e., may have been used in tests previously) - called Avatars. In International Standards they are referred to as subversive capture subjects[20]. It may not be necessary to utilise a real or simulated identity depending on the Target of Evaluation.

The presentation of a capture subject should also record the facial orientation - typically at indices up to 15º of centre, between 15º and 30º of centre and greater than 30º of centre. By default, <15º of centre should be used as test methodology. It is important to note, however, that the operational capability of age assurance technologies may need testing at much wider orientations - for instance, some are still designed to be effective at 90ºorientation (i.e., a profile shot of the subject.

The capture device[21] is the equipment or system that we are going to utilise to collect the signal from the capture subject to perform the test. A capture device could be:

[19] ISO/IEC 2382-37:2017 - Information technology — Vocabulary — Part 37: Biometrics, 3.7.3
[20] *Ibid*, 3.7.17
[21] *Ibid*, 3.4.1

- Integrated/Purpose Built in an age assurance technology
- A smart device or connected device (like a mobile phone)
- A Web Camera
- A Microphone/Telephone (Audio Only Testing)
- A Scanner

9.2 Presentation Attack Detection

Presentation attack detection is the process of determining if an Age Assurance system is susceptible to being 'spoofed'.

This can involve the presentation of attack instruments such as:

- Pseudo Identities
- Mannequins
- Masks
- False Identity Documents
- False Instruments
- Tamper Evident Instruments
- Genuine Instruments that have been amended
- Disfigured Instruments

Biometric presentation attack is set out in international standards BS ISO/IEC 30107-3:2017 - Information technology — Biometric presentation attack detection - Part 3: Testing and Reporting.

When a non-living object that exhibits human traits (an "artifact") is presented to a camera or biometric sensor, it is called a "spoof." Photos, videos on screens, masks, and dolls are all common examples of spoof artifacts. When biometric data is tampered with post-capture, or the camera is bypassed altogether, that is called a "bypass." A deepfake puppet injected into the camera feed is an example of a bypass. There are no NIST/NLVAP lab tests available for PAD Level 3, or Levels 4 & 5 bypasses, as those attack vectors are missing from the ISO 30107-3 standard and thus all associated lab testing.

TABLE 8 - PRESENTATION ATTACK DETECTION - ARTEFACT TYPES

Artefact Type	Description
Level 1	Hi-res paper & digital photos, hi-def challenge/response videos and paper masks.
Level 2	Commercially available lifelike dolls, and human-worn resin, latex & silicone 3D masks
Level 3	Custom-made ultra-realistic 3D masks, wax heads, etc
Level 4	Decrypt & edit the contents of a 3D FaceMape to contain synthetic data not collected from the session, have the Server process and respond with Liveness Success.
Level 5	Take over the camera feed & inject previously captured video frames or a deepfake puppet that results in the AI responding with "Liveness Success.ë

More recently, the European Union Agency for Cyber Security (ENISA) has published an analysis of threats to remote identity proofing systems[22]. It highlights attacks that are very viable, yet still are not acknowledged by the ISO 30107-3 PAD standard. These attack vectors include Level 4 & 5 attacks, like Deepfake Puppets and Video Injection.

As age assurance systems become more broadly deployed through information society services, it will be necessary to continuously review and address threats associated with both simple presentation attack, but also much more sophisticated attacks which will become prevalent and easily accessible to young people seeking to circumvent age assurance systems.

9.3 Document Authenticity

Presentation attacks utilising false identity documentation or records are affected by the assessment of the capability to detect documents and extract age attributes.

For authentication, documents should be classified (scored) according to their inherent features that are designed to provide detectable security features.

TABLE 9 - CLASSIFICATION OF DOCUMENT AUTHENTICITY SECURITY FEATURES

Document Type	Description
Tier 1	No material security features available, and no fraud evaluation can be performed. Extraction only. Documents in this tier sometimes include hand-written documents.
Tier 2	A low security document where only basic fraud checks can be performed and confidence in authenticity (based on a digital photo) is low. • No cross-comparison possible due to missing Machine-Readable Zone (MRZ) or barcode; and/or • The documents may not have consistent template format and/or fonts.
Tier 3	Documents in this tier lack advanced security features and are easier to execute fraud attacks, but still carry sufficient security features to enable automated verification using data cross-comparison, checksums, and other logical checks. Documents have a consistent template format and font within a version. Documents in this tier SHALL Include one or more of the following features: • machine readable zone (MRZ) • barcode Tier 3 also SHOULD meet requirements for Tier 2.

