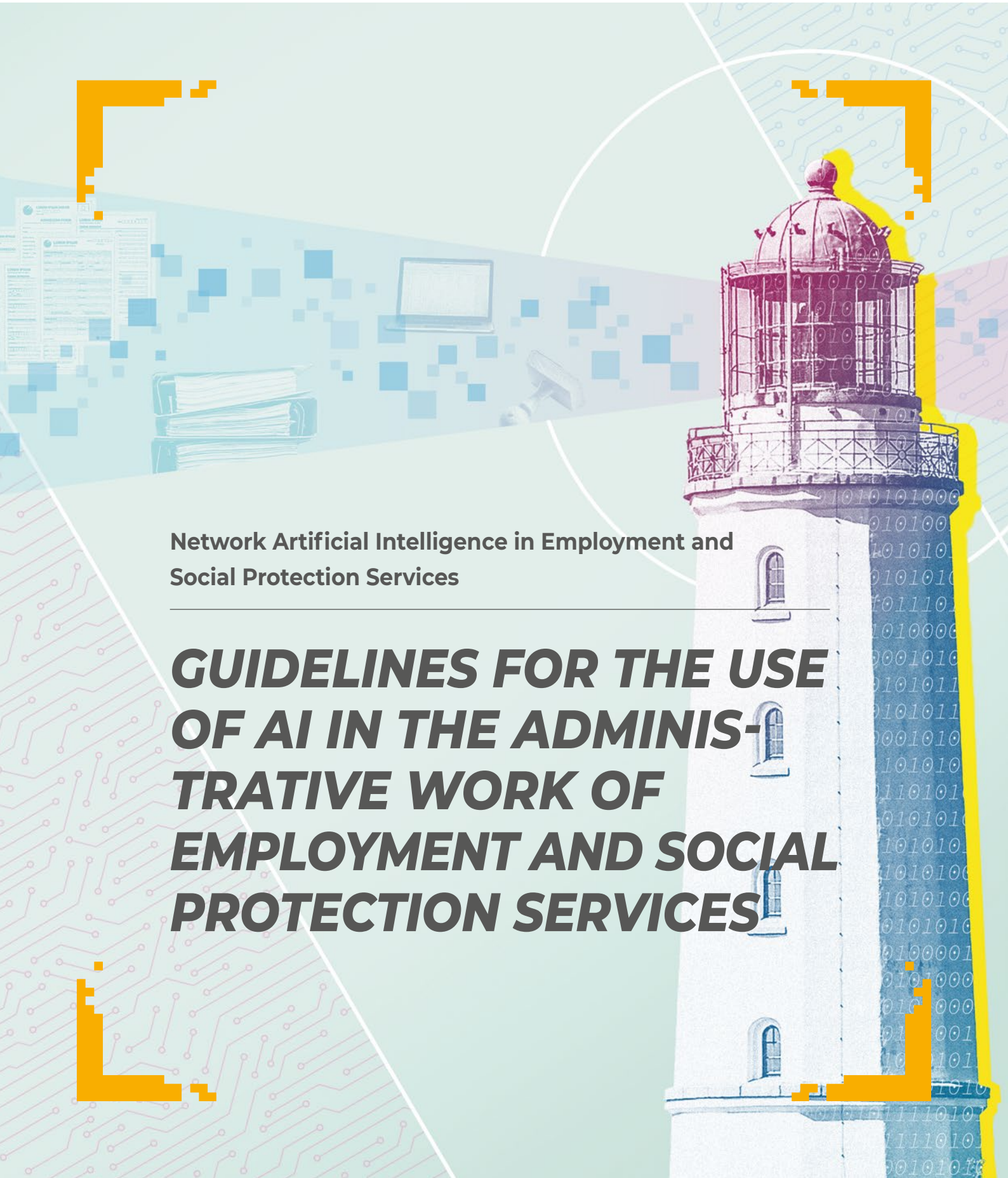
Federal Ministry
of Labour and Social Affairs

DENK FABRIK DIGITALE ARBEITS-GESELLSCHAFT



Network Artificial Intelligence in Employment and Social Protection Services

GUIDELINES FOR THE USE OF AI IN THE ADMINISTRATIVE WORK OF EMPLOYMENT AND SOCIAL PROTECTION SERVICES

This publication was developed by the Network AI in Employment and Social Protection Services. Member organisations of the Network are:

