10.3 Attachment 3: Framework for the assessment of responsible AI

For assessments of whether the principles of responsible AI are safeguarded in the AI systems in case investigations, we use a framework of audit questions based on Auditing machine learning algorithms[1] and Audit framework for algorithms[2].

Each question is linked to one of the ethical principles for responsible AI as the main principle, or as *governance and accountability* that relates to several ethical principles.

10.3.1 Audit questions assessed in the AI systems

Governance and accountability	
Process management	Quality of the documentation: Are all aspects of the development, implementation and use of the AI system adequately documented?
Process management	Does the algorithm have a clearly defined purpose?
Process management	Has the entire process surrounding the algorithm been documented? (AI life cycle management)
Process management	Is there an agreed and documented policy on quality and performance objectives for the algorithm?
Process management	Is the algorithm monitored at regular intervals (at least in relation to availability, performance/quality, safety, compliance with current legislation and regulations)?
Resource management	Does the organisation have sufficient access to high-quality expertise?
Resource management	Have roles, tasks, responsibilities and powers (including ownership) been defined and have these been applied in practice?
Risk management	Is there an assessment of the risks regarding the use of the algorithm at fixed (periodic) intervals? Which of the following aspects have been assessed in a risk analysis, and what were the results? - the application - the surrounding IT environment - data preparation and - the processes mapped by the application

[1] A framework for audit of AI systems based on machine learning, developed by the supreme audit institutions of Finland, Germany, the Nederlands, the UK and Norway in 2020 (updated 2023), with corresponding Excel verktøy.
[2] A framework developed by the Netherlands Court og Audit in context of the audit Understanding algorithms in 2021, and applied in the audit An audit og 9 algorithms used by the Dutch government in 2022[2]. This framework does not focus on machine learning, but includes simpler AI systems as well.

Management of external contributions	When outsourcing parts or activities related to the algorithm, have arrangements for product and knowledge transfer been made and documented with the external parties involved?
Management of external contributions	For procured AI-systems, have arrangements with the supplier been made and documented to ensure frequent updates, timely information from the supplier about security or performance issues?
Information and data management	Are the data sources, technical/operational criteria for data selection, and legal basis for data collection and processing sufficiently documented?
Model development and usage	Were the various stakeholders and 'end users' of the algorithm involved in the development process?
Privacy	
Process management	Does the AI system follow the principle of privacy by design?
Data source	Is the legal basis for processing of personal data documented? Is data processing grounded in statutory duty?
Data source	Is the processing of (special categories of) personal data by the algorithm in line with the original purpose?
Information and data management	Is the retention period for the data defined and a data lifecycle policy in place?
Information and data management	Have the controller and the data processor for the algorithm and the data used been designated?
Data processing	Is there evidence of data minimalisation? Is the data processing reasonable, relevant and limited (proportionality)?
Data processing	Is processing of personal data recorded in a maintained, digital record?
Data processing	Has a Data Protection Impact Assessment been performed (if applicable)?
Data processing	Have arrangements been made for informing, either pro-actively or on request, individuals whose data is processed or used (in relation to both data and algorithm)?
Data processing	Is there a publicly available privacy policy that covers the data and algorithms used?
Model development and usage	Is there evidence of automated decision-making, and if so, does this comply with relevant legislation and are people impacted by the outcome of the model enabled to demand human intervention and contest the decision?

Model development and usage	Has an assessment been made of whether there is evidence of profiling and whether this is permitted?
Equal treatment	
Process management	How do you guarantee model fairness? Does the AI system follow the principle of fairness by design?
Data source	Is there no undesirable bias in the data source?
Data source	Is the data used for development representative of the application for which the algorithm is used?
Management of external contributions	For procured AI-systems or outsourced developments, is sufficient documentation obtained about data sources to evaluate applicability to the data and way of usage in the organisation?
Model development and usage	Have safeguards been put in place to avoid any bias resulting from the choices made in relation to the model?
Model development and usage	Is there discrimination due to the data and model used?
Transparency	
Process management	How do you guarantee model transparency? Does the AI system follow the principle of transparency by design?
Data processing	Is the logic of the algorithm, its impact and the data used sufficiently clear to data subjects?
Process management	Does the algorithm have a clearly formulated and explainable purpose that is shared by the owner, developer and user, that enables a multidisciplinary approach?
Model development and usage	Is information about the use of an AI model, a description of the underlying data (as far as possible) and of its method for analyzing the data available to stakeholders? Has the operation of the model or algorithm, including its limitations (i.e. what it can and cannot do) been communicated to all stakeholders?
Model development and usage	Is the algorithm explainable and has an attempt been made to strike a balance between the models' explainability and performance?
Model development and usage	Is there documentation describing the design and implementation of the algorithm?
Model development and usage	Has a record been made of the reasons underlying the choices made in the design and implementation of the algorithm?

