English

Procurement guide for secure AI components

Version July 2022

exploreAI

Procurement guide for secure AI Components

- Preamble

With the progressive development of AI and its fields of application, many companies and organisations will be increasingly confronted with serious and less serious offers for AI components. **The following guidelines are intended to help assess the quality of AI services offered and focus on the security of AI components.** The questions and criteria described here were derived from the research conducted as part of the FFG-funded ExploreAI project and from the risks that arise in connection with AI solutions in order to reduce these risks.

Please note:
The procurement guide described here is no substitute for a human evaluation of the offer and the general usefulness of an AI must be evaluated individually for the respective use case. The focus here is on avoiding generic errors that could jeopardize the security and thus the profitability of the AI as well as the reputation of the respective organisation. Examples of such errors include the purchase of a supposed AI solution that does not contain any AI at all, the unintentionally agreed return of models to the manufacturer, and equally unintentionally agreed remote access.
In addition, ethical components are deliberately not dealt with in this guide, as these must be clarified by the organisation in the context of the respective use case.

Before using this guide, ask yourself the following questions:

… Does the application even need to be AI-based?

… Who is sitting across from you?

… How technically proficient is this person? Are you talking to the right person to assess the safety of the AI components?

… How sensitive is the use case for which the AI solution is to be deployed?

… How sensitive is the data that would have to be used for the AI solution?

… Where should AI be used?

… How important is the procurement scenario? Do you have to decide on a supplier in a short time or do you have room for maneuver?

For use:

▨ ▨ ▨

Questions and indicators can and should be used in each use case

▣ ▣ ▣

Questions and indicators apply only to a specific use case

Procurement Guide

- Overview

	Aspect	Reason	Indicators
General aspects	Categorization of AI	Fraud prevention, assessment of interlocutors and internal risk processes.	Existence of AI Explainability
	Use & Maintenance	Assessment of the effort to establish a technological basis for the AI component.	Integration capability Training effort Attack Surface
	Service / Cloud / On-Premise	Assessment of the control that the company has over the data and processes in the long term.	Control Privacy Compliance
	Source Code	Functionality guarantee (does the component do what it should or unintentionally more?) or further development possibilities	Tasks Non-tasks Extensibility
	Interfaces	Analysis of the connection to external systems and integration capability in the own processes	Flexibility Requirements Robustness
	Audit & Control	Assessment of the manipulation security against internal and external attackers	Degree of protection Transparency Traceability
	Privacy	Applies to the assessment of the processing of the data, if sensitive data is used	Applicability Methods Error-proneness
Data & Sources	Training data	Basis for any AI and origin of various problems (e.g. backdoors)	Data control Backdoors Model stability
	Bias	Avoidance of AI prejudices	Awareness Counterstrategies Relevance

Procurement Guide

- Detail view I

Categorization of AI

Reason

Fraud prevention, assessment of interlocutors and internal risk processes.

Indicators

Existence of AI

Explainability

1. What type of AI is used?

Machine Learning

i. Unsupervised Learning
ii. Supervised Learning
iii. Reinforcement Learning
iv. (Deep Learning)

Natural Language Processing

Automation for processes or robotics (no AI!)

Use & Maintenance

Reason

Assessment of the effort to establish a technological basis for the AI component.

Indicators

Integration capability

Training effort

Attack Surface

1. Can the tasks (training & processing) be parallelized?
2. Are GPUs supported?
3. Is the AI running in a container, virtual machine or physically?

Container

Virtual Machine

physical

4. Are open source libraries / functions used?
5. Are certifications of the software or the manufacturer available?
6. Which programming languages or frameworks are used?

Phyton

Tensorflow

…

7. What additional skills are required on the part of the users?
8. Are updates included?
9. Is there a clear process for updates (security, models, ...)?

Procurement Guide

- Detail view II

10. For individual AI modules:

- How does pipeline integration including scheduling, automation and the like work?
- What metadata is revealed?
- What meta information is needed?
- What integration options exist with regard to metadata (e.g. for A&C)?

11. For black box applications:

- What control options exist?
- How are these secured against unauthorized access?
- What metadata do they disclose without authorization?
- What metadata do they need from the connected systems?

Service / Cloud / On-Premise

Reason	Assessment of the control that the company has over the data and processes in the long term		
Indicators	Control	Privacy	Compliance

1. Is it a service, a cloud solution or an on-premise solution?

- Service
- Cloud solution
- On-premise

2. Especially for on-premise: what are the system requirements?

- Is only the software supplied or is it a complete system?
- Do maintenance accesses exist for the manufacturer? How is patching done?
- Can data (training or processing) or the models be accessed in the course of patching?

