Auditing machine learning algorithms

A white paper for public auditors

by the Supreme Audit Institutions of Finland, Germany, the Netherlands, Norway and the UK

14 October 2020

Contents

Abstract

This paper discusses audits of machine learning (ML) algorithms by Supreme Audit Institutions (SAIs). The paper aims to help SAIs and individual auditors to perform audits on ML algorithms that have been applied by government agencies. It is designed for auditors with some knowledge of quantitative methods. Expert level knowledge of ML-models is not assumed.

We include an audit catalogue - a set of guidelines including suggested audit topics based on risks, as well as methodology to perform audit tests. The paper is accompanied by an Excel helper tool that sums up and guides through different parts of the audit.

Chapter 1

Executive summary

In 2017, the Supreme Audit Institutions (SAIs) of Brazil, Finland, Germany, the Netherlands, Norway and the UK entered into a ‘Memorandum of Understanding Data Analytics’ (MoU). Recognising the fact that digitisation and datafication changes the way governments work and hence requires SAIs to develop new methodologies and practices to ensure effective and efficient audits, they agreed to cooperate on the topic of data analytics by sharing knowledge, working experiences and code. In their annual conference in 2019, hosted by the Finnish SAI, the members agreed to jointly develop this paper for audits of artificial intelligence (AI) applications.

AI systems based on machine learning (ML) models are increasingly used in the public sector to improve services and reduce costs. ML is a field of computer science dealing with methods to develop (‘learn’) rules from input data to achieve a given goal. An ML model is the resulting set of rules, which may be used to make predictions on data previously unknown to the model. However, new technologies tend to be accompanied by new risks. While dedicated legislation is still underway, both international and European guidelines have been proposed (references are indicated within square brackets in the text) that emphasise the need for control mechanisms and audits. The AI community is addressing issues linked to ethical principles and the negative social impact of AI applications. A recent publication on trustworthy AI development, co-authored by a broad collaboration of researchers from both academia and industry, recommends conducting and funding third-party audits of AI systems.[1]

While data protection authorities are working on dedicated guidelines and can take on a supervisory role for personal data protection, many of the risks linked to AI applications are not related to personal data. For example, opaque ML models that might automate and institutionalise unequal treatment could damage trust in our auditee organisations and, by extension, in our democratic

[1] See [9]

institutions. Therefore it will become increasingly necessary that SAIs are able to audit applications based on ML algorithms in both compliance and performance audits. Several MoU member SAIs are currently performing case studies or pilot audits to develop a generic methodology for audits of AI applications.

This paper summarises what we believe are key risks connected to the use of ML in public services, and suggests an audit catalogue[2] that includes methodological approaches for audits of AI applications. These suggestions are based on contributions from the MoU member SAIs that stem from their respective experiences with ML audits and audits of other software development projects. The SAIs of Germany, the Netherlands, Norway and the UK are the main authors of this paper.

Depending on the application, audits of ML algorithms are usually performed as special cases of performance or compliance audits. ML models tend to be embedded in wider IT infrastructure. Therefore, elements from IT audit are often included. The proposed audit areas include the data understanding and model development process, the performance of the model and ethical considerations such as explainability and fairness. This paper is based on the broadly used 'cross-industry standard process for data mining' (CRISP-DM - see Chapter 3) that includes all phases of an AI application's lifecycle - from business understanding to deployment and continued operation.

The main chapters of this paper focus on auditing the ML component. Appendix One discusses how to assess other steps of an AI application's lifecycle and provides tips on how SAIs could create well-balanced audit teams to enhance the efficiency of their audit work on AI applications. We also include a helper tool that auditors can use to prepare their audits. It provides a host of suitable audit questions that auditors may draw upon. Auditors can select steps along the CRISP-DM cycle based on their risk assessment and get suggestions for suitable audit evidence and contacts within the auditee organisation.

We identified the following main general problem areas and risks:

- Developers of ML algorithms will often focus on optimising specific numeric performance metrics. As a result, there is a high risk that requirements of compliance, transparency and fairness are neglected.
- Product owners within the auditee organisation might not communicate their requirements well to ML developers, leading to ML algorithms that could, in a worst case scenario, increase costs and make routine tasks more time-consuming.
- Auditee organisations often lack the resources and competence to develop ML applications internally and thus rely on consultants or procure ready-made solutions from commercial businesses. This increases the risk of

[2] By audit catalogue we mean a set of guidelines including both the suggested content of audit topics based on risks, as well as methodology to perform respective audit tests.

using ML without the understanding necessary both for ML-based production/maintenance and compliance requirements.

- There is significant uncertainty among public-sector entities in the MoU member states about the use of personal data in ML models. While the data protection agencies have begun to issue guidelines, organisational regulatory structures are not necessarily in place and accountability tends to be unclarified.

Auditors need specific training in the following areas of expertise in order to perform meaningful assessments of AI applications and to give appropriate recommendations:

- Auditors need a good understanding of the high-level principles of ML algorithms and up-to-date knowledge of the rapid technical developments in this field - this is sufficient to perform a baseline audit by reviewing the respective documentation of an ML-system.
- For a thorough audit that includes substantial tests, auditors need to understand common coding languages and model implementations, and be able to use appropriate software tools.
- ML-related IT infrastructure often includes cloud-based solutions due to the high demand on computing power. Therefore, auditors need a basic understanding of cloud services for this kind of audit work.

This paper reaches the following conclusions and recommendations for SAIs:

- SAIs should be able to audit ML-based AI applications in order to fulfil their statutory mission and to assess whether use of ML contributes to efficient and effective public services, in compliance with relevant rules and regulations.
- ML audits require special auditor knowledge and skills, and SAIs should build up the competence of their auditors.
- The ML audit catalogue and helper tool proposed in this paper have been tested in our case studies and may be used as templates. They are living documents and thus should be refined by application to more cases and to more diverse cases, and consistently updated with new AI research results.
- SAIs should build up their capacities to perform more ML audit work.
- The authors hope that the guidance and good practices provided within this paper, alongside the audit helper tool, will enable the international audit community to begin auditing ML.

Chapter 2

Introduction

Artificial intelligence systems based on *machine learning (ML)* models are currently under rapid development, with many successful applications - so far predominantly in the private sector. Public sector entities are beginning to develop and implement *ML* algorithms in the provision of public services. The goal is a more efficient public administration with improved, possibly personalised services at lower costs.

Development and implementation of *ML* algorithms leads to new challenges, including: the use of personal data versus privacy rights; inexplicable and therefore unjustifiable decisions; or potentially institutionalised discrimination by algorithmic bias. If an algorithm is not properly tailored to the objective and its environment, it can result in higher workload, delays and frustrated staff. Usage of a carelessly developed algorithm in public services can thus lead to obscured inefficiency, damaged trust in the authorities and be detrimental to a well-functioning public sector. Both internal control mechanisms and external audits are needed to ensure the proper use of ML and prevent these dangerous side effects.

The first principles and guidelines to address *AI*-related risks have been developed both internationally and on a national level in several countries, and are likely to result in relevant legislation in the near future.[1] Independent third-party auditing is not only recommended in the context of the EU's General Data Protection Regulation (GDPR) and following interpretations, [2] but for all AI systems affecting fundamental rights; the EU's Ethics Guidelines for Trustworthy Artificial Intelligence [4] points out the need for the system to be lawful,

[1] Several countries are in the process of producing standards for AI similar to the EU's guidelines [2], the European Commission has announced a legislative proposal [15], and the Organisation for Economic Co-operation and Development has already launched recommendations that were accepted by G20 ministers as guiding principles for trustworthy AI [18]. See further [8] for an overview of existing AI laws and policies.

[2] For example, see the EU's *Guidelines on Automated individual decision-making and Profiling* [23]

ethical and robust. It further lists accountability, including auditability, as one requirement for trustworthy AI, and explicitly states the necessity for independent internal and external audits. The topics of fairness, transparency and accountability of AI are extensively discussed in the global research community. [3] Although it is not obvious how to facilitate the auditability of ML algorithms, the necessity is widely acknowledged.[4] The idea of specialised, licensed AI system auditors has been put forward [9].

This paper outlines potential audits of *AI* systems by Supreme Audit Institutions (SAIs), covering risks related to the use of *ML* models in government agencies as well as possible tests to gain audit evidence. It further includes an *auditablity checklist* which summarises the minimum prerequisites an auditee organisation should retain from the ML implementation phase to enable any subsequent audit.

An audit of *ML* algorithms can have components of both performance audit and *compliance audit.* It should always include a risk assessment of the related IT system, potentially leading to a wider IT audit that includes the *AI* system. ML algorithms are typically not used as stand-alone software, but rather they are embedded in a pipeline of procedures as part of a wider IT infrastructure. The focus in this paper lies in the audit of the ML component.