Model development and usage	Has the quality of the model been documented, including performance metrics or other performance indicators?
Robustness and security	
Prosesstyring	Have the objectives set for the AI system been accomplished and has the application achieved the intended purposes?
Risikostyring	How does the application respond to faulty or manipulated datasets?
ITGC - backups	What back-up practices are established for the AI system, i.e. what kind of back-up, frequency, storage location, and testing practice?
ITGC - logging	What kind of logging practice is applied for the algorithm and its runtime environment (including databases), and how are the logs monitored?
ITGC - tilgangsstyring	How are privileged users administered in the operation environment of the AI system? (Describe at least creation, change, deactivation, periodic review, overview/register of authorizations with person responsible for approval, do administrators have normal user accounts for non-admin usage, to what extend are unpersonal accounts used?)
ITGC - tilgangsstyring	What kind of users have access to the AI system (person/application, level of privileges) and what authentication method is used? If multi-factor authentication is used, what kind of and to what extend wrt. systems, intern/external access and user groups? What are the requirements for passwords, and how are these requirements implemented, and are there different requirements for different user groups? If encryption keys are used, are they bound to identifiable owners, and where are the keys stored?
ITGC - endringshåndtering	What kind of change management of the AI system is applied, including evaluation of consequences and authorization? Which requirements exists for change of code or the algorithm/model?
ITGC - Management of external contributions	For procured AI systems, have agreements been made with the supplier for sufficient logging, backup, change management and access control, where this is managed by the supplier? Is information sent between customer and supplier by the system sufficiently secured?
Management of external contributions	For procured AI-systems, have arrangement been made with the supplier for continuous support of the system for a reasonable assumed lifetime of the application, or have

	alternatives been explored in case of end of support from the supplier?
Process management	Is the algorithm's purpose operationalized and translated into practical features with respect to the model and data? Which particular tasks or aspects of operational management is the algorithm intended to support?
Process management	Have arrangements been made for the maintenance and management of the algorithm? Is the model updated at regular intervals, including with respect to current legislation?
Model development and usage	What training / validation / test datasets do you use? How were they generated or selected, how are they updated during the life cycle of the system? Have safeguards been put in place regarding the quality of the choices made in relation to training and test data? Are training, test and validation data processed separately? Is the amount of data sufficient for this separation method?
Model development and usage	Was the selection of hyper-parameters supported by arguments and evidence?
Model development and usage	Did you benchmark the performance of your model against any alternative methods/models? Please specify.
Model development and usage	Is the model's output monitored?
Management of external contributions	For procured AI systems, is documentation from the supplier available about the quality of the model, including performance metrics or other performance indicators? Which performance metrics are obtained from the supplier?

Some of the more comprehensive audit questions are operationalized in the next section.

10.3.2 Operationalization of the framework

Each audit question in the framework is assessed as one of three levels, similar to the system for process capability or maturity level in COBIT 2019[3], but simplified to only three maturity levels: low, medium or high.

Since not all the requirements in the framework are equally relevant for all AI systems, an assessment of low or medium may be sufficient for some AI system. Therefore, we use the term activity level in the description of results.

[3] Capability Maturity Model Integration® (CMMI)-based process-capability scheme, from 0 to 5, COBIT 2019 Framework: Governance and management objectives, ISACA

Based on these individual assessments, a summarized activity level per ethical principle is calculated as a percentage of the requirements met in the category, with a weighting of 0, 0.5 or 1 for low, medium or high levels in each requirement, respectively.

10.3.3 Governance principles considered in the framework

Some of the governance principles considered in the framework are comprehensive and are therefore broken down into several aspects in the following.

AI life cycle management

AI life cycle management is a management system for managing the entire AI development process, from planning to operation or decommissioning. A plan for such process management includes planning time and resources for work on the AI system's purpose and requirements, design and development, quality assurance and testing, production setting and maintenance. For supplier systems, design and development have been replaced by procurement and, potentially, customization. Compared to traditional software, the risks associated with AI are larger in the testing and maintenance phase, since for example machine learning models are known to be susceptible to unwanted changes in performance over time. The focus points are therefore documented time and resource planning for

1. development
2. monitoring av the AI system’s results
3. maintenance and updates, including retraining av models

In the context of monitoring and maintenance, resource planning includes that responsibilities are clearly defined. Note that for the audit question concerning AI life cycle management, only time and resource planning related to the performance of the AI system is assessed. Ideally, AI life cycle management should be coupled with privacy, equal treatment and transparency *by design*, and corresponding behavior of the AI system should be included in development, monitoring, and maintenance. In order to assess these aspects separately, we have narrowed the scope of life cycle management to performance and consider ethical principles *by design* separately (see next section).