3. On-premise: Is something delivered back to the manufacturer (data / benchmarks ...)

- Problem: Sidechannels / information about amount of data and frequency of use and integration

3. Cloud/SAAS data security issues:

- How is the confidentiality of the data with the cloud / SAAS provider ensured?
- How is the data transmission secured?
- What access do operators and software developers have to the data to be evaluated?

Procurement Guide

- Detail view III

3. Cloud/SAAS data security issues II:

Sidechannels and metainformation:

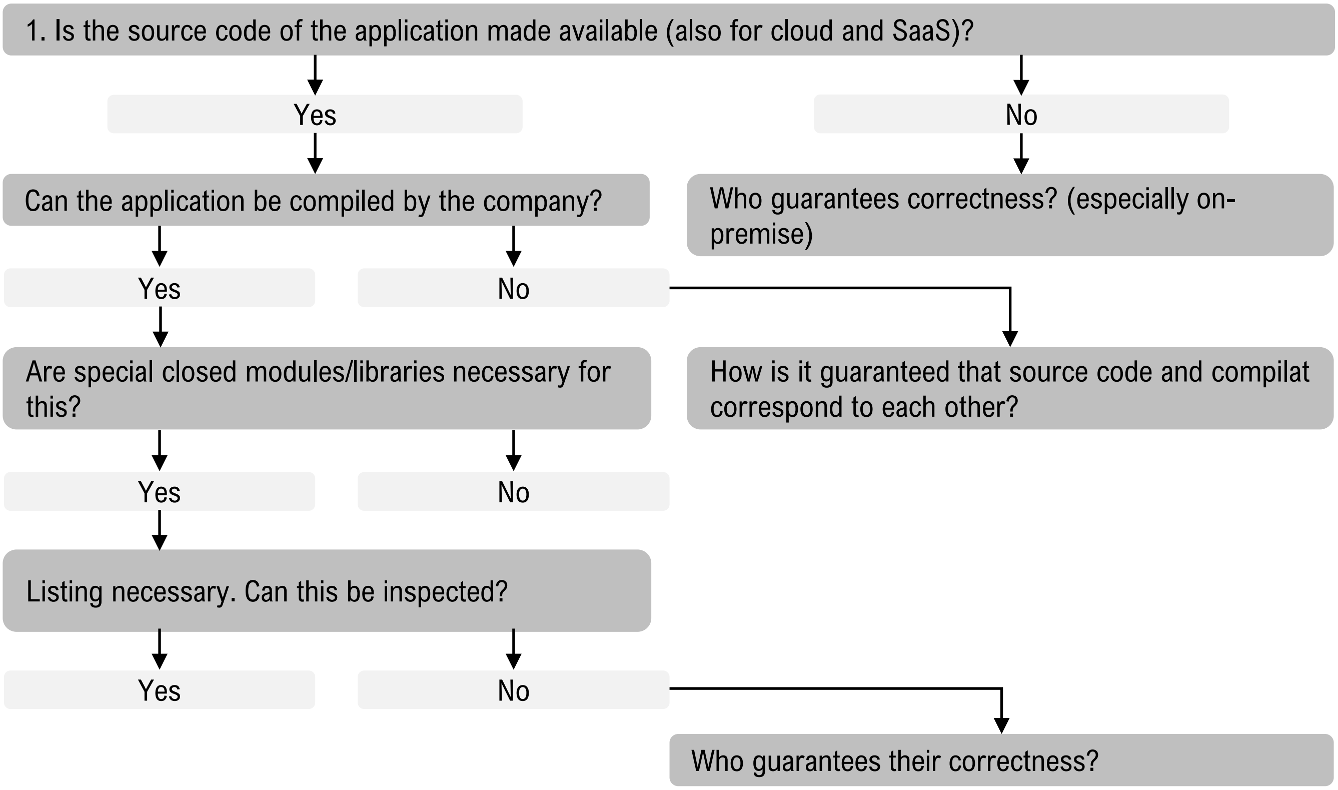
a. Can external parties detect any use taking place / intensification of use by the company?
b. Does the operator recognize more intensive use by the company?
c. What meta-information is generated by the operator and the manufacturer during use (e.g., usage times, number of data records, etc.)?

For virtualization:

a. Can the container be assigned to a physical machine or a geo-location?
b. What sealing measures between the containers have been put in place?

