# From AI Principles to AI Assurance: an Online Safety Case Study

**Miranda Cross**[1], **Andreas Gutmann**[2], **Ismini Psychoula**[2], **Pedro Friere** [3]

[1]UK Office of Communications
[2]UK Office of Communications, University College London
[3]Aston University
Riverside House, 2A Southwark Bridge, London, UK SE1 9HA
miranda.cross@ofcom.org.uk, ismini.psychoula@ofcom.org.uk

## Abstract

Principles-based frameworks for AI assurance have been proposed for various AI/ML use cases, focusing on aspects such as ethical design, trustworthiness, and safety. However, translating these high-level principles into actionable, objective criteria for auditing, particularly by third parties, remains challenging. Our analysis shows this is due to the inherent subjectivity of principles, the need for vertical frameworks tailored to specific AI/ML applications, and the unreliability of information gathered during the assurance process. In this paper, we present a case study on how to develop and operationalise a principles-based framework for AI assurance aimed at assessing the 'accuracy' of child sexual exploitation (CSEA) and terrorism detection technologies in the context of online safety. The proposed assurance framework addresses a requirement in the UK's 2023 Online Safety Act to create an 'accreditation' scheme specifically for CSEA and terrorism detection technologies. We discuss the critical challenges for operationalising such principles-based frameworks for assurance, particularly in relation to ensuring transparency, reliability, and consistency in audits. We also map potential issues which remain for effectively assessing and auditing AI/ML technologies, informing the development of future research agendas which further research and development of robust standards for assurance, particularly in sociotechnical contexts.

## Introduction

The omnipresence of artificial intelligence (AI) and machine learning (ML) technologies has led to their increased integration into the daily lives of individuals, as well as their use within a variety of private and public organizations (Mikalef et al. 2022). As the speed and scale of adoption has increased, so has the development of the field of 'Responsible AI' (Sadek et al. 2024), and the need to standardise assessment and assurance of AI/ML systems, which involves measuring, evaluating, and communicating the trustworthiness of AI systems in practice using measurable, objective criteria (UK Department for Science, Innovation, and Technology 2024). Assurance can be conducted internally or by external parties, and poorly assured systems may pose increased risks of bias and error, leading to significant consequences. For example, biased AI systems in contexts like law enforcement can perpetuate discrimination or result in incorrect decisions that affect people's lives and livelihoods.

Typically, assurance frameworks are based on either formal standards, such as those set by the International Organization for Standardization (ISO) or the Institute of Electrical and Electronics Engineers (IEEE), or on normative principles. For AI/ML assurance, few formally adopted standards (from ISO or IEEE) exist, and where they do, they are generally domain-specific or focused on risk management. AI/ML assurance frameworks can be developed using normative principles, which can also be specifically tailored towards assessing the sociotechnical implications of AI/ML technology and its use cases. However, using broad principles alone for assurance purposes is difficult because they are not necessarily measurable using standardised metrics, and are often assessed subjectively. Advancements are needed to create assurance frameworks that can effectively measure and report on these types of normative, sociotechnical principles in the AI/ML context. In this paper, we contribute a case study of the process for creating an operationalisable, standardised assessment based on normative principles. We detail the process of developing a scored assessment for the 'accuracy' of technology developed to detect terrorism and child sexual abuse and exploitation (CSEA) content– a requirement set out in the UK's 2023 Online Safety Act. We collaborated with the UK's regulator Ofcom, which will consider our findings in their accreditation scheme. The process of operationalising a framework for assessing the concept of 'accuracy' by creating constituent principles, and then developing objective, scorable statements is detailed in Section 3. In Section 4, we discuss the challenges we encountered with operationalising these principles, and the issues we anticipate arising when operationalising similar types of assessments in the future. In Section 5, we present the lessons learned from this case study and the applicability of this method to other AI assurance contexts.

## Related Work

### Principles-Based Frameworks for the Governance and Assessment of AI

As governments and organisations around the world seek to standardise the assessment of artificial intelligence (AI) and machine learning (ML) technologies for safety, fair-

ness, and trustworthiness, a consensus around the development of high-level, flexible principles-based frameworks has emerged across multiple governance domains— including AI 'trustworthiness' (Li et al. 2023), AI 'responsibility' (Akbarighatar 2024), and AI 'ethics' (Floridi and Cowls 2019; Jobin, Ienca, and Vayena 2019). Over a hundred of these frameworks have been published thus far, originating from AI/ML developers and their organisations, civil society, academia, and governments (Prem 2023; Percy et al. 2021; Morley et al. 2020; Canca 2020) A commonality among these frameworks is the use of high-level 'principles' as an organising mechanism for goals the AI system in practice should achieve, generally including concepts such as transparency, fairness, privacy, and safety (Smit, Zoet, and van Meerten 2020).

