



**Certification for Ethical and Regulatory Transparency
in Artificial Intelligence**

D6.1: TEMPLATE FOR DEVELOPING TESTING PROTOCOLS FOR A GIVEN AI APPLICATION



**Co-funded by
the European Union**

Project funded by



**Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra**

Swiss Confederation

**Federal Department of Economic Affairs,
Education and Research EAER
State Secretariat for Education,
Research and Innovation SERI**

Work package	WP6
Task	Task 6.1
Due date	30/06/2026
Submission date	03/08/2026
Deliverable lead	IDIAP
Version	1.0
Authors	Sushil BHATTACHARJEE (IDIAP), Olena Denysenko (UT), Natael Couturier (RAL), Lea El Khoury (RAL)
Reviewers	Martin HAYFORD (UT), Pierre GACON (IPS), Claude BAUZOU (IPS), Annalisa FRANCO (University of Bologna), Davide MALTONI (University of Bologna)
Security reviewer	Pierre GACON (IPS)
Abstract	Under certain circumstances, an AI product aspiring to enter the EU marketplace may be subjected to a conformity assessment to the EU AI Act. Such a conformity assessment is performed by a Notified Body. In some cases, the Notified Body may ask an independent, accredited testing lab to perform a technical evaluation of the AI product. This document provides guidelines for testing labs (both, independent accredited labs as well as in-house testing labs of technology providers), on how to construct a comprehensive test plan for such conformity assessment. The test plan should cover the five categories of tests: functional testing, adversarial testing, fairness testing, explainability testing, and cybersecurity testing. In this document we describe how to define the scope and the objectives for each category of testing, to prepare an adequate test plan for the AI product in question.
Keywords	Functional testing, adversarial testing, fairness testing, explainability testing, cybersecurity testing, test-report template.

Document Revision History

Version	Date	Description of change	List of contributor(s)
V0.1	24/06/2026	First draft submitted for review	IDIAP
V0.2	30/06/2026	Draft submission	IDIAP, IPS, UT
V1.0	03/08/2026	Final cleaned version	IDIAP

Grant Agreement No: 101189650 | Topic: HORIZON-CL4-2024-DATA-01-01

Call: HORIZON-CL4-2024-DATA-01. | Type of action: HORIZON-IA

REVIEWERS COMMENTS

Dr. Pierre GACON	IPS	Date of review:24/06/2026
The document provides a methodology to define a test plan to evaluate the conformity of an AI system to some of the requirement from the AI Act. The methodology is, in particular, illustrated by a biometric use case. The mapping between the AI Act and the content could be improved. As security officer, I do not see any issue with the proposed content.		

Prof. Martin HAYFORD	UT	Date of review: 24/06/2026
Provided comments by email. He suggested adding a section giving an overview of the EU AI Act, as relevant to compliance testing. This section has been added, as a contribution from UT.		

Dr. Claude Bauzou	IPS	Date of review:
Asked for clarification of the entity doing the test, and provided detailed feedback on individual paragraphs as margin comments.		

The University of Bologna (UBO) is not a member of the CERTAIN consortium. They -- specifically, Prof. Annalisa Franco and Prof. Davide Maltoni, members of the Biometric System Lab. at the UBO -- have been invited by IPS as external reviewers of this deliverable, as they have vast experience in testing biometrics systems. They have offered insightful comments, that have helped improve the clarity of this document.

Dr. Annalisa Franco	Univ. Of Bologna (UBO)	Date of review: 28/07/2026
Provided detailed feedback, by email, on several sections. The document has been updated to take her comments into account. Her main comments were as follows: It is important to adhere to the standardized biometric vocabulary; sometimes the terms used in the text do not comply with the standard. Section 4 contains high-level concepts (objectives and certification criteria) that should be introduced earlier and integrated with the preceding sections. Some sections currently mix aspects of other testing categories (e.g. the Functional Testing section includes aspects of adversarial and fairness testing). These topics should be more clearly separated to improve the structure and readability. The proposed certification workflow should be explained more clearly, preferably with a diagram or flowchart, as it is not sufficiently clear from the current description:		

The AI provider should specify the certification targets according to the intended application and propose the datasets, operational boundaries, and evaluation metrics.

The certification body should review the submitted documentation, proposed targets, datasets, and metrics, and, where necessary, negotiate revisions or additional requirements with the provider.

Once the certification scope, targets, and evaluation methodology have been agreed upon, targeted testing should be conducted using the predefined datasets and metrics. Based on the results, the certification body either grants the certification or requests further modifications before certification can be issued.

Prof. Davide Maltoni	UBO	Date of review: 28/07/2026
Provided comments by email, on several sections of the document. The document has been updated to take his comments into account.		

DISCLAIMER



Co-funded by
the European Union

Project funded by



Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Economic Affairs,
Education and Research EAER
State Secretariat for Education,
Research and Innovation SERI

The CERTAIN project received funding from the European Union's Horizon Europe Research and Innovation Programme under Grant Agreement No 101189650. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them. This work has received funding from the Swiss State Secretariat for Education, Research and Innovation (SERI).

COPYRIGHT NOTICE

© 2025 – 2027 CERTAIN

Project funded by the European Commission in the Horizon Europe Programme		
Nature of the deliverable:	R (report)	
Dissemination Level		
PU	Public, fully open, e.g. web (Deliverables flagged as public will be automatically published in CORDIS project's page)	☐

*

TABLE OF CONTENTS

Reviewers comments.....3

DISCLAIMER.....5

Copyright notice.....5

1 INTRODUCTION.....12

2 CONFORMITY ASSESSMENT UNDER THE EU AI ACT15

2.1 What is conformity assessment 15

2.2 When is a third-party assessment required? 15

2.3 The Annex VII Procedure 16

2.4 Relevance to this document 17

3 TYPES OF TESTING18

3.1 Functional Testing.....18

3.2 Adversarial Testing18

3.3 Fairness Testing19

3.4 Explainability TESTING19

3.5 Cybersecurity Testing.....20

4 TEST SCOPE AND OBJECTIVES.....22

4.1 functional Testing22

4.2 Adversarial Testing27

4.3 Fairness Testing28

4.4 Explainability Testing.....30

4.5 Cybersecurity Testing.....33

4. INTEGRITY OF GROUND TRUTH FOR TEST DATA.....42

5. TEST REPORT TEMPLATE43

6. CONCLUSIONS51

EXECUTIVE SUMMARY

AI products marketed in the EU must comply with the EU AI Act, which classifies AI products into five categories by risk: *Unacceptable risk*, *High risk*, *Limited risk*, *Low risk*, and *General purpose AI*. Products in the Unacceptable risk category are banned in the EU. For AI products in the High-risk category, the EU AI Act envisages mandatory evaluation by a Notified Body (NB) — formally called third-party conformity assessment. National Market Surveillance Authorities (MSA) may also require specific AI products in other categories to undergo third party conformity assessment under certain circumstances.

The aim of conformity assessment is to verify that the AI product meets the safety and fundamental rights requirements mandated under the EU AI Act. Conformity assessment consists of three primary components: the *core compliance requirements* (what is being assessed), the *procedural pathways* (how it is assessed), and the *post-assessment deliverables* (the final compliance proof).

As laid out in Article 43, two procedural pathways are possible: (1) internal control by the AI system provider, or (2) assessment by a NB. In the second case (where conformity assessment is performed by a NB) the NB may determine that an evaluation, of the capabilities of the AI product in question, by a testing lab (an AL) is also required as part of the conformity assessment. *This process of evaluation of an AI product by an independent AL is the scope of the current document.*

More specifically, the purpose of this document is to serve as a guide for testing labs on how to create a comprehensive test plan for assessing the compliance of an AI product to the EU AI Act. The test plan should cover five categories of tests:

- functional tests, that assess whether the AI product performs according to the behaviour expected under normal operating conditions;
- adversarial tests, that evaluate the resilience of the AI product to possible attacks;
- fairness tests, that evaluate whether the AI product behaves in the same way (without any bias) for different demographic categories, considering several factors such as gender, age and ethnicity;
- explainability tests, that determine whether the AI product has sufficient transparency to support human supervision and oversight, as well as the legal right to explanation of any affected person; and
- cybersecurity tests, that evaluate whether the system can withstand malicious attempts to compromise its operation, manipulate its outputs, or undermine the reliability of its decisions.

For each testing category, the lab should define the scope and objectives of the required tests. The testing scope clearly outlines the boundaries of the tests. Auditors use the scope document to determine the domain of inputs for which the AI product is expected to show expected behaviour. The test objectives are useful in designing the exact tests for each purpose. Test objectives are derived by mapping legal requirements from various articles in the EU AI Act to quantifiable metrics. In this document we show how to define the scope and objectives for the five testing categories.

The document also includes a section on ensuring the integrity of ground-truth for test data, as well as a section on how to create a test report, closing with an example test report template for evaluating a face-recognition product to be deployed in a border-control scenario.

LIST OF TABLES

TABLE 1: GLOSSARY9

TABLE 2: ACRONYMS9

TABLE 3: FIVE CATEGORIES OF TESTS TO BE INCLUDED IN A TEST PLAN FOR ASSESSING CONFORMANCE OF AN AI PRODUCT TO THE EU AI ACT. 13

TABLE 4: EXAMPLES OF SPECIFIC BOUNDARIES (PERFORMANCE LIMITS) ON RELEVANT METRICS SHOULD BE PRESCRIBED IN THE TESTING PROTOCOL. THE TABLE SHOWS OF EXAMPLES RELEVANT FOR TESTING A FACE-RECOGNITION SYSTEM TO BE USED IN A BORDER-CONTROL APPLICATION. 25

TABLE 5: MAPPING VARIOUS ARTICLES OF THE EU AI ACT TO THE VARIOUS OBJECTIVES OF FUNCTIONAL TESTING..... 25

TABLE 6: MAPPING EU LEGAL CONCEPTS TO MEASURABLE OBJECTIVES FOR FAIRNESS TESTING. 29

TABLE 7: MAPPING EXPLAINABILITY MANDATES IN THE EU AI ACT TO QUANTIFIABLE METRICS FOR EXPLAINABILITY METRICS..... 31

TABLE 8: MAPPING EU AI ACT MANDATES TO OBJECTIVES OF CYBERSECURITY TESTING. 35

TABLE 9: LIST OF MANDATORY SECTIONS OF A TEST REPORT..... 43

TABLE 10: EXAMPLE TEST REPORT TEMPLATE FOR TESTING A FACE RECOGNITION PRODUCT FOR A BORDER-CONTROL APPLICATION..... 44

GLOSSARY AND ACRONYMS

The list of terms, group of words, expression used in this document are explained in Table 1 below.

Table 1: *Glossary*

Term, expression	Definition
F1 Score	A metric used in machine learning to characterize the performance of a classifier. It is expressed as the harmonic mean of the precision and recall of the classifier being evaluated. $F1 = 2 * P * R / (P + R)$, where P represents the estimated Precision of the classifier, and R represents its Recall.
Precision	A metric used in machine learning when evaluating classifiers. For binary classifiers, it is expressed as the ratio: 'True Positives' / ('True Positives' + 'False Positives'). It gives the proportion of correct results among all the test samples identified as 'Positive' by the classifier.
Recall	A metric used in machine learning when evaluating classifiers. For binary classifiers, it is expressed as the ratio: 'True Positives' / ('True Positives' + 'False Negatives'). It gives the proportion of the test samples correctly identified as 'Positive' by the classifier, among all the Positive samples in the test dataset.
Test Oracle	A rule, or a baseline, used to determine whether the target of Evaluation (TOE) passed or failed a test. This may be strict threshold defined for a specific metric, or may be defined in statistical terms (e.g., confidence bounds).
Testing Lab	In this document, this is the entity that carries out conformity tests in the context of the AI act. This may be an internal testing lab of the AI provider, or this could be an independent accredited lab (AL).

The list of acronyms used in this document and their meaning is explained in Table 2 below.

Table 2: *Acronyms*

Acronym	Meaning
AL	'Accredited Lab.' A testing lab that has been accredited by the relevant National Accreditation Body, to evaluate products wishing to enter the EU marketplace.
APCER	'Attack Presentation Classification Error Rate.' A metric used in adversarial evaluations in biometrics. Defined in the ISO/IEC 30107-3 standard, it refers to the proportion of presentation-attacks that are falsely classified by a PAD system as <i>bona fide</i> presentations.
API	'Application Programming Interface.'
CCTV	Closed-Circuit TeleVision (video camera system)

Acronym	Meaning
eMRTD	Electronic Machine Readable Travel Document
FMR	‘False Match Rate.’ A metric used in biometrics evaluations. Defined in the ISO/IEC 19075-1 standard, it refers to the proportion of zero-effort impostor samples that are falsely declared by a biometrics system as true matches. This metric is expressed as a percentage.
FNMR	‘False Non-Match Rate.’ A metric used in biometrics evaluations. It is defined in the ISO/IEC 19795-1 standard, and refers to the proportion of genuine samples that are falsely declared by a biometrics system as zero-effort-impostors. This metric is expressed as a percentage.
FR	‘Face Recognition.’ An algorithm or system designed to identify a person based on a face-image or face-video of the person.
FTE	‘Failure To Enroll.’ A metric used in evaluation of biometric products. It is defined in the standard ISO/IEC 19795-1, and refers to the proportion of the population for whom the biometrics system fails to complete the enrolment process.
IAPAR	‘Impostor Attack Presentation Acceptance Rate.’ A metric used in the evaluation of biometrics products. It is defined in the standard ISO/IEC 30107, and refers to the proportion of presentation attack samples that were incorrectly accepted by the biometrics system (the full system, not just the PAD sub-system).
ICAO	‘International Civil Aviation Organization.’ An industry body that sets, among other things, the requirements for face-images in biometric passports.
ICT	Information and Communication Technologies.
IEC	International Electrotechnical Commission
ISO	International Organization for Standardization
JPEG	Joint Photographic Experts Group, an image and image compression standard
LIME	‘Local Interpretable Model-agnostic Explanations.’ It is a testing technique used to understand why a machine learning model makes a specific prediction. It works by building a simple, interpretable local surrogate model around a single data point to approximate complex black-box decisions.
NB	‘Notified Body.’ It is an independent, third-party organization designated by a EU member state to assess the conformity of regulated products against specific EU legislation before they are placed on the market.
MRZ	‘Machine Readable Zone.’ It appears in most modern identity documents. It is a specific zone that contains text and symbols printed in a specific way, so that machines can easily scan and decode the text.
ODD	Operational Design Domain.
OOD	Out Of Distribution.
PAD	‘Presentation Attack Detection.’ An algorithm or system (combining both hardware and software components) to detect presentation attacks, which are a form of adversarial attack on biometrics systems.