Source Code

Reason	Functionality guarantee (does the component do what it should or unintentionally more?) or further development possibilities		
Indicators	Tasks	Non-tasks	Extensibility



1. Is the source code of the application made available (also for cloud and SaaS)?

Yes → Can the application be compiled by the company?

- Yes → Are special closed modules/libraries necessary for this?
  - Yes → Listing necessary. Can this be inspected?
    - Yes
    - No → Who guarantees their correctness?
  - No
- No → How is it guaranteed that source code and compilat correspond to each other?

No → Who guarantees correctness? (especially on-premise)

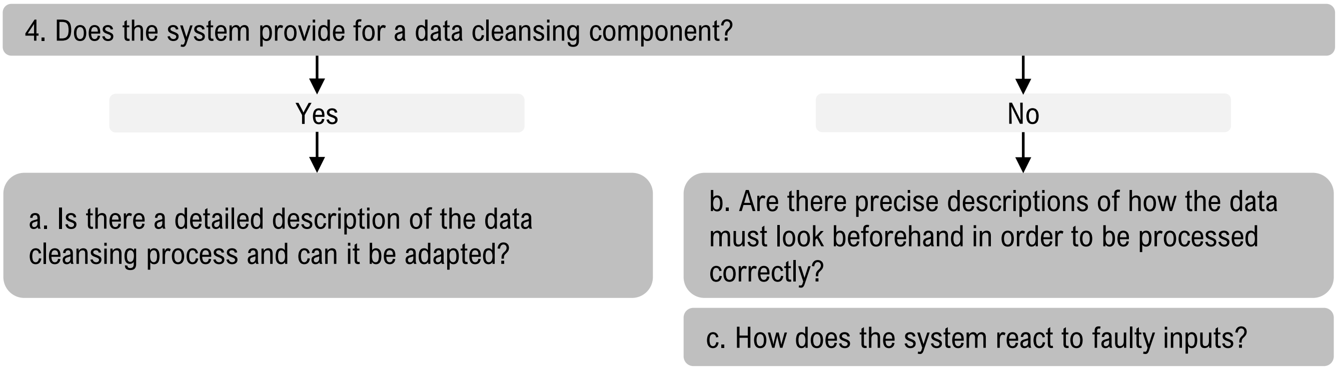
Procurement Guide

- Detail view IV

Interfaces

Reason	Analysis of the connection to external systems and integration capabilities in processes		
Indicators	Flexibility	Requirements	Robustness

1. Which standardized interfaces exist?
2. What information must / can be made available to other systems?
3. Security of the interfaces - can they be locked?
4. Does the system provide for a data cleansing component?



Yes

a. Is there a detailed description of the data cleansing process and can it be adapted?

No

b. Are there precise descriptions of how the data must look beforehand in order to be processed correctly?

c. How does the system react to faulty inputs?

5. How does the system handle conflicting sensor information?
6. Are there methods for detecting manipulated data streams in place?
7. What is specifically returned in the data output?
8. Can mechanisms of fingerprinting or watermarking of output data be used?

Integrity Protection, Audit & Control

Reason	Assessment of the manipulation security against internal and external attackers		
Indicators	Degree of protection	Transparency	Traceability

1. Are models and data sets versioned?
   - If applicable, training and processing?
   - Where is this information stored?
   - What meta information is provided?
   - In the case of off-premise systems, can this data be exported on-premise?

Procurement Guide

- Detail view V

2. What measures are there to protect against tampering?

- In terms of data?
- In terms of the system?
- In terms of models?

3. Is there an audit & control system in place (transparency in data processing)?

- Is there GUI support for this?
- Does an alerting or automated analysis option exist?
- Only for data changes, or also for data extractions (SELECTs)?
- Can data be handled differently depending on the workflow (configurability of the A&C system)?

4. Can changes to data / system / model be tracked?

5. What log data / A&C data is provided?

- How to access the data?
- For Off-Premise:

  a. Who can access the data?
  b. Where is it stored?
  c. Is the data encrypted?

- Is a connection to typical evaluation tools (e.g. Splunk) available?