The suggested audit model is based on the literature given in the bibliography, as well as on the experiences of the authoring SAIs with audits of IT systems in general and, in particular pilot audits of *ML* applications. It is thus focused on the most commonly used *AI* systems and those encountered in the pilot audits, and should eventually be updated with more audit experience and the results of new research where appropriate.

Chapter 3 is structured into five sections aligning with five audit topics. It suggests an 'audit catalogue' that specifies the relevant considerations with audit questions and risks for each point. A detailed list of practical audit tests and suggested contacts within the auditee organisations is given at the end of each section.

- Appendix One *Classic IT audit components in ML/AI context* summarises IT audit components relevant to *ML.*
- Appendix One *Personal data and GDPR in the context of ML and AI* includes risks related to personal data and violation of GDPR.
- Appendix One *Equality and Fairness measures in classification models* gives an overview of the most relevant equality and fairness measures for classification applications.

[3] For example see the *ACM FAccT* conference series on fairness, accountability and transparency [3]

[4] For example researchers at Google have proposed a framework for internal algorithmic auditing [7] composed of five stages that aim to mitigate related risks before the deployment of the ML system.

- An auditability checklist of minimum requirements for an auditee using *ML* can be found in Appendix One *Auditablity checklist*.

The ML audit helper tool (in Excel) is available as a seperate file that accompanies this paper.

Chapter 3

ML audit catalogue

While AI applications can cause issues of a different nature compared to other software or IT systems, the audit of ML models can be structured according to the cross-industry standard process for data mining (CRISP-DM)[11] as it reflects a standard development process of ML models (even if not explicitly employed by the ML developers). IT auditors familiar with CRISP-DM can thus perform a high-level review without expert knowledge of ML and assess whether ML experts need to be consulted for further tests by following these seven phases:

i) Business understanding

ii) Data understanding

iii) Data preparation

iv) Modelling and development

v) Evaluation of the model before deployment

vi) Deployment and the accompanying change management processes

vii) Operation of the model and performance in production.

While each of these phases are important (especially the business understanding and data understanding of the auditee organisation), phases i) and vi) may, by and large, be audited in the same way as software development projects, and are therefore combined here into Section *3.1 Project management & governance.* Phases ii) and iii) are combined into Section *3.2 Data* [1].

[1] Keep in consideration that without solid business and data understanding by the auditee organisation, ML systems are set up for failure - that is, they may not improve or reach a business goal or the data may be misunderstood in the context of the business goal.

This paper focuses on issues that are particular to AI applications, with a special consideration of values such as transparency of an ML algorithm's decisions, the equality and fairness of these decisions, and autonomy as well as accountability. While these aspects should be considered in every step of the ML development process,[2] the suggested audit catalogue places those considerations in the evaluation step. *Section 3.5* is thus not considered as one step in a linear process, but rather in a loop accompanying the other steps, notably after the deployment of the application. ML models tend to be frequently (sometimes regularly or even continuously) re-trained as more data becomes available. Consequently, this section is described last (unlike the standard CRISP-DM evaluation step performed before deployment).

ML audits may be performed in variable depths, requiring different levels of technical expertise from auditors, and different levels of access to the underlying technical components:

(1) The audit baseline consists of reviewing the documentation and ensuring that all key components are addressed, relevant risks are identified and mitigation strategies are in place.

(2) Close inspection of the data and a review of the code can give higher confidence in the accuracy of the documentation, in particular with regard to the specifications of the ML model.

(3) Reproduction of (parts of) the model training, testing, scoring and performance measures might be necessary to understand and verify details of the model, its performance and reproducibility, as well as its fairness implications. This might include tests of the model's behaviour with manipulated data. This requires infrastructure that is suitable for conducting such verification. The costs and potential benefits of this approach have to be taken into account when considering it.

(4) Development of suitable alternatives to the model can be advantageous to highlight deficiencies and how they can be prevented. As with (3), careful consideration is required before deciding to develop alternative models, as doing so may lead to responsibilities that are not in accordance with the SAIs statutory tasks. Furthermore, the costs and benefits of this approach should be considered.

Auditors should assume that the data, or extracts of the data, are available in addition to the documentation. The code and the model itself might however be in a format that is not accessible to auditors.[3]. In that case, close cooperation

[2] Compare the concepts of fairness by design, transparency by design and privacy by design.

[3] While many ML models are developed in *python* or *R*, there are many other software writing options. Thus, it is unreasonable to request that auditors be familiar with every coding language and software framework. Costs related to the acquisition of additional software and/or computing power might create additional problems, but these could be circumvented by the auditors working on the auditee organisation's premises.

with the auditee organisation is necessary to perform tests and/or demonstrations on their premises, supervised by auditors with a sound understanding of the topic.

Auditors who would like to perform all levels of review should have a deep technical understanding of the subject matter in order to give recommendations for improvements. Where the auditee organisation lacks expertise, any shortcomings should be recognised and recommendations offered. Depending on a risk assessment prior to the audit, it may suffice for the auditors to review the respective documentation and/or focus only on selected phases of the CRISP-DM cycle.

In case of proprietary models where neither the source code nor the detailed model specifications are known to (or verifiable by) the auditee organisation, it remains their responsibility to provide extensive documentation that proves compliance to the proper standards of public administration processes.

3.1 Project management & governance

We have already discussed the notion that in undertaking an audit of an ML project auditors would start with a baseline review of project documentation. Regardless of the depth of audit being undertaken, a review of documents and understanding the context surrounding the project is at least as important as undertaking more detailed investigative work into the model itself.

There are issues that are common to the review of other (software development) projects and with any projects where a statistical model is used to support decision making more broadly. The following considerations are loosely based on the UK National Audit Office's framework for reviewing models [20], extended to meet the challenges of AI applications.

3.1.1 Misalignment/ diversion from project objectives

Low technical understanding within management and limited expertise with practical issues in the technical staff can lead to miscommunication and wrong expectations. Auditors should look for indications that the model is tailored to the project's objective(s):

- Has the model been developed in collaboration with the project owner? For example:
  - Are requirements captured and documented into a specification?
  - Is consideration made for the relative importance of different types of error (false positives/ false negatives)?
  - Are assumptions listed and agreed?

  - Is there an agreed quality assurance plan throughout the model development process?
  - Is there evidence that the model's project owner has influenced the development of the model to meet expectations?
  - Is the level of transparency required well defined in planning? There is active research in explainable ML, so while we may find new techniques to 'explain' models previously deemed as *black boxes*, the issue we should identify is to what extent the model that was delivered meets the requirements of the project.

- Is there a forum within the auditee organisation for people with relevant technical expertise, outside of the development process, to challenge the development and use of model outputs? This means both an independent internal control unit, and a forum with clear accountability for complaints from external users or data subjects.

Risks:

- Project fails to deliver on stated objectives.
- Misunderstanding between project owners and developers resulting in wasted or poorly focused effort.
- Project meets documented requirements, but stakeholders are dissatisfied and, in practice, their desired outcomes are not realised.

3.1.2 Lack of business readiness/ inability to support the project

Knowledge about a particular model is often concentrated in few staff members with high ML competence. Miscommunication between these ML developers and either the users of the model (such as case workers) or the IT staff responsible for maintenance in production, can lead to inefficient implementation of the ML algorithm and failure of the project. To mitigate against these risks, auditors should consider:

- Are roles and responsibilities documented?
- Training of end-users: Has the potentially probabilistic nature of model predictions been properly explained to end users (such as case workers)? Has a policy or guidance for the interaction between AI system and human workers communicated, such as the authority and accountability regime to arbitrate in case of disagreement between human end-users and decisions or recommendations of the AI system?
- What processes are in place for succession planning/handover when a key person leaves the project? Similarly, what processes are in place for the handover from the development project to operation and maintenance in production?

Risks:

- Transition from development project into the business as usual process is dysfunctional.
- Inability to support project on an ongoing basis.
- End users are unable to understand/challenge model outcomes leading to non-transparent or unfair decisions.

3.1.3 Legal and ethical issues

There are additional laws and regulations applicable to ML algorithms when considered alongside standard IT systems. Possible issues strongly depend on the type of model and application context.