Acronym	Meaning
PAI	Presentation Attack Instrument
PKI	'Public Key Infrastructure.' A PKI certificate is a digital credential providing cryptographic proof of identity.
QMS	Quality Management System
SHAP	'SHapley Additive exPlanations.' It is a framework, based on Game Theory, for Explainable AI testing.
TOE	'Target Of Evaluation.' Refers to the exact version of the product being evaluated. The evaluation results are valid only for this version of the product.
WSQ	Wavelet Scalar Quantization: an image compression algorithm

1 INTRODUCTION

This document is concerned with a specific and legally defined process under the EU AI Act (Regulation EU 2024/1689)¹, namely, conformity assessment of AI systems.

The AI Act uses a risk-based approach to regulate AI products in the European Union market. It categorizes AI products under five risk levels: *Unacceptable risk*, *High risk*, *Limited risk*, *Low risk*, and *General purpose AI (GPAI)*. Products in the Unacceptable risk category are banned in the EU. For AI products in the other four categories, the AI Act lays out the conditions dictating conformity assessment. For AI products in the High-risk category, conformity assessment is mandatory before the product may be placed in the EU market. Conformity assessment is also required for AI products that may carry systemic risk.

Conformity assessment verifies that the AI product meets the safety and fundamental rights requirements mandated under the EU AI Act. The process consists of three primary components: the *core compliance requirements* (what is being assessed), the *procedural pathways* (how it is assessed), and the *post-assessment deliverables* (the final compliance proof).

Core compliance requirements set out the topics that are to be covered in the conformity assessment. Typically, these include evaluation of the risk management system, data governance, technical documentation, logging, user-information and transparency, human oversight, accuracy, robustness and cybersecurity,

As laid out in Article 43 of the AI Act, two procedural pathways are possible: internal control by the AI system provider, or assessment by a Notified Body (NB). Though not a legally distinct category in the EU AI Act, there is also a third entity that can participate in the lifecycle of a conformity assessment — an Accredited Testing Lab, or Accredited Lab (AL). Accredited Labs do not possess the regulatory designation to carry out conformity assessments or issue regulatory certificates of compliance, but they do have the various necessary clearances to perform the testing required for the conformity assessment process. NBs may be ALs themselves, or ALs can be subcontracted to perform tests — either by the AI technology provider before the submission of a conformity assessment application to the NB, or by the NB itself in the case further testing or verification of testing is required.

Post-assessment deliverables, triggered by the successful completion of the assessment include three final bureaucratic components required for legal market entry: EU declaration of conformity, CE Marking, and registration in a centralized publicly available database managed by the EU.

As we see from the core compliance requirements, results of a rigorous test may need to be included in the conformity assessment report for certain AI products. Such testing would cover aspects such as accuracy, robustness, fairness, explainability (human oversight) and cybersecurity of the product. In some cases, the conformity assessment tests will need to be performed by third-party testing lab (AL); in other cases, the AI system provider may perform the testing internally, and provide the results to the NB.

This document focuses on the construction of an adequate test plan for conformity assessment. Such a test plan may be created either by an AL for third-party conformity assessment or by AI technology providers performing internal testing. Specifically, we propose a template for constructing test plans for products in various AI applicable domains. The wide variety of tests pertinent to the evaluation of an AI product can be grouped into five categories. **Erreur ! Source du renvoi introuvable.** lists the five categories, and summarizes the general purpose of each test category.

¹ The EU AI Act: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>

Table 3: Five categories of tests to be included in a test plan for assessing conformance of an AI product to the EU AI Act.

	Testing Type	Purpose
1	Functional Testing	Verifies the capability of the product
2	Adversarial Testing	Ensures that the product is safe
3	Fairness Testing	Verifies that it works uniformly for everyone
4	Explainability Testing	Ensures human supervisors understand how it works
5	Cybersecurity Testing	Verifies that the system is resilient against unauthorized access, data poisoning, and malicious manipulation

The general idea of each of these five testing categories is as follows.

Functional testing verifies the claimed capabilities of the Target of Evaluation (TOE). It evaluates the accuracy and reliability of the TOE. In the legal context, functional testing proves that the AI system is fit for its intended purpose under normal operating conditions².

AI systems are vulnerable to a range of attack vectors. Standard functional testing does not evaluate the robustness of an AI system to attacks.

Adversarial testing is used to evaluate the security and robustness of the TOE to attacks. Such tests simulate malicious exploits and severe edge cases to verify that the AI system cannot be manipulated into generating incorrect results.

Fairness testing verifies whether the behaviour of the TOE is uniform across different demographic factors. AI models frequently inherit or amplify biases hidden in their training data. Data governance mandates require developers to monitor, detect, and mitigate these biases before deployment^{3 4 5 6}. Fairness testing analyses the outputs produced by the TOE across diverse demographic groups to ensure equal treatment and compliance with fundamental EU human rights laws⁷.

Explainability testing verifies that humans can correctly interpret the functioning and behaviour of an AI system. AI systems should support sufficient transparency to facilitate human oversight. For example, if an AI system denies someone a loan or a job, human operators must be able to interpret and explain *why* the AI system

² <https://www.testinglab.net/artificial-intelligence-and-robotics-product-testing>

³ <https://www.testim.io/blog/whats-ai-testing/>

⁴ <https://www.centercode.com/blog/beta-testing-ai-ml-products>

⁵ <https://nexos.ai/blog/ai-ethics/>

⁶ <https://cyberfortgroup.com/demystifying-the-eu-ai-act-what-you-need-to-know/>

⁷ <https://www.appen.com/ai-regulatory-compliance>

made that recommendation^{8 9}. Black-box AI systems prevent effective human oversight. The EU AI Act imposes strict transparency obligations, requiring that high-risk AI systems be designed so that human overseers can understand how the AI system arrives at its outputs^{10 11}. Explainability testing aims to determine whether the decision-making logic of the TOE can be translated into clear, traceable documentation, enabling users to exercise their legal right to an explanation.

Cybersecurity testing is providing objective evidence that the implemented security measures are adequate for the intended purpose and deployment environment of the AI system. It evaluates whether the system can withstand malicious attempts to compromise its operation, manipulate its outputs, or undermine the reliability of its decisions.

The key for preparing a suitable test plan is to first define the testing scope and the testing objectives. The scope definition clearly delineates the boundaries of the tests. The scope document informs auditors about the domain of inputs for which the AI product has been evaluated. If the AI product passes the compliance assessment, then auditors may assume that within the specified scope of the tests, the product complies with the relevant articles of the AI Act.

Testing objectives help to construct the exact tests to be included in the test plan. The objectives are determined by mapping relevant legal mandates in the AI Act to measurable quantities, as well as performance thresholds.

We start in S15ection 2 with a detailed discussion about the relevance of conformity assessment in the AI Act, followed by a overview, in Section 3 of the types of tests to be covered in conformity testing. The main body of this document is contained in Section 4, where we discuss in detail how to develop the scope and testing objectives for the five categories of tests. As a running example, we have used a hypothetical face-recognition (FR) product deployed in a border-control application to illustrate the various scopes and lists of objectives. (The reason for a biometrics product as an example is that such biometrics products usually fall into the high-risk category, which typically require mandatory conformity assessment.) In Section 4 we discuss methods that testing labs can employ to maintain and verify the integrity of ground-truth information for the test datasets. The essential components of a comprehensive test report are discussed in Section 5. We conclude the report with a summary in Section 66.

⁸ <https://dxc.com/insights/knowledge-base/paper/responsible-ai>

⁹ <https://emildai.eu/the-eu-ai-act-and-its-adherence-to-the-european-convention-on-human-rights/>

¹⁰ <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>

¹¹ <https://www.kiteworks.com/risk-compliance-glossary/eu-ai-act/>

2 CONFORMITY ASSESSMENT UNDER THE EU AI ACT

This section provides a brief overview of the relevant legal framework for readers who may not be familiar with the specific provisions of the Act that govern conformity assessment. Readers seeking a broader explanation of the EU AI Act and its implications are advised to look at CERTAIN Project's Deliverable D3.1: Comprehensive Legal Analysis And Map ¹², which covers the legal landscape of the AI Act in more detail.

2.1 WHAT IS CONFORMITY ASSESSMENT

Conformity assessment is the formal procedure by which a provider of a high-risk AI system demonstrates that the system meets the requirements set out in Chapter III, Section 2 of the AI Act. These requirements cover risk management, data governance, technical documentation, transparency, human oversight, as well as accuracy, robustness and cybersecurity. A conformity assessment must be completed before a high-risk AI system is placed on the EU market or put into service, and must be repeated whenever a previously assessed system undergoes a substantial modification.

2.2 WHEN IS A THIRD-PARTY ASSESSMENT REQUIRED?

Article 43 of the AI Act governs which procedure applies, and the answer depends heavily on which category of high-risk AI system is in question (as listed in Annex III of the AI Act). For high-risk AI systems falling under point 1 of Annex III (biometrics, including remote biometric identification, biometric categorisation, and emotion recognition) the provider may, in principle, choose between the two routes. However, this choice is only available where the provider has applied the relevant harmonised standards or common specifications. Where this is not the case, the provider must follow the third-party route (Annex VII). Specifically, Annex VII becomes mandatory for point 1 systems where:

- no harmonised standard exists for the relevant domain, and no common specification is available;
- the provider has not applied the harmonised standard, or has applied only part of it;
- relevant common specifications exist but have not been applied; or
- the applicable harmonised standard has been published with a restriction, in which case third-party assessment is required only for the restricted part. In practice, since harmonised standards under the AI Act are still being developed in many domains, this is currently a significant and common trigger for third-party involvement.

For all other high-risk AI systems – that is, those falling under points 2 to 8 of Annex III, covering domains such as critical infrastructure, education, employment, access to essential services, law enforcement, migration and border management, and the administration of justice – the Act mandates the internal control procedure (Annex VI) without any provision for third-party assessment. In other words, third-party conformity assessment under Annex VII is, at present, essentially confined to biometric AI systems under point 1 of Annex III.

Independently of system category, a new conformity assessment is required whenever a previously assessed high-risk AI system undergoes a substantial modification – a change affecting its compliance with the Act or its intended purpose. The same rules above determine whether the new assessment must again involve a NB.

¹² <https://certain-project.eu/wp-content/uploads/2026/02/CERTAIN-WP3-D3.1-comprehensive-legal-analysis-and-data-map-V1.0.pdf>

2.3 THE ANNEX VII PROCEDURE

Where third-party conformity assessment is triggered, the procedure to be followed is set out in Annex VII of the Act. It comprises two parallel components, both initiated by applications that the provider lodges with a NB of its choosing.

The first component is an assessment of the provider's QMS. The application must include, among other things, the technical documentation for each AI system covered by the QMS and documentation covering all aspects of the QMS required under Article 17. The NB assesses whether the QMS satisfies those requirements and notifies the provider of its decision, with reasoned conclusions. The approved QMS must continue to be implemented and maintained by the provider, and any intended change to it must be brought to the attention of the NB for re-examination.

The second component is an assessment of the technical documentation for the specific AI system the provider intends to place on the market. The NB examines this documentation and, where relevant, is granted access to the datasets used for training, validation, and testing the AI system. It may require the provider to supply further evidence or carry out further tests. Where it is not satisfied with the provider's own testing, it may itself carry out further tests. In exceptional cases – only after other reasonable means of verification have been exhausted, and upon a reasoned request – the NB may also be granted access to the trained model itself, subject to applicable intellectual property and trade secret protections.

If the AI system is found to be in conformity with the relevant requirements, the NB issues a Union technical documentation assessment certificate, identifying the provider, the conclusions of the examination, and any conditions of validity. If the system is not in conformity, the NB refuses to issue the certificate and provides detailed reasons. Where the deficiency concerns the quality of training data specifically, the refusal must address this explicitly, and retraining will be required before a new assessment can be sought. Any subsequent change to the AI system that could affect its compliance must be reported to the NB that issued the certificate, which would then determine whether the change requires a full new conformity assessment or can be addressed through a supplement to the existing certificate. Finally, the NB conducts periodic surveillance audits – with access to the provider's premises and relevant information – to verify that the approved QMS continues to be properly maintained, and may carry out additional testing of certified systems during these audits.

Under the regulatory framework of the EU AI Act, a fundamental legal and functional distinction exists between Notified Bodies (NB) and accredited testing laboratories (AL). Notified Bodies are independent, third-party organizations officially designated by EU Member States under Articles 30–34 to conduct formal conformity assessments and issue the legal certificates necessary for high-risk AI systems to bear the CE mark. In contrast, accredited testing laboratories – typically operating under standards such as ISO/IEC 17025 – are technical facilities equipped to perform specialized measurements, evaluations, and performance tests, but they lack the legal authority to issue regulatory certificates or determine final compliance.

During the lifecycle of a conformity assessment, these entities interact through both direct provider engagement and official oversight. AI providers frequently contract accredited testing laboratories voluntarily during the pre-compliance phase to generate objective performance and bias reports that substantiate their technical documentation. When formal assessment begins, Notified Bodies leverage their own qualified internal testing capacities or outsource specific operations under Article 33 of the EU AI Act (Subsidiaries of notified bodies and subcontracting). Under this provision, a Notified Body may subcontract specialized technical testing tasks to external accredited laboratories—provided the subcontractor meets the required competence standards, the relevant notifying authority is informed, and the agreement of the AI provider is secured. Regardless of whether in-house testing capacity or external subcontractors are utilized, the Notified Body maintains absolute legal responsibility for the work performed, the final conformity assessment, and the certification decision.

The content of this deliverable directly addresses accredited testing labs, whether it is an NB with the resources and expertise to perform the necessary tests in-house, or another AL that has been contracted by the AI provider or subcontracted by the NB itself.