Privacy

Reason	Applies to the assessment of the processing of the data, if sensitive data is used		
Indicators	Applicability	Methods	Error-proneness

1. Privacy Aware Machine Learning:

- Is the data anonymized?
- How does de-anonymization (trick question) work?
- Are the data pseudonymized (ATTENTION: dissimilar to anonymization, specifically legal)?
- Are approaches to Federated Learning applied? (NOTE: area is currently undergoing a lot of change)

Procurement Guide

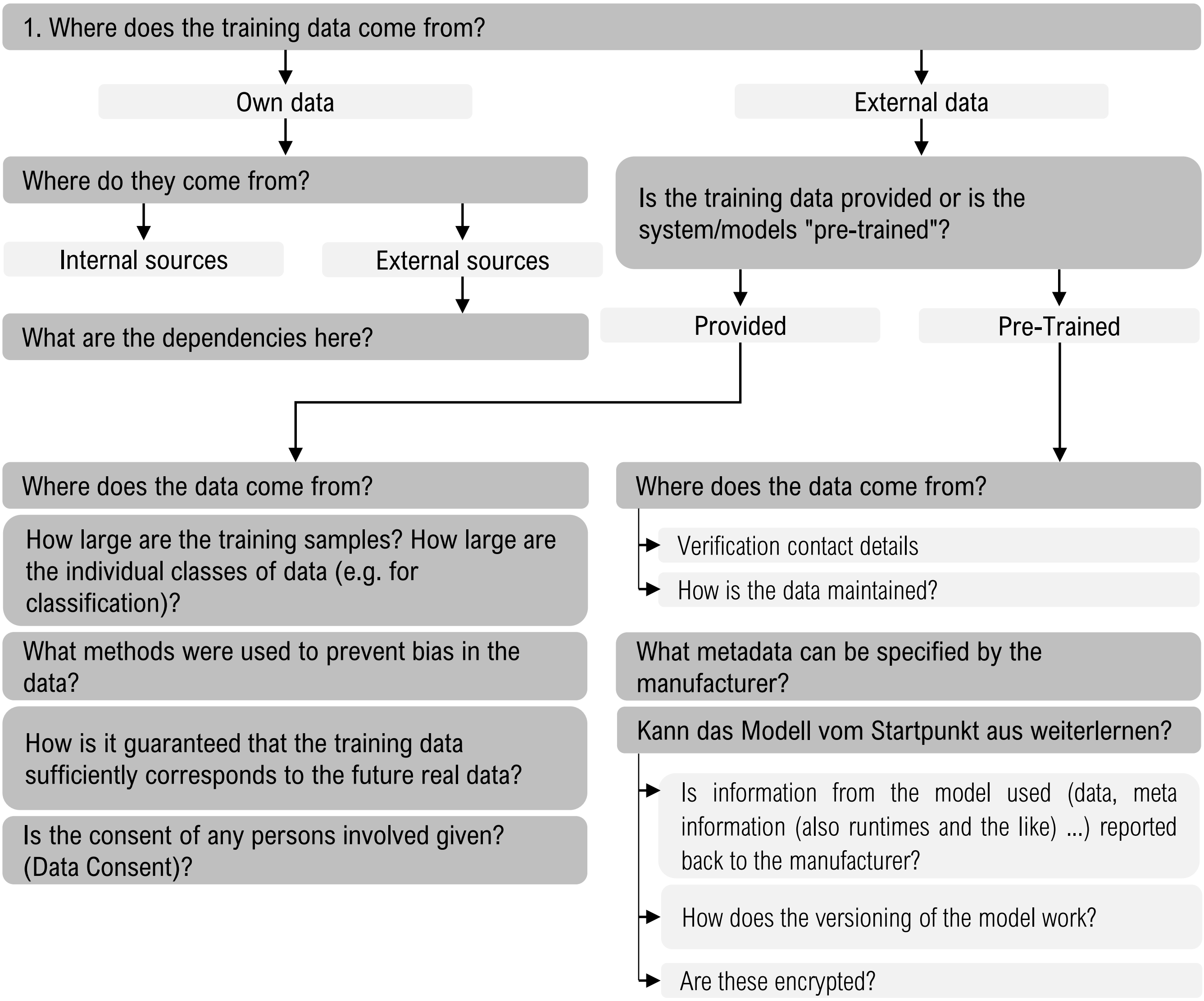
- Detail view VI

2. Are the effects of data anonymization assessable?
   - Data quality of the evaluation? Performance?

3. Which aspects of the GDPR are taken into account?

Training and processing data

Reason	Basis for any AI and origin of various problems (e.g. backdoors)		
Indicators	Data control	Backdoors	Model stability



1. Where does the training data come from?

Own data

Where do they come from?

Internal sources

External sources

What are the dependencies here?

External data

Is the training data provided or is the system/models "pre-trained"?

Provided

Pre-Trained

Where does the data come from?

How large are the training samples? How large are the individual classes of data (e.g. for classification)?

What methods were used to prevent bias in the data?