- What laws and regulations have to be considered? This includes
  - Normal operation. For example what type and level of transparency/explainability is needed: is a global understanding of tendencies enough (global explainability), or do single decisions need to be justifiable to the extent that advice can be given about how citizens can change the outcome (such as to get approval for support)? Are data subjects informed about the processing of their data by an AI system, and/or an automated process (if that is the case)?
  - Possible side effects of a perfectly well-operating system: For example reinforcement of existing structures, under- or overexposure. Details depend on the type of AI application. for example a recommender system used to suggest relevant job advertisements to unemployed citizens might concentrate on certain career paths, missing out on the potential for non-standard retraining.
  - Possible side effects due to model imperfections: For example, a biased model that discriminates on protected characteristics.

3.1.4 Inappropriate use of ML

Another risk not unique to ML projects but notable due to the current 'hype' around such technology is the risk that some auditee organisations might apply ML techniques not because they add value but instead due to a desire to be seen to be using cutting edge technology. While we are positive about the potential of ML to add value in a broad spectrum of applications, we must also be clear in our assessment of these applications when there is a risk of negative outcomes to the general public. Instances of this risk should be identifiable from an understanding of the project objectives.

In evaluating this, auditors should ensure:

- the component of the solution that ML is applied to is clearly identifiable, justifiable and separable from the other surrounding business logic. This avoids the tendency for optimistic project planning to treat ML as a 'blackbox' that can solve any and all business problems;
- in planning the project, the problem statement is well defined and gives experts scope to experiment. More specifically avoiding statements like 'We will use deep learning to do X' instead, focus on the outcome of the project, and how its success will be measured.
- there is clear evidence that management has identified that their chosen ML model is a necessary improvement over current methodologies.

Risks:

- Project objectives are not realised.
- Overly expensive or complicated solutions to otherwise simple problems.

3.1.5 Transparency, explainability and fairness

Public administration has to be transparent in the sense that the decision making process should be justifiable and to some extent understandable by the general public. Further, citizens usually have the right to explanations of decisions that impact them. The concepts of 'transparency by design' and 'fairness by design' incorporate considerations along these lines in every step of the development of the AI system (and rightfully so); the audit of these aspects is explored in Section 3.5 Evaluation. *Section 3.5 Evaluation*

3.1.6 Privacy

If personal data or proxy variables for personal attributes are used, the EU's General Data Protection Regulation (GDPR) and/or national privacy laws apply and auditors should consider if the ML application is the least intrusive option to satisfy the objective[4] and whether all features related to personal data contribute enough to performance to justify their use.

Additionally, it might be necessary, depending on the ML application, to consider the disclosure risk - this can occur when personal data has been used to train the model, this personal information is encoded in the model and it can be possible to reconstruct parts of the dataset.[5]

Risks:

- Violation of data protection regulations

[4] See Appendix One *Personal data and GDPR in the context of ML and AI* for a summary of relevant GDPR rules.

[5] For example in the context of diagnosis codes or crime convictions, reconstructing which person was part of the training dataset can already be revealing personal information.

3.1.7 Autonomy and accountability

Decisions with legal or similarly significant effects made by ML using personal data are not allowed to be fully automated - citizens have the right to human involvement (with some exceptions: see article 22 of GDPR). Hence the auditor must evaluate the method of human involvement and ensure the system includes the ability to execute this right, including sufficient information being communicated to the affected person.

In ML-assisted decisions, where a human is responsible for the decision but uses ML as one (possibly the main) source of information, the discretion of that person should be evaluated, examining their ability to decide against the algorithm's advice. Additionally, auditors should examine the possible consequences if that decision turns out to be wrong.

In particular, it must be clarified which real person or legal entity bears responsibility for AI-autonomous or AI-assisted decisions. Two separate questions need to be answered in this context [25]: (1) Who is responsible for harm caused by the ML algorithm performing as expected? (2) Who is responsible in the case of failure?

This can become particularly challenging if a third party has developed the ML system.

Risks:

- Automated processing of personal data without the knowledge or consent of the affected persons
- The condition of a 'human in the loop' is not realised, or only formally
- Unclear roles and responsibilities

3.1.8 Risk assessment: Project management and governance

Aspect	**Roles**												Tool
Aspect	**Product owner**	**User**	**Helpdesk**	**Chief information officer**	**Project leader**	**Data analyst**	**Data engineer**	**Developer**	**Controller**	**IT security**	**Data protection official**	**Budget holder**	**Helper tool reference**
Overall responsibility: Governance	x			x	x								A1, A7
Communication with data subjects (where applicable)				x							x		A7.007, A7.009
Policy for human-AI interactions	x				x								A7, A6.004, A6.009
Quality assurance plan	x				x			x					A7.003
Strategy for development/maintenance				x			x	x		x			A6, A7

Table 3.1: Aspects and contact persons: Project management and governance

3.2 Data

Data quality is central to the performance of ML models. Like with any other quantitative modelling, when data is not representative, outcomes tend to be biased. The model may perform well for those characteristics that are represented well in the data, but it may underperform for those characteristics that are underrepresented (for example, a facial recognition algorithm trained on data with pictures of a certain ethnicity will perform well for phenotypes sufficiently represented and badly for phenotypes that are underrepresented). Hence, scrutinising data quality for issues that cause problems in regular (statistical) modeling remains important. Examples of these central issues for data-analysis are: data reliability, population representativeness of the data and disclosure of personal data.

However, there are new issues concerning data that are specific to ML modelling. One of the more well known issues is the lack of separation between training and testing/validation data. When a part of the data is used both for shaping the model (during the training phase) and verifying the performance of the model (during the testing or validation phase), the performance metrics of the model will be inflated. This is called 'overfitting' and leads to a loss of performance of the model on new data, such as in production. ML models can also underperform when the historical data used to train/retrain the model is no longer representative. When data is collected and used for training/retraining the model can also be 'poisoned'. Poisoning occurs when an agent outside of the developers of the ML model introduces corrupted or manipulated data into the training set. Several ways to achieve this come to mind when one considers ML models that use user input for retraining. Intended effects can be to introduce bias by systematically influencing training data, or to let the ML model underperform by introducing 'noise' into the training data.

A final important new data issue that comes with the use of ML models is *leakage*, or target leakage. This occurs when the training data contains information that is not available to the ML model during prediction. This information is usually contained in the variables that are used in the model, we will addres these in that section.

Risks:

- Insufficient separation between training and testing/validation data
- Mis-treatment of personal data (for example purpose limitation violation, lack of control over access and timely deletion, lack of transparency)
- Biased, unreliable or not representative training data
- Poisoning - adversarial introduction of bad-quality data
- Target leakage

3.2.1 Personal data and GDPR in the context of ML and AI

In most countries, special rules apply to the use of personal data. While the definition of ‘personal data’ and the accompanying laws are country-specific, this white refers to the EU’s GDPR [22] as it applies in many countries either directly (in EU member states), via a European Economic Area-agreement, or due to processing of EU citizen’ data.

National data protection authorities are working on GDPR interpretations and guidance for practitioners, and the Norwegian data protection authority (Datatilsynet) has summarised the most important challenges around the use of personal data in ML algorithms in a dedicated report [1]. Relevant considerations are summarised in Appendix *Personal data and GDPR in the context of ML and AI.* The main risks identified are related to purpose limitation, data minimisation, proportionality and transparency.

The responsibility to guarantee compliance with data protection laws lies with the authority that develops and uses the ML algorithm. For auditors, a review of the respective documentation should suffice (in particular the data protection impact assessment, where appropriate). On suspicion of violation of data protection laws, the case could possibly be forwarded to data protection authorities.

Auditors should pay attention to data used in different development steps. Their considerations should also include data that is not used in the final model but that was nonetheless considered and tested during the model development phase. For example, a feature importance test could also indicate that personal data was used.

3.2.2 Risk assessment: Data

Aspect	Roles												Tool
Aspect	**Product owner**	**User**	**Helpdesk**	**Chief Information Officer**	**Project leader**	**Data analyst**	**Data engineer**	**Developer**	**Controller**	**IT security**	**Data protection oficial**	**Budget holder**	**Helper tool reference**
Overall responsibility: Data	x				x		x	x					A2, A3
Data acquisition method					x	x	x						A2
Group representation and potential bias (raw data)	x				x	x		x					A2
Data quality (raw data)							x						A2.004
Database structure							x						A3
List of variables used						x		x					A3
Personal data and data protection					x			x			x		A2.008

Table 3.2: Aspects and contact persons: Data

3.2.3 Possible audit tests: Data

- Verify that test and validation data are used and stored separately.
- Define important population characteristics and test whether they are sufficiently represented in the data
- Verify that stored training data is not outdated.
- Verify compliance with GDPR.

3.3 Model development

ML developers often focus on optimising a certain performance metric. Defining the best-suited metric might be challenging in itself due to unavoidable trade-offs that include value judgements such as weighting consequences of false positives versus false negatives. There are, however, many more aspects of model development that ML applications at first impression have in common with other IT development projects, but incorporate additional challenges when it comes to ML models.