2.4 RELEVANCE TO THIS DOCUMENT

The test plan guidelines and templates developed in this deliverable are directly oriented toward the Annex VII procedure described above. The five categories of tests addressed here – functional testing, adversarial testing, fairness testing, explainability testing, and cybersecurity testing – map to the substantive compliance requirements of Chapter III, Section 2 that a NB must examine when assessing an AI system's technical documentation. The running example used throughout this document, a face-recognition system deployed in a border-control context, falls under point 1 of Annex III and therefore sits squarely within the scope of systems for which third-party conformity assessment under Annex VII applies.

3 TYPES OF TESTING

In this section, we discuss each of the five types of testing in more detail.

3.1 FUNCTIONAL TESTING

Functional testing validates that the TOE performs its intended tasks reliably and accurately under standard operating conditions. It ensures the AI system meets the performance claims made in its user documentation and technical specifications. For AI products, functional testing verifies the claimed accuracy, precision, and alignment to stated business requirements, using standard, expected, and clean real-world inputs. Functional testing focuses four key areas:

- **Performance of Core Algorithm:** Estimating standard metrics such as Precision, Recall, F1-Score, False Match Rates (FMR), False Non-Match Rates (FNMR) and so on.
- **Boundary Values:** Testing system behaviour at the limits of standard operational specifications (e.g., maximum permitted head tilt, or lowest acceptable camera resolution in the field of biometrics).
- **API and Pipeline Integrity:** Verifying that data formats, sensor inputs, and output messages flow correctly through all software hooks without performance drops.
- **Regression Tracking:** Running tests with identical datasets against new software-patches to confirm that software updates do not break previously stable functions^{13 14}.

3.2 ADVERSARIAL TESTING

Adversarial testing evaluates how resilient the AI model is against intentional, malicious manipulation or exploitation designed to trick, break, or extract data from the system^{15 16}. Inputs involved in adversarial testing may include corrupted data, perturbed inputs, and inputs derived by reverse-engineering. Fuzz testing, where randomized inputs are used for stress testing (to test for vulnerabilities, hallucinations, and system crashes), is also part of adversarial testing. Adversarial testing evaluates the resilience of the TOE to several kinds of attacks. Some examples of attacks in different AI application domains are listed below.

- **Presentation Attacks:** Testing the system against presentation attacks designed to deceive the TOE (e.g., physical 3D facial masks or digital adversarial clothing patterns).
- **Backdoor Attacks:** Evaluating how well the system resists or flags contaminated training data updates intended to cause intentional model decay.
- **Stress and Error Resilience:** Ingesting random, broken, or heavily corrupted input streams (fuzzing) to confirm the system fails safely into a locked state rather than breaking open^{17 18}.

¹³ Tormos B, Pla B, Sánchez-Márquez R, Carballo JL. Explainable AI Using On-Board Diagnostics Data for Urban Buses Maintenance Management: A Study Case. *Information*. 2025; 16(2):74. <https://doi.org/10.3390/info16020074>

¹⁴ <https://www.idslogic.co.uk/functional-testing/>

¹⁵ Mylrea M, Robinson N. Artificial Intelligence (AI) Trust Framework and Maturity Model: Applying an Entropy Lens to Improve Security, Privacy, and Ethical AI. *Entropy* (Basel). 2023 Oct 9; 25(10):1429. doi: 10.3390/e25101429.

¹⁶ <https://smartdev.com/ai-model-testing-guide/>

¹⁷ Ilari, L., Tiribelli, S. & Caruso, F. From AI security to ethical AI security: a comparative risk-mitigation framework for classical and hybrid AI governance. *AI Ethics* 6, 91 (2026). <https://doi.org/10.1007/s43681-025-00935-x>

¹⁸ Pareek, C.S. From Detection to Prevention: The Evolution of Fraud Testing Frameworks in Insurance through AI. *Journal of Artificial Intelligence, Machine Learning & Data Science*, 2023, 1(2): 1-7.

3.3 FAIRNESS TESTING

Fairness testing systematically evaluates the AI system for discriminatory patterns, performance disparities, or inequitable outcomes across protected demographic characteristics¹⁹. Fairness testing is based on demographically diverse datasets, spanning variations in gender, skin tone, race, age, and physical disabilities^{20 21}. Specifically, fairness testing should cover the following four areas:

- **Demographic Parity:** Assessing whether the system passes or rejects users at approximately equal proportional rates (in a statistical sense) across distinct population subsets.
- **Equalized Performance Metrics:** Measuring if specific errors (such as high FMR for a biometrics product) skew disproportionately toward a specific group (e.g., checking if the model misidentifies biometric samples with darker skin tones more frequently than those with lighter skin tones).
- **Algorithmic Bias Auditing:** Reviewing the statistical distribution of internal weights or feature selections to flag hidden proxy variables for protected attributes.
- **Data Under-representation Tests:** Evaluating the performance of the TOE on minority groups to determine if small sample sizes in initial datasets caused performance bottlenecks^{22 23}.

3.4 EXPLAINABILITY TESTING

Explainability testing verifies that the AI system can generate interpretable, verifiable, and transparent justifications for its decisions so human supervisors can audit and safely trust the output²⁴. Tests in this category involve edge-case queries, typical real-world transactions, and standard verification samples paired with downstream diagnostic tools. They focus on:

- **Feature Attribution Validation:** Testing if feature importance tools (e.g., SHAP, LIME, or saliency maps) reliably highlight the correct decision drivers (such as checking if a FR model actually captures facial information rather than correlations with background lighting).
- **Consistency and Fidelity:** Ensuring the generated explanation closely tracks the internal logic of the model and does not change when identical query conditions are rerun.
- **Human Comprehensibility:** Running interface testing to confirm that system logs and alerts are written in plain language that human operators can quickly parse during an intervention^{25 26}.

¹⁹ Jones, E., Hardalupas, M., Agnew, W. Under the Radar? Examining the evaluation of Foundation Models. 2024, May 30. <https://www.adalovelaceinstitute.org/wp-content/uploads/2024/09/Ada-Lovelace-Institute-Under-the-radar-230924.pdf?v=1727110909>

²⁰ Kuravi, L. (2025). Predictive Analysis, Explainable AI, and Fairness Auditing: How Digital Transformation Reshapes Bank Models Using the Bank Marketing Dataset. Preprints. <https://doi.org/10.20944/preprints202512.0447.v1>

²¹ Singhal A, Neveditsin N, Tanveer H, Mago V. Toward Fairness, Accountability, Transparency, and Ethics in AI for Social Media and Health Care: Scoping Review. *JMIR Med Inform*. 2024;12:e50048. Published 2024 Apr 3. doi:10.2196/50048

²² Ainhize Barrainkua, Paula Gordaliza, Jose A. Lozano, and Novi Quadrianto. 2025. Preserving the Fairness Guarantees of Classifiers in Changing Environments: A Survey. *ACM Comput. Surv.* 2025 June; 57(6) Article 137, doi: <https://doi.org/10.1145/3637438>

²³ OECD. Catalog of Tools & Metrics for Trustworthy AI. <https://oecd.ai/en/catalogue/metrics>

²⁴ What is explainability? <https://www.moveworks.com/us/en/resources/ai-terms-glossary/explainability>

²⁵ Lago MA, Zamzmi G, Eich B, Delfino JG. Evaluating Explainability: A Framework for Systematic Assessment of Explainable AI Features in Medical Imaging. *Bioengineering (Basel)*. 2026;13(1):111. Published 2026 Jan 16. doi:10.3390/bioengineering13010111

²⁶ Kalele A., Subbiah R., Operationalizing AI: A Game of Trust and Ethics. TCS White paper. <https://www.tcs.com/what-we-do/pace-innovation/white-paper/building-trustworthy-ai-five-pillars>

For explainability testing, care should be taken to ensure that the explanation logs capture the necessary metadata required for safety audits and compliance requirements.

3.5 CYBERSECURITY TESTING

Cybersecurity testing evaluates whether the system can withstand malicious attempts to compromise its operation, manipulate its outputs, or undermine the reliability of its decisions. Consequently, cybersecurity testing plays a central role in the conformity assessment of high-risk AI systems by providing objective evidence that the implemented security measures are adequate for the intended purpose and deployment environment of the AI system.

The EU AI Act explicitly recognises cybersecurity as an essential characteristic of trustworthy AI. According to Article 15, high-risk AI systems shall be designed and developed in such a way that they achieve an appropriate level of accuracy, robustness and cybersecurity throughout their entire lifecycle. In particular, technology providers shall ensure that the AI system remains resilient against attempts by unauthorised parties to alter its use, behaviour or performance by exploiting system vulnerabilities. These requirements extend beyond conventional information security and encompass attacks that specifically target machine learning algorithms, training data, inference mechanisms and model deployment environments.

Unlike traditional software systems, AI systems introduce additional attack vectors resulting from the data-driven nature of machine learning models. The behaviour of an AI model is largely determined by the quality of the training data, the optimisation process and the statistical relationships learned during model development. Consequently, attackers may attempt to manipulate any stage of the AI lifecycle, including data collection, model training, model deployment and operational inference. Successful attacks may not only compromise the confidentiality, integrity or availability of the system but may also modify the AI system's behaviour without altering its software implementation. Such attacks may result in incorrect predictions, unsafe decisions, discriminatory outcomes or complete loss of confidence in the AI system.

For this reason, cybersecurity testing within the context of AI conformity assessment cannot be limited to conventional penetration testing or vulnerability scanning. Instead, the assessment shall consider both the supporting ICT infrastructure and the AI-specific components that collectively determine the behaviour of the system. These components include, but are not limited to, the training and validation datasets, feature engineering pipeline, trained models, inference services, software dependencies, communication interfaces, deployment infrastructure, monitoring mechanisms and update procedures. The testing lab should therefore evaluate cybersecurity from a system-wide perspective, recognising that vulnerabilities affecting any of these components may ultimately influence the reliability and safety of the AI system.

Cybersecurity testing should be performed as an integral part of the overall conformity assessment process. Rather than being considered an isolated evaluation activity, it complements the results obtained through functional testing, adversarial testing, fairness assessment and explainability evaluation. Functional testing demonstrates that the AI system fulfils its intended purpose under normal operating conditions, whereas cybersecurity testing determines whether this functionality can be preserved under malicious conditions. Similarly, adversarial testing focuses primarily on evaluating the robustness of machine learning models against carefully crafted perturbations, while cybersecurity testing encompasses a broader range of threats including attacks against the supporting software infrastructure, communication interfaces, access control mechanisms and deployment environment. Together, these complementary assessments provide a comprehensive evaluation of the AI system's trustworthiness.

The cybersecurity assessment should begin with the definition of an appropriate threat model describing the operational context of the AI system. This threat model should identify the critical assets requiring protection, the potential adversaries, their capabilities and objectives, the possible attack surfaces and the foreseeable misuse scenarios relevant to the intended purpose of the AI system. Establishing such a threat model enables the testing laboratory to select appropriate cybersecurity tests that are proportionate to the risks associated with the specific AI application rather than applying a generic security checklist. For example, an AI-based biometric identification system deployed in a public environment presents significantly different threats from a medical diagnostic system operating within a secured hospital infrastructure. Consequently, cybersecurity testing should always be adapted to the operational context and risk profile of the AI system under evaluation.

The assets considered during cybersecurity testing typically include both conventional ICT assets and AI-specific assets. Conventional assets comprise computing infrastructure, operating systems, communication networks, databases, authentication mechanisms and application programming interfaces (APIs). AI-specific assets include training datasets, validation datasets, feature extraction pipelines, machine learning models, model parameters, prompts in the case of generative AI systems, inference engines, confidence scores and generated outputs. Protecting these assets is essential to maintaining the integrity and reliability of the AI system throughout its operational lifecycle.

Cybersecurity testing should address both conventional cybersecurity threats and attacks specifically targeting machine learning technologies. Conventional threats include among other unauthorized access, privilege escalation, software vulnerabilities, insecure communication channels, denial-of-service attacks, malware infection and software supply chain compromise. AI-specific threats could include adversarial example attacks intended to manipulate model predictions during inference, data poisoning attacks targeting the training process, model extraction attacks attempting to reconstruct proprietary models through repeated queries, model inversion attacks seeking to recover sensitive information from trained models, membership inference attacks revealing whether specific data records were included during training, and prompt injection or indirect prompt manipulation attacks applicable to large language models and generative AI systems. Depending on the intended purpose of the AI system, additional domain-specific threats may also require consideration.

The objective of cybersecurity testing is not to demonstrate the complete absence of vulnerabilities, as no complex software system can be considered entirely secure. Instead, the objective is to obtain sufficient objective evidence that the implemented security controls reduce cybersecurity risks to an acceptable level consistent with the intended purpose of the AI system and the applicable requirements of the AI Act. The testing laboratory should therefore evaluate whether appropriate preventive, detective and corrective security measures have been implemented by the provider and whether these measures remain effective under realistic attack conditions.

The cybersecurity assessment should follow a structured methodology that combines documentation review with technical testing activities. Initially, the testing laboratory should examine the technical documentation provided by the AI system provider, including the system architecture, software bill of materials, security architecture, deployment configuration, access control mechanisms, update procedures, logging capabilities and incident-response processes. This documentation review enables the laboratory to understand the security assumptions made during system development and to identify components requiring detailed technical evaluation.

Following the documentation review, practical cybersecurity testing should be performed. These activities may include vulnerability assessment, secure configuration analysis, penetration testing, dependency analysis, API security testing, cryptographic verification, authentication and authorisation testing, resilience evaluation and AI-specific robustness assessments. Whenever appropriate, cybersecurity testing should also verify the effectiveness of monitoring mechanisms, audit logging, integrity verification procedures and secure update mechanisms implemented by the provider. The combination of these activities provides a comprehensive understanding of the AI system's ability to maintain secure operation throughout its lifecycle.

The outcome of cybersecurity testing should contribute directly to the overall conformity assessment. The testing laboratory should document the scope of the assessment, the threat model considered, the cybersecurity tests performed, the identified vulnerabilities, the associated risks, the effectiveness of the implemented mitigation measures and the residual risks remaining after mitigation. This information forms part of the technical evidence supporting the conformity assessment and enables the relevant NB to determine whether the AI system satisfies the cybersecurity requirements established by the AI Act.

4 TEST SCOPE AND OBJECTIVES

The section covers the process of defining the scope and the testing objectives for the five categories of tests. Where necessary, examples, addressing the compliance assessment of a hypothetical FR system, are given to illustrate certain concepts. Here we use the term 'Target of Evaluation' (TOE), to refer to the specific version of the AI product being evaluated.