[22] Remote ID Proofing — ENISA (europa.eu)

Document Type	Description
Tier 4	Documents of this tier are highly secured documents with state-of-the-art security features. Documents in this tier SHALL include one or more of the following technologies: • optically variable ink (OVI), holograms, watergrams • guilloche (e.g., intricate and subtle patterns of thin interwoven lines) • tactile laser engraving • micro printing • ghost image Tier 4 also SHOULD meet requirements for Tier 3.
Tier 5	Documents of this tier are highly secured documents with state-of-the-art security features. Documents in this tier SHALL include one or more of the following technologies: • embedded chip technology (e.g., contact card, RFID, NFC)

9.4 Ambient Lighting

It is important to note that the performance of electronic detection devices, such as smartphone cameras, webcams or scanners, are susceptible to diminished performance in different ambient lighting conditions.

The ambient lighting can have a significant impact on the efficacy of the data capture, so tests should be carried out under controlled lighting conditions. The lighting can be directed ambient to the presentation object (i.e., the person being age estimated) or the detection device (i.e., the camera) or both.

The following ambient lighting choices should be considered:

- Bright LED Gantry (such as may be found in a retail shop) - around 700 lux
- Sodium Low Level (such as may be found in a pub or restaurant) - around 70 lux
- Strobe Lighting (such as may be found in an entertainment venue)
- Ultraviolet Lighting (such as may be used in a scanner detection devices)
- Multi Colour Lighting (such as may be emitted by a gaming machine)
- Outdoor Daylight
- Outdoor Nightlight

In addition to the effect of lighting on the presentation object, there can be adverse effects of lighting[23] on the detection device, caused by issues like:

- Glare - which occurs when one part of the visual field is much brighter than the average brightness to which the detection device is adapted
- Colour effects - which occurs when the detection device is lit by different artificial light sources, or by daylight under changing sky conditions, may appear to vary in colour

[23] ISO 8995-1:2002 - Lighting of workplaces — Part 1: Indoor

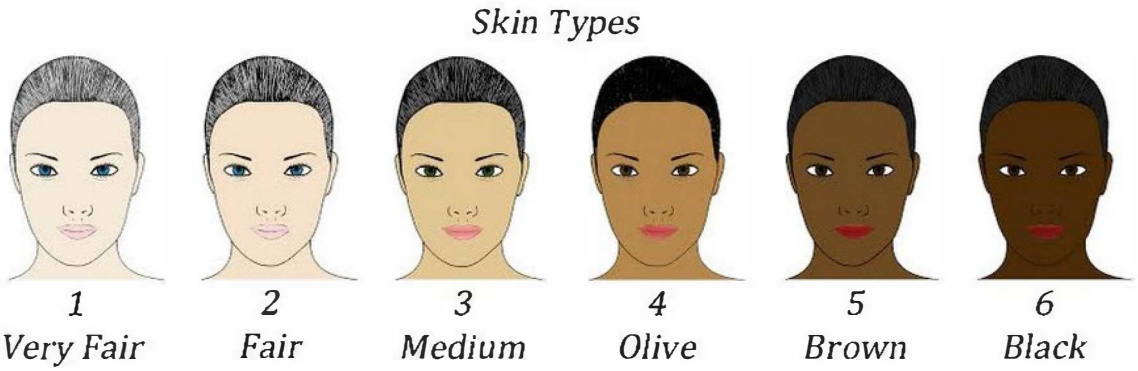
- Under monochromatic light sources - such as low-pressure sodium discharge lamps, colours will not be identifiable a detection device may not perform properly
- Stroboscopic effects - can confuse detection devices. When the magnitude of these oscillations is great, Presentation Attack Instruments will appear to be stationary or moving in a different manner. This is called the stroboscopic effect.
- Flicker - Light modulation at lower frequencies (about 50 Hz or less) which is visible to most people, is called flicker. Detection Devices can be sensitive to flicker, and it is especially detectable at the edges of the visual system's field of view.
- Veiling reflections - are high luminance reflections which overlay the detail of the Presentation Attack Instrument. Such reflections may be sharp-edged or vague in outline, but regardless of form they can affect Detection Device performance.
- Infrared and ultraviolet radiation - Some lamp designs also produce significant emissions at infrared and ultraviolet wavelengths, both of which are invisible; some Detection Devices also rely upon Infrared and ultraviolet radiation.

We do not believe that ambient temperature, humidity, pressure or other climatic conditions have a material impact on the efficacy of the Target of Evaluation.

