## 1 Annex 22: Artificial Intelligence

2 **Reasons for changes:** Not applicable (new annex).

## 3 Document map

## 1. Scope

This annex applies to all types of computerised systems used in the manufacturing of medicinal products and active substances, where Artificial Intelligence models are used in critical applications with direct impact on patient safety, product quality or data integrity, e.g. to predict or classify data. The document provides additional guidance to Annex 11 for computerised systems in which AI models are embedded.

The document applies to machine learning (AI/ML) models which have obtained their functionality through training with data, rather than being explicitly programmed. Models may consist of several individual models, each automating specific process steps in GMP.

The document applies to static models, i.e. models that do not adapt their performance during use by incorporating new data. The use of dynamic models which continuously and automatically learn and adapt performance during use, is not covered by this document, and should not be used in critical GMP applications.

The document applies to models with a deterministic output which, when given identical inputs, provide identical outputs. Models with a probabilistic output which, when given identical inputs, might not provide identical outputs are not covered by this document and should not be used in critical GMP applications.

Following the above, the document does not apply to Generative AI and Large Language Models (LLM), and such models should not be used in critical GMP applications. If used in non-critical GMP applications, which do not have direct impact on patient safety, product quality or data integrity, personnel with adequate qualification and training should always be responsible for ensuring that the outputs from such models are suitable for the intended use, i.e. a human-in-the-loop (HITL) and the principles described in this document may be considered where applicable.

## 2. Principles

2.1. *Personnel*. In order to adequately understand the intended use and the associated risks of the application of an AI model in a GMP environment, there should be close cooperation between all relevant parties during algorithm selection, and model training, validation, testing and operation. This includes but may not be limited to process subject matter experts (SMEs), QA, data scientists, IT, and consultants. All personnel should have adequate qualifications, defined responsibilities and appropriate level of access.

2.2. *Documentation*. Documentation for activities described in this section should be available and reviewed by the regulated user irrespective of whether a model is trained, validated and tested in-house or whether it is provided by a supplier or service provider.

2.3. *Quality Risk Management* Activities described in this document should be implemented based on the risk to patient safety, product quality and data integrity.

## 3. Intended Use

3.1. *Intended use*. The intended use of a model and the specific tasks it is designed to assist or automate should be described in detail based on an in-depth knowledge of the process the model is integrated in. This should include a comprehensive characterisation of the data the model is intended to use as input and all common and rare variations; i.e. the input sample space. Any limitations and possible erroneous and biased inputs should be identified. A process subject matter expert (SME) should be responsible for the adequacy of the description, and it should be documented and approved before the start of acceptance testing.

3.2. *Subgroups*. Where applicable, the input sample space should be divided into subgroups based on relevant characteristics. Subgroups may be defined by characteristics like the decision output (e.g. ‘accept’ or ‘reject’), process specific baseline characteristics (e.g. geographical site or equipment), specific characteristics in material or product, and characteristics specific to the task being automated (e.g. types and severity of defects).

3.3. *Human-in-the-loop*. Where a model is used to give an input to a decision made by a human operator (human-in-the-loop), and where the effort to test such model has been diminished, the description of the intended use should include the responsibility of the operator. In this case, the training and consistent performance of the operator should be monitored like any other manual process.

## 4. Acceptance Criteria

4.1. *Test metrics*. Suitable, case dependent test metrics, should be defined to measure the performance of the model according to the intended use. As an example, suitable test metrics for a model used to classify products (e.g. ‘accept’ or ‘reject’) may include, but may not be limited to, a confusion matrix, sensitivity, specificity, accuracy, precision and/or F1 score.

4.2. *Acceptance criteria*. Acceptance criteria for the defined test metrics should be established by which the performance of the model should be considered acceptable for the intended use. The acceptance criteria may differ for specific subgroups within the intended use. A process subject matter expert (SME) should be responsible for the definition of the acceptance criteria, which should be documented and approved before the start of acceptance testing.

4.3. *No decrease*. The acceptance criteria of a model, should be at least as high as the performance of the process it replaces. This implies, that the performance should be known for the process which is to be replaced by a model (see Annex 11 2.7).

## 5. Test Data

5.1. *Selection*. Test data should be representative of and expand the full sample space of the intended use. It should be stratified, include all subgroups, and reflect the limitations, complexity and all common and rare variations within the intended use of the model. The criteria and rationale for selection of test data should be documented.

5.2. *Sufficient in size*. The test dataset, and any of its subgroups, should be sufficient in size to calculate the test metrics with adequate statistical confidence.

5.3. *Labelling*. The labelling of test data should be verified following a process that ensures a very high degree of correctness. This may include independent verification by multiple experts, validated equipment or laboratory tests.

5.4. *Pre-processing*. Any pre-processing of the test data, e.g. transformation, normalisation, or standardisation, should be pre-specified and a rationale should be provided, that it represents intended use conditions.

5.5. *Exclusion*. Any cleaning or exclusion of test data should be documented and fully justified.

5.6. *Data generation*. Generation of test data or labels, e.g. by means of generative AI, is not recommended and any use hereof should be fully justified.

## 6. Test Data Independency

6.1. *Independence*. Effective measures consisting of technical and/or procedural controls should be implemented to ensure the independency of test data, i.e. that data which will be used to test a model, is not used during development, training or validation of the model. This may be by capturing test data only after completion of training and validation, or by splitting test data from a complete pool of data before training has started.