3.3.1 Development process and performance

An important aspect of the development process, and consequently of the final model, is reproducibility. Reproducibility can be tested by a simple review of the documentation, if that is deemed sufficient by the auditors. If any information on reproducibility is lacking, unclear or contradictory, and the process is not fully automated, reproducing the model and/or its predictions may be required. This can be a time-consuming exercise as it might include an iterative process with updated tests based on new information obtained by additional dialogue with the auditee organisation.

A well-structured and commented codebase (according to standards of the chosen coding language) and extensive documentation of all hardware and software used, including versions (also of, for example, R/python modules and libraries), is not only a prerequisite for reproducibility, but also for long-term reliability, maintenance and possible succession or handover to new staff.

The effect of variables that are used as ‘features’ of the model can be investigated, both in number as well as nature, to avoid unnecessarily complicated models that are prone to overfitting. Most features can be interpreted as private data depending on their use and context, and some can be proxies for protected variables, hence careful attention is necessary to ensure compliance with data protection regulations (for examples, see Possible audit tests (above)).

The type of ML algorithm that is chosen should be well motivated. If a hard-to-explain *black box* model is used, it should be documented that this is justified by a significantly better performance compared to a *white box* model. If auditors doubt that this is the case, they could train a simple *white box* model to test any (or the lack of) performance differences[6].

Given that an appropriate performance metric is chosen that fits the objective of the application (see Section *3.1*), it can be assumed that the model's performance is well documented. Auditors should nevertheless verify that the reported performance is accurate and consistent, in particular if the granularity of input data to the model is not identical to that of the reported performance (for example the model operates on applications/issues/periods while the reported performance is defined on customers).

A comparison of the performance on training data versus test data is a standard procedure to test how well the model generalises to new data. If cross-validation was used during training, or trustworthy independent test data is not available, auditors might decide to test the performance on synthetic data[7]. If the performance on production data is very different from the test performance, the reason might be an overfitted model, or substantial differences in the production data compared to the training/test data. The latter might occur when the use of a well-performing model is extended to areas it was not originally designed for.

Risks

- Irreproducible and/or incomprehensible predictions
- Dependencies on unspecified hardware, installation or environment variables (for example default model parameters dependent on the use of GPU vs CPU, undocumented software version)
- Use of unnecessary data (correlated or unpredictive variables); if personal data is used unnecessarily, additional violation of GDPR may occur (data minimisation principle)
- Overfitting (model does not generalise well to new data) or model bias/underfitting (oversimplified; model does not describe the data well), inappropriate metric not targeting the application objective

- Unnecessarily complicated model used because of convenience from earlier use, or personal preference rather than performance (for example, *black box* model that is not significantly better than a *white box* model)
- Model optimised for inappropriate performance metric. For example, use of the inappropriate metric because of personal preference or convenience because of earlier implementation rather than the application objective
- Code can only be executed or understood by a single person or a small group of people

[6] Note that while *black box* models often have the potential to outperform *white box* models, this potential might not be realised if the *black box* model is not sufficiently optimised. If its performance has further been judged to be sufficient and a *white box* model can achieve similar performance, the simpler model should be used to comply with explainability requirements.

[7] For example data produced with the synthpop package in R or similar. Synthetic data is data that is produced artificially but with the same features as real data.

3.3.2 Cost-benefit analysis

The performance of the ML model in production should be compared to the previously used system (that is the respective performance metric without ML or with an older model). If the main objective of the ML application is to reduce costs, the savings with ML can be compared to the development plus maintenance costs (presumably staff or consultancy costs). Otherwise, depending on the application objective, classical cost-efficiency, cost-utility or cost-benefit analyses can be appropriate.

Risks:

- ML used for the sake of using ML, without improving the service
- Inefficient spending and inappropriate use of the auditee organisation's budget

3.3.3 Reliability

Reproducibility, which is a mandatory condition for reliability, is already discussed above but there the focus was on the ML component alone. Dependencies on other parts of the pipeline can influence the reliability of the model's performance, in particular if the model runs automatically in real-time mode. Auditors should consider all possible variations in the input to the model (intentional and unintentional) to assess the behaviour of the model under these circumstances. Another aspect of reliability in the long term is the in-house competence in the auditee organisation to maintain the model (in particular, in performance monitoring, retraining and, where appropriate, re-optimisation).

Risks:

- Performance expected from development not reached in production, or degrading over time
- Untrustworthy prediction if unintended input data is given to the algorithm
- High maintenance costs due to lack of in-house competence

3.3.4 Quality assurance.

Quality assurance in the context of ML algorithms can be viewed as three-fold, with separate implementations for data quality, code quality and model quality. Data quality and code quality are not specific to audits of algorithms and we therefore do not discuss them further here. However, model quality in this context refers to the model's performance under different circumstances and over time.

- Data quality: How is data quality ensured internally?
- Code quality: A code review by auditors may not be feasible (over what is done for reproducibility), hence a test of the internal code quality assurance is advisable (for example, version control, unit tests, code review).
- Model quality: Are sufficient performance tests in place including
  - tests for overfitting: Does the model generalize well to unseen data?
  - retraining frequency: Is the model frequently retrained to accommodate changes in the data (for example, demographic changes), changes in policies or legislation, orupdated objectives?

  - is the data available for retraining biased by previous model predictions?

Risks:

- Unstable or erroneous results (for example, dependencies on data types that can change)
- Unknown performance
- Degrading performance over time
- Model reinforcement loop (for example, if one is retraining on data selected by the model)

3.3.5 Risk assessment: Model development

In order to assess the various risks explained in the sections above, documentation of the following aspects should be reviewed:

- Coding language, hardware, software versions (including all libraries)
- Data transformations
- Definition of performance metric(s) based on the project objective(s)
- Choice of ML algorithm type (including *black box* versus *white box*)
- *Hyperparameter* optimisation and final values (incl. used default values)
- Performance on different datasets
- Comparison with the previous system
- Privacy considerations (where applicable)
- Safety considerations (where applicable)
- Internal quality assurance

Aspect	Roles												Tool
Aspect	**Product Owner**	**User**	**Helpdesk**	**CIO**	**Project leader**	**Data Analyst**	**Data Engineer**	**Developer**	**Controller**	**IT Security**	**Data Protection Official**	**Budget Guy**	**Helper Tool Reference**
Overall responsibility: Model development					x		x						A4
Hardware and software specifications						x	x	x		x			
Data transformations, choice of 'features'					x	x			x				
Choice of performance metric(s)					x	x							
Optimisation process					x	x		x					A4.011
Choice of ML algorithm type (including black box versus white box)					x	x		x					
Code quality assurance					x			x					
Maintenance plan					x			x					
Model quality assurance	x				x								
Cost-benefit analysis	x				x							x	

Table 3.3: Aspects and contact persons: Model Development

3.3.6 Possible audit tests: Model development

- Reproduction of the model (with given parameters), possibly on a subset of the training data or independent (possibly synthetic) data.
- Reproduction of the model prediction/score with (a) the model and/or (b) a best reproduction of the model.
- If not done internally: test of the feature importance.[8]
- If there is a suspicion that too many/unnecessary features are used: re-train the model with less features to show the significant performance difference (or lack of them).[9]
- If personal data is used as features: re-train the model with less / no personal data to quantify the trade-off between performance and personal data protection (see also suggestions from ICO in [19]).
- If not done internally: test for overfitting, performance train versus test/validation versus production data.
- If a *black box* model is used: train a simple *white box* model and compare the performance (if not done internally).
- Cost-benefit analysis.

If the code and model are not available:

- Reproducibility can only be verified completely in the same software, as the training of the model depends on the implementation in the respective library. Hhowever, if a similarly performing model of the same type with the same parameters can be trained in a different software, it can be used for further tests.
- Correlations between features, and between each feature and the model prediction, can be analysed by external tools[10] using only the data (including predictions).
- Retraining with less or different features can be done by the auditee organisation.

3.4 Model in production

Appropriate monitoring of the model's performance over time depends on the implementation of the model: automatic real-time scoring requires continuous monitoring with automated tests to ensure stable model performance, while

[8] Most ML libraries have built-in methods to test and visualise feature importance. A useful tool for explorative data analysis in R is the DataExplorer library, which can be used to easily test the features by their relation to the model predictions, feature correlations and principal component analysis.

[9] Note that features that do not contribute much to the performance alone can be important in combination with other features (feature combination possibilities depend on the type of model)

[10] E.g. with the R package "DataExplorer"

a manual setup with cyclical retraining or re-optimisation naturally includes performance checks in each cycle. In any case, a mechanism to flag possible changes in the performance over time should be in place.