9.5 Data subject skin tone

Biometric age estimation systems can be adversely affected by inherent skin tone bias. This is all dependent on the range of training images that are used. We utilise the Fitzpatrick Scale[24] 1 - 6 to determine the skin tone of our presentation attack assets. All our assets are assigned a skin tone score.

TABLE 10 - FITZPATRICK SCALE OF SKIN TONE TYPES



9.6 Sample size and breakdown

To calculate a sufficient sample size when testing an age estimation or verification technology, the objective of the assessment must be defined. This would typically reflect how the

[24] Fitzpatrick, T. B. (1975). "Soleil et peau" [Sun and skin]. *Journal de Médecine Esthétique* (in French): 33-34

technology would be deployed and what metric is being used to assess its accuracy. Some illustrative examples are given below for both an age estimation and verification technology.

Age Estimation Technology

If the technology is being deployed to estimate the ages of teenagers, for example, the objective of the test would be:

What is the MAE of an age estimation technology for those who are 13-18 years old?

Here, the primary accuracy measure is MAE to a sample size formula for estimating a population mean can be used to calculate the sample size. The formula is as follows:

$$N = \frac{N \cdot X}{(N+X-1)},$$

where,

$$X = \frac{Z_{\alpha/2}^2 \cdot \sigma^2}{MOE^2},$$

and $Z_{\alpha/2}$ is the critical value of the Normal distribution at $\alpha/2$ (e.g., for a confidence level of 95%, α is 0.05 and the critical value is 1.96), MOE is the margin of error, σ^2 is the population variance, and N is the population size. Note that a Finite Population Correction has been applied to the sample size formula.

This sample size calculation provides the recommended number of samples required to estimate the true population mean (in this case the MAE) with the required margin of error and confidence level.

The margin of error is the level of precision required. This is the plus or minus number that is often reported with an estimated mean and is also called the confidence interval. It is the range in which the true population mean is estimated to be. Note that the actual precision achieved after you collect your data will be more or less than this target amount, because it will be based on the population variance estimated from the data and not your expected variance.

The confidence level is the probability that the margin of error contains the true mean. If the study was repeated and the range calculated each time, you would expect the true value to lie within these ranges on 95% of occasions. The higher the confidence level the more certain you can be that the interval contains the true mean.

The population size is the total number of distinct individuals in your population. In this formula we use a finite population correction to account for sampling from populations that are small. If your population is large, but you do not know how large, you can conservatively use 100,000. The sample size does not change much for populations larger than 100,000.

The population variance tells you how the data points in a specific population are spread out. It is the average of the distances from each data point in the population to the mean, squared. An estimate of the expected variance is required for the calculation and may be obtained from previous tests carried out on the technology.

The table below shows how the sample size changes as the inputs change (assuming a population size of 100,000). The larger the sample size, the more certain you can be that the estimates reflect the population, so the narrower the confidence interval. However, the relationship is not linear, e.g., doubling the sample size does not halve the confidence interval.

Confidence Level	Margin of Error	Population Variance		
		4	9	16
90%	0.25	173	389	688
95%		246	551	974
99%		423	947	1671
90%	0.5	44	98	173
95%		62	139	246
99%		107	239	423
90%	1.0	11	25	44
95%		16	35	62
99%		27	60	107

For example, for a technology in this deployment setting that has an expected MAE variance of 9 years (or standard deviation of 3 years), a sample size of 139 would be needed to achieve a margin of error of 0.5 years with 95% confidence (i.e., to estimate the MAE within plus or minus half a year with 95% confidence), but the sample size would need to increase to 551 for a margin of error of 0.25 years.

Note that if the population variance was underestimated, for example, then for the same sample size, the actual margin of error calculated from the sample would then be larger (the confidence interval would be greater than plus or minus the margin of error stated in the sample size calculation).

Once a sample size has been calculated, the test subjects it is made up with should reflect its deployment and therefore, in the above example, be made up of 13- to 18-year-olds and the breakdown of characteristics should be representative of the population in relation to age, gender and skin tone (e.g., the proportion of females to males should be approximately 50/50).

Age Verification Technology

If the technology is being deployed based on scenario 1 (identifying if a person is over an age threshold), then the objective of the test would be:

What is the false positive rate (FPR) of an age verification technology for those who are +/- 5 years of the age threshold?

Note that the +/- 5 years could be amended to an age range most suitable to the deployment.