6.2. *Data split*. If test data is split from a complete pool of data before training of the model, it is essential that employees involved in the development and training of the model have never had access to the test data. The test data should be protected by access control and audit trail functionality logging accesses and changes to these. There should be no copies of test data outside this repository.

6.3. *Identification*. It should be recorded which data has been used for testing, when and how many times.

6.4. *Physical objects*. When test data originates from physical objects, it should be ensured, that the objects used for the final test of the model have not previously been used to train or validate the model, unless features are independent.

6.5. *Staff independency*. Effective procedural and/or technical controls should be implemented to prevent staff members who have had access to test data from being involved in training and validation of the same model. In organisations where it is impossible to maintain this independency, a staff member who might have had access to test data for a model, should only have access to training and validation of the same model when working together (in pair) with a colleague who has not had this access (4-eyes principle).

## 7. Test Execution

7.1. *Fit for intended use*. The test should ensure that a model is fit for intended use and is ‘generalising well’, i.e. that the model has a satisfactory performance with new data from the intended use. This includes detecting possible over- or underfitting of the model to the training data.

7.2. *Test plan*. Before the test is initiated, a test plan should be prepared and approved. It should contain a summary of the intended use, the pre-defined metrics and acceptance criteria, a reference to the test data, a test script including a description of all steps necessary to conduct the test, and a description of how to calculate the test metrics. A process subject matter expert (SME) should be involved in developing the plan.

7.3. *Deviation*. Any deviation from the test plan, failure to meet acceptance criteria, or omission to use all test data should be documented, investigated, and fully justified.

7.4. *Test documentation*. All test documentation should be retained along with the description of the intended use, the characterisation of test data, the actual test data, and where relevant, physical test objects. In addition, documentation for access control to test data and related audit trail records, should be retained similarly to other GMP documentation.

## 8. Explainability

8.1. *Feature attribution*. During testing of models used in critical GMP applications, systems should capture and record the features in the test data that have contributed to a particular classification or decision (e.g. rejection). Where applicable, techniques like feature attribution (e.g. SHAP values or LIME) or visual tools like heat maps should be used to highlight key factors contributing to the outcome.

8.2. *Feature justification*. In order to ensure that a model is making decisions based on relevant and appropriate features and based on risk, a review of these features should be part of the process for approval of test results.

## 9. Confidence

9.1. *Confidence score*. When testing a model used to predict or classify data, the system should, where applicable, log the confidence score of the model for each prediction or classification outcome.

9.2. *Threshold*. Models used to predict or classify data should have an appropriate threshold setting to ensure predictions or classifications are made only when suitable. If the confidence score is very low, it should be considered whether the model should flag the outcome as ‘undecided’, rather than making potentially unreliable predictions or classifications.

## 10. Operation

10.1. *Change control*. A tested model, the system it is implemented in, and the whole process it is automating or assisting should be put under change control before it is deployed in operation. Any change to the model itself, the system, or the process in which it is used, including any change to physical objects the model is using as input, should be documented and evaluated to determine if the model needs to be retested. Any decision not to conduct such retest should be fully justified.

10.2. *Configuration control*. A tested model should be put under configuration control before being deployed in operation, and effective measures should be used to detect any unauthorised change.

10.3. *System performance monitoring*. The performance of a model as defined by its metrics should be regularly monitored to detect any changes in the computerised system (e.g. deterioration or change of a lighting condition).

10.4. *Input sample space monitoring*. It should be regularly monitored whether the input data are still within the model sample space and intended use. Metrics should be defined for monitoring any drift in the input data.

10.5. *Human review*. When a model is used to give an input to a decision made by a human operator (human-in-the-loop), and where the effort to test such model has been diminished, records should be kept from this process. Depending on the criticality of the process and the level of testing of the model, this may imply a consistent review and/or test of every output from the model, according to a procedure.

**Glossary**

Artificial Intelligence – 'AI system' means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments;

Deep Learning – Approach to creating rich hierarchical representations through the training of neural networks with many hidden layers

Feature – A pattern in data that can be reduced to a simpler higher-level representation

LIME – Local Interpretable Model-Agnostic Explanations; a technique that approximates any black box machine learning model with a local, interpretable model to explain each individual prediction.

Machine Learning – Machine learning refers to the computational process of optimising the parameters of a model from data, which is a mathematical construct generating an output based on input data. Machine learning approaches include, for instance, supervised, unsupervised and reinforcement learning, using a variety of methods including deep learning with neural networks.

Model – Mathematical algorithms with parameters (weights) arranged in an architecture that allows learning of patterns (features) from training data

Overfitting – Learning details from training data that cannot be generalised to new data

SHAP – Shapley Additive Explanations; an explainable AI (XAI) framework that can provide model-agnostic local explainability for tabular, image, and text datasets

Static – Frozen model: A model where all parameters have been finally set, not allowing further adaption to new data.

Test dataset – The "hold-out" data that is used to estimate performance of the final ML model.

Training dataset – The data used to train the ML model.

Validation dataset (in AI) – The dataset used during model development, to inform on how to optimally train the model from training data. size smaller than the training set