4.1 FUNCTIONAL TESTING

The aim of functional testing is to assess whether the TOE exhibits the expected behaviour — the predefined, acceptable output or performance of the AI system for a given set of inputs, as determined by its business requirements, safety guards, and design specifications.

4.1.1 Scope

The documented scope of functional testing proves to an auditor: *"We have tested this AI system strictly within these exact test data distributions, under these specific environmental conditions, using these precise mathematical metrics. In the real world the product may behave differently outside of these boundaries."*

For an AI product, defining the scope implies establishing the mathematical and operational boundaries within which the probabilistic behaviours of the TOE are validated as safe and accurate. The scope defines the exact sandbox where your lab²⁷ guarantees the test results are valid. Operational boundaries (the input space) refer to the precise, measurable limits of the physical world and data properties under which the performance of the TOE is guaranteed. If an input falls inside these parameters, the system must process it accurately. If the input lies outside these parameters, it is explicitly *out of scope*, and the TOE is expected to reject the input.

There are four dimensions to defining the scope for functional testing.

1. **The System Boundaries:** Define exactly where the responsibility of the lab begins and ends. Document whether you are testing the raw model weights, a containerised API, or the full end-to-end software/hardware stack. This is also referred to as the **Architectural Scope**.
2. **The Operational Design Domain:** Specify the precise conditions under which the AI is claimed to work. This includes acceptable input ranges, environmental factors (lighting, noise), and user demographics. Anything outside this domain is explicitly marked as "out of scope." This is also called the **Operational Scope**.
3. **The Data Space and Coverage Boundaries:** Define the characteristics of the test datasets. Under ISO 17025, you must prove that the test data is completely isolated from the developer's training data, free from leakages, and statistically representative of real-world distributions. The term **Data Scope** refers to this framework.
4. **The Test Oracle and Boundaries of Success:** Define the exact performance, safety, and fairness metrics that will be measured. Avoid open-ended exploration; explicitly list the statistical confidence intervals required for validation. This is also referred to as the **Metric Scope**.

4.1.1.1 System Boundaries for Functional Testing

The scope definition should isolate the AI functions under evaluation from external software ecosystem layers.

²⁷ In this document, the expressions 'the lab' and 'the testing lab' refer to both, an AL or a testing lab internal to the AI technology provider.

- **Identify Core Tasks:** Delineate whether you are testing raw model inference (e.g., probabilities or scores), deterministic application wrappers (e.g., hardcoded decision thresholds), or edge-hardware interactions.
- **Declare Exclusions:** Explicitly list infrastructure dependencies (e.g., cloud network latency, external database uptime) as out-of-scope to isolate the AI's actual functional performance²⁸.

4.1.1.2 Operational Design Domain (ODD)

AI functional testing should not be open-ended; it should be constrained to the rated environment of the product.

- **Physical and Environmental Limits:** Define the exact physical parameters required for the function to execute reliably (e.g., illumination, temperature, presentation angles).
- **User/Input Parameters:** Clarify the exact profile of compliant inputs the system is certified to handle. State what variations (e.g., corrupted data packets, extreme anomalies) drop the system into error-handling mode rather than normal processing.

4.1.1.3 Data Space and Coverage Boundaries

Unlike classical software where code coverage (e.g., branch coverage) is sufficient, AI requires defining input and output data space coverage.

- **Test Data Segregation:** Define the scope of test datasets to guarantee they are statistically independent of the developer's training and validation datasets (preventing data leakage).
- **Distribution Bounds:** Specify the required diversity, stratification, and edge-case distributions within your evaluation test suites²⁹.

4.1.1.4 Test Oracle and Metric Boundaries

As AI systems are inherently probabilistic, the scope must establish how a "correct" function is mathematically evaluated when an exact string-to-string or Boolean match does not exist³⁰.

- **Statistical Oracles:** Define the relationships or golden datasets that serve as the ground truth.
- **Statistical Tolerance:** Bound the acceptable statistical margins error-rates instead of absolute Boolean outputs^{31,30}.

4.1.1.5 Example

To illustrate the various aspects of scope-specification discussed above, here is a sample Functional Scope Specification for a FR system deployed in border-control application.

Functional System Boundaries

- In scope:
 - The face detection and alignment function from the live camera stream.
 - The face-biometric feature extraction algorithm.
- The 1:1 and 1:N template comparison algorithm.

²⁸ https://www.nist.gov/system/files/documents/2020/12/15/340_1_mansfield_panel_ibpc.pdf

²⁹ https://www.data-spaces-symposium.eu/wp-content/uploads/2024/03/AI-Testing-and-Data-Spaces-consolidated_final.pdf

³⁰ ISO/IEC TR 29119-11:2020: Software and systems engineering — Software testing — Part 11: Guidelines on the testing of AI-based systems <https://www.iso.org/obp/ui/en/#iso:std:iso-iec:tr:29119:-11:ed-1:v1:en>

³¹ ISO/IEC 19795-1 Information technology – Biometric Performance testing and reporting. Part 1 – Principles and framework. <https://www.iso.org/standard/73515.html>

- The processing stage that extracts the ICAO-compliant photo from the e-Passport chip^{32 33}.
- Out of scope:
 - The PKI certificate validation of the digital signature in the passport.
 - The downstream mechanical gate release trigger and turnstile motor functions.

Operational Design Domain Boundaries

- **Physical and Environmental Constraints:**
 - Controlled indoor lighting between 300 lux and 600 lux at the camera focal plane.
 - Camera sensor mounted at a fixed height range of 1.2m to 2.1m.
 - Subject positioning constrained to a distance of 0.8m to 1.4m from the camera lens.
- **Subject Pose and Presentation Constraints:**
 - Subject head rotation restricted to a maximum yaw, pitch, and roll of +/- 15°.
 - Neutral facial expression (mouth closed, no smile)^{31 34}.
- **Exclusions:** Outdoor border checkpoints, processing of infants (under four years old), or subjects with severe facial bandages or other trauma. (These would, by default, go directly to a border officer lane for manual processing).

Data Space and Coverage Boundaries

The testing data scope specifies the exact characteristics of the evaluation and testing framework used by the lab:

- **Test-Data Independence:** A curated, zero-effort dataset of image pairs (live capture simulation vs. token image) completely isolated from the vendor³². (Equivalently, for in-house conformity assessment by the vendor's own quality-assurance (QA) group, the QA lab should take care to maintain strict segregation between test data and training data.)
- **Demographic Stratification:** The dataset must be uniformly distributed across two genders, three distinct age cohorts, and across all six categories of the Fitzpatrick Skin Type scale³¹.
- **Presentation Attacks:** A separate test suite of PAIs categorised as Levels 1 and 2 in biometrics literature³⁵ (e.g., 2D paper printouts and electronic display replays).

Metric Boundaries and Test Oracles

The success of the TOE in functional testing is quantified using performance metrics calculated at pre-specified confidence levels, according to ISO/IEC 19795-1³¹. Some examples illustrating how to specify such boundaries are show in Table 4.

³² FIDO Biometrics Requirements <https://fidoalliance.org/specs/biometric/requirements/Biometrics-Requirements-v1.1-fd-20190606.pdf>

³³ Facial recognition in security systems: How they work and what they could contribute to the security posture <https://www.ericsson.com/en/blog/2021/5/facial-recognition-security-system>

³⁴ Storer M, Urschler M, Bischof H, Birchbauer JA. On combining classifiers for assessing portrait image compliance with ICAO/ISO standards. *Proc BIOSIG*. 2008. www.tugraz.at/fileadmin/user_upload/Institute/ICG/Images/team_bischof/mib/paper_pdfs/2008StorerBIOSIG.pdf

³⁵ ISO/IEC 30107-3: Information Technology — Biometric Presentation Attack Detection. Part 3: Testing and reporting. <https://www.iso.org/standard/79520.html>

Table 4 Examples of specific boundaries (performance limits) on relevant metrics should be prescribed in the testing protocol. The table shows of examples relevant for testing a face-recognition system to be used in a border-control application.

	Functional Attribute	Metric Definition	Target Boundary
1	Security	False Match Rate (FMR)	<= 0.001% (1 in 100,000 unauthorized entries allowed)
2	Operational Efficiency	False Non-Match Rate (FNMR)	<= 0.5% (1 in 200 valid travellers falsely rejected)
3	System Usability	Failure to Enrol (FTE) / Failure to Acquire (FTA)	<= 0.1% of valid presentations fail to generate a template
4	Demographic Fairness	Bias Disparity Ratio : (FNMR _{Max} / FNMR _{Min})	<= 1.20 variance across all demographic cohorts

4.1.2 Objectives

To verify conformity, every functional testing objective must be aligned with specific articles of the AI Act³⁶. Below, we expand on the five relevant articles, listed in Table 5.

Table 5: Mapping various articles of the EU AI Act to the various objectives of functional testing.

	Article in EU AI Act	Topic
1	Article 9(2) and 9(4): Risk Management	Test risk-mitigation measures (does the TOE include features designed to prevent harm?)
2	Article 10(3): Data Ingestion and Parsing	Test that TOE can handle the intended (normal) inputs

³⁶ Yannick Caballero Cuevas, The principles for interpreting the requirements for high-risk AI systems. 2025, April 7, by the Centre de droit bancaire et financier <https://cdbf.ch/en/1406/>

3	Article 13: Transparency	Verify "Instructions of Use"
4	Article 14: Human Oversight	Test Override and Kill-Switch
5	Article 15(1) and 15(2): Accuracy and Performance Verification	Specify metrics to log accuracy and performance

From the articles identified in Table 5, the following testing objectives may be derived.

- **Article 9(2) and 9(4) —Risk Management:** The test-objective is to demonstrate that the guard-rails built into the TOE function correctly.
- **Article 10(3) — Data Ingestion and Parsing:** The test-objective is to test that the TOE can handle and parse intended inputs correctly.
- **Article 13 — Transparency:** Testing must confirm that the system operates strictly within the performance parameters documented in its official "Instructions for Use".
- **Article 14 — Human Oversight:** Functional tests must prove that the user-interface of the TOE allow human operators to understand outputs, spot anomalies, and override or shutdown the system.
- **Article 15 — Accuracy:** the objective is the demonstrate that the TOE achieves the claimed accuracy, based on the relevant metrics.

4.1.2.1 Core Objectives in Functional Testing

Functional testing ensures an AI system performs its intended tasks according to specified requirements. For high-risk AI products, objectives should cover the standard software metrics taking into account statistical variance, data drift, and real-world harms. In general, functional testing should address four objectives:

- **Define Deterministic Pass/Fail Thresholds:** AI outputs are probabilistic. Convert these into binary outcomes by setting strict statistical acceptance criteria (e.g., "FMR $\leq$ 0.01% for a FNMR $\leq$ 1%").
- **Map Objectives to Harm Scenarios:** Align every functional test objective with a specific risk or compliance failure mode.
- **Test Operational Boundaries:** Explicitly define objectives to find where the model's performance degrades. Use out-of-distribution (OOD) data.
- **Quantify Regression Limits:** Establish strict limits for performance degradation when updating models or retraining pipelines.

4.1.2.2 Example

Once again, we use example of a face recognition system deployed in a border control scenario, to illustrate the definition of functional testing objectives. Automated border control gates (e-Gates) fall under the category of high-risk AI. Functional testing must prove the system accurately verifies identity while maintaining airport throughput and security.

1. Core Competency Objectives

Verify that the AI accurately matches a traveller's live face to their e-Passport chip image.

- **Objective:** Achieve a False Match Rate (FMR) of less than 0.001% to prevent unauthorized entry.
- **Objective:** Maintain a False Non-Match Rate (FNMR) below 0.5% to avoid excessive manual identity-verification for travellers, which not only cause additional inconvenience to travellers, but also increase operational costs.
- **Objective:** Ensure the system handles 1:1 verification within a maximum latency of 2.0 seconds^{37 38}.

2. Operational Boundary and Environmental Objectives

Verify that the system functions reliably under varying physical border conditions.

- **Objective:** Maintain target FNMR thresholds across specified lux levels (e.g., low light at 150 lux to bright glare at 2000 lux).
- **Objective:** Verify accurate matching across diverse traveller heights (e.g., 1.2m to 2.0m) using automated camera adjustment.
- **Objective:** Evaluate matching degradation when travellers wear minor obstructions like medical masks, reading glasses, or headscarves.

3. Data Invariance and Robustness Objectives

Verify that the system is resilient to real-world image quality issues.

- **Objective:** Pass verification when passport photos suffer from common print degradation, compression artifacts, or slight geometric rotations.
- **Objective:** Successfully reject low-resolution, blurred, or motion-corrupted live camera frames before attempting a comparison.

4.2 Adversarial Testing

Adversarial testing assesses the robustness of the TOE by actively attempting to "break," confuse, or trick the model. When setting boundaries for high-risk AI adversarial testing, use a structured methodology to ensure compliance with laboratory standards. In this section we will discuss the aspects to keep in mind when elaborating the scope and objectives for adversarial testing of an AI product.

4.2.1 Scope

The scope defines *what* is being tested, the environmental constraints, and the operational boundaries. The scope for adversarial testing should clearly *establish the threat model*. For example, should the evaluation be based on white-box testing (full access to AI model), gray-box testing (knowledge about model architecture), or black-box testing (access only to inputs and outputs).

- **Specify Perturbation Limits:** One kind of adversarial testing involves partially modifying (perturbing) a 'normal' input to test whether the perturbation may cause the TOE to generate a wrong result. For example, if the TOE is designed to recognize objects in images, some pixels in an input image of an object, X, may be perturbed, to make the TOE to identify the input as a different object, Y. The scope of such testing should quantify the maximum allowed manipulation. In digital inputs (images, audio signals), such bounds on allowed perturbation are often specified in terms of distance-metrics (L_p -norm bounds for pixel shifts, semantic bounds for lighting or rotation) to ensure tests simulate realistic, plausible attack vectors.

³⁷ Sarier, N.D. Best of two worlds: Efficient, usable and auditable biometric ABC on the blockchain. Computer Standards & Interfaces. 2025 Vol.92. Doi: <https://doi.org/10.1016/j.csi.2024.103916>
<https://www.sciencedirect.com/science/article/pii/S0920548924000850>

³⁸ <https://ibeta.com/wp-content/uploads/2019/06/PalmSecure-FPro-Biometric-Subsystem-Certification-Report-v1.0.pdf>

- **Document Exclusions:** Clearly state out-of-scope testing vectors (e.g., physical facility breaches or classic network DDoS attacks) to prevent scope creep³⁹.