FPR has been identified here as the primary accuracy measure (and therefore the measure used to calculate the sample size) as it is identified that minimising false positives is critical in deployment, but we do recommend several measures be calculated from the sample (e.g.,

overall accuracy, PPV etc.). The formula to calculate the sample size to estimate the FPR is as follows:

$$n_{total} = \frac{n_{FPR}}{(1-prev)},$$

where:

$$n_{FPR} = \frac{Z^2_{\alpha/2}\big(p(1-p)\big)}{d^2}$$

and $Z^2_{\alpha/2}$ is the critical value of the Normal distribution at $\alpha/2$ (e.g., for a confidence level of 95%, α is 0.05 and the critical value is 1.96), MOE is the margin of error, p is the estimated FPR, and $prev$ is the prevalence.

This sample size calculation provides the recommended number of samples required to estimate the true FPR with the required margin of error and confidence level.

The margin of error and confidence level were both defined above.

An estimate of the expected FPR is required for the calculation and may be obtained from previous tests carried out on the technology. If it is unknown 50% can be used as a conservative estimate.

The prevalence is the expected proportion of the population to be estimated to be greater than the age threshold.

The table below shows how the sample size changes as the inputs change (assuming a prevalence of 50%; I.e., that half of the sample will be greater than the threshold and half less than the sample). The larger the sample size, the more certain you can be that the estimates reflect the population, so the narrower the confidence interval. However, the relationship is not linear, e.g., doubling the sample size does not halve the confidence interval.

Confidence Level	**Margin of Error**	**Estimated FPR**		
		20%	**10%**	**5%**
90%	**2%**	2142	1212	321
95%		3028	1714	910
99%		5172	2942	1564
90%	**3%**	958	540	286
95%		1358	766	406
99%		2332	1320	698
90%	**5%**	346	196	104
95%		492	278	146
99%		846	478	252

For example, for a technology in this deployment setting that has an expected FPR of 10%, a sample size of 766 would be needed to achieve a margin of error of 3% with 95% confidence (i.e., to estimate the FPR within plus or minus 3% with 95% confidence), but the sample size could reduce to 278 for a margin of error of 5%.

Note that if the FPR was underestimated, for example, then for the same sample size, the actual margin of error calculated from the sample would then be larger (the confidence interval would be greater than plus or minus the margin of error stated in the sample size calculation).

Once a sample size has been calculated, the test subjects it is made up with should reflect its deployment and therefore, in the above example, be made up of test subjects of which 50% are up to 5 years over the age threshold and 50% up to 5 years under the age threshold, and the breakdown of characteristics should be representative of the population in relation to age, gender and skin tone (e.g., the proportion of females to males should be approximately 50/50).

9.7 Repeatability and Reproducibility of Testing

Repeatability is the closeness of the agreement between the results of successive measurements of the same measure, when carried out under the same conditions of measurement.[25]

For the findings of a study to be reproducible means that results obtained by an experiment or an observational study or in a statistical analysis of a data set should be achieved again with a high degree of reliability when the study is replicated.

The term reproducible research refers to the idea that scientific results should be documented in such a way that their deduction is fully transparent. This requires a detailed description of the methods used to obtain the data and making the full dataset and the code to calculate the results easily accessible.

A risk associated with this is the predictability of the testing resulting in a design culture aimed at passing the test rather than delivering the efficacy of the system. All age assurance testing should include for outliers, i.e., the presentation of artefacts that are deliberately outside the parameters of the test to ensure that the target of evaluation has not been designed, set up or configured merely to pass the test.

It is suggested that the approach to age assurance testing should require:

- Estimates of the repeatability, reproducibility and trueness of the method in use, obtained by collaborative study as described in ISO 5725-2[26], be available from published information about the test method in use.
- The conformity assessment body confirming that its implementation of the test method is consistent with the established performance of the test method by checking its own bias and precision.

[25] JCGM 100:2008. Evaluation of measurement data - Guide to the expression of uncertainty in measurement, Joint Committee for Guides in Metrology
[26] ISO 5725-2:2019 - Accuracy (trueness and precision) of measurement methods and results — Part 2: Basic method for the determination of repeatability and reproducibility of a standard measurement method

- Any influences on the measurement results that were not adequately covered by the collaborative study be identified and the variance associated with the results that could arise from these effects be quantified.

9.8 Certification

The ICO should regard testing and reporting of age assurance technologies by unaccredited conformity assessment bodies with a degree of scepticism. Although first party testing (i.e., self-assessment) has a role to play where vendors of age assurance technologies provide their own marketing materials, white papers or internal test transparency, these should always be accompanied by independent 3rd party verification, validation and certification.

The ICO should consider 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 as described in s.2.2 above to maintain its own confidence and controls in age assurance certification criteria.