Similar principles-based frameworks for the assessment and assurance of AI/ML technologies have been created by governmental or inter-governmental organisations, including: the Organisation for Economic Cooperation Development (OECD) for AI trustworthiness (OECD AI 2024), the UK's Information Commissioner's Office (ICO) for data protection (UK Information Commissioner's Office 2024), the European Union's High-Level Expert Group on AI's Ethics Guidelines for Trustworthy Artificial Intelligence (European Commission, High-Level Expert Group on AI 2020), and the EU Digital Services Act (DSA) for online harms (European Commission 2022). This is only a sampling of published frameworks—further work has also been done to establish principles for risk management of AI (notably, the US National Institute for Standards and Technology's AI Risk Management Framework) (NIST 2022), and the governance of the use of public-sector AI tools (see for example, in Canada (Government of Canada 2024)). These frameworks differ along many lines—for example, whether they are statutory (e.g., EU DSA) or voluntary (e.g., OECD), designed to assess risk or ethics, applicable across multiple domains or a singular domain, and the degree to which the principles are general or specific (Díaz-Rodríguez et al. 2023; Smit, Zoet, and van Meerten 2020; Morley et al. 2020; Floridi et al. 2018; Hagendorff 2020; Li et al. 2023). Principles-based frameworks can be understood as a kind of 'scaffolding' upon which to build assurance— many of these frameworks are technology and domain agnostic, and so have an inherent flexibility in their application. These frameworks are useful as an organising mechanism, but as we will discuss in Section 3, a key challenge for practitioners and developers of AI/ML technologies is to translate such principles into practical, measurable metrics, technical testing thresholds, and practices, especially for the purposes of assurance and trust building (see (Prem 2023)).

## Aims of AI Assurance and Assessment

The purpose of many of the principles-based frameworks mentioned above is to guide 'assurance', the process of measuring, evaluating and communicating the trustworthiness or performance of AI systems in practice (Boer, de Beer, and van Praat 2023). According to the UK Government's Introduction to AI Assurance white paper, *'AI assurance processes can help to build confidence in AI systems by measuring and evaluating reliable, standardised, and accessible evidence about the capabilities of these systems [including] whether they will work as intended, hold limitations, and pose potential risks, as well as how those risks are being mitigated to ensure that ethical considerations are built-in throughout the AI development lifecycle.'* (UK Department for Science, Innovation, and Technology 2024) Assurance can be assessed internally or by external parties, and there is a growing demand for principles-based frameworks to be translated into scorable assessments or benchmarks (Percy et al. 2021). When communicated, assurance results also form a key component of building trust between users/subjects and developers/deployers of AI/ML systems .

Self-assessments for principles-based frameworks have previously been built: for example, the European Union's Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment translates seven key principles of trustworthy AI into a checklist of questions for developers to verify against (European Commission, High-Level Expert Group on AI 2020). Similarly, AI Verify, a framework and software toolkit developed by the Singaporean Government and major technology companies, seeks to help developers validate the performance of AI systems against a set of 11 governance principles through standardised tests and process checks (AI Verify Foundation 2024). These frameworks are intended for flexible use on a wide variety of technologies– organisations are meant to select elements relevant to the AI model being tested as they see fit. Notably, the two described assurance processes are self-assessments, and thereby rely on the data inputted by developers directly, avoiding the issue of data access third party auditors face (Raji et al. 2022). However, given the alignment gap between (some) developers' profit motivations and their desire for risk mitigation, self-assessments may not always be appropriate or sufficient for ensuring high levels of confidence in assurance, and may indeed lead to inadequate assurance processes which provide false confidence in the technology's trustworthiness and performance (Goodman and Trehu 2022).

When considering third-party (external) assurance, responsible computing practitioners have focused less on holistic sociotechnical assurance, and instead focused on a specific facet of either the organisation (i.e. risk management) or the technology (i.e. bias) (Raji et al. 2022; Nonnecke 2024). Significant energy—particularly in a regulatory context— has been dedicated to systematising risk management audits; for example, the European Union's development of a harmonised standard on risk management (led by CEN-CENELEC) forms a key part of the EU AI Act's risk mitigation strategy (Soler et al. 2024; Pouget 2023). Other codified examples of external assessments include 'bias audits' for in-scope Automated Employment Decision Tools (AEDT) algorithms required by the New York City Local Law 144 of 2021 (Maurer 2024). However, third-party assurance faces the substantial issue of information asymmetry— even auditors directly contracted to conduct an assessment can run into challenges with information verification, as access to a model and its documentation and underlying training data needs to be provided by the corresponding de-

velopers. Below, we will elaborate further on the challenges for external, socio-technical AI assessment at scale.

## Challenges for AI Assurance and Assessment

Discussion on the challenges of AI assurance has been explored at length in recent publications, particularly around the capacity and positionality of auditors (Costanza-Chock, Raji, and Buolamwini 2022), lessons learned from assurance in other fields (such as finance) (Raji et al. 2022), and the multitude of dimensions that an AI audit can take on (Bogen et al. 2025). However, when considering mandated assurance processes within an established legislative framework, such as that presented by the European Union's Digital Services Act (DSA) or AI Act (AIA), or the UK's Online Safety Act (OSA) (Terzis, Veale, and Gaumann 2024), three key challenges for external assurance are relevant to our discussion:

1. Objective, standardised assessment of subjective, socio-technical components of trustworthiness
2. Applicability of assurance frameworks and processes to different technology types
3. Minimisation of conflict of interest and information obfuscation