4.2.2 Objectives

Objectives justify *why* specific tests are included in the evaluation, and establish quantitative metrics of success. There are three main objectives in adversarial testing:

- **Anchor to Core Security Targets:** Align objectives directly with the security criteria outlined in high-risk regulations like the EU AI Act (e.g., robustness, cybersecurity, accuracy)^{40 41 42}.
- **Determine Performance Degradation Margins:** Set acceptable performance floors. For example, specify that accuracy must not fall below a certain percentage under a defined adversarial noise budget.
- **Measure Attack Success and Robustness Rates:** Quantify your goals using standardized lab metrics, such as the maximum percentage of successful bypasses an attacker can achieve.

4.3 FAIRNESS TESTING

The definition of testing scope and testing objectives requires translating the abstract concept of fairness into measurable targets. The scope definition specifies the concrete boundaries of fairness assessment, and the testing objectives specify how the compliance of the TOE to the AI Act, in terms of uniform behaviour for different demographic sections of human society, is assessed.

4.3.1 Scope

As before, the scope of fairness testing should clearly set out what is being evaluated, who it impacts, and where the boundaries lie. Unlike functional testing and adversarial testing, fairness testing may rely not only on analysis of test results, but also audits of model-training process. The audit should cover all aspects of training and data-processing in the AI model. Fairness testing should explicitly cover the following three dimensions:

1. **Data and Processing Pipeline:** The scope for data and pipeline elements must define the boundaries of what enters, processes within, and exits the AI system. It should explicitly cover these four lifecycle areas.

Test data should be representative of the demographic diversity of the target market. Auditors of the relevant NB or MSA should also verify that the training data includes adequate demographic diversity. Fairness testing should take into account whether the TOE uses, or is derived from, specific open-source AI models, including foundation models or APIs of commercial products. Data sampling strategies, including timeframes, historical cut-off dates for data, and frequency should also be covered.

The scope of fairness testing should cover the various downstream API responses, scores generated by the AI system, as well as final automated decisions. This aspect of the scope covers preprocessing stages (data cleaning rules, missing value imputation, and data normalization methods that may skew

³⁹ <https://www.abileneacademy.ch/en/answers/aims-scope-context-analysis>

⁴⁰ <https://verityai.co/blog/eu-ai-act-conformity-assessment-banks>

⁴¹ <https://www.tuv.com/world/en/artificial-intelligence-for-products.html>

⁴² <https://cognisys.co.uk/eu-ai-act-high-risk-systems-how-to-classify-your-ai/>

data statistics), feature engineering steps, data labelling (annotation) guidelines, and data partitioning policies.

Audits for fairness testing should also examine the architecture of the underlying AI models, including policies for setting hyperparameters, as well as the embedding spaces of the models, decision thresholds that may require per-group tuning to enhance fairness, as well as feature importance ranking to see whether the model gives excessive importance to sensitive attributes.

Raw score distributions, confidence intervals, and categorical labels output by the TOE should be examined in the scope of fairness testing. The use of logged operational data recycled for retraining the model should also be inspected, as such feedback may amplify inherent biases.

2. **Operational and Environmental Context:** Fairness audits should study the influence on deployment environments, including cameras, microphones and other sensors, sensor calibration policies, ambient lighting and other relevant environmental conditions (e.g., background noise, network latencies) that may affect inputs to the AI system.

Regional demographics, such as local dialects and accents, and other geographical variations among deployment sites should be studied in the scope of fairness testing.

Study of human-in-the-loop aspects, such as user interfaces, operator dashboards, and how human overrides affect the final decision pipeline should also form part of fairness testing.

3. **Demographic and Population Boundaries:** Fairness testing should be adapted to the relevant *protected attributes* — explicit traits such as age, gender identity, race, ethnicity, religion, and disability status. Other areas of focus are *intersecting identities* (combinations of traits that compound bias, such as older age combined with darker skin tones), *vulnerable subgroups* (non-native language speakers, children, elderly populations, or individuals with atypical physical traits), and *user personas* (system operators, primary end-users, and bystanders incidentally captured by the AI sensor).

4.3.2 Objectives

To outline the objectives for fairness testing in the context of the EU AI Act, we start by mapping the relevant legal mandates (pertaining to Article 10 of the EU AI Act) to quantifiable metrics, as shown in Table 6. **Erreur ! Source du renvoi introuvable.** This mapping can lead to specific objectives, such as “verify that error rate disparities do not cross legally actionable thresholds.” For example, specify that the ratio of True Positive Rates between protected groups must remain between 0.80 and 1.25.”

Table 6: Mapping EU legal concepts to measurable objectives for fairness testing.

	Legal Concept	Technical Metric
1	Direct discrimination (discrimination against specific groups).	Demographic parity (statistically similar presence in training data)
2	Indirect discrimination (neutral rules, biased impact)	Equalized odds (statistically similar error rates per group)

The objectives must require separate metric tracking for: *core demographics* (gender, racial or ethnic origin, religion or belief, disability, age, and sexual orientation), and *intersecting subgroups*. The objectives must explicitly mandate testing for compound vulnerabilities (e.g., assessing error rates specifically for elderly females of colour, rather than evaluating age and gender in isolation).

Articles 10(3) and 10(4) of the AI Act dictate that datasets must possess appropriate statistical properties and reflect the specific geographical, contextual, and behavioural settings of deployment. **Erreur ! Signet non défini.** These mandates lead to the following objectives:

- Define a minimum threshold for demographic representation in the validation/testing datasets. For instance, specify that the test dataset must match the regional census data of the target EU member state deployment site within a +/- 2% margin.
- Define baseline performance targets across distinct contextual conditions (e.g., ensuring a biometrics model achieves fairness thresholds under even low-light or outdoor European weather conditions).

The EU AI Act provides the legal basis (under Article 10(5)) allowing system providers to process special categories of personal data (e.g., race, health data) *strictly* for bias detection and correction. **Erreur ! Signet non défini.**⁴³. This permission can be translated into the following two objectives:

- Mandate the use of secure, pseudonymised ground-truth demographic data during internal lab testing to verify model fairness, rather than relying on proxy variables. **Erreur ! Signet non défini.**
- Build in verification metrics to prove that this sensitive data is strictly isolated and deleted once the de-biasing phase is completed.

Following Article 14 (Human Oversight), high-risk AI systems must be designed so human operators can counter (override) automation bias. This requirement leads to the objective that tests should verify that the user-interface does not cause human operators to disproportionately accept incorrect AI decisions for specific demographic groups.

4.4 EXPLAINABILITY TESTING

In this section we discuss the development of the scope and objectives for explainability testing. Here the scope of the test suite helps to clarify how the term ‘explainability’ is interpreted for the TOE, whereas the objectives help to define specific criteria for evaluating the explainability offered by the TOE.

4.4.1 Scope

The scope for explainability testing defines *which layers* of the AI system are audited, *who* receives the explanations, and *which components* are responsible for generating them. It must be bounded across the following four dimensions^{44 45 46}.

1. **Target Audience Personas:** These include end-users and affected persons (people impacted by the decisions of an AI system) who require plain-language justifications^{47 48} (e.g., rejected loan applicants); system operators / domain experts who may need granular, feature-level insights to validate or override

⁴³ AI Bias vs. Data-Privacy: Can the EU's laws find balance? <https://data-privacy-office.eu/ai-bias-vs-data-privacy-can-the-eus-laws-find-balance/>

⁴⁴ Kirat Th., Tambou O., Do V., Tsoukiàs A. Fairness and explainability in automatic decision-making systems. A challenge for computer science and law. EURO Journal on Decision Processes. 2023 Volume 11. Doi: <https://doi.org/10.1016/j.ejdp.2023.100036>. <https://www.sciencedirect.com/science/article/pii/S2193943823000092>

⁴⁵ Buhrmester V., Muench D., Bulatov D. and Arens M. Evaluating Explainability in Safety-Critical ATR Systems: Limitations of Post-Hoc Methods and Paths Toward Robust XAI. 2026 <https://arxiv.org/abs/2605.05748>

⁴⁶ Zhang Y. et al. Decision Information Meets Large Language Models: The Future of Explainable Operations Research 2025. <https://arxiv.org/abs/2502.09994>

⁴⁷ Ramachandran D. et al. Transparent AI: The Case for Interpretability and Explainability. 2025. <https://arxiv.org/html/2507.23535v1>

⁴⁸ <https://verityai.co/blog/financial-ai-explainability-requirements-regulations>

system outputs⁴⁹; compliance officers, auditors, and members of other regulatory, legal, and internal risk teams who require comprehensive, reproducible documentation of the logic in the model; and even AI developers who need deep architectural diagnostics to troubleshoot, debug, and optimize model behaviour.

- 2. Methodological and Algorithmic Approaches:** The scope should cover testing of intrinsic explainability — for inherently interpretable models (e.g., decision trees, linear regression) — to *post-hoc* explanations using surrogate models (e.g., LIME, SHAP) applied to black-box systems^{50 51 52}. Explainability testing should also include tests for global explanations (for the final result of an AI system) as well as local explanations, corresponding to the outputs of distinct subsystems (e.g., the recognition subsystem and the PAD subsystem of an FR product).
- 3. Pipeline and Component Boundaries:** Input feature attribution, that maps which input feature contributed the most to the final result, should be included in explainability testing. Similarly, testing in this category should be able to trace a specific result back to relevant segments of the training data, to historical baseline datasets, or to the current input.
- 4. Out-of-Scope Elements:** The scope definition should clearly set out what is not covered by explainability testing. Specifically, testing for clarity of explanation should not result in exposure of proprietary code, or other commercially sensitive intellectual property. Testing should also focus on *post-hoc* explanation of data, instead of examining the methods used to collect data in its raw form.

4.4.2 Objectives

With explainability testing, the primary objective is to demonstrate that the TOE is sufficiently transparent, such that deployers may interpret its outputs, understand its limits, and exercise meaningful human oversight. Determining the objectives of explainability testing starts with a mapping of legal mandates for explainability in the EU AI Act to quantifiable explainability metrics. The relevant mappings⁵³ are given in Table 8. As shown in the table, four articles in the EU AI Act are relevant to explainability testing. In this table we discuss the four articles in order of significance, instead of number order of the articles.

Table 7: Mapping explainability mandates in the EU AI Act to quantifiable metrics for explainability metrics.

	Article	Legal Mandate	Explainability Metrics
--	---------	---------------	------------------------

⁴⁹ Lago MA, Zamzmi G, Eich B, Delfino JG. Evaluating Explainability: A Framework for Systematic Assessment of Explainable AI Features in Medical Imaging. *Bioengineering (Basel)*. 2026;13(1):111. Published 2026 Jan 16. doi:10.3390/bioengineering13010111

⁵⁰ <https://practical-ai-act.eu/latest/engineering-practice/explainability/>

⁵¹ Zhang K., Wang D., Lin F., Xie J., Zhou W. A comprehensive review of explainable artificial intelligence in healthcare methods, evaluation, and clinical integration. *iScience* 2026 Vol.29(3). DOI: <https://doi.org/10.1016/j.isci.2026.115026>

⁵² Khan, F.S., Mazhar, S.S., Mazhar, K. et al. Model-agnostic explainable artificial intelligence methods in finance: a systematic review, recent developments, limitations, challenges and future directions. 2025. *Artificial Intelligence Review*. 58, 232. <https://doi.org/10.1007/s10462-025-11215-9>

⁵³ Damen V, Wiersma M, Aydin G, van Haasteren R. Explainable AI for EU AI Act compliance audits. *Maandblad voor Accountancy en Bedrijfseconomie*. 2025. 99(4): 231-242. <https://doi.org/10.5117/mab.99.150303>
<https://mab-online.nl/article/150303/list/9/>

1	Article 13: Transparency and Interpretability	High-risk AI systems must be transparent enough to enable deployers to interpret outputs and use the system appropriately.	- Feature Importance / Attribution Scores: Measure quantitative impact of individual input features on result (e.g., SHAP values or LIME). - Explanation Fidelity: Quantify how well the explanation approximates the underlying black-box model (e.g., R^2 scores for surrogate models). - Completeness and Conciseness Metrics: Measure percentage of the model-variance explained by the provided rationale versus the brevity of the explanation.
2	Article 14: Human Oversight	Systems must be designed so natural persons can oversee their functioning, override decisions, or disregard the output.	- Time to Intervention: Measure latency between the AI output, the explanation delivery, and the human oversight action (depending on intended purpose of TOE). - Explanation Understandability and Accuracy: Use testing panels to measure user comprehension and ensure the provided confidence / probability scores are mathematically faithful to actual uncertainty of model. - False Discovery/Negative Rates in Human Overrides: Test whether explanations cause "automation bias" (humans blindly trusting the AI) or "alert fatigue" (humans needlessly overriding accurate AI).
3	Article 9: Risk Management	Providers must ensure pre-market testing to mitigate risks to health, safety, and fundamental rights.	- Concept Drift or Covariate Shift: Verify that explanations remain stable even if the input data distribution shifts over time. - Robustness of Explanations: Test if the explanation is susceptible to adversarial perturbations. Measure the variance in feature attribution scores when minor noise is added to the inputs.
4	Article 86: Right (of Affected Persons) to Explanation	Affected persons have the right to obtain a clear and meaningful explanation of the AI's role in a decision.	Plain-Language Translation Accuracy: Assess if complex mathematical explanations (e.g., tensors, gradient maps) can be consistently and accurately translated into text templates understandable to lay persons.

4.5 CYBERSECURITY TESTING

Cybersecurity testing is performed to determine whether an AI system maintains an acceptable level of security throughout its operational lifecycle and continues to satisfy the cybersecurity requirements established by the EU AI Act when exposed to realistic attack scenarios. The purpose of this activity is not to certify that the AI system is immune to cyberattacks, but rather to provide objective evidence that the provider has implemented appropriate technical and organisational measures to reduce cybersecurity risks to an acceptable level and that foreseeable attacks do not compromise the intended operation of the AI system.