How is it guaranteed that the training data sufficiently corresponds to the future real data?

Is the consent of any persons involved given? (Data Consent)?

Where does the data come from?

- Verification contact details
- How is the data maintained?

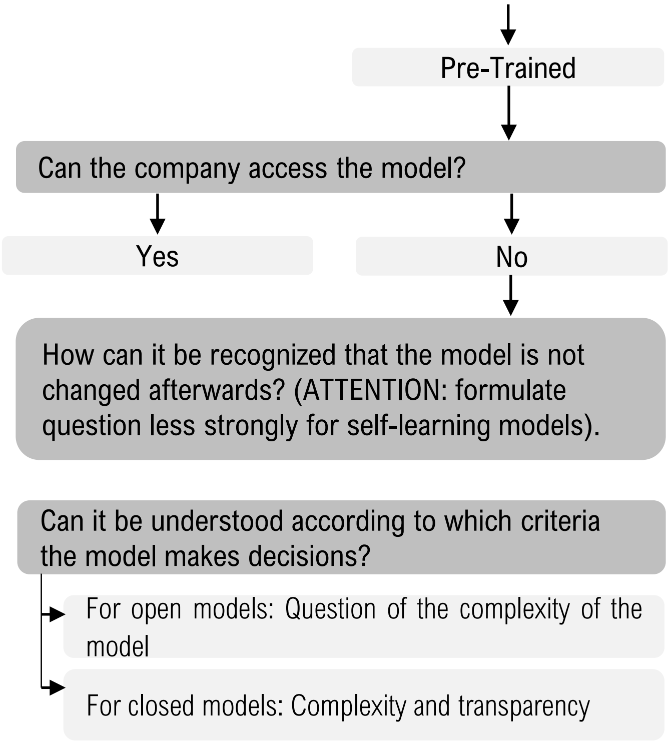
What metadata can be specified by the manufacturer?

Kann das Modell vom Startpunkt aus weiterlernen?

- Is information from the model used (data, meta information (also runtimes and the like) ...) reported back to the manufacturer?
- How does the versioning of the model work?
- Are these encrypted?

Procurement Guide

- Detail view VII



Pre-Trained

Can the company access the model?

Yes

No

How can it be recognized that the model is not changed afterwards? (ATTENTION: formulate question less strongly for self-learning models).

Can it be understood according to which criteria the model makes decisions?

For open models: Question of the complexity of the model

For closed models: Complexity and transparency

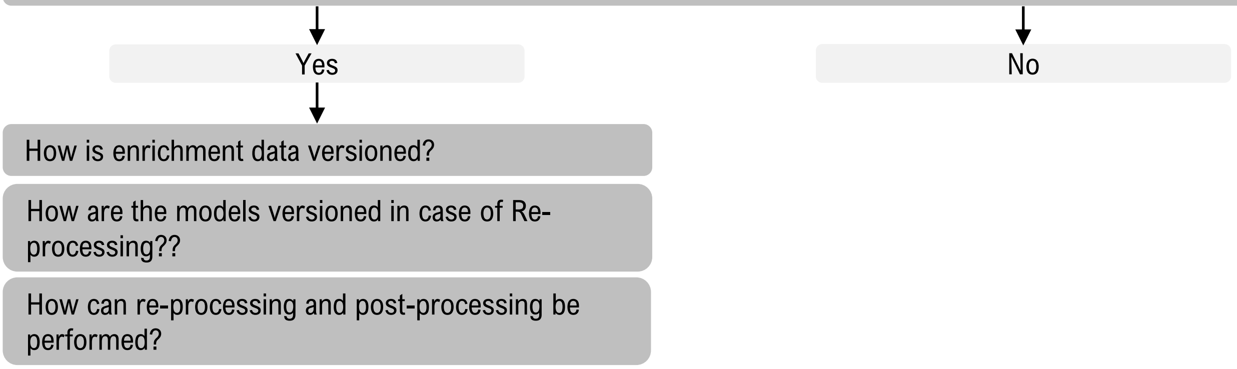
2. How strong is the sensitivity to data quality (stability of algorithms)?

Bias			
Reason	Avoidance of AI prejudices		
Indicators	Awareness	Counterstrategies	Relevance

1. What measures are taken against bias in the data?

2. How were the risks assessed in the case of the specific application?

3. What measures have been taken? Do "re-processing" and "post-processing" capabilities exist?



Yes

No

How is enrichment data versioned?

How are the models versioned in case of Re-processing??

How can re-processing and post-processing be performed?