**Objective Assessment of Subjective Principles** Though the principles for assessment frameworks may be prescribed, either in legislation or by adopting existing frameworks such as those presented by Smit (2020), a key part of the assurance process is translating subjective principles into auditable, operationalised duties that an auditee is assessed to have complied or not complied with. For 'subjective' principles, particularly those where experts present different opinions on best practices (see, for example, the computational definition of fairness) or where there are trade-offs, creating an operationalisable framework is not an easy task (Lam et al. 2024; Morley et al. 2020). In addition, subjective principles are highly context-dependent—as Mittelstadt [2019] states: 'Norms and requirements [for assurance] cannot be deduced directly from mid-level principles without accounting for specific elements of the technology, application, context of use, or relevant local norms.' (Mittelstadt 2019) In the case of competing norms (e.g., transparency and security), objective assessment is difficult because no universal hierarchy of which principles to prioritise exists for AI (Mittelstadt 2019). There is also the question of who creates the 'scorecard' for the operationalisation of principles—is it the developers, or the assurance body? If left to developers, the risk of corporate capture of independent assessments is high (Young, Katell, and Krafft 2022)—however, outside entities may lack the technical expertise to correctly operationalise high-level principles in context (Raji et al. 2022).

**Technology Agnosticism** As stated above, the successful application of an assessment of high-level principles is inherently dependent on the context of the design and use of a technology— for example, adequate 'fairness' of a loan origination AI/ML system can be vastly different than when a similar AI/ML system is used for triaging customer support. Even where the context of deployment is the same, different underlying technologies can present differently in terms of test results and metrics which aim to measure the same principle (Akbarighatar 2024). This presents a substantial issue for external assurance at scale— even within one specific deployment environment, how is it possible to fairly assess compliance of different technologies without creating bespoke criteria for each technology? Technology agnosticism has been a constant issue for the development of 'horizontal' AI standards by international technical standards bodies, including those tasked with creating harmonised standards for the AIA (Pouget 2023).

**Information Reliability** The question of *who* conducts assurance checks has significant impacts on the reliability of the assurance mechanism. The propensity for 'auditwashing', wherein firms create false assurance of their or others' products by subjecting them to inadequate auditing, which can also obfuscate problematic or illegal practices, is significant (Goodman and Trehu 2022). The aforementioned legally enforceable assurance processes (in the DSA, AIA, and OSA) have required a combination of first-party (e.g., risk assessment) and third-party assurance[1] (e.g., the EU DSA's provision for the annual independent audit of Very Large Online Platforms) (Terzis, Veale, and Gaumann 2024). However, as stated previously, third-party assurance presents a trade-off between information reliability and access.

In a third-party audit, information can be verified and considered reliable by a source outside of the organisation responsible for the technology (avoiding conflicts of interest), but third-party auditors are often given less access to information than those within the organisation. Full access to a model, often called 'white-box auditing', may require a firm to provide access to ostensibly sensitive information, and may often require extensive non-disclosure, intellectual property sharing, data sharing, and other contractual agreements to be put in place (Koshiyama et al. 2024). Protecting proprietary information, including commercially sensitive details on training data and model weights, from leakage is often of tantamount importance to firms in the rapidly-developing area of AI/ML—however, this is at direct odds with the goals of a third-party audit (Hasan et al. 2022), as 'all audit systems provide some form of privileged access to auditors' [Raji et al., 2022, p. 564]. Even where auditors are given full access, the complexity and explainability of different kinds of AI/ML technologies vary widely, while other systems may not be able to be evaluated due to their 'black-box' nature, as auditors cannot trace where imperfection occurs within the system (Mökander et al. 2021). The motivations of auditors themselves are also important

[1]For definition of audit types, see Raji, et al [2022]: *'"First party" audits are conducted by a company of its own products. Many AI ethics teams can be conceived of as fulfilling a first-party audit role. "Second party" audits are performed by a contractual counterparty, or an entity hired by that contractual counterparty. Second party audits typically ensure compliance with contract terms . . . "Third party" audits are conducted by ostensibly independent parties engaged specifically to conduct the audit, typically subject to pre-determined auditing standards.'* (p. 558)

to ensuring information reliability– if a third-party auditor is directly contracted by the firm, there is a significant risk of 'opinion-shopping,' where firms look for the least stringent auditors or those most likely to provide a favourable outcome (Lam et al. 2024). Ensuring the reliability of both the information presented to auditors, as well as the information presented from the outcome of an audit, is crucial to the trust-building process of assurance.

## Case Study: Online Safety Technology Accreditation

Given the challenges presented above, the question of how a regulator or assurance body successfully creates an assurance process for AI/ML technologies, especially without codified standards, remains. Answering this question is of particular importance when audit/assurance processes, as well as their scope and purpose, are mandated by law as they are in the DSA, OSA, and AIA.