Whilst the terms 'accreditation' and 'certification' are often used interchangeably, they are two distinct steps on the quality assurance ladder. Certification is the assessment of whether a management system, product (such as an age assurance technology), or person meets the criteria laid out in a generic quality standard or scheme. Accreditation, sitting on the rung above, is the determination of the competence of the certification body to perform specific activities under a recognised international or national standard or scheme.

Working together, certification acts as the third-party endorsement of an organisation's systems, products or personnel whilst accreditation is an independent third-party endorsement of that certification body's competence. Just as end-user organisations seeking certification must demonstrate to a certification body that they conform to the criteria of the relevant standard, in turn certification bodies must demonstrate their competence, consistency and integrity to a National Accreditation Body (NAB) such as UKAS to be accredited. In other words, if certification bodies are 'the checkers' then UKAS's role as the UK's sole government-appointed NAB is to 'check the checkers'.

Whilst it is not mandatory for certification bodies to be accredited by a NAB, those that are accredited are able to demonstrate that they have been rigorously assessed by an independent authority against internationally recognised standards. Non-accredited certification bodies are not subject to this independent scrutiny.

Impartiality is a key component of achieving accredited status, meaning accredited certification bodies cannot offer both consultancy and certification services. Although this does not necessarily mean that a non-accredited certification body is not a competent, impartial and capable organisation, it does mean that it will have difficulty demonstrating it possesses these qualities in a universally accepted way.

Accreditation from a recognised NAB helps generate confidence in the competence of accredited certification bodies and, in turn, the competence of organisations that have been certified by accredited certification bodies. As a result, a growing number of organisations in both the public and private sectors are requiring their suppliers' management systems to be certified. Increasingly, these procurers will only accept certificates issued by a certification body that has been accredited by a recognised NAB. In addition to becoming the expected

industry norm for many sectors, accredited certification also offers market differentiation and shows credible evidence of best practice. Businesses with non-accredited certificates run the significant risk of being excluded from the tendering process and/or losing ground in an increasingly competitive marketplace.

Recognised NABs that are members of international accreditation organisations such as IAF, ILAC and EA are subject to regular reviews by a cross section of their peers. This added layer of scrutiny provides assurance in the competence of the NAB, and in turn, increases confidence in the certificates issued by any certification bodies it accredits.

In addition to providing access to domestic contracts, accredited certification can open doors to a worldwide marketplace. Thanks to UKAS being a signatory to IAF, ILAC and EA multilateral agreements, accredited certificates are recognised in over 100 different economies, delivering a truly global "accredited once, accepted everywhere" service. Non-accredited certificates do not offer this level of competitive advantage, either at home or abroad.

The UK government shares the industry's view over the relative merits of using accredited and non-accredited conformity assessment services. In addition to its Conformity Assessment and Accreditation Policy in the UK, government's guidance on accreditation and conformity assessment also states that: "the only 'authoritative statement' of competence, that has public authority status - providing the last level of control in the conformity assessment chain is from the UK's national accreditation body, UKAS." Government therefore specifies the use of UKAS accredited organisations where testing, inspection or certification is required to demonstrate compliance with national legislation or guidance.

Having a sole NAB is important for the UK, as it provides certainty for regulators, accredited bodies and businesses using accredited services both in the UK and around the world. In line with a significant number of its international peers, UKAS is appointed by, but operates independently of government. This arrangement allows UKAS accreditation to underpin the UK's quality infrastructure by working entirely in the public interest, free from commercial pressures and impartially from both government and the organisations it accredits.

9.9 Depth of Evaluation

Depth of evaluation describes the levels to which a conformity assessment body undertakes testing and analysis to be able to provide confidence that the age assurance technology will work as intended. In general, the more independent assessment and checking that is undertaken by a conformity assessment body, the greater the level of confidence that can be achieved. However, this must be balanced in a proportionate manner against risk and costs of conformity assessment.

This report proposes using the Common Criteria Evaluation Methodology for the purposes of evaluating the efficacy, security and reliability of Age Assurance Systems. This is an existing, widely adopted, methodology.

The Common Criteria (ISO/IEC 15408-1, ISO/IEC 15408-2 and ISO/IEC 15408-3) and the Common Evaluation Methodology (ISO/IEC 18045) are relevant standards for independent security evaluation of IT products. The independent evaluation and certification of IT products

according to these standards is widely used in many different areas. The Common Criteria standard is defined in three parts:

- ISO/IEC 15408-1 contains the “introduction and general model”.
- ISO/IEC 15408-2 contains the “security functional components”.
- ISO/IEC 15408-3 contains the “security assurance components”.