Ethical principles by design

The OECD recommends[4] implementing mechanisms and safeguards to ensure that ethical principles are adhered to throughout the life cycle of an AI system. The principle of trustworthy AI *by design* entails considering privacy, fairness, transparency and security at all stages of the AI system's life cycle.

For example, the Norwegian Data Protection Authority (*Datatilsynet*) defines data protection *by design* as the principle whereby a technical system or solution is developed in such a way that privacy is safeguarded. The Authority emphasizes that a prerequisite for this is sufficient expertise in the underlying data protection principles.[5] Similarly, the Norwegian Digitalisation Agency (*Digdir*) states that information security *by design* includes incorporation into business processes and projects from the outset while considering the entire life cycle of ICT solutions.[6]

In the AI audit framework, we look for indications of privacy, fairness, and transparency *by design* as governance principles. This includes at least an evaluation of relevant principles and associated risks, as well as planning and implementation of relevant measures.

[4] OECD, Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449
[5] Datatilsynet. (2023, 27, juli). *Innebygd personvern og personvern som standard*. https://www.datatilsynet.no/rettigheter-og-plikter/virksomhetenes-plikter/innebygd-personvern-og-personvern-som-standard/
[6] Digdir. (u.å.). Innebygd informasjonssikkerheit. Access 23.05.2024 from https://www.digdir.no/informasjonssikkerhet/innebygd-informasjonssikkerheit/2146

We do not assess security *by design*, since general security aspects are handled by the company's general security management independently of the AI systems, and none of the AI systems in the case study are exposed to new, AI-specific security risks such as *poisoning* or *adversarial attacks*. Furthermore, none of the three in-house developed AI systems fall within the classification of high-risk AI systems[7], according to the risk classification in the proposed EU AI Act. There is no danger to the life and health of natural persons if these AI systems fail. The fourth AI system is an exception to this. It falls under the classification of high-risk due to its use in medical diagnostics, but here AI-specific security measures are closely linked to monitoring of the system's performance, which is covered under AI lifecycle management.

Privacy by design

In order to fulfil the requirements for privacy by design according to Article 25 of the General Data Protection Regulation, agencies may refer to the Norwegian Data Protection Authority's guidance on privacy by design and privacy by default. The guidance describes key elements of the privacy principles, including examples for implementation. The privacy principles that public agencies should consider in the context of AI systems are legality, fairness, transparency, purpose limitation, data minimization, accuracy, storage limitation, integrity and confidentiality, and accountability.

The privacy principles of fairness and openness thus overlap with the governance principles of fairness and transparency by design. To minimize overlap in our assessment, the requirements for meeting fairness or transparency as one of nine privacy principles are less stringent than those we consider for the respective principle by design. These principles are not limited to the context of privacy but are also evaluated in accordance with both statutory and non-statutory principles for administrative procedure in public administration.

Fairness by design

Fairness has many definitions, and its meaning depends on context. A possible practical method for examining fairness is to compare whether equal cases are treated equally. It is often difficult to determine "equal" for both the initial situation, a treatment and the resulting situation in order to achieve individual justice. Therefore, group-based fairness is often used, which compares the treatment or results of different groups of people who are equal in all relevant aspects and should therefore be treated equally.

The Equality and Anti-Discrimination Act (*likestillings- og diskrimineringsloven*) defines a set of human characteristics which are forbidden as a reason for differential treatment. In public administration, decisions shall not entail unjustified differential treatment, without limitation to characteristics protected under the Equality and Anti-Discrimination Act.

One possible reason for AI systems performing differently for different groups of people can be historical differences that models learn from the data. Another common challenge in the development of machine learning systems is that some groups are underrepresented in the training data, which leads to inferior model performance for these groups.[8] This form of bias is referred to as representation bias.