In the following sections, we present a case study for designing and operationalising a principles-based framework to assess technologies developed to detect terrorism and child sexual abuse and exploitation (CSEA) content. This assurance process is mandated by the UK OSA, but the legislation only goes so far as to say that these technologies should be 'accredited against minimum standards of accuracy', creating a gap between the expected outcome and the assurance process. The proposed approach for developing an assurance process breaks down 'accuracy' into principles, and then further into assessable objectives, which are then scored based on evidence provided in response to questions. Breaking down each principle into discrete, scorable objectives helps to address the issue of subjective assessment of principles-based frameworks described above, allowing these types of frameworks to be used as a basis for assurance and/or accreditation. Technology agnosticism is addressed through a holistic assessment, the scoring of which allows for flexibility between different technologies. Finally, information reliability is addressed through a fully independent assurance body which assesses presented evidence against clear objectives.

### Section 121 of the UK Online Safety Act, Minimum Standards of Accuracy, and Accreditation

The following approach has been developed to satisfy assurance in the context of measuring 'accuracy' of CSEA and terrorism detection technologies under the UK Online Safety Act 2023 (OSA) (UK Government 2023). Ofcom, the regulator responsible for the OSA, has additional powers to tackle two categories of illegal content: terrorism and child sexual abuse and exploitation (CSEA). Under section 121 of the Act, Ofcom has the power to issue a notice to the provider of a particular Part 3 service where considered 'necessary and proportionate' to deal with terrorism or CSEA content (or both) by requiring a regulated service provider to:

1. use technology that has been accredited ('accredited technology'), by Ofcom or another person appointed by Ofcom to identify and/or prevent individuals from encountering terrorism content communicated publicly; and/or
2. use accredited technology to identify and/or prevent individuals from encountering CSEA content communicated publicly or privately.

Before using this power, technologies must be 'accredited' against 'minimum standards of accuracy' published by the UK Secretary of State (an office held by certain senior ministers in the UK government.) Ofcom is responsible for providing advice to the Secretary of State for what these minimum standards should be, as well as setting up (or nominating a third party to set up) an accreditation scheme for technologies against these standards.

While developing the proposals for the Minimum Standards of Accuracy for these technologies to be accredited against, a significant challenge arose— 'accuracy' is not defined in the OSA. Our starting point was to consider the field of statistical accuracy– however, our view is that the minimum standards of accuracy should consider accuracy in its widest sociotechnical sense, rather than seeking to set a single (or set of) metrics based on statistical measurements of accuracy. For example, we found that given breadth and complexity of technologies that could potentially be accredited, we could not feasibly set a single (or set of) numerical thresholds for statistical 'accuracy'. In addition, statistical accuracy (defined as correct classifications/total classifications) is not sufficient when considering terrorism/CSEA content detection classification because it cannot by itself adequately reflect the real-world performance of a technology, particularly in the case of low-prevalence harms like CSEA and terror. Because different kinds of technologies would likely not be tested on the same dataset, the measurement of statistical 'accuracy' would also not be comparable across technologies.

This challenge led us to propose a principles-based approach to accuracy assessment in this context, which would account for further socio-technical factors throughout the development cycle of the technology which impact its accuracy in detecting terrorism/CSEA content. However, as discussed above, a principles-based approach to assessment is not traditionally used for evaluation of technologies, let alone accreditation, because of the subjectivity of assessing qualitative, normative principles. The sections below discuss the process of 1. Developing a bespoke, principles-based framework for this specific use case that is technology-agnostic; 2. Operationalising the principles into scorable objectives, which are assessed via questions, and 3. Developing a fair, objective scoring system for assessment.

### Developing a Principles-Based Framework for Terror/CSEA Detection Technologies

There are a broad range of technologies potentially in scope of this power—potential technologies to be accredited could include (but are not limited to): hashing; keyword matching; uniform resource locator (URL) detection; image-, text-, audio- or behavioural-based machine learning and AI classifiers; and rule-based technologies (UK Office of Communications 2024). Some technologies are designed to match

data to known, previously identified terrorism and CSEA content, while others are designed to also identify and flag previously unseen terrorism and CSEA content. These technologies identify content through a variety of mechanisms, such as machine learning, hard-coded rules, and mathematical optimisation. They also change and update frequently, in line with evolving harms in the terror and CSEA landscape. (UK Office of Communications 2024)

Where other bodies have attempted to codify technical assessments for a similarly broad group of technologies, they have typically opted to utilise principles-based frameworks for flexibility and breadth (Smit, Zoet, and van Meerten 2020). In this case, a principles-based framework should account for the factors beyond statistical accuracy that can provide a comprehensive understanding of a technology's accuracy in deployment. One example of this is the effect of biases— while a technology might perform well in a controlled test today, biases in training data could lead to decreased accuracy over time or in different use cases (Bogen et al. 2025). We have adapted the proposed principles from pre-existing frameworks, such as those described in (Floridi and Cowls 2019; Li et al. 2023; Díaz-Rodríguez et al. 2023) but have tailored the chosen principles to the very specific use case of Technology Notice accreditation.

We have identified four principles—technical performance, fairness, robustness and maintainability— as crucial to ensuring the accuracy of identification technologies. These principles help address the socio-technical factors that statistical performance testing often overlooks.