If the model is retrained or redeveloped based on the outcome of previous predictions, this feedback loop needs to be designed such that no additional bias is introduced.

The performance metric used when optimising the model reflects policy decisions, (for example, in prioritising sensitivity over precision). As such policies can change, the performance metric must be adjusted when updating the model.

Risks:

- Performance degrading over time (for example, due to change in demographics)
- Increased model bias
- Obsolete choices embedded in the model
- Model is ‘repurposed’ over time, and predictions begin to be used out of context.

3.4.1 Risk assessment: Model In production

Aspect	**Roles**												Tool
Aspect	**Product Owner**	**User**	**Helpdesk**	**CIO**	**Project leader**	**Data Analyst**	**Data Engineer**	**Developer**	**Controller**	**IT Security**	**Data Protection Official**	**Budget Guy**	**Helper Tool Reference**
Overall responsibility: Model in production	x				x			x					A6
Data update and monitoring					x	x		x					A6.001
Model re-training					x	x		x					
Automation, system architecture, interface to other systems	x			x									A6.003
Long-term quality assurance	x												A6.006
Performance control in production	x											x	A6.006

Table 3.4: Aspects and contact persons: ModelInProduction

3.4.2 Possible audit tests: Model in production

- Verify that the population for which the model is used in production is (still) sufficiently represented in the training data.
- Obtain the code of the production version of the model.
- Compare performance in production to expectation from development.
- Review monitoring of development of performance and input data distributions.

3.5 Evaluation

The evaluation step includes the choice of a 'best' model and the decision to deploy a model (or not), based on performance indicators that were derived from the project's objective. In addition to the performance directly related to the business objective, such as better services or reduced costs, compliance with regulations as well as possible risks and side effects have to be evaluated before deployment. Even after deployment, any change in the AI system should entail a new evaluation, including passive changes like demographic shifts in the input data.

The performance optimised in all the aspects described in Section *3.3 Model development* is weighted against the aspects described here. The relevance and relative importance of the respective audit parts depend on the type of model and on the application.

3.5.1 Transparency and explainability

Regulations for public administration usually include the requirement for transparent procedures and substantiation of decisions where individuals are concerned. Use of ML in decision-making can render these requirements difficult to fulfil, in particular when a *black box* model is used. However, there are several approaches to understand the behaviour of ML algorithms: [11]

- The **global** relationship between features and model predictions can be tested by visualising correct and false predictions dependent on single features. This approach can serve as a first, coarse understanding, but it neglects correlations and nonlinear behaviour.

  Other standard tools for global explainability readily implemented in various *R* and *python* packages include partial dependence plots (PDP), individual conditional expectation (ICE) and accumulated local effect (ALE). These usually require the model to be available, as they use different averaging procedures. If the model is not available to the auditors, they

[11] For example, see reference [16] for an instructive overview

can either implement the first-mentioned approach themselves, or ask the auditee organisation to provide suitable plots.

- In order to explain a particular model prediction (for example, in context of a user complaint and the necessity to justify a particular decision), **local** explainability is necessary, where 'local explainability' is defined as the need to explain the influence of single features in a specific point of parameter space (that is, for a specific user or case). Most common methods include *LIME* and *Shapley values* and are available in standard libraries in *R* and *python*. Motivation for the chosen method and its applicability should be documented.[12] These methods use local approximations of the model and hence need the model to be available. It is not constructive for auditors to test the model behaviour with these methods. The main objective here is to test that the auditee organisation has implemented respective methods to be able to substantiate ML based decisions.

Risks:

- No understanding of the model's predictions
- No understanding of the effects of the different input variables
- The administrative unit is not able to explain and justify decisions made by or with support of the ML model

3.5.2 Equal treatment and fairness

The training data for ML models may incorporate demographic disparities which are learned by the model, and then reinforced if the model predictions are used for decisions that impact the same demographics. The most common sources for such disparities are the training data and the training procedure. The data is influenced by the measurement procedure and variable definitions (for example, when a model that is supposed to find the best candidate in a recruitment process is trained on data from candidates passing previous recruitment criteria). Since ML models usually improve if more data is available, their performance is often best on the majority group, while they can give significantly worse results for minorities.

Fairness in ML has become an increasingly important topic in the last few years. There is no common standard for ML fairness, instead, many different definitions and metrics apply. The most tangible class of fairness definitions is group based fairness, which requires ML models to treat different groups of people in the same way and thus fits well to requirements of equal treatment in anti-discrimination laws. At the same time, group based fairness is easy to test by looking at the performance of a ML model separately for different groups.

[12] For example, see [6] and [14] for problems with explanations with *LIME* or *Shapley values*, respectively.

For classification models, the relevant metrics are based on the confusion matrix, which shows (in the most simple case of binary classification) how many true/false cases are classified correctly. [13]

It is important for auditors to understand that if the true distribution of classes is not the same between two groups represented in the data, it is impossible to satisfy all fairness criteria[14] (see Appendix One *Equality and Fairness measures in classification models* for details). This mathematical fact is most easily understood when considering fairness focusing on equal *treatment* (procedural fairness, equality of opportunity) versus equal *impact* (minimal inequality of outcome).

Auditors have to assess whether or not a model sufficiently satisfies equality requirements by defining relevant groups, testing which criteria are violated to what extent, and considering the consequences in the respective application of the model.

Risks:

- Reinforcement of inequalities that were picked up from the training data
- Worse performance for minorities
- Unequal treatment based on protected variables, in the worst case discrimination of groups defined by gender, religion, nationality etc.

3.5.3 Security

AI systems naturally face the same security issues related to physical infrastructure as other IT systems. Due to the massive amount of data and computing power needed for the development of ML models, and sometimes also the deployed system, the security of distributed and possibly cloud-based computing infrastructure tends to be relevant. Privacy protection might be particularly challenging if the data is processed or temporarily stored in countries with different regulations.

ML applications can bear new additional security risks, in particular when they run automatically in real-time applications. Poisoning was mentioned in Section *3.2*, as well as disclosure of personal data (in Section *3.1*). Similarly, there might be a disclosure risk for industrial secrets, which in the context of public administration can relate to information about infrastructure or safety procedures that should not be publicly available.

For image recognition models, the risk of adversarial attacks has to be considered. Adversarial attacks are data modifications that are carefully designed to

[13] See Appendix One *Equality and Fairness measures in classification models* for an overview of the most common equality and fairness metrics, and Appendix Two *Model evaluation terms* for an introduction to the confusion matrix.

[14] Except for the unrealistic case of a perfect model with 100% accuracy.

trick the algorithm, for example, small stickers placed on a roadside stop sign so that the image recognition system in a self-driving car falsely identifies it as a speed limit sign[12] [17].

These examples are by no means exhaustive. The variety of ML model types and applications makes it difficult to foresee all potential failures and attack vectors, even for ML developers.[15]

Risks:

- Security risks depend on the application: disclosure of personal data or of other confidential information, poisoning of the model, adversarial attacks

3.5.4 Risk assessment: Evaluation

The above mentioned risks can be assessed by reviewing the documentation of the internal evaluation, including

- internal risk assessments related to relevant security risks;
- a comparison of different models (including a defined weighting of performance, transparency, fairness and safety aspects);
- approaches to explain model behaviour;
- approaches to minimise bias;
- (where applicable) compliance with privacy laws; and
- the model's deployment and retraining strategy documents.

[15] See reference [9] for suggestions of control measures: While 'bug bounties' might be less appropriate in the public sector, an independent internal third party challenging the application ('red team exercise') could be feasible for comparatively large auditee organisations. Smaller organisations might apply organisational incentives for independent individuals to be vigilant about their AI applications, raising issues when necessary.

Aspect	**Roles**												Tool
Aspect	**Product owner**	**User**	**Helpdesk**	**Chief information officer**	**Project leader**	**Data analyst**	**Data engineer**	**Developer**	**Controller**	**IT security**	**Data protection official**	**Budget holder**	**Helper tool reference**
Overall responsibility: Evaluation	x				x								A5, A6
Evaluation method	x				x			x					A5.001
Comparison of different models	x				x			x					A5.002, A5.003
Approaches to explain model behaviour	x	x			x			x	x				A6.013, A6.014
Bias tests	x				x			x	x				A6.021
Compliance to privacy laws (where applicable)	x				x		x		x		x		A6.011
Safety risks assessment and mitigation strategies	x				x		x	x	x	x	(x)		A5.004, A6.015- A6.018
Communication with data subjects (where applicable)	x	x											A7.009
Policy for human-AI interactions	x	x			x								A6.004, A6.008, A6.009
Quality assurance plan	x				x	x	x	x					A6.019, A6.020, A7.004)
Monitoring of the model performance in production	x				x	x		x					A6.006
Strategy for development/maintenance	x				x			x					A6.001

Table 3.5: Aspects and contact persons: Evaluation

3.5.5 Possible audit tests: Evaluation

In the evaluation phase, auditors should pay close attention to the grounds on which the project owner declared their acceptance of the deliverables of the ML development project.