German Occupational Accident Insurance Fund for the Building Trade (BG BAU)

German Occupational Accident Insurance Fund for the Energy, Textile, Electricity and Media Industries (BG ETEM)

German Occupational Accident Insurance Fund for non-state Institutions within the Health and Welfare Service Sectors (BGW)

German Occupational Accident Insurance Fund for the Retail and Logistics Sector (BGHW)

German Occupational Accident Insurance Fund for the Woodworking and Metalworking Industries (BGHM)

German Occupational Accident Insurance Fund for the Food and Hospitality Industry (BGN)

German Occupational Accident Insurance Fund for the Raw Materials and Chemicals Industries (BG RCI)

German Occupational Accident Insurance Fund for the Transport, Mail Logistics and Telecommunications Sectors (BG Verkehr)

Federal Employment Agency (BA)

Federal Office for Social Security (BAS)

Federal Institute for Occupational Safety and Health (BAUA)

Federal Ministry of Labour and Social Affairs (BMAS)

German Social Accident Insurance (DGUV)

German Statutory Pension Insurance (Deutsche Rentenversicherung)

German Pension Insurance for Mining, Rail and Sea (DRV KBS)

Artists' Social Insurance Fund (KSK)

Social Insurance Fund for Agriculture, Forestry and Horticulture (SVLFG)*

German Occupational Accident Insurance Fund for the Federal Civil Service and Railway (UVB)

German Occupational Accident Insurance Fund for the Administrative Sector (VBG)

* Within the Network, the SVLFG (Social Insurance Fund for Agriculture, Forestry and Horticulture) also acts on behalf of the Supplementary Pension Funds for Employees in Agriculture and Forestry (ZLF VVaG and ZLA).

The Network AI is a project managed by the Policy Lab Digital, Work & Society at the Federal Ministry of Labour and Social Affairs, which provides funding and coordinates the project with the support of the iRights.Lab.

BG BAU
Berufsgenossenschaft der Bauwirtschaft

BG ETEM
Energie Textil Elektro Medienerzeugnisse

BGW
Berufsgenossenschaft für Gesundheitsdienst und Wohlfahrtspflege

BGHW
Berufsgenossenschaft Handel und Warenlogistik

BGHM
Berufsgenossenschaft Holz und Metall

BGN
Berufsgenossenschaft Nahrungsmittel und Gastgewerbe

BG RCI
Berufsgenossenschaft Rohstoffe und chemische Industrie

BG Verkehr
Verkehrswirtschaft Post-Logistik Telekommunikation

Bundesagentur für Arbeit

Bundesamt für Soziale Sicherung

baua:
Federal Institute for Occupational Safety and Health

Federal Ministry of Labour and Social Affairs

DGUV
Deutsche Gesetzliche Unfallversicherung
Spitzenverband

Deutsche Rentenversicherung
Bund

IFA
Institut für Arbeitsschutz der Deutschen Gesetzlichen Unfallversicherung

iRights.Lab
Think Tank für die digitale Welt

Knappschaft Bahn See
sozial kompetent für mich!

KÜNSTLER SOZIALKASSE

SVLFG

UVB
Unfallversicherung Bund und Bahn

VBG
Ihre gesetzliche Unfallversicherung

Table of contents

1. Introduction

Artificial Intelligence (AI) is being used in more and more areas of daily life. AI applications also offer enormous potential for public sector administrative functions. Millions of applications for pensions, social assistance, unemployment benefits, and information requests are made every year to the relevant public authorities. AI systems can help the staff in the employment and social protection service offices to carry out their work, make processes more efficient, and reduce processing times. Among other things, the importance of modern, effective public administration was highlighted recently by the coronavirus pandemic. But many administrative agencies at federal and state level are already short of staff. Demographic change will further exacerbate this situation: as the baby boomer generation reaches retirement age, approximately ten million people will be applying for a pension and these applications will all have to be processed. At the same time, some of these people will also be retiring from the public authorities themselves. This creates significant challenges for the welfare state, which innovative AI-based solutions can help to overcome.