1. **Technical Performance**, which refers to the testing and reporting of a technology's ability to perform against specified metrics and technical requirements. The focus of this principle is on whether technologies can perform well at detecting terrorism and/or CSEA content in testing conditions
2. **Fairness**, which refers to the ability of a technology to avoid unfair bias and make equal and accurate decisions across different groups of people. We recognise that some bias may be inherent. However, identifying, reporting and mitigating for, or eliminating harmful bias is essential for technology to be, and remain, accurate.
3. **Robustness**, which refers to the ability of a technology to perform reliably and maintain functionality under various conditions, including unexpected or challenging scenarios. Technology that is not robust can result in a total compromise of accuracy, as performance becomes unreliable or ineffective when the conditions under which the technology operates change.
4. **Maintainability**, which refers to the ability of a technology to be modified, repaired, or updated to ensure its continued accuracy and performance over time, including in response to new threats. This is particularly relevant for terrorism and CSEA content detection, as malicious actors frequently change tactics to bypass detection. Without maintainability, the accuracy of technology could degrade over time.

These principles are mutually exclusive for this given deployment use case. To develop this framework, we undertook an iterative process of refining existing principles with the specific use case in mind. The selection of four principles also reflects a balance between comprehensiveness and practicality. Each principle addresses a distinct and critical dimension of technologies for detecting illegal content. Technical Performance addresses the fundamental question of whether the technology can perform its primary task effectively. This principle serves as the baseline requirement; if a system cannot effectively detect illegal content, other principles become irrelevant. Fairness focuses on whether the system works equally well for everyone, which is crucial for public trust, legal compliance, and preventing discriminatory outcomes. Robustness answers the question of whether the system will work reliably in real-world scenarios, bridging the gap between controlled lab testing and practical deployment, and is essential for ensuring consistent performance under varying conditions (for example, in relation to adversarial scenarios common in illegal content detection.) Maintainability addresses the long-term question of whether the system will continue to work over time. This principle is critical for adapting to evolving threats and ensuring sustainable implementation.

## Operationalising Principles via Objective Statements

Although the principles above are the basis of the assessment of the accuracy of technologies, they are not sufficient on their own for purposes of assessment without a robust and objective evaluation framework. To develop this, we have proposed breaking down each principle into a set of objectives, which technology developers would need to provide evidence for meeting. This evidence can then be audited and scored independently against set questions. The objectives are statements about how the technology has been developed, trained, or tested, and are intended to be scored against. This is like approaches adopted in other AI evaluation frameworks such as AI Verify. The objectives are intended to be technology agnostic, and applicable across the many kinds of terrorism and CSEA content detection technologies that could potentially be accredited. We have structured the proposed framework as a top-down hierarchy (see Figure 1), which is structured as:

- **Principles**: which define the general goal to be achieved by the technology.
- **Objectives**: statements which set specific outcomes that should be achieved in the development or testing of the technology.
- **Questions**: which provide details on the processes that should be undertaken to ensure the objectives have been met.
- **Evidence Levels**: which provide practical guidance on the information needed to meet the requirements presented by each question.
- **Score**: based on the presented evidence, a way to quantify the degree to which an objective has been met (has not been met, has been partially met, has been fully met).

### Scoring the Assessment

For each proposed question, we recommend the development of a scoring rubric based on the evidence provided by the applicant. This evidence will be independently evaluated by Ofcom or a nominated third party. A score will then be assigned to each objective according to the following system:

- **Five (5) points**: Robust and comprehensive evidence demonstrating the objective has been met.
- **One (1) point**: Limited evidence showing the objective has been partially met.
- **Zero (0) points**: No evidence that the objective has been met.

This scoring system ensures that a higher weight is given to robust, comprehensive evidence, and guarantees that solutions with limited evidence across all objectives will not pass the assessment. The evidence provided for the proposed questions cumulatively determine the overall score for each objective. Each principle score will then be weighted before being combined to form an overall score for the assessment. The principles of technical performance, robustness, and fairness will be given equal weight (30% each), while maintainability, considered slightly less critical in this context (given technology will be periodically reviewed for re-accreditation), will be weighted at 10%.

## Analysis

### Development of Principles of Accuracy

Evaluating technologies that address sensitive and subjective issues, such as detecting terrorism and CSEA content, requires a clear, balanced, and effective framework. In our initial drafting of the principles-based framework, based on our research we found that existing definitions for accuracy (particularly statistical accuracy) were not sufficiently comprehensive to address a technology's accuracy in real-world conditions. To address this, we needed to create our own principles-based framework for 'accuracy' by considering the sociotechnical issues that can impact a technology's ability to 'accurately' flag CSEA/terror content. The rationale for selecting the 4 principles presented in the previous section (technical performance, fairness, robustness, and maintainability) was grounded in feasibility and alignment with the current state of the art for the technologies being assessed.

First, the principles were designed to be **verifiable within reasonable constraints**. The objective statements created underneath each principle must be able to be proven or disproven, and ideally incorporate clear evaluation metrics such as precision, recall, and false-positive rates (for technical performance and robustness) or predictive parity (for fairness). For each objective, the developer of the technology would need to reasonably be able to provide evidence to show that they have met the proposed standards. Additionally, the framework was developed with critical considerations in mind, such as the availability of appropriate test data for illegal content, measuring performance on evolving threats, balancing detection rates with false-positive risks, and ensuring the solution functions effectively in real-world conditions. In addition, the principles were designed to be **ambitious yet achievable** given current technological capabilities. The framework necessitates testing which can feasibly be done during technology development and testing with readily available tools, but also encourages continuous improvement as technology and organisational security and risk mitigation measures continue to develop. This ensures the framework remains forward-looking while remaining practical for developers.