In addition, one should look for the following:

- Test the global model behaviour for each feature, using tools like PDP, ICE or ALE.
- Test applicability and implementation of local explainability methods like LIME or Shapley values.
- Determine which laws and policies apply in the ML application context (protected groups, affirmative action).
- Calculate group-based equality and fairness metrics based on the data, including predictions.[16]
  If auditors suspect the use of proxy variables, independent data sources might be necessary to show the correlations and define relevant groups.
  Suggested metrics:[17]
  - Disparity in: Prevalence, predicted prevalence, precision, false positive rate, true positive rate, negative predictive value.
  - Fairness metrics: Statistical parity (also called demographic parity), equalized odds (also called disparate mistreatment), sufficiency (also called predictive rate parity).
- Test the model predictions with one feature changed (for example, change all men into women while keeping all other features the same). If the model is not available, the auditee organisation can test the model's performance on similarly manipulated data
- Calculate the performance when less/no personal data is used.

[16] For example, use the *python* package 'aequitas'

[17] See Appendix One *Equality and fairness measures in classification models* for details

Chapter 4

Summary and conclusions

Public authorities and government entities have started to develop and put into production ML algorithms in order to possibly improve services and lower costs. This new technology comes with new challenges. SAIs need to be able to assess and audit ML applications, and several countries have started pilot projects.

This paper presents some of the issues and risks of ML applications in the public sector, and suggests suitable audit methods. Structured into five audit areas focusing on governance, data, model development, the model in production, as well as evaluation that includes ethical aspects, the audit catalogue guides auditors through a typical ML development process. The auditing techniques described here were applied in case studies to assess the utility and trustworthiness of ML applications, as well as the efficiency and effectiveness of the implementation and operation of such applications.

ML models are used for a large variety of topics, with different risks applying to different model classes. Therefore, this audit catalogue could be refined and extended by applying it to more diverse and more complex ML models. As the field of ML is still evolving in line with new research developments, this audit catalogue needs to be updated regularly.

This paper is accompanied by an ML audit helper tool that enables auditors to choose from a host of questions and create a tailor-made questionnaire that is suitable for their specific audit. Auditors can select steps of the ML development process, possibly structured along the CRISP-DM cycle, that they would like to study. They are provided with recommendations for suitable questions as well as hints on which interview partners might be suitable to answer these questions and what audit evidence one should expect from the auditee organisation. The authors hope to enable the international audit community to begin auditing ML systems with the guidance and good practices provided within this paper and the audit helper tool.

Appendix One

Classic IT audit components in ML/AI context

While ML systems are an emerging technology, and their broad usage by public entities is just beginning, they share key features with regular software.

Notably, their development cycle is similar, which is why well-established standards, such as the CRISP-DM cycle that was introduced in Chapter 3, may be applied to break down the lifecycle of such an application into several phases that can be audited all at once or selectively.

Thus, auditors may apply the same techniques and audit questions that would be appropriate for regular IT performance and compliance audits for some of these phases. Risks that may occur in regular software development projects are also of relevance to ML algorithm development, however, one has to keep the risks that are unique to ML applications in mind (a selection of these risks is included in Sections *3.1* to *3.5*).

Therefore, it might be suitable to either audit ML systems in teams that are composed of specialist auditors, IT auditors and data scientists, or to focus on one component (data science or classical IT audit) while keeping the other component in mind. Teams that lack the necessary experience in data science might be well-advised to rely on the expertise of their data science colleagues, while teams that lack experience in classical IT audits would be well-advised to delegate certain aspects of the audit to more experienced IT auditors. This well-balanced approach guarantees that no aspects of the system are left unaudited due to a potential lack in knowledge or tools.

The first audit of a system with ML components that was performed by the Bundesrechnungshof, the German SAI, successfully applied the approach detailed above: the audit team was composed of a specialist auditor with good knowledge of the auditee organisation and two technical auditors (one with an IT background, one with a background in natural sciences). The auditors were recruited from their respective audit units to form the ML audit team for this specific audit.

The audit of the following phases might benefit from auditors with an IT or specialist background:

i) Business understanding (see Section *3.1*)

ii) Deployment and change management (see Section *3.1*)

iii) Operation (see Section *3.4*)

However, the following phases of the CRISP-DM cycle might be more suited to an audit by data scientists with a background in ML:

iii) Data preparation (see Section *3.2*)

iv) Modelling (see Section *3.3*)

v) Evaluation (see Section *3.5*)

The 'data understanding' phase (see Section *3.2*), might benefit from a combined approach as it requires a technical understanding of the data as well as an understanding of the business goal that is supposed to be reached by the application of the ML algorithm on the data.

The ML audit helper tool that is included with this paper offers a host of questions that are suitable to all phases of the CRISP-DM cycle, for specialist auditors and IT auditors, as well as for data scientists.

Personal data and GDPR in the context of ML and AI

Datatilsynet, the Norwegian data protection authority, summarised the most important challenges relating to the use of personal data in ML algorithms in a dedicated report [1]. Relevant considerations for the auditor are:

- **Purpose limitation**: Personal data may only be collected for a specific, expressly stated purpose. Any further processing has to be compatible with the original purpose, with some limited exceptions for scientific research. Use of ML in decision-making on new data has to be included in the purpose statement. When an ML algorithm is trained on historical data (possibly collected before the ML project started), further processing of this kind must be covered by the original purpose. In some cases (for example, medical applications), the development of the ML algorithm might be considered to be scientific research.

- **Data minimisation**: The use of personal data has to be limited to what is necessary to fulfil the purpose it was collected for. This is challenging during the development of ML, where data is often used in training to later test its impact on performance. Even if personal data is not used in the final algorithm, this testing procedure already counts as processing of personal data and thus is protected under the EU's General Data Protection Regulation (GDPR). For auditors it is therefore important to review both the variables used in the final algorithm and their importance for the performance, as well as the development process.
- **Proportionality**: The data minimisation principle also restricts the degree of interference with a person's privacy. The amount and nature of the data used has to be proportionate to the purpose and the least invasive for the data subject (for example, facial recognition to measure school attendance is considered out of proportion even when consent has been obtained [10]).
- **Transparency**: Explainable processes and decisions are a general requirement for public services, not limited to the use of personal data, but of even higher importance if personal data is involved. This aspect is treated in detail in the proposed ML audit methodology.
- If the use of ML poses a high risk to a person's rights and freedoms, then a data protection impact assessment (DPIA) is mandatory. A DPIA is also required in the following cases [1, 21]:
  - Profiling (and similar) with significant effect, where profiling is defined as "any automated processing of personal data with the objective to evaluate personal aspects about a natural person, in particular predictions of a natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements" [23]
  - Large scale use of sensitive data.
  - Systematic monitoring of a publicly accessible area on a large scale
- The responsibility to guarantee compliance with GDPR, including all of the points above, lies with the authority developing and using the ML algorithm. A review of the development documentation should suffice for auditors.

Equality and fairness measures in classification models

The performance of classification models is usually evaluated based on the confusion matrix and derived metrics. In a binary classification problem, one class is defined as the positive outcome. For example [5], suppose a model is being built to predict whether a a criminal will reoffend after being released from prison, the positive outcome would be that they do indeed re-offend.

Expanding the criminal reoffending example, the true positives are cases where the model correctly identifies that a former convict will reoffend, and the true positive rate (TPR) is the fraction of cases where our algorithm correctly identifies a reoffender out of all the reoffenders within the population. The false positive rate (FPR) therefore describes the fraction of cases where the algorithm predicts reoffending, but the individual is law-abiding.

Where TPR and FPR evaluate the model prediction for a given actual outcome, another important aspect is the evaluation of the actual outcome against a model's prediction: The positive predicted value (PPV), also known as precision, describes the fraction of positive predictions that are correct, while the negative predictive value (NPV) describes the fraction of correct negative predictions.

In our example, these two metrics could be summarised as:

- TPR: How likely is it that a released prisoner who does reoffend is given a positive prediction (correctly identified),

versus

- PPV: How likely is it that a released prisoner actually reoffends when a positive prediction has been be given.