Unlike traditional software security assessments, cybersecurity testing within AI conformity assessment requires evaluating both conventional ICT security and AI-specific security properties. While many cyberattacks target the surrounding infrastructure, AI systems introduce additional attack surfaces associated with machine learning models, training datasets, inference mechanisms and automated decision-making processes. Consequently, the cybersecurity assessment should consider the AI system as an integrated socio-technical system rather than focusing exclusively on the machine learning model or the supporting software infrastructure.

The methodology described in this section provides guidance for conformity assessment bodies and independent testing laboratories on how to define the scope of cybersecurity testing, identify relevant threat scenarios, perform technical security evaluations and document objective evidence supporting the conformity assessment. The methodology is intentionally risk-based and technology-neutral, allowing it to be applied to different categories of AI systems while remaining aligned with the requirements of Article 15 of the AI Act.

Cybersecurity testing should be planned at the beginning of the conformity assessment process, once sufficient technical documentation has been made available by the provider. The results obtained from cybersecurity testing should be analysed together with those from functional, adversarial, fairness and explainability testing to produce a comprehensive evaluation of the AI system. In many situations, weaknesses identified during cybersecurity testing may explain failures observed during other testing activities. For example, an unexpected degradation of model performance may result from unauthorised modification of model parameters, while abnormal prediction behaviour may be caused by successful adversarial manipulation of the inference process. Integrating the different testing activities therefore enables a more accurate interpretation of the observed results and supports a more reliable conformity assessment.

4.5.1 Scope

Cybersecurity testing is performed to determine whether an AI system maintains an acceptable level of security throughout its operational lifecycle and continues to satisfy the cybersecurity requirements established by the EU AI Act when exposed to realistic attack scenarios. The purpose of this activity is not to certify that the AI system is immune to cyberattacks, but rather to provide objective evidence that the provider has implemented appropriate technical and organisational measures to reduce cybersecurity risks to an acceptable level and that foreseeable attacks do not compromise the intended operation of the AI system.

Unlike traditional software security assessments, cybersecurity testing within AI conformity assessment requires evaluating both conventional ICT security and AI-specific security properties. While many cyberattacks target the surrounding infrastructure, AI systems introduce additional attack surfaces associated with machine learning models, training datasets, inference mechanisms and automated decision-making processes. Consequently, the cybersecurity assessment should consider the AI system as an integrated socio-technical system rather than focusing exclusively on the machine learning model or the supporting software infrastructure.

The methodology described in this section provides guidance for conformity assessment bodies and independent testing laboratories on how to define the scope of cybersecurity testing, identify relevant threat scenarios, perform technical security evaluations and document objective evidence supporting the conformity

assessment. The methodology is intentionally risk-based and technology-neutral, allowing it to be applied to different categories of AI systems while remaining aligned with the requirements of Article 15 of the AI Act.

Cybersecurity testing should be planned at the beginning of the conformity assessment process, once sufficient technical documentation has been made available by the provider. The results obtained from cybersecurity testing should be analysed together with those from functional, adversarial, fairness and explainability testing to produce a comprehensive evaluation of the AI system. In many situations, weaknesses identified during cybersecurity testing may explain failures observed during other testing activities. For example, an unexpected degradation of model performance may result from unauthorised modification of model parameters, while abnormal prediction behaviour may be caused by successful adversarial manipulation of the inference process. Integrating the different testing activities therefore enables a more accurate interpretation of the observed results and supports a more reliable conformity assessment.

4.5.2 Objectives

The primary objective of cybersecurity testing is to determine whether the AI system maintains an appropriate level of security throughout its operational lifecycle and whether its cybersecurity characteristics remain consistent with the essential requirements established by Article 15 of the EU AI Act. The assessment seeks to provide objective evidence that the implemented technical and organisational measures effectively protect the AI system against foreseeable cyber threats and that successful attacks do not result in unacceptable degradation of its intended functionality, safety or regulatory compliance.

Cybersecurity testing should therefore verify not only the existence of security controls but also their effectiveness under realistic operational conditions. The objective is to evaluate whether the provider has correctly identified the relevant cybersecurity risks, implemented appropriate mitigation measures and continuously maintains the security posture of the AI system throughout deployment and operation. Testing should also determine whether residual risks remain acceptable with respect to the intended purpose of the AI system and its expected operating environment.

One of the principal objectives is to verify the integrity of the AI system. Integrity refers to the ability of the system to prevent unauthorised modification of software components, machine learning models, datasets, configuration files and operational parameters. Since the behaviour of AI systems depends heavily on the integrity of training data and model parameters, even minor modifications may significantly affect prediction accuracy or alter the decision-making process. Cybersecurity testing should therefore determine whether mechanisms such as cryptographic integrity verification, secure model storage, authenticated software updates and configuration management adequately protect the AI system against tampering.

A second objective is to evaluate the availability and operational resilience of the AI system. High-risk AI systems frequently support safety-critical or business-critical processes in which interruptions of service may have significant operational or societal consequences. The cybersecurity assessment should therefore evaluate whether the AI system continues to operate within acceptable performance limits when subjected to resource exhaustion, communication failures, denial-of-service attacks or partial infrastructure failures. Where complete service continuity cannot be guaranteed, the testing laboratory should verify that appropriate fail-safe mechanisms and recovery procedures have been implemented to minimise operational impact.

Cybersecurity testing should also assess the confidentiality of sensitive assets associated with the AI system. These assets may include proprietary machine learning models, training datasets containing personal or commercially sensitive information, authentication credentials, cryptographic material and operational logs. The objective of the assessment is to verify that appropriate access control mechanisms, encryption techniques and identity management procedures prevent unauthorised disclosure of confidential information. Particular attention should be paid to attacks such as model inversion and membership inference, which may indirectly reveal sensitive information even when conventional access control mechanisms remain effective.

Another important objective is to evaluate the AI system's resilience against AI-specific attacks. Unlike conventional software, machine learning models may be manipulated through carefully crafted inputs that exploit statistical properties of the learning algorithm rather than software implementation flaws. Cybersecurity testing should therefore determine whether the AI system remains sufficiently robust against adversarial examples, poisoning attacks, model extraction, evasion attacks and other techniques specifically targeting

machine learning systems. The assessment should focus not only on the immediate success of such attacks but also on their impact on the reliability, safety and trustworthiness of the AI system.

The assessment should further verify the effectiveness of authentication, authorisation and access control mechanisms protecting both administrative functions and operational services. Unauthorised access to model repositories, deployment infrastructure or administrative interfaces may enable attackers to modify system behaviour, extract proprietary information or disable critical security controls. Cybersecurity testing should therefore evaluate identity management procedures, privilege separation, credential protection, multi-factor authentication where applicable and the enforcement of the principle of least privilege across the AI system.

Traceability and accountability constitute additional objectives of cybersecurity testing. The AI Act places significant emphasis on technical documentation and record-keeping throughout the lifecycle of high-risk AI systems. Consequently, the testing laboratory should verify that the AI system generates sufficient audit information to support incident investigation, forensic analysis and regulatory oversight. Logging mechanisms should accurately record relevant security events, administrative actions, model updates and operational anomalies while ensuring the integrity and confidentiality of the collected records. The assessment should also determine whether logged information is sufficient to reconstruct significant security incidents and support post-incident analysis.

Cybersecurity testing should additionally evaluate the security of software maintenance and update procedures. AI systems frequently evolve after deployment through software updates, model retraining or adaptation to changing operational conditions. Although these activities are necessary to maintain system performance, they may also introduce new vulnerabilities or unintentionally alter the behaviour of the AI system. The assessment should therefore verify that update mechanisms preserve software authenticity, maintain version traceability and prevent unauthorized deployment of modified models or software components. Where continuous learning or online adaptation is implemented, additional controls should be evaluated to ensure that learning processes cannot be manipulated by malicious actors.

The testing laboratory should also assess the effectiveness of monitoring and incident response capabilities implemented by the provider. Continuous monitoring enables early detection of abnormal system behaviour, attempted intrusions and operational anomalies that may indicate ongoing cyberattacks. Cybersecurity testing should therefore determine whether monitoring mechanisms generate meaningful alerts, support timely incident response and provide sufficient information for security operators to investigate suspicious activities. Where automated response mechanisms are implemented, their behaviour should also be evaluated to ensure that mitigation actions do not introduce additional operational risks.

Finally, cybersecurity testing should determine whether the overall security posture of the AI system remains appropriate for its intended purpose and operational context. Individual security controls should not be evaluated in isolation but rather as components of an integrated security architecture supporting trustworthy AI. The testing laboratory should therefore consider the combined effectiveness of preventive, detective and corrective measures when determining whether the AI system satisfies the cybersecurity requirements of the AI Act. The outcome of this evaluation should provide a well-justified basis for the subsequent conformity assessment and support the final conclusion regarding the cybersecurity compliance of the AI system.

The cybersecurity objectives established in this section directly influence the design of the testing strategy described in the following subsection. Once the security objectives have been defined, the testing laboratory can identify the relevant threat scenarios, establish the threat model and select the cybersecurity tests that provide the most appropriate evidence for demonstrating compliance with the AI Act.

The mapping of the legal mandates from the various articles of the AI Act to specific testing objectives is presented in Table 8.

Table 8: Mapping EU AI Act mandates to objectives of cybersecurity testing.

CI	Topic	The Legal Mandate	Test Objectives	Quantitative Metrics
----	-------	-------------------	-----------------	----------------------

1	Cybersecurity (Art. 15)	High-risk AI systems shall achieve an appropriate level of cybersecurity throughout their lifecycle.	Verify that the AI system maintains an appropriate cybersecurity posture during operation.	Attack Success Rate (ASR), Mean Time to Detect (MTTD), Mean Time to Recover (MTTR), ...
2	Robustness against malicious manipulation (Art. 15)	AI systems shall be resilient against attempts by unauthorised third parties to alter their use, outputs or performance by exploiting system vulnerabilities.	Evaluate the system's resilience against attacks intended to manipulate its behaviour or outputs.	Adversarial Attack Success Rate, Replay Attack Success Rate, Model Extraction Accuracy, Inference Success Rate, ...
3	Performance consistency (Art. 15)	AI systems shall perform consistently with respect to accuracy, robustness and cybersecurity throughout their lifecycle.	Verify that cyberattacks do not cause unacceptable degradation of the AI system's intended performance.	Accuracy degradation under attack, False Acceptance/False Rejection increase, Service Availability, Response Time Degradation, ...
4	Fault tolerance and resilience (Art. 15)	AI systems shall be resilient regarding errors, faults or inconsistencies that may occur within the system or the environment.	Assess whether the AI system continues operating safely during failures or cyber incidents.	Recovery Time, Service Availability, Failure Recovery Rate, Successful Failover Rate, ...
5	Record-keeping (Article 12)	High-risk AI systems shall be designed and developed with logging capabilities enabling the automatic recording of events relevant for identifying situations that may lead to risks or substantial modifications throughout the system's lifecycle.	Verify that cybersecurity-relevant events are automatically recorded, that logs are sufficiently complete to support incident investigation and post-market monitoring, and that they are protected against unauthorised modification or deletion.	Logging coverage, Log completeness, Log integrity protection rate, Timestamp accuracy, ...

4.5.3 Test Methodology Example

The following section describes the usual ideal test methodology, however this is not written in stone, cybersecurity tests should be stay flexible and should be adapted according to what is discovered. What is presented below is an example of how cybersecurity tests can be planned, organised and carried out during the conformity assessment activities of an AI system.

Following the planning phase, the testing laboratory should execute the cybersecurity assessment according to the established test plan. The objective of this phase is to obtain objective, repeatable and verifiable evidence demonstrating whether the AI system satisfies the cybersecurity requirements established by Article 15 of the EU AI Act. The testing methodology should adopt a layered approach, progressively evaluating the AI system from documentation and architectural design through to technical implementation and operational behaviour under realistic attack scenarios. This progressive methodology enables the testing laboratory to

identify vulnerabilities at different levels of the system while maintaining traceability between identified risks, executed tests and the resulting conformity assessment.

The assessment should begin with a review of the security architecture implemented by the provider. This activity aims to verify that cybersecurity has been incorporated into the design of the AI system rather than added as an afterthought. The testing laboratory should evaluate whether the provider has implemented an appropriate security architecture, identified the principal assets requiring protection and established suitable trust boundaries between internal and external components. Particular attention should be given to the interfaces between machine learning components and conventional software services, as these interfaces frequently constitute attractive attack surfaces. The review should also determine whether the provider has documented security assumptions, operational constraints and residual risks that may influence the interpretation of subsequent testing activities.

Once the architectural review has been completed, the testing laboratory should verify the integrity of the software and AI artefacts that constitute the system under evaluation. The integrity assessment should determine whether machine learning models, training datasets, configuration files, software packages and deployment artefacts can be modified without detection. The laboratory should verify the use of cryptographic integrity mechanisms, digital signatures, secure repositories and authenticated software distribution processes where appropriate. Secure update procedures should also be evaluated to ensure that modifications introduced after deployment preserve the integrity of the AI system and cannot be exploited to introduce malicious models or software components.

The next stage consists of evaluating the exposure of the AI system to known software vulnerabilities. Conventional vulnerability assessment techniques should be applied to the operating systems, application frameworks, software dependencies, container images and network services supporting the AI application. The objective is to identify publicly known vulnerabilities, insecure configurations or obsolete software components that could provide attackers with an entry point into the AI system. Where software bills of materials (SBOMs) are available, they should be reviewed to verify the provenance and maintenance status of software dependencies and to identify components affected by publicly disclosed vulnerabilities.

Following vulnerability assessment, penetration testing should be conducted to determine whether identified weaknesses can be exploited under realistic operational conditions. Penetration testing should focus on the interfaces exposed by the AI system, including web applications, APIs, management consoles, authentication mechanisms and communication services. The objective is not only to determine whether unauthorised access can be obtained but also to evaluate the potential impact of successful exploitation on the behaviour of the AI system. Whenever vulnerabilities are successfully exploited, the testing laboratory should document the attack path, the affected assets, the resulting impact on system operation and the effectiveness of any implemented detection or mitigation mechanisms.

In parallel with conventional penetration testing, the cybersecurity assessment should evaluate threats that specifically target machine learning systems. These attacks exploit the statistical characteristics of AI models rather than implementation defects within conventional software components. Consequently, dedicated AI security evaluations should be performed to determine whether the learning algorithm remains sufficiently robust against manipulation.