### Creating a Rubric for Objective Evaluation of Presented Evidence

In terms of assessing principles, our method of using levels of evidence and providing a rubric for third-party assessors (in this case, Ofcom or a nominated third-party) is informed by the need for objective and scorable criteria, as discussed in Section 2. Our approach ensures that principles are not only clearly defined but also objectively measurable, enhancing the overall reliability of the assessment process. This objectivity is crucial for maintaining the integrity of assessments and ensuring that they can be consistently applied across different harm types and technology types. One of the key challenges was determining how to assess principles effectively given the reliance on a developer directly providing information to a third-party assessor—to do this, we broke down each principle into objective statements, and then created questions to establish whether a statement is true or false. Under each question, we established levels of evidence and provided examples of relevant evidence and/or documentation we would expect to receive. This creates a 'rubric' for the information presented to the assessor, making principles objectively scorable.

However, this case study is different from previous assurance processes required by legislation, in that developers do not have to apply for accreditation unless they choose to do so, and thus the burden to present verifiable and truthful information falls upon them if they choose to apply. In this case, the developer is responsible for collating evidence and providing it in an accessible format to the accrediting body. If this is not done, the application is unlikely to be successful, with the only legal consequence being that the technology would not be one that Ofcom could require the use of via a Technology Notice. This contrasts with how some other legally-required audit processes operate—such as those under NY Local Law 144, where employers are prohibited outright from using technologies which have not been certified under an independent bias audit. Accreditation under Section 121 is also conducted by a single regulator or nominated third-party, and so the risk of 'opinion-shopping' is mitigated as developers cannot choose a different assessor with less stringent requirements—this is a function of the law, rather than the chosen mechanism for assessment.

### Technology Agnosticism and Assessment Scope

Ensuring that assessments are applicable across multiple technology types proved to be a significant challenge during the operationalisation of principles into scorable objectives and questions. For example, a technology based around

cryptographic hashing, which could potentially be eligible to apply for accreditation given the existence of hash lists for both known CSEA and terror content, would rely on an algorithm which matches hashes of photos/videos rather than any content contained in the photos/videos themselves. Developers of hash-matching technology often do not have access to the underlying content—they cannot assess the demographic representativeness, for example, of hash lists, as this information is not encoded in the photo/video's cryptographic hash. Thus, the testing which developers of hash-matching technologies undertake to assess the fairness and bias of their products must be different than that undertaken by developers of ML classifiers, who do have access to information about demographics and other characteristics. To address this challenge at the objectives level, we proposed the exemption of scoring for certain objectives where providers do not have visibility into the training data.

The variety of technologies that may feasibly apply for accreditation also influenced the development of the proposed assessment scoring mechanism— in-scope technologies are developed and tested in many different ways, and some have higher implementation costs (for compute, etc.) than others. In the proposal for scoring the assessment, recognising that these are intended to be *minimum* standards of accuracy, we sought to ensure that technologies which are effective, yet potentially not at the state-of-the-art level with regards to performance/fairness/robustness/maintainability are still able to be assessed, rather than prioritising passing solutions that are only feasible for large companies with substantial infrastructure and budgets to implement (e.g., highly complex and accurate solutions).

Additionally, we had to ensure that the assessment's scope did not expand to a point where it was infeasible to carry out. Had we attempted to design an assessment process which was scoped to include assessments for every, or even a majority of scenarios that may be encountered when detecting CSEA/terror content, we risked making the assessment so complex that no developer would undertake an application. To combat this, we have proposed objectives and questions that are generalisable across a multitude of different use cases, and allowed for flexibility in scoring evidence only for applicable questions.

## Conclusion

This paper has demonstrated the development and operationalisation of a principles-based framework for AI assurance, specifically targeting the accuracy of child sexual exploitation (CSEA) and terrorism detection technologies under the UK Online Safety Act (2023). The approach outlined in this case study holds significant relevance beyond the immediate context, offering valuable insights for other sociotechnical AI/ML systems. Given the widespread integration of AI across various sectors and the corresponding need for assurance and assessment, the process of developing a bespoke principles-based framework, and operationalising such framework into a scorable, objective assessment can be adapted to different domains, helping to ensure that AI systems are able to be assessed and assured in a manner which promotes holistic trustworthiness, safety, and ethical design.

The replicability of the presented approach to operationalisation is noteworthy– while the proposed framework was tailored to meet specific policy goals related to online safety, the process of operationalising high-level principles into objective, scorable criteria is broadly applicable. The frameworks upon which assessment are built are most effective when they are bespoke, addressing the unique challenges and requirements of the specific AI applications they are designed to assess. In other domains which require assessment, a similar process for building a bespoke principles-based framework could potentially be undertaken with an eye towards feasibility, proportionality, and alignment in that specific sector. As well, the underlying methodology of translating principles into measurable metrics and creating robust assessment processes can be replicated across different contexts and sectors.