A common approach to fairness is to demand different groups of people to be treated in the same way according to either of these indicators, with the first indicator leading to measures categorised as disparate treatment/mistreatment, procedural fairness or equality of opportunity, and the second classified as disparate impact, distributive justice or minimal inequality of outcome.

A third approach to group-based fairness is to demand that the fraction of predicted positives (predicted prevalence) is independent of any group affiliation, irrespective of possible differences in the actual fraction of positives (prevalence) in these groups.

In order to calculate measures for equal/unequal treatment of different groups of people by a model, and to assess the extent of such disparities, the following metrics are helpful:

- Prevalence : fraction of actual positives
- Predicted prevalence : fraction of predicted positives
- False positive rate: fraction of false positives in all real negatives
- True positive rate, also called recall : fraction of true positives in all real positives

- Precision, also called positive predicted value: fraction of true positives in all predicted positives
- Negative predictive value: fraction of true negatives in all predicted negatives

Based on these, the following group fairness metrics can be calculated:

- **Statistical parity** (also called demographic parity): The predicted prevalence is the same between groups - that is the probability for positive or negative prediction is equal
- **Equalised odds** (also called disparate mistreatment): Same TPR and same FPR - that is the probability for a positive prediction given a positive or negative truth is equal
- **Sufficiency** (also called predictive rate parity): Same PPV and same NPV - that is the probability of a real positive or negative given a predicted positive or negative is equal

It is important for auditors to understand that in the common case of different prevalence in different groups, no imperfect model can satisfy any two of the three fairness metrics at the same time. It is therefore important to take the prevalence (often called base rates) into account when assessing the seriousness of violations of these fairness principles, as well as the magnitude of the difference and (obviously) the practical implications in the specific ML application.

The group-fairness metrics discussed here have the advantages that they can easily be calculated and that they reflect anti-discrimination legislation. Other fairness concepts auditors should be aware of include:

- **Fairness through unawareness**: The naïve idea that an algorithm cannot discriminate with respect to a personal attribute if that attribute is not given to the model is too simplistic for ML applications used in public services, as it neglects correlations with attibutes contained in the data.

- **Individual fairness**: Focusing on individual cases, this approach demands similar cases to be treated in a similar way by the model. It is much more challenging to calculate a metric for individual fairness compared to group fairness, as the notion of 'similarity' needs to be defined in appropriate distance measures in both the feature space (model input) and the prediction space (model output).

- **Counterfactual fairness**: This approach tries to determine the influence of a personal attribute on the prediction by changing that attribute plus all correlated attributes. It can thus help to analyse the reasons and mechanisms behind a potential bias rather than just to observe and quantify it. It is, however, unclear how to implement this approach, as one needs to make sure all relevant variables and correlations are correctly taken into account, and define a causal graph that relates them.

A good overview of these and more fairness concepts are given in [24].

Auditability checklist

This part of Appendix One summarises the prerequisites an auditee organisation should be able to provide for audits as described in Chapter *3 ML audit catalogue*.

A list of **contact persons** with the knowledge areas and roles (see the *glossary* for definitions) as described in Table 4.1 is helpful, and the respective responsibilities should be defined within the auditee organisation.

Responsibility	**Example role title**
Domain knowledge, relevant performance metrics, practical implications	Requisitioning unit: Project owner / product owner
User of the AI system	Processing official, case worker
User support	Process hotline, helpdesk
IT support, developer support	Chief information officer
Project management	Project leader
Raw data quality and understanding	Data engineer
ML model details	Developer
Internal audit	Controller
IT security	IT security officer
Data protection	Commissioner for data protection and privacy
Budget	Budget holder, Budgetary commissioner

Table 4.1: Roles and responsibilities to be defined by the auditee organisation

Note that several roles may be filled by a team or person. If external consultants are used, a proper handover, including sufficient knowledge (and role) transfer, has to be ensured by the auditee organisation.

The audit areas described in the audit catalogue require **documentation** that covers the aspects summarised in Table 4.2.

In order to enable technical tests, **code** and **raw data** necessary to reproduce the model should be accessible, as well as the **model** itself. Note that where providing copies is not appropriate or feasible, *accessibility* may be fulfilled by available staff able to run/rerun code, transform data and score the model, as requested by the auditors.

Audit area	**Audit question/aspect**
Governance, project management	Roles and responsibilities (defined and communicated)
	Context evaluation: relevant laws and regulations (including required level of transparency), risk assessment (including side efects) and mitigation strategies
	Objectives and measure(s) for success
	Quality assurance plan
	Maintenance, development and succession strategy
	Communication with stakeholders ('customers' such as a ministry, users, data subjects)
	Independent control unit
	Human-AI interaction policy
	Autonomy and accountability
	...
Data	Data acquisition method
	Group representation and potential bias (raw data)
	Data quality (raw data)
	Database structure
	List of variables used
	Personal data and data protection
	...
Model development	Hardware and software specifications
	Data transformations, choice of 'features'
	Choice of performance metric(s)
	Optimisation process
	Expectation for unseen data
	Choice of ML algorithm type (including black box versus white box)
	Code quality assurance
	Maintenance plan
	Model quality assurance
	Cost-benefit analysis
	...
Model in production	Data update and monitoring
	Model re-training
	Automation, system architecture, interface to other systems
	Long-term quality assurance
	Performance control in production
	...
Evaluation	Evaluation method
	Comparison of different approaches
	Transparency and explainability approaches
	Bias and fairness tests
	Security risks and mitigation strategy
	...

Table 4.2: Required documentation

Appendix Two

This glossary adds details on the terminology used throughout this paper. A special focus is placed on the technical terms concerning algorithms and artificial intelligence, as well as an explanation of the roles that contribute to the development and operation of a system for algorithm-based decision-making.

Abbreviations

AI Artificial intelligence

ALE Accumulated local effect

CPU Central processing unit

CRISP-DM Cross-industry standard process for data mining [11]

GDPR General Data Protection Regulation [22]

GPU Graphics processing unit

ICE Individual conditional expectation

ICO Information commissioner's office

ML Machine learning

PDP Partial dependence plot

SAI Supreme Audit Institution

Technical terminology

Artificial intelligence (AI):

In this white paper, *artificial intelligence* refers to a human-made system that is capable of intelligent behaviour in the sense that it develops the rules of its behaviour itself, based on a given environment (input data) and a given goal. Machine learning (see below) can be one part of such an AI system.

Black box model:

The black box metaphor is usually used for systems where the internal mechanisms are unknown, and only input and output can be observed. An ML model is thus referred to as a 'black box' model when it is not possible (or too difficult to be feasible) to explain to a human how a given input data leads to the output data provided by the model (prediction). Common examples include neural networks and ensemble models. Note that there are different interpretations of 'explainable' in this context: In this paper, a theoretical possibility to calculate the model output manually from a large number of internal model parameters is not deemed to be a sufficient 'explanation' if that procedure does not lead to a set of rules that is comprehensible by humans. Approaches to make the relation between model input and output more comprehensible by adding an additional, separate algorithm (such as *lime*, *shapley values*) do not turn a black box model into a white box model.

Compliance audit

An independent review of an organisation's adherence to regulatory guidelines. (See [13] for details.)

Hyperparameter

Hyperparameters are used to control the learning process in machine learning. These parameters are controlled by the coder, unlike the values of the model parameters which are derived via training.

Machine learning (ML)

Machine learning is a field of computer science dealing with methods to develop ('learn') rules from input data to achieve a given goal. A **machine learning algorithm** is a certain programmatic implementation of a strategy to find such rules (for example, with a neural network, a logistic regression or a decision forest). A **machine learning model** is the resulting set of rules (encoded in model parameters), inherently including the type of the ML algorithm. The ML model can be used to make *predictions* on data previously unknown to the ML model.

Performance audit

In the case of an SAI, a performance audit is an independent evaluation of a policy, programme or institute of a country's central government. The aim of this specific type of audit is to assess the efficiency and effectiveness of the spend resources. In general there is an adherence to previously agreed methodology and an objective and systematic execution of the evaluation.

White box model

In the context of ML, the ‘white box’ metaphor is used to emphasise that the model outcome can be explained to humans given the input data and the model parameters. Common examples are logistic regression and (single) decision trees.

Train, validation and test data

Initially training data is used to fit an optimal performing ML model. The validation data is then used to provide an evaluation of the model fit whilst tuning the hyperparameters. The test data is used last in order to provide a final evaluation of the model’s performance.

Model evaluation terms

Confusion matrix: In the context of a binary classification algorithm a confusion matrix shows a cross-tabulation of how many of the actual positive/negative results the model has predicted to be positive/negative.

False positive: A false positive indicates an instance where a ML model predicts a positive outcome, but the prediction is wrong because a negative outcome occurs.