One important activity consists of assessing the resilience of the AI model against adversarial manipulation during inference. The testing laboratory should generate adversarial inputs representing realistic perturbations that preserve the semantic content of the original input while attempting to alter the model's prediction. The objective is to evaluate whether such perturbations significantly reduce the reliability of the AI system and whether implemented defence mechanisms maintain acceptable performance under attack conditions. The testing methodology should consider both white-box and black-box attack scenarios whenever the required information about the model is available.

Where the AI system relies on supervised learning, the assessment should also evaluate its resilience against data poisoning attacks. These attacks attempt to influence the behaviour of the trained model by introducing malicious or misleading samples into the training dataset. Although complete retraining may not always be feasible during conformity assessment, the testing laboratory should review the provider's data governance procedures, dataset validation mechanisms and quality assurance processes to determine whether appropriate measures have been implemented to prevent or detect malicious modifications of training data. Where retraining can be reproduced, representative poisoning scenarios may also be evaluated to quantify their impact on model behaviour.

The methodology should further evaluate the confidentiality of machine learning models and associated datasets by assessing the system's resistance to model extraction, model inversion and membership inference attacks. These attacks may enable an adversary to reconstruct proprietary models, recover sensitive information contained within training datasets or determine whether specific records were used during model development. The testing laboratory should evaluate whether repeated interaction with the AI system allows an attacker to infer information beyond the intended outputs and whether appropriate technical measures have been implemented to limit information leakage.

For AI systems providing externally accessible interfaces, API security should receive particular attention. APIs frequently constitute the primary interaction point between users and AI services and therefore represent a significant attack surface. The assessment should verify authentication and authorisation mechanisms, rate limiting, input validation, session management and error handling. Particular care should be taken to identify opportunities for excessive information disclosure, insecure direct object references, injection attacks or misuse of inference endpoints that could facilitate model extraction or service disruption.

The cybersecurity assessment should also evaluate the effectiveness of monitoring, logging and incident detection capabilities. Security monitoring constitutes an essential element of operational resilience, enabling providers to identify malicious activities, investigate security incidents and implement timely corrective actions. The testing laboratory should therefore verify that relevant security events are logged, that logs cannot be modified without authorisation and that monitoring mechanisms provide sufficient visibility into both conventional cyberattacks and AI-specific anomalies. Where automated detection mechanisms have been implemented, representative attack scenarios should be used to evaluate their detection capability and the timeliness of generated alerts.

Throughout the execution of cybersecurity testing, all observations should be systematically documented to ensure complete traceability. For every identified vulnerability or security weakness, the testing laboratory should record the affected component, the executed test procedure, the observed behaviour, the impact on the AI system, the associated regulatory requirement and any supporting evidence. Maintaining this traceability facilitates subsequent analysis, enables reproducibility of the assessment and supports the preparation of the final conformity assessment report.

Although the individual testing activities described above evaluate different aspects of the AI system, they should not be interpreted independently. The overall objective of the methodology is to assess the effectiveness of the provider's cybersecurity posture as a whole. Consequently, the testing laboratory should analyse the combined results of architectural review, vulnerability assessment, penetration testing, AI-specific security evaluation and operational resilience testing before determining whether the AI system satisfies the cybersecurity requirements established by the AI Act. This integrated evaluation provides a more realistic representation of the system's security than isolated testing activities and supports a more robust conformity assessment.

The execution phase concludes with the consolidation of all collected evidence into a structured assessment record. This record provides the basis for the scoring and classification process described in the following section, where the testing laboratory analyses the identified findings, determines their significance for regulatory compliance and formulates the final cybersecurity conclusions supporting the conformity assessments.

4.5.4 Scoring and Classification of Cybersecurity Test Results

The objective of the scoring and classification phase is to determine whether the evidence collected during cybersecurity testing demonstrates compliance with the cybersecurity requirements of the AI system or whether additional mitigation measures are required before conformity can be established. Consequently, the assessment should consider not only the existence of vulnerabilities but also their potential impact on the safety, robustness and trustworthy operation of the AI system.

The evaluation should begin by consolidating the results obtained from all cybersecurity testing activities. Since different testing activities may identify related weaknesses affecting the same system component, the testing laboratory should correlate the collected evidence before performing the final assessment. This integrated analysis provides a more realistic understanding of the overall cybersecurity posture than evaluating each finding independently.

Each identified vulnerability should be analysed to determine its likelihood of exploitation, the assets affected, the operational consequences of a successful attack and its impact on the intended functionality of the AI system. Particular attention should be given to vulnerabilities that directly influence the behaviour of the machine learning model, compromise the integrity of training data or enable unauthorised modification of inference results. While certain software vulnerabilities may have limited operational consequences, weaknesses affecting the AI decision-making process may significantly alter the behaviour of the system and therefore require a higher level of attention during conformity assessment.

The evaluation should also consider the effectiveness of the security controls implemented by the provider. The presence of vulnerability does not necessarily imply that the AI system fails to satisfy the cybersecurity requirements of the AI Act if appropriate preventive, detective or corrective measures reduce the associated risk to an acceptable level. For example, an identified software vulnerability may present limited practical risk if the affected service is isolated within a protected network segment, continuously monitored and subject to strong authentication controls. Conversely, even relatively simple vulnerabilities may represent unacceptable risks if they affect externally accessible components supporting safety-critical AI functions.

When evaluating AI-specific attacks, the testing laboratory should assess both the technical success of the attack and its practical impact on the behaviour of the AI system. An adversarial attack that slightly modifies prediction confidence without affecting the final decision may represent a lower operational risk than an attack that consistently alters classification outcomes. Similarly, model extraction attacks should be evaluated according to the amount of information recovered, the effort required to reproduce the original model and the potential consequences of successful extraction for both the provider and end users. The evaluation should therefore focus on the operational significance of the observed behaviour rather than solely on the technical feasibility of the attack.

The cybersecurity evaluation should further determine whether the provider has implemented appropriate monitoring and incident response capabilities capable of detecting and responding to the identified threats. During testing, the laboratory should assess whether simulated attacks generate appropriate security events, whether these events are accurately recorded and whether the implemented monitoring mechanisms provide sufficient information to support incident investigation. The ability to detect and respond to cybersecurity incidents constitutes an essential element of operational resilience and should therefore be considered alongside preventive security controls during the conformity assessment.

An important aspect of the evaluation consists of analysing the cumulative effect of multiple vulnerabilities. Individual weaknesses that appear insignificant when considered independently may collectively create unacceptable cybersecurity risks when combined through realistic attack paths. The testing laboratory should therefore evaluate potential attack chains that exploit several vulnerabilities in sequence to compromise critical AI assets or modify system behaviour. This holistic evaluation reflects realistic attack scenarios more accurately than isolated vulnerability analysis and provides a more representative assessment of the overall cybersecurity posture of the AI system.

Following the technical evaluation, each finding should be classified according to its potential impact on regulatory compliance. Findings that compromise the integrity of machine learning models, enable manipulation of AI decisions or significantly reduce operational robustness should generally be considered more critical than vulnerabilities affecting non-essential supporting components. This regulatory perspective ensures that the evaluation remains aligned with the objectives of conformity assessment rather than conventional penetration testing.

The evaluation should conclude with an assessment of the residual cybersecurity risk remaining after consideration of all implemented mitigation measures. Residual risk represents the level of risk that continues to exist despite the provider's security controls and should be evaluated with respect to the intended purpose of the AI system, its deployment environment and the potential consequences of successful attacks. Although the complete elimination of cybersecurity risk is generally not achievable, the residual risk should remain compatible with the intended use of the AI system and should not compromise its safety, robustness or regulatory compliance.

The outcome of the scoring and classification phase forms the basis of the cybersecurity conclusions included in the conformity assessment report. These conclusions should clearly state whether the evidence collected during testing demonstrates that the AI system satisfies the applicable cybersecurity requirements, whether additional corrective actions are required or whether unresolved cybersecurity weaknesses prevent a positive conformity assessment. The conclusions should be supported by objective evidence and provide sufficient

justification to enable authorities and other stakeholders to understand the rationale underlying the final assessment.

This should also contain enough elements to be able to reproduce everything and being able to find again the same behaviours and results.

4.5.5 Example of Cybersecurity Testing of a Biometric Identification System

To illustrate the proposed methodology, this section presents an example of cybersecurity testing performed on a high-risk biometric identification system intended for access control. The objective of the assessment is to determine whether the system satisfies the cybersecurity requirements by evaluating its resilience against both conventional cybersecurity threats and attacks specifically targeting machine learning models.

The biometric identification system consists of several interconnected components that collectively perform identity verification. The system receives facial images through a camera interface, preprocesses the acquired images, executes inference using a deep learning model and compares the resulting feature representation with biometric templates stored in a secure database. Following successful identification, the system communicates the decision to an external access control application responsible for granting or denying access to protected facilities. Supporting components include authentication services, system logs, administrative interfaces and secure communication channels between the different software modules.

Prior to initiating the cybersecurity assessment, the testing laboratory reviews the technical documentation provided by the AI system provider. The documentation includes the system architecture, intended purpose, software dependencies, deployment environment, risk assessment, security architecture and software update procedures. This review enables the laboratory to identify the principal assets requiring protection and to establish the scope of the cybersecurity evaluation. In this example, the critical assets include the facial recognition model, biometric templates, training datasets, inference service, communication interfaces, authentication mechanisms and audit logs.

Following the documentation review, the testing laboratory establishes or reuses an existing threat model describing the potential adversaries and attack scenarios relevant to the biometric identification system. The threat model assumes that an external attacker may attempt to bypass the authentication process, manipulate model predictions, extract sensitive biometric information or disrupt system availability. In addition, insider threats are considered, including privileged users attempting to modify model parameters, alter biometric templates or disable security monitoring mechanisms. These assumptions provide the basis for selecting representative cybersecurity tests that reflect realistic operational conditions.

The first phase of technical testing focuses on the security architecture and deployment environment. The laboratory verifies that communication between system components is encrypted, administrative interfaces require strong authentication and role-based access control has been correctly implemented. Software dependencies are analysed to identify publicly disclosed vulnerabilities, while configuration reviews verify that unnecessary services have been disabled and security settings comply with recognised cybersecurity best practices. The laboratory also evaluates the provider's software update mechanism to ensure that only authenticated and integrity-protected software packages can be deployed.

Once the infrastructure assessment has been completed, penetration testing is performed against externally accessible interfaces. The testing laboratory evaluates the resilience of the authentication mechanisms, API endpoints and administrative interfaces against common attack techniques such as authentication bypass, privilege escalation, injection attacks and session manipulation. Where vulnerabilities are identified, the laboratory determines whether exploitation could affect the operation of the biometric identification process or compromise sensitive information stored by the system.

The assessment then proceeds to AI-specific cybersecurity testing. Adversarial examples are generated by introducing carefully controlled perturbations to facial images while preserving their visual appearance to human observers. These modified images are submitted to the biometric recognition model to determine whether prediction results can be manipulated without significant perceptual changes. The laboratory records changes in recognition accuracy, confidence scores and false acceptance or false rejection rates. Where implemented, adversarial defence mechanisms are evaluated by comparing system performance before and after activation of the corresponding protection mechanisms.

To evaluate the confidentiality of sensitive information, the testing laboratory performs representative model extraction and membership inference experiments. Model extraction testing seeks to determine whether repeated interaction with the prediction interface allows reconstruction of the deployed machine learning model or approximation of its decision boundaries. Membership inference testing evaluates whether an attacker can determine whether a particular biometric record was included within the training dataset. These assessments enable the laboratory to evaluate the effectiveness of implemented privacy protection mechanisms and the potential exposure of sensitive biometric information.

The integrity of the biometric templates and machine learning model is subsequently evaluated. The testing laboratory verifies whether unauthorized modifications of model parameters, template databases or configuration files can be detected through cryptographic integrity mechanisms or monitoring systems. Simulated attempts to replace the deployed model or alter stored biometric templates are performed in a controlled environment to evaluate whether appropriate security controls prevent or detect such modifications before operational decisions are affected.

The laboratory also evaluates the resilience of the system against denial-of-service scenarios. Representative overload conditions are generated by increasing the number of authentication requests submitted to the inference service. The assessment measures whether the system maintains acceptable response times, whether resource management mechanisms prevent service degradation and whether critical security functions remain operational throughout the attack. Where failover mechanisms have been implemented, their effectiveness is also evaluated by simulating partial infrastructure failures.

Throughout the assessment, all security events generated during testing are collected and analysed. The laboratory verifies that authentication failures, administrative actions, model updates, anomalous inference requests and detected attack attempts are accurately recorded within the audit logs. Particular attention is given to the integrity of logging mechanisms and the ability of the provider to reconstruct security incidents using the available evidence. The assessment also evaluates whether monitoring systems generate timely alerts and whether sufficient information is available to support incident investigation and forensic analysis.

Following completion of the technical testing activities, the laboratory consolidates the collected evidence and evaluates the identified findings with respect to the cybersecurity. Each vulnerability is analysed according to its likelihood of exploitation, operational impact and influence on the safety, robustness and reliability of the biometric identification system. Where appropriate, mitigation recommendations are formulated and the provider may implement corrective actions before the affected components are re-evaluated. The final cybersecurity assessment concludes whether the implemented security measures provide an appropriate level of protection against foreseeable cyber threats and whether the biometric identification system satisfies the cybersecurity requirements necessary to support a positive conformity assessment.

4. INTEGRITY OF GROUND TRUTH FOR TEST DATA

Ensuring the integrity of ground-truth labels for test data is critical to successful testing. If the test labels contain errors, the performance metrics derived from the tests will be inaccurate, rendering the validation useless⁵⁴. Implementing processes for ensuring and verifying the quality of ground-truth is an important aspect of testing, although these processes are not directly related to the construction of a test protocol.

For data collected manually, it is important to have digital tools that support the data collection process. Such tools should be designed to minimize the chances of human error during the data labelling process.

Sometimes annotation may occur after data collection is completed. Instead of relying on automated annotation scripts, multiple-blind human annotation should be used. The recommended best practice here is to assign the same data sample to several people for independent, manual annotation. For a given data sample, the annotations or labels produced by the multiple annotators may be combined into a single ground-truth using predefined rules (such as majority consensus), combined with suitable tie-breaking policies.