Looking ahead, several concerns must be addressed to ensure the continued effectiveness and relevance of principles-based AI assurance frameworks. For the purposes of this case-study, the rapidly evolving landscape of online harms, particularly with the advent of generative AI content, poses new challenges that the presented framework must be able to adapt to. However, the legislation does not enable Ofcom to revise the minimum standards of accuracy after they are approved and published by the Secretary of State—rather, the OSA reserves this ability for the Secretary of State, and would require Ofcom to submit new advice. This risks limiting the ability of the framework to respond rapidly to changing technologies and evolving harms in the CSEA/terror landscape. We have sought to mitigate this risk by proposing a flexible, principles-based approach, and we recognise there are also likely to be benefits for parties seeking accreditation in having a stable minimum standard of accuracy that is not subject to frequent change. Future policymakers may consider how to strike the right balance between certainty for stakeholders and building sufficient flexibility into legislation which prescribes assessment to ensure that any frameworks or assessments can remain relevant over time. In addition, the lack of qualified auditors and assessors (as identified by Raji et. al. (2022)) remains a critical issue, as the effectiveness of any assurance process relies heavily on the expertise and integrity of those conducting the assessments. For the presented case study, if the proposed framework is adopted following consultation, significant practical decisions around implementation will need to be made about who is conducting these assessments, and the resources required to operationalise such an accreditation regime. In a broader sense, the need for increased resourcing and professionalisation of algorithmic auditors remains a pressing concern in the AI/ML assessment space.

While the operationalisation process of a principles-based framework presented in this paper offers a robust and adaptable approach to AI assurance, ongoing efforts are needed to address emerging challenges and ensure that these frameworks remain effective and relevant in the face of technological advancements and evolving societal needs.

## Acknowledgements

This paper is based on the Ofcom Technology Notices Consultation (2025). However, the paper represents the views and opinions of the authors and does not necessarily represent statements of concluded Ofcom policy.

## References

AI Verify Foundation. 2024. AI Verify Foundation - Building Trustworthy AI. https://aiverifyfoundation.sg/. Accessed: 2025-07-24.

Akbarighatar, P. 2024. Operationalizing responsible AI principles through responsible AI capabilities. *AI and Ethics*.

Boer, A.; de Beer, L.; and van Praat, F. 2023. Algorithm Assurance: Auditing Applications of Artificial Intelligence. In Berghout, E.; Fijneman, R.; Hendriks, L.; de Boer, M.; and Butijn, B.-J., eds., *Advanced Digital Auditing: Theory and Practice of Auditing Complex Information Systems and Technologies*, 149–183. Cham: Springer International Publishing. ISBN 978-3-031-11089-4.

Bogen, M.; Bankston, K.; Deshpande, C.; Joshi, R.; Radiya-Dixit, E.; and Winecoff, A. 2025. Assessing AI: Surveying the Spectrum of Approaches to Understanding and Auditing AI Systems. https://cdt.org/insights/assessing-ai-surveying-the-spectrum-of-approaches-to-understanding-and-auditing-ai-systems/.

Canca, C. 2020. Operationalizing AI ethics principles. *Commun. ACM*, 63(12): 18–21.

Costanza-Chock, S.; Raji, I. D.; and Buolamwini, J. 2022. Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 1571–1583. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9352-2.

Díaz-Rodríguez, N.; Del Ser, J.; Coeckelbergh, M.; López de Prado, M.; Herrera-Viedma, E.; and Herrera, F. 2023. Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. *Information Fusion*, 99: 101896.

European Commission. 2022. The EU's Digital Services Act. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en. Accessed: 2025-07-24.

European Commission, High-Level Expert Group on AI. 2020. Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment | Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment. Accessed: 2025-07-24.

Floridi, L.; and Cowls, J. 2019. A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*, 1(1). Publisher: The MIT Press.

Floridi, L.; Cowls, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; Dignum, V.; Luetge, C.; Madelin, R.; Pagallo, U.; Rossi, F.; Schafer, B.; Valcke, P.; and Vayena, E. 2018. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, 28(4): 689–707.

Goodman, E. P.; and Trehu, J. 2022. AI Audit Washing and Accountability.

Government of Canada. 2024. Guiding principles for the use of AI in government - Canada.ca. https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/principles.html. Accessed: 2025-07-24.

Hagendorff, T. 2020. The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, 30(1): 99–120.

Hasan, A.; Brown, S.; Davidovic, J.; Lange, B.; and Regan, M. 2022. Algorithmic Bias and Risk Assessments: Lessons from Practice. *Digital Society*, 1(2): 14.

Jobin, A.; Ienca, M.; and Vayena, E. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9): 389–399.

Koshiyama, A.; Kazim, E.; Treleaven, P.; Rai, P.; Szpruch, L.; Pavey, G.; Ahamat, G.; Leutner, F.; Goebel, R.; Knight, A.; Adams, J.; Hitrova, C.; Barnett, J.; Nachev, P.; Barber, D.; Chamorro-Premuzic, T.; Klemmer, K.; Gregorovic, M.; Khan, S.; Lomas, E.; Hilliard, A.; and Chatterjee, S. 2024. Towards algorithm auditing: managing legal, ethical and technological risks of AI, ML and associated algorithms. *Royal Society Open Science*, 11(5): 230859. Publisher: Royal Society.