The Common Evaluation Methodology is a companion document to the Common Criteria standard and defines the minimum actions to be performed by an evaluator to conduct a Common Criteria evaluation, using the criteria and evaluation evidence defined in the Common Criteria standard.

The depth of evaluation set out in the Common Criteria explores four elements of testing:

- Analysis of coverage (ATE_COV) - The developer is required to demonstrate that the tests which have been identified include testing of all of the functions as described in the functional specification for the age assurance system. The analysis should not only show the correspondence between tests and age assurance functions, but should provide also sufficient information for the evaluator to determine how the functions have been exercised. This information can be used in planning for additional evaluator tests. Although at this level the developer has to demonstrate that each of the functions within the functional specification has been tested, the amount of testing of each function need not be exhaustive.
- Depth of coverage (ATE_DPT) - The testing of the high level design of the age assurance subsystems provide a high-level description of the internal workings of the system. Testing at the level of the subsystems, in order to demonstrate the presence of any flaws, provides assurance that the subsystems have been correctly realised. This depth of coverage particularly provides assurance for multi-faceted systems, those utilising the waterfall technique, or those enabling the gathering of permutations and combinations of age assurance over time.
- Ordered functional testing (ATE_FUN) - The objective is for the developer to demonstrate that all age assurance functions perform as specified. The developer is required to perform internal testing and to provide test documentation. In this component, an additional objective is to ensure that testing is structured such as to avoid circular arguments about the correctness of the portions of the functionality being tested.
- Independent testing (ATE_IND) - The intent is that the developer should provide the evaluator with materials necessary for the efficient reproduction of developer tests. This may include such things as machine-readable test documentation, test programs, etc. In this component the evaluator must repeat all of the developer's tests as part of the programme of testing.

9.10 Regulatory Options and Tolerance Levels

A significant challenge for both regulators and relying parties is comparison of different age assurance systems to derive a feel for 'what good looks like'. Although the outputs of the system can be effectively measured, the outcomes are not realised unless set in the context of appropriate tolerances.

This report has not identified specific tolerances, nor suggested tolerances. The setting of the tolerances is a matter for regulators and should be done after undertaking appropriate consultation, consideration of the views of diverse stakeholders and setting a fair, proportionate and risk-based approach to tolerances. The method of measurement as to whether or not age assurance systems are within those tolerances is set out in this report. The precise quantum of those tolerances should, however, be ultimately a public policy decision.

There are two options for setting tolerances:

- A tolerance that is a factor of the inputs or outputs, such that it is not fixed in time, but as technology improves and accuracy increases, so the tolerances narrow;
- A tolerance that is an arbitrary determination fixed in time, but kept under review, such that the age assurance service providers know what it is and that it is set for the whole market place.

We did not find any suitable approaches to setting a tolerance level that is a factor of the inputs or outputs. In many respects, the age assurance systems and the marketplace is not sufficiently developed to have this kind of dynamic tolerance applied. As such, our recommendations are based on setting an arbitrary determination based on the state-of-the-art and a reasonable assessment of what good looks like.

The working draft of ISO/IEC 27566 contains some suggested tolerances, but it is important to emphasise that these have not been subject to widespread consultation and comment. It is inevitably the case that age assurance service providers would press for the widest possible tolerances and conversely, campaign groups, child protection professionals and others may well press for the narrowest possible tolerances.

There is no right or wrong level, but a level of tolerances selected by the Regulator, should at least provide some certainty and be based upon an appropriate risk appetite.

The suggested levels of tolerance for each of the five proposed levels of confidence are as follows:

TABLE 11 – SCHEMATIC: LEVELS OF TOLERANCE TO BE APPLIED TO EACH LEVEL OF CONFIDENCE IN AGE ASSURANCE

Asserted	Basic	Standard	Enhanced	Strict
• None - we conclude that for self asserted age assurance, there should be no level of tolerance applied. • In all cases, such self-assertion should be treated with zero trust.	• Mean Absolute Error (within 3 years of the target age) • Standard Deviation (within 2 deviations so that the absolute errors are +/- 3.92 years) • False Positive Rate (FPR) less than 15% • Outcome Error Parity within 3% across protected characteristics • Evaluation Assurance Level (depth of testing) - Level 1	• Mean Absolute Error (within 2 years of the target age) • Standard Deviation (within 1.5 deviations so that the absolute errors are +/- 2.94 years) • False Positive Rate (FPR) less than 10% • Outcome Error Parity within 3% across protected characteristics • Evaluation Assurance Level (depth of testing) - Level 2 • Authentication Assurance Level 1 (LoA1) • Liveness detection error (less than 5%)	• Mean Absolute Error (within 1.5 years of the target age) • Standard Deviation (within 1.25 deviations so that the absolute errors are +/- 2.45 years) • False Positive Rate (FPR) less than 5% • Outcome Error Parity within 3% across protected characteristics • Evaluation Assurance Level (depth of testing) - Level 3 • Authentication Assurance Level 2 (LoA2) • Liveness detection error (less than 3%)	• Mean Absolute Error (within 1 year of the target age) • Standard Deviation (within 1 deviation so that the absolute errors are +/- 1.96 years) • False Positive Rate (FPR) less than 1% • Outcome Error Parity within 3% across protected characteristics • Evaluation Assurance Level (depth of testing) - Level 4 • Authentication Assurance Level 3 (LoA3) • Liveness detection error (less than 1%)