Algorithmic and statistical sanity checks can be implemented to programmatically catch human errors, typos, and logical contradictions before the test data is finalized. This may include deterministic validation processes, such as cross-checking the status of a subject — minor or adult — against the age information in the MRZ in the subject's passport. Reliability metrics, such as Cohen's Kappa scores⁵⁵, may be computed continuously during the annotation process. Any sudden drop in this score signals that the annotators are probably guessing, and that the data or the instructions for annotation are not clear.

More rigorous best practices recommend blind *infiltration*, based on the use of a 'gold standard' dataset — a secret, small, well curated dataset for which the ground-truth annotation has been completely validated. From time to time, some samples from this dataset are mixed into the work assigned to manual annotators, to check the efficacy of the annotators. An alarm is triggered if an annotator does not label a golden data-sample correctly.

To maintain continuous traceability and label provenance, ground-truth labels should be treated in the same way as software source code. Every change must leave a digital trail. Structured metadata logs, cryptographic hashing, and immutable version control may be applied to track changes in annotation.

⁵⁴ Omer Nahum O., Calderon N., Keller O., Szpektor I., Reichart R. Are LLMs Better than Reported? Detecting Label Errors and Mitigating Their Effect on Model Performance . 2025. <https://arxiv.org/abs/2410.18889>

⁵⁵ Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1), 37-46. <https://journals.sagepub.com/doi/10.1177/001316446002000104>

5. TEST REPORT TEMPLATE

A test report template helps to standardize how the testing lab documents AI product evaluations, ensuring compliance, reproducibility, and clarity for the various stakeholders.

A number of specific sections must appear in every test report. These mandatory sections are listed in Table 9 **Erreur ! Source du renvoi introuvable.** The table only presents the relevance of each section to the test report. In general, each section will include several topics.

Table 9: List of mandatory sections of a test report.

I1	Section	Purpose
1	Document Control and Administration	Ensures traceability, accountability, and reproducibility of the test results, and supports the legal and operational foundations of the evaluation process.
2	TOE Identification	Identifies the precise version of the AI product under test. It isolates the software artifact, preventing ambiguity, product switching, or unapproved modifications during and after lab evaluation.
3	Product Intent and Operational Context	Clarifies that the test results are applicable to a specific TOE for a specific set of use-cases.
4	Test Data Profile	Summarizes the physical characteristics, size, and integrity of the testing environment. It proves that the evaluation is statistically sound, realistic, and completely independent of the AI developer's training process.
5	Test Data Quality and Bias Audit	Ensures the integrity, fairness, and demographic neutrality of the evaluation process. Demonstrates that the test suite is balanced and that the test results will be trustworthy across all populations.
6	Adversarial Testing	Proves that the TOE will not fail catastrophically when faced with real-world fraud or unexpected edge cases.
7	Performance Metrics Estimates	Establishes a baseline expectation of how the AI algorithm <i>should</i> perform before the actual test execution. It serves as a scientific benchmark to ensure the test design is valid and that the final results are mathematically trustworthy.
8	Cybersecurity testing	Description of each test performed. The description should be detailed enough to allow the reproducibility of the tests performed. The description for each test should include the inputs used, the steps of the test, the tools used, the code used or created if any, the output obtained, the potential attack calculation and the conclusion of the test.

9	Pass/Fail Criteria	Converts statistical estimates into a clear decision on whether the AI product passed the evaluation and is safe and effective for deployment.
---	--------------------	--

Finally, in Table 10 **Erreur ! Source du renvoi introuvable.** we use a specific (hypothetical) test case, to illustrate the structure of a suitable test report. In our example, the TOE is face-recognition (FR) system to be deployed at e-gates in European airports.

Table 10: Example Test Report Template for testing a face recognition product for a border-control application.

ol		
1	Document Control and Administration	
a	Protocol Version	v2.4.0-EGATE-ALGO
b	Date of Evaluation	12 June 2026
c	Lead Evaluator	Dr. M. Jackson, Senior Biometrics Auditor
d	Lab Accreditation ID	STS 12345
2	Target of Evaluation (TOE) Identification	
a	TOE Name	IPS Core Matcher
b	TOE Version	v4.12.2-SDK
c	TOE Type	Facial Recognition Software Development Kit (1:1 Verification Core)
d	Developer/Provider	IDEMIA Public Security Ltd.
e	Unique Identifier	SHA-256: 8f3b9c1e7a4d6f5e2b3c4a5b6f7e8a9b0c1d2e3f4a5b6c7d8e9f0a1b2c3d4e5f

3	Product Intent and Operational Context	
a	Intended Purpose	High-throughput 1:1 biometric verification matching live e-gate camera captures against electronic passport (eMRTD) chip images
b	Expected Use-Cases	Automated border control passport checkpoints for international arriving and departing passengers
c	Out-of-Scope Uses	1:N surveillance crowd scanning, identification from long distances, or matching low-resolution CCTV footage
d	Target Population	Global air travellers aged 12 to 85 spanning diverse nationalities, ethnicities, and skin tones
4	Test Data Profile	
a	Data Sources	Anonymized datasets collected by test lab; synthetic images.
b	Data Volume	500,000 image pairs (1,000,000 total images)
c	Reference Images	JPEG/WSQ passport chip photos (ISO/IEC 19794-5 token format)
d	Probe Images	E-gate camera captures simulated under varied indoor lighting environments.
e	Data Splitting	Evaluation set 100% isolated from the developer's SDK training pipeline.
f	Storage Location	volumes/secure_bio_test/egates_2026_v2/
5	Test Data Quality and Bias Audit	
a	Data Diversity Matrix	Balanced sample across the Fitzpatrick Skin Phototype Scale (Types I-VI), gender (51% Male, 49% Female), and age brackets (12–30, 31–60, 61+).
b	Identified Biases	Historical reference data features a 12% higher density of Fitzpatrick Types II and III relative to Types V and VI. Passport photos can present variations in age gaps (up to 10 years between passport issuance and e-gate capture).

c	Mitigation Actions	The lab applied synthetic illumination balancing and used synthetic data to augment under-represented demographic pairs to ensure statistically equal representation during verification scoring.
d	Data Completeness	0.02% of data excluded due to corrupt ICAO headers or severe motion blur rendering the image unreadable by standard baseline tools
6	Adversarial Testing	
a	Attack Vectors Tested	Presentation Attack Detection (PAD), morphing attacks, digital adversarial eye-wear perturbations, and printed 2D/3D high-resolution masks
b	Scenario Descriptions	
	Scenario A: Morphing	Blended passport photos combining two distinct identities at a 50/50 split, to test algorithm classification thresholds.
	Scenario B: Illumination Extremes	Probe images subjected to severe overhead airport skylight glare and under-exposure simulator filters.
	Scenario C: Age Progression	Probe images matched against passport reference photos taken 8–10 years prior.
c	Defence Mechanism Audit	Evaluated the TOE's deep feature embedding distance, to flag anomalous textures or non-human geometric patterns indicative of deepfakes or physical artifacts.
d	Adversarial Testing: Presentation Attacks	
	Level 1 PAIs	
	Species 1.1	High-resolution 2D glossy-paper prints of passport-holder faces.
	Species 1.2	Digital video replay displays using an iPad Pro
	Level 2 PAIs	
	Species 2.1	Curved printed masks (matte photo paper, with holes cut-out for eyes, nose, lips)
	Species 2.2	Face images digitally enhanced before printing on glossy photo paper.

7	Performance Metrics Estimates: Pre-test estimates derived from vendor claims			
Metric Category		Metric Name	Estimated / Baseline Value	Confidence Interval (95%)
FR Accuracy		False Match Rate (FMR)	0.0001	+/- 0.00002
FR Accuracy		False Non-Match Rate (FNMR)	0.008	+/- 0.0005
FR Fairness		Gini Aggregation Rate for Biometric Equitability (GARBE)	<= 0.2 for a global FMR <= 0.00001 (1 in 100,000)	0.21 - 0.29
PAD Security		Attack Presentation Classification Error Rate (APCER)	0.015	0.0110 - 0.0190
PAD Usability		Bona Fide Presentation Classification Error Rate (BPCER)	0.02	0.0170 - 0.0230
Combined System		Imposter Attack Presentation Accept Rate (IAPAR)	0.005	0.0020 - 0.0080
Robustness to Morphs		Morphing Attack Detection Rate	0.95	+/- 0.012
Computing Efficiency		Template Extraction Latency	85 ms	82 - 88 ms
8	Cybersecurity test			
Test #1				
Name of the test			Replay Attack Against the Decision API	
Input for the test			Legitimate verification messages exchanged between the face recognition service and the e-gate controller. Fifty replay attempts were performed using previously captured authenticated messages.	

Description of the test (test steps, tools used, script developed if any)	The objective of the test was to determine whether previously valid authentication responses could be replayed to obtain repeated gate-opening decisions. The following steps were performed: 1. Legitimate communication between the face recognition service and the e-gate controller was captured using Wireshark. 2. The exchanged HTTP requests were analysed and modified using Burp Suite. 3. A custom Python replay script was developed to automatically resend previously valid authentication responses while modifying selected parameters (timestamps and transaction identifiers). 4. The replayed requests were submitted to the decision API and the corresponding HTTP responses, and gate decisions were recorded. 5. Audit logs and monitoring alerts were analysed to verify whether the replay attack was detected and prevented.
Output obtained	12 of 50 replayed messages were accepted, resulting in repeated gate-opening commands. All attempts were logged, but only one monitoring alert was generated.
Attack calculation	Replay Success Rate: $12/50 = 24\%$ Successful replay demonstrates the absence of effective transaction replay protection, despite message authentication.
Conclusion of the test	Fail. The system correctly rejected forged requests but remained vulnerable to replay attacks. Single-use transaction identifiers and replay protection mechanisms should be implemented before conformity can be demonstrated.
Test #2	
Name of the test	Adversarial Evasion Attack Against the Face Recognition Model
Input for the test	100 enrolled identities, 1,000 genuine verification attempts, 150 adversarially modified facial images generated from 30 identities.
Description of the test (test steps, tools used, script developed if any)	The objective of the test was to evaluate whether adversarial perturbations could cause the face recognition system to incorrectly authenticate an individual. The following steps were performed:

		1. A baseline recognition campaign was conducted using genuine facial images to establish the nominal recognition performance. 2. Adversarial facial images were generated using a custom Python script based on OpenCV and the Adversarial Robustness Toolbox (ART). 3. The generated images were presented to the operational e-gate under normal environmental conditions. 4. For each attempt, the predicted identity, similarity score, accept/reject decision and response time were recorded. 5. Audit logs and security monitoring events were analysed to determine whether the attack was detected.		
Output obtained		9 out of 150 adversarial attempts resulted in an incorrect acceptance (6%). Only 2 successful attacks triggered a security alert. All attempts were logged.		
Attack calculation		Attack Success Rate = 9/150 = 6% . This exceeds the provider's acceptance criterion of <1% , indicating insufficient robustness against adversarial examples.		
Conclusion of the test		Fail. The facial recognition model remains vulnerable to adversarial manipulation. Additional defences (adversarial detection, stronger liveness checks and multi-frame verification) should be implemented before reassessment.		
9	Pass/Fail Criteria: The TOE must satisfy all thresholds listed below under simulated operational conditions to be certified for airport deployment.			
Evaluated Metric		Minimum Acceptable Threshold	Actual Measured Result	Status (Pass/Fail)
FMR		<= 0.001	<Insert Result>	<Pending>
FNMR		<= 0.01 at FMR = 10 ⁻⁴	<Insert Result>	<Pending>
FTE		<= 0.0005	<Insert Result>	<Pending>
Demographic Differential (Max Δ FNMR)		<= 0.003 between subgroups	<Insert Result>	<Pending>
APCER – Level 1		<= 0.01	<Insert Result>	<Pending>
APCER – Level 2		<= 0.03	<Insert Result>	<Pending>

BPCER	≤ 0.1	<Insert Result>	<Pending>
IAPAR	≤ 0.001	<Insert Result>	<Pending>
Morph Attack Detection (MAD) Rate	≥ 0.9 accuracy	<Insert Result>	<Pending>
Template Matching Compute Time	≤ 150 ms per comparison	<Insert Result>	<Pending>

6. CONCLUSIONS

The EU AI Act includes provision for independent evaluation of a high-risk AI product under a variety of circumstances. For example, a biometrics system used for remote identification or categorization must undergo evaluation by a third-party independent lab (must be an AL). High-risk AI systems in other application domains, such as critical infrastructure, education, or employment may be required to furnish an assessment report by a third-party if the provider has been deemed to not apply European harmonized standards, or operates in a domain where such harmonized standards are not available. Similarly, an AI system that has already been cleared for compliance with the EU AI Act would be required to undergo third-party assessment if substantial modifications are made to the AI product.

To carry out such an independent assessment, the testing lab should first prepare a test plan. In this document we lay out the principles to be followed when preparing a test plan to evaluate a specific AI product (or family of similar products).

Five suites of tests should be created, to assess the compliance of an AI product to the EU AI Act, namely, functional tests, adversarial tests, fairness tests, explainability tests, and cybersecurity tests. Functional testing verifies the claimed capabilities of the TOE and demonstrates that the TOE is fit for its intended purpose under normal operating conditions. Adversarial tests probe the robustness of the TOE to possible attacks, and thus evaluate the security of the AI product. Fairness testing verifies that the TOE behaves uniformly across different demographic factors, and neither favours nor discriminates against any specific demographic group. Explainability tests verify that the TOE is sufficiently transparent, as required by the EU AI Act, to facilitate human oversight, as well as the right to an explanation (of an affected party). Finally, the purpose of cybersecurity testing is to verify that the TOE is resilient to unauthorized access, and capable of functioning securely throughout its lifecycle. We recommend that cybersecurity testing should be performed first, as failures in tests of other types may be explained by cybersecurity shortcomings.

In this report we emphasize that the construction of a comprehensive test plan requires first defining both the scope as well as the objectives of the necessary tests. The scope and objectives lead directly to the design of appropriate tests. The core of this document is a set of guidelines about defining the scope and the objectives for the five sets of tests. We show how the objectives for the various sets of tests can be derived by mapping the legal mandates specified in the AI Act to quantifiable metrics. Where necessary, we illustrate the scope and objective definition with specific examples. In this report we have considered a hypothetical face-recognition (FR) product as the TOE, and have created examples for testing such a product.

The report concludes with guidelines about how to structure a test report, including a sample test report template for evaluating a FR product to be deployed in a border-control scenario.