Common definitions of fairness

The methodology for built-in anti-discrimination protection in guidelines[9] developed by the Equality and Anti-Discrimination Ombud's (*Likestillings- og diskrimineringsombud*, LDO) is based on **group fairness**. This way of defining fairness is easy to relate to the law, and at the same time it is easy to calculate fairness

[7] High risk AI after categorisation in the EUs proposal for tha AI Act, article 6 ond Annex III.
[8] Storås, A. M., Prabhu, R., Hammer, H. L. & Strümke, I. (2022). Bias og kvantitativ analyse innen velferd: opphav til skjevheter og relasjon til utfallsrettferdighet. *Tidsskrift for velferdsforskning, 25*(3). https://www.idunn.no/doi/10.18261/tfv.25.3.3
[9] Likestillings- og diskrimineringsombudet. (2023). *Innebygd diskrimineringsvern : En veileder for å avdekke og forebygge diskriminering i utvikling og bruk av kunstig intelligens*. https://ldo.no/globalassets/_ldo_2019/_bilder-til-nye-nettsider/ki/ldo.-innebygd-diskrimineringsvern.pdf

measures[10]. Therefore, we also use this fairness definition. Typical fairness measures for a classification system are based on the number of misclassifications in different groups, for example how many women are incorrectly selected for a control compared to men.

An alternative that has already been mentioned is **individual fairness**, which means that equal cases are treated equally. The definition is less used, because it is challenging to measure similarity of input data and similarity of result.

Counterfactual fairness is more commonly employed. It measures whether a trait or variable is the cause of the outcome by performing the same calculations with the inverse trait. For instance, testing the outcome of a machine learning model for the same individual when only the gender variable has been altered in the input data. This method is useful for exploring the mechanisms behind differential treatment. However, it is often difficult to correctly address correlated variables (for example, a variable such as vocational education is related to gender). The term **counterfactual explanation** refers to a related methodology that aims at explaining the cause of a machine learning model's outcome by identifying the minimal change in input data that alters the outcome. This method can also be suitable for uncovering whether unjustified reasons affect a prediction.

These three approaches to assessing fairness can be considered standard, with the aforementioned technical challenges of individual and counterfactual fairness. A fourth occasionally used approach, termed **fairness through unawareness**, considers an AI system to be fair if none of the characteristics prohibited as grounds for differential treatment are present in the input data. This definition is **not** considered sufficient, since it has been shown in several cases that correlations with included variables can lead to discrimination. Consequently, the LDO's guide for built-in protection against discrimination emphasize that even if personal data is removed, the system may find patterns in the data such that information that coincides with the removed characteristics is given unlawful weight.

Equal treatment in enterprise oversight

Fairness definitions usually target fairness in the treatment of natural persons. However, the requirements for objectivity, soundness and proportionality in public administration are not limited to natural persons but apply to the entire administrative procedure. [11] Similar definitions and methods can thus be used to assess equal treatment of enterprises. The main difference is which characteristics of an enterprise can lead to unjustified differential treatment.

The possible risk of unjustified differential treatment of enterprises must be assessed for each AI system

Transparency by design

Transparency in AI consists of openness and explainability. Openness pertains to the provision of relevant information to stakeholders, whilst explainability of AI systems refers to the understanding of how the initial situation (input) leads to a particular result (outcome). Explainability can thus be a prerequisite for openness where such an understanding is relevant. For instance, individual decisions must be justified in accordance with section 24 of the Public Administration Act (openness), and section 25 stipulates that the grounds must contain both the relevant rules and the factual circumstances upon which the decision is based (explanation).

In the assessment of transparency by design, we investigate how openness and explainability is planned for and implemented. Furthermore, we examine whether there is evidence that transparency is a stated objective.

[10] It can be challenging to define the most relevant definition of fairness to fulfill, which is necessary since it is often impossible to fulfill all, jf. *impossibility theorem of fairness*.
[11] Jf. ulovfestede prinsipper i norsk forvaltningsrett, NOU 2019: 5 (2019). Justis- og beredskapsdepartementet. Kapittel 11.7.

In AI development, it is critical to consider, among other things, how to balance the consideration of accurate and high-performance models with the ability to understand and explain the decisions made by the AI system.[12]

Robustness and safety

Safety in relation to artificial intelligence includes, among other factors, information security, human safety and secure use of AI. Information security concerns protection of information from unwanted access (confidentiality), the availability of information when its needed (accessibility) and safeguarding against manipulation (integrity). To prevent AI systems from potentially causing harm, they must be rendered technically secure and robust by, among other things, adding safeguards against manipulation and misuse. Artificial intelligence must be built upon systems with technically robust solutions that mitigate risk and ensure the systems function as intended.

[12] Raz, A., Heinrichs, B., Avnoon, N., Eyal, G. & Inbar, E. (2024). Prediction and explainability in AI: Striking a new balance? *Big Data & Society, 11*(1) https://doi.org/10.1177/20539517241235871.