In our view, such tolerances do not need to be set in regulation. In fact, there is good reason for them not to be set in the constrictors of regulation. The most appropriate place for them to be set is in international standards and that work is underway. However, in the interim, suitable guidance issued by a regulator, such as the ICO or OFCOM or similar, would have a conforming affect on the marketplace and start to see a convergence of age assurance service providers around appropriate testing, certification and transparency of results.

10. Conclusions

Our report sets out a structured approach to the measurement of age assurance technologies. We have provided a series of definitions that enables standardisation across the sector, including clarity around 'age assurance' (including both 'age estimation' and 'age verification'), 'levels of assurance' and the technical terms that will be applicable to the recommended measurement approach.

There is more work to be done on this. As the Online Safety Bill progresses through Parliament, the statutory definition of age assurance (in current clause 189 of the Bill) may evolve. Similarly, the preliminary work item at 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 a statistical sense, we conclude that two separate approaches to measurement are required. One relating to 'continuous age assurance' such as age estimation techniques; and one relating to 'binary age assurance' such as age verification techniques. However, it is also worth noting that combinations and permutations of these can also be appropriate and the consequences of those require further research and understanding.

Each of the two approaches, however, is capable of measurement.

For continuous approaches to age assurance (the formulae are described in more detail in the body of the report):

Mean Absolute Error (MAE)

- $MAE = \frac{\sum_{i=1}^{n}|(p_i - o_i)|}{n}$
- The central value of the absolute errors of the sample.

Standard Deviation (SD)

- $SD_{AE} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(AE_i - MAE)^2}$
- The amount of variation or spread over the distribution of absolute errors in the sample.

These two measures taken together provide an effective means of measurement of age estimation systems - note - the current common practice of just stating the mean absolute error is, in our view, inadequate.

For binary approaches to age assurance:

True Positive Rate (TPR)

- $TPR = \frac{TP}{TP+FN}$
- Is the sensitivity of the technology's ability to correctly detect people who are over the age threshold.

False Positive Rate (FPR)

- $FPR = \frac{FP}{FP+TN}$
- Is the technology's probability of false alarm (i.e., incorrectly identifying someone as being over the age threshold).

Positive Predictive Value (PPV)

- $PPV = \frac{TP}{TP+FP}$
- The PPV is the proportion of the sample correctly identified as being over the age threshold given that they have been predicted as being over the age threshold.

These three measures taken together provide an effective means of measurement of age verification systems. We have set out why the current common practice of just stating the false positive rate is, in our view, inadequate.

This assessment of the most appropriate measurements and/or indicators of accuracy, both alone and in combination, are designed to deliver a practical result that can be universally applied across all methods of age assurance but are as simple to understand and explain as possible.

We go further to suggest, as is currently proposed in ISO/IEC 27566, that these be characterised in even simpler terms for different levels of confidence:

We conclude that there are suitable methods available to test the full range of age assurance techniques, which should focus on adapting the common criteria for IT security evaluation to the development of testing for age assurance systems. This needs to include the selection of appropriate sample sizes and undertaking presentation attack detection testing.

Our report includes a detailed analysis of measurement uncertainties or bias, how these can be measured, appropriate tolerances to be applied and their impact upon the statements of efficacy of age assurance systems. With regard to tolerances, the report stops short of actually setting a tolerance, but it does propose some tolerances that may be considered suitable. We conclude that the actual setting of the tolerances is a matter for regulators and should be done after undertaking appropriate consultation, consideration of the views of diverse stakeholders and setting a fair, proportionate and risk-based approach to tolerances. The method of measurement to see whether or not age assurance systems are within those tolerances is set out in this report. Ultimately, however, it is for public policy decision-makers to determine those tolerances.

11. Recommendations

The research brief asked for recommendations for the ICO to take forward consideration of how it should approach the assessment and measurement of age assurance technologies.