Nonetheless, the employment and social protection service offices carry special responsibilities when they make use of AI. These agencies process highly sensitive data, and their decisions and services have a direct impact on members of the public, who are often experiencing a particularly challenging life situation. AI applications are already being used in some areas by the Employment and Social Protection Services, such as for the automatic recognition of enrolment certificates or certificates of study by the Federal Employment Agency, or the identification of promising cases for recourse claims by the German Occupational Accident Insurance Fund for the Energy, Textile, Electricity and Media Industries (BG ETEM). The development and use of AI within the employment and social protection services is thus still at the beginning; it is nevertheless very important to agree at the outset on fundamental rights and obligations, values, and principles, and to develop policies and practical guidelines for its use within employment and social protection service functions.

The lawful, ethical, and value-driven use of AI forms the foundation for exploiting its potential for the benefit of society and the harness of modern public services administration. The EU Artificial Intelligence Regulation, a draft of which has been presented and is currently being negotiated (COM(2021)206), will in future provide a specific framework for the use of AI. There are already binding rules on individual aspects of the topic, in the EU General Data Protection Regulation (GDPR), for instance, and in German social and administrative law. Recommendations like those formulated in the Ethics Guidelines for Trustworthy AI from the High-Level

Expert Group on Artificial Intelligence set up by the European Commission, the report by the German Data Ethics Commission, the final report by the AI Study Commission of the German Bundestag, the German federal government's AI strategy, the Hambach Declaration on Artificial Intelligence by the German Data Protection Conference, or the Recommendation of the OECD Council on Artificial Intelligence all influence the current debate about an operational framework for AI. There are also concrete contributions from civil society, such as the Algo.Rules from the Bertelsmann Foundation in cooperation with iRights.Lab, the Impact Assessment Tool for Automated Decision-Making Systems by AlgorithmWatch on behalf of the canton of Zurich, or the concept paper "Artificial Intelligence (AI) for Good Work" by the German Trade Union Confederation.

The Network Artificial Intelligence in Employment and Social Protection Services has drawn up these self-committing guidelines for the use of AI in Employment and Social Protection Services as a supplement to the existing framework and to provide guidance and practical information for using AI in public administrative functions in accordance with applicable law. Given that the use of AI within the Employment and Social Protection Services is increasing as well, it is important to provide practical guidance now, before regulation is adopted at the European level.

These guidelines offer practical guidance for public authorities. They include a brief introduction to the topics, recommendations, questions to ask, and checklists. Their intention is to provide support for project managers who are designing, developing, and operating AI systems with the aim of helping ensure processes are human-centric. Decision-makers, employee representatives, users, and developers can also use them to find out about the principles of value-based AI design and thereby better perform their respective roles. The guidelines are further addressed to the general public and thus to people who are potentially affected by AI-based decisions. The values, principles, and recommendations underlying the use of AI are presented transparently to them as well, which creates the basis for trust and acceptance.

The first section of the guidelines covers the fundamental values for the use of AI on which the Network has agreed. They consist of the seven value pairs, "Human-centricity & Common good", "Fairness & Non-discrimination", "Explainability & Transparency", "privacy & protection of personal rights", "Security/Safety & Robustness", "Human oversight & Responsibility", and "Ecological sustainability & Conservation of resources". Later chapters look in more detail at the four key areas of designing AI systems for use in employment and social protection services. The guidelines are thus intended to help:

- design human-centric introduction processes and define the objectives for the AI application together with stakeholders (see Chapter 3),
- assess the impact of the planned AI application at an early stage and systematically evaluate potential risks for different groups of individuals, but also for society as a whole (see Chapter 4),
- ensure good data quality and avoid bias (see Chapter 5), and
- achieve transparency about the use, the objectives, and the operational process of the AI applications and enable explainability (see Chapter 6).

This is intended to facilitate the introduction of AI-based innovations in employment and social protection services. At the same time, the aim is also to ensure that such innovations are compatible with the value foundation as well as the values, principles, and quality requirements described in the other chapters. The guidelines have been drawn up primarily with regard to machine learning systems. However, they should be applied to non-learning systems, too, and thus for all algorithmic decision-making systems (ADM systems). The focus here is always on embedding the technology in its socio-technical context and on its effects on members of the public, employees, and society.[1]

[1] When AI systems are used for research purposes, i.e. as experimental scientific systems, as is the case at the Federal Institute for Occupational Safety and Health, these guidelines are going to be assessed for implementability in the respective research setting and tested in the application if possible. The fundamental values (see Chapter 2) represent a common understanding of fundamental rights, values, and principles for the use of AI and are therefore also a binding framework for research activities.

Drafting and ongoing development of the guidelines: participatory, inter-agency, and practice-oriented

The