Lam, K.; Lange, B.; Blili-Hamelin, B.; Davidovic, J.; Brown, S.; and Hasan, A. 2024. A Framework for Assurance Audits of Algorithmic Systems. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 1078–1092. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.

Li, B.; Qi, P.; Liu, B.; Di, S.; Liu, J.; Pei, J.; Yi, J.; and Zhou, B. 2023. Trustworthy AI: From Principles to Practices. *ACM Comput. Surv.*, 55(9): 177:1–177:46.

Maurer, R. 2024. New York City AI Law Is a Bust. https://www.shrm.org/mena/topics-tools/news/technology/new-york-city-ai-law. Accessed: 2025-07-24.

Mikalef, P.; Conboy, K.; Lundström, J. E.; and Popovič, A. 2022. Thinking responsibly about responsible AI and 'the dark side' of AI. *European Journal of Information Systems*, 31(3): 257–268. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/0960085X.2022.2026621.

Mittelstadt, B. 2019. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1(11): 501–507. Publisher: Nature Publishing Group.

Morley, J.; Floridi, L.; Kinsey, L.; and Elhalal, A. 2020. From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Science and Engineering Ethics*, 26(4): 2141–2168.

Mökander, J.; Morley, J.; Taddeo, M.; and Floridi, L. 2021. Ethics-Based Auditing of Automated Decision-Making Systems: Nature, Scope, and Limitations. *Science and Engineering Ethics*, 27(4): 44.

NIST. 2022. NIST AI RMF Playbook. https://www.nist.gov/itl/ai-risk-management-framework/nist-ai-rmf-playbook. Accessed: 2025-07-24.

Nonnecke, B. 2024. Mandated Third-Party AI Audits are Coming—Addressing AI's Socio-Technical Challenges Will Be Key | TechPolicy.Press. https://techpolicy.press/mandated-thirdparty-ai-audits-are-coming-addressing-ais-sociotechnical-challenges-will-be-key. Accessed: 2025-07-24.

OECD AI. 2024. AI Principles Overview. https://oecd.ai/en/principles. Accessed: 2025-07-24.

Percy, C.; Dragicevic, S.; Sarkar, S.; and d'Avila Garcez, A. 2021. Accountability in AI: From principles to industry-specific accreditation. *AI Communications*, 34(3): 181–196. Publisher: SAGE Publications.

Pouget, H. 2023. What will the role of standards be in AI governance? https://www.adalovelaceinstitute.org/blog/role-of-standards-in-ai-governance/. Accessed: 2025-07-24.

Prem, E. 2023. From ethical AI frameworks to tools: a review of approaches. *AI and Ethics*, 3(3): 699–716.

Raji, I. D.; Xu, P.; Honigsberg, C.; and Ho, D. 2022. Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 557–571. Oxford United Kingdom: ACM. ISBN 978-1-4503-9247-1.

Sadek, M.; Kallina, E.; Bohné, T.; Mougenot, C.; Calvo, R. A.; and Cave, S. 2024. Challenges of responsible AI in practice: scoping review and recommended actions. *AI & SOCIETY*.

Smit, K.; Zoet, M.; and van Meerten, J. 2020. A Review of AI Principles in Practice. *PACIS 2020 Proceedings*.

Soler, G. J.; De, N. S.; Bassani, E.; Sanchez, I.; Evas, T.; André, A.-A.; and Boulangé, T. 2024. Harmonised Standards for the European AI Act. https://publications.jrc.ec.europa.eu/repository/handle/JRC139430. Accessed: 2025-07-24.

Terzis, P.; Veale, M.; and Gaumann, N. 2024. Law and the Emerging Political Economy of Algorithmic Audits. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 1255–1267. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.

UK Department for Science, Innovation, and Technology. 2024. Introduction to AI Assurance. https://assets.publishing.service.gov.uk/media/65ccf508c96cf3000c6a37a1/Introduction_to_AI_Assurance.pdf. Accessed: 2025-07-24.

UK Government. 2023. Online Safety Act 2023. https://www.legislation.gov.uk/ukpga/2023/50. Publisher: Statute Law Database; Accessed: 2025-07-24.

UK Information Commissioner's Office. 2024. A guide to the data protection principles. https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-protection-principles/a-guide-to-the-data-protection-principles/. Accessed: 2025-07-24.

UK Office of Communications. 2024. Technology Notices to deal with terrorism content and/or CSEA content Consultation on policy proposals for minimum standards of accuracy for accredited technologies, and guidance to providers. https://www.ofcom.org.uk/siteassets/resources/documents/consultations/category-1-10-weeks/consultation-technology-notices/main-document/technology-notices-consultation.pdf?v=388881. Accessed: 2025-07-24.

Young, M.; Katell, M.; and Krafft, P. 2022. Confronting Power and Corporate Capture at the FAccT Conference. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 1375–1386. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9352-2.