False positive Rate: The false positive rate represents the proportion of predicted positive outcomes that are incorrect.

True positive: In the context of predictions made by a classification model, a true positive represents a case where the model predicts a ‘positive’ outcome, and the real outcome is also positive.

True positive rate: The true positive rate represents the proportion of positive outcomes that were predicted as positive by an ML model.

Roles

The following list attempts to summarise the responsibilities in an ML project that can be relevant in an audit. It is provided to help auditors identify contacts within the auditee’s organisation.
Comparatively small auditee organisations might combine several of the roles

		Actual	
		Positive	**Negative**
Predicted	**Positive**	**True positive** PPV (precision): $\frac{TP}{TP+FP}$, TPR (recall) : $\frac{TP}{TP+FN}$	**False positive** FPR: $\frac{FP}{FP+TN}$ (FDR: $\frac{FP}{TP+FP}$)
	Negative	**False Negative (FN)** FNR: $\frac{FN}{TP+FN}$ (FOR: $\frac{FN}{TN+FN}$)	**True Negative (TN)** TNR: $\frac{TN}{TN+FP}$ NPV: $\frac{TN}{TN+FN}$

Table 4.3: Confusion matrix for binary classification

below into a single person or team.

Some roles can furthermore be taken on by external consultants; however, in the case of the audit happening after the consultants' assignment is finished, internal personnel should have acquired the knowledge of the respective roles. It is thus the responsibility of the auditee to ensure sufficient documentation of any work done by external consultants.

Budgetary commissioner

This person is responsible for the budget of the auditee organisation and thus for any spending on ML algorithm software development projects or consulting. They are the authority on whether the development and operation of such an algorithm is a worthwhile use of the auditee organisation's budget and should be able to provide all budgetary information on the ML algorithm's development and operation.

Chief information officer (CIO)

The CIO of the auditeeo organisation is responsible for all its IT and thus should be informed about all ML algorithms already in operation and all projects that are developing such algorithms.

Commissioner for data protection and privacy

This is the chief data protection official of the auditee organisation. They must be informed of any concerns about the use of personal data by the ML algorithm. Their role is to ensure that the algorithm adheres to data privacy laws and regulations, such as the EU's GDPR.

Controller

The person who audits projects and checks for adherence to governance principles.

Data analyst

The person who analyses and works with the data that is to be fed to the ML algorithm. They are responsible for data understanding and should be closely involved with the development process. They assist the product owner, by translating their demands into specifications and requirements for the developers.

Data engineer

The person responsible for technical aspects the raw data (data warehousing, data quality, access control) as well as understanding of the raw data and sources. They are also responsible for data provision and data management.

Developer

The person/people who produce the ML model according to the specifications and requirements that were agreed upon with the product owner (and train the model, for models that require a training phase). The are responsible for transformations of the raw data to the final variables used by the model ('feature engineering'), and they are closely involved with the data analysts and engineers, the project leader and the product owners.

IT security officer

This the chief IT security official of the auditee organisation. They must be informed about any and all IT security aspects of the development and operation of the ML algorithm.

Process hotline/user helpdesk

The team that is tasked with providing support for users/processing officials. They should be able to answer all questions that arise during the routine operation of the software.

Project leader

The person responsible for all project management/project governance topics. They hould be able to provide any required project management documents.

Project owner/product owner

The team or unit within the auditee organisation that is responsible for the task that now should be supported or automated with an ML algorithm. They decide on which performance measures are required from the ML algorithm and the acceptance of the deliverables at the end of an ML development project

User/processing official

The person or unit that is supposed to use the results of the ML algorithm for their job. They are a deciding factor for the success of ML projects as they have to understand the suggestions or results from the ML algorithm and apply them to their (routine) tasks.

Subject matter expert

This is a generic term for someone with expert knowledge in a specific domain.

Bibliography

[1] The Norwegian Data Protection Authority (Datatilsynet). *Artificial intelligence and privacy.* 2018. URL: https://www.datatilsynet.no/globalassets/global/english/ai-and-privacy.pdf.

[2] European Parliament Research Service (EPRS). *EU guidelines on ethics in artificial intelligence: Context and implementation.* 2019. URL: https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/640163/EPRS_BRI(2019)640163_EN.pdf.

[3] *ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT).* URL: https://facctconference.org/.

[4] High-Level Expert Group on AI. *Ethics Guidelines for Trustworthy Artificial Intelligence.* 2019.

[5] A. Feller et al. "A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear." In: *The Washington Post* (2016). URL: http://www.cs.yale.edu/homes/jf/Feller.pdf.

[6] C. Molnar et al. *Limitations of Interpretable Machine Learning Methods.* 2020. URL: https://compstat-lmu.github.io/iml_methods_limitations/.

[7] I. D. Raji et al. *Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing.* Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pages 33–44. DOI: https://doi.org/10.1145/3351095.3372873. URL: https://dl.acm.org/doi/proceedings/10.1145/3351095.

[8] J. Gesley et al. *Regulation of Artificial Intelligence in Selected Jurisdictions.* 2019. URL: https://www.loc.gov/law/help/artificial-intelligence/regulation-artificial-intelligence.pdf.

[9] M. Brundage et al. *Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims.* 2020. URL: https://arxiv.org/abs/2004.07213.

[10] Swedish Data Protection Authority. *Supervision pursuant to the General Data Protection Regulation (EU) 2016/679 - facial recognition used to monitor the attendance of students.* 2019. URL: https://www.datainspektionen.se/globalassets/dokument/beslut/facial-recognition-used-to-monitor-the-attendance-of-students.pdf.

[11] A. Clark. *The Machine Learning Audit–CRISP-DM Framework.* ISACA Journal. 2018. URL: https://www.isaca.org/resources/isaca-journal/issues/2018/volume-1/the-machine-learning-auditcrisp-dm-framework.

[12] K. Eykholt et al. *Robust Physical-World Attacks on Deep Learning Models.* 2017. URL: https://arxiv.org/abs/1707.08945.

[13] INTOSAI. *ISSAI 400 – Fundamental Principles of Compliance Auditing.* URL: https://www.intosai.org/fileadmin/downloads/documents/open_access/ISSAI_100_to_400/issai_400/issai_400_en.pdf.

[14] I. E. Kumar et al. *Problems with Shapley-value-based explanations as feature importance measures.* 2020. URL: https://arxiv.org/pdf/2002.11097.pdf.

[15] U. von der Leyen. Letter of the European Commission President-elect, then candidate for president of the European Commission, to the members of the European Parliament, about political guidelines in the case of her election, 15.07.2019. 2019. URL: https://g8fip1kplyr33r3krz5b97d1-wpengine.netdna-ssl.com/wp-content/uploads/2019/07/190714-Letter-Candidate-RENEW-1.pdf.

[16] C. Molnar. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable.* 2019. URL: https://christophm.github.io/interpretable-ml-book/.

[17] N. Morgulis et al. *Fooling a Real Car with Adversarial Traffic Signs.* 2019. URL: https://arxiv.org/abs/1907.00374.

[18] OECD. *Recommendation of the Council on Artificial Intelligence.* 2019. URL: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

[19] Information commissioner's Office. *AI auditing framework.* blog series. URL: https://ico.org.uk/about-the-ico/news-and-events/ai-auditing-framework/.

[20] UK National Audit Office. *Framework to review models.* 2016. URL: https://www.nao.org.uk/report/framework-to-review-models/#:~:text=National%20Audit%20Office%20report%3A%20Framework%20to%20review%20models&text=This%20framework%20provides%20a%20structured%2C%20flexible%20approach%20to%20reviewing%20models.&text=The%20framework%20is%20based%20on,HM%20Treasury%20and%20international%20standards..

[21] Information commissioner's Office website. URL: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/data-protection-impact-assessments-dpias/when-do-we-need-to-do-a-dpia/.

[22] The European Parliament and of the Council of the European Union. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).* 2016. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[23] The Article 29 Working Party. *Guidelines on Automated individual decision-making and Profiling for the purpose of Regulation 2016/679.* The European Data Protection Board (EDPB), which replaced the Article 29 WP in May 2018, endorsed these guidelines during its first plenary meeting. 2018. URL: https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053.

[24] S. Verma and J. Rubin. *Fairness Definitions Explained.* ACM/IEEE International Workshop on Software Fairness. 2018. URL: http://fairware.cs.umass.edu/papers/Verma.pdf.

[25] M. Wieringa. *What to account for when accounting for algorithms.* FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Pages 1–18. 2020. DOI: https://doi.org/10.1145/3351095.3372833. URL: https://dl.acm.org/doi/proceedings/10.1145/3351095.