Our recommendations are as follows:

1. The Commissioner should support the development of national and international standards for the assessment of age assurance, most notably the work by ISO/IEC on 27566 - Age Assurance Systems - Framework. This work will, eventually, lead to an international recognised approach to defining, applying and testing these technologies.

2. The Commissioner should apply measures of **efficacy** to age assurance systems based upon whether the output is continuous (i.e., age estimation) or binary (i.e., age verification).

3. For continuous age assurance, the Commissioner should expect to see conformity test reports showing:

a) the **Mean Absolute Error (MAE)** is the most useful overall measure of the efficacy of the system, but this should always be stated with its distribution of errors (the **Standard Deviation (SD)**) to a 95% confidence interval. The MAE shown on its own has the potential to be a misleading indicator; and
b) the **Mean Absolute Error Parity (MAEP)** across protected characteristics including skin tone and gender as a minimum

4. For binary age assurance, the Commissioner should expect to see conformity test reports showing:

a) the overall accuracy of a system (i.e., the proportion of the sample that have been correctly classified) should be reported, however, this should only be stated with both the sensitivity of the system (the **True Positive Rate (TPR)**), the likelihood of 'false alarm' (the **False Positive Rate (FPR)**) and the ability of the system to correctly predict values (**Positive Predictive Value (PPV)**) should be stated; and
b) the **Positive Predictive Value Parity (PPVP)** ensuring that the precision of the age assurance technology is equivalent between different population subgroups.

5. The Commissioner should 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. This report provides some suggestions for tolerance levels, but ultimately, these are for a regulator (and by consequence) legislators and courts, to determine or recommend.

6. The Commissioner should consider further research on the implications for combinations and permutations of age assurance techniques, especially in the context of a Trust Framework involving the combination of multiple approaches potentially by multiple operators in the marketplace acting interoperably to provide an age assurance output. The Commissioner

should consider how these could contribute to elevating the level of confidence in the age assurance output, but also how it could impact upon the handling of contra-indicators.

7. The Commissioner should consider how the use of their powers under Articles 57 (1) (n) and 58 (3) (f) pursuant to Article 42 (5) of UK GDPR could be developed to maintain oversight and approval of conformity assessment of age assurance techniques used for the purpose of age-appropriate design applications and for demonstrating conformity with the processing of personal data. As a minimum, the approval of certification criteria should consider:

a) the test protocols applied to secure repeatability and reproducibility of age assurance testing results
b) the identification and controls associated with the data capture subjects and data capture devices
c) the approach to both human and document presentation attack detection (spoofing)
d) the ambient lighting under which testing was undertaken
e) the assessment of 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.

8. The Commissioner should provide supplementary guidance to the Opinion on Age Assurance and the Age-Appropriate Design Code on the measurement and reporting of age assurance technologies to ensure upholding information rights, whilst taking into account the need for an open, fair and comparable marketplace in the provision of such technologies to relying parties.

Bibliography

Journals, Articles and Learned Works

Herriman, D. S. (2003). Using the Common Criteria for IT Security Evaluation. Auerbach Publications

Fitzpatrick, T. B. (1975). "Soleil et peau" [Sun and skin]. *Journal de Médecine Esthétique* (in French): 33-34

Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute. https://doi.org/10.5281/zenodo.3240529

Joint Committee for Guides in Metrology Evaluation of measurement data - Guide to the expression of uncertainty in measurement (GUM) JCGM. 100:2008. http://www.bipm.org/utils/common/documents/jcgm/JCGM_100_2008_E.pdf

Bland, M. (1987). An Introduction to Medical Statistics.

Information Commissioner's opinion: Age Assurance for the Children's Code, 14 October 2021

Government's Good Practice Guide (GPG45) Identity Proofing of an Individual

Standards and Normative References

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

ISO-JCGM 200, 2008 International vocabulary of metrology — Basic and general concepts and associated terms (VIM)

ISO/IEC 2382-37:2017 - Information technology — Vocabularye– Part 37: Biometrics

ISO 8995-1:2002 - Lighting of workplaces — Part 1: Indoor

ISO 9000:2015 - Quality management systems — Fundamentals and vocabulary

ISO/IEC 15408-1:2009 - Information technology — Security techniques — Evaluation criteria for IT security — Part 1: Introduction and general model

ISO/IEC 29115:2013, Information technology - Security techniques - Entity authentication assurance framework

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

ISO/IEC 30108-1:2015 - Information technologye– Biometric Identity Assurance Servicese– Part 1: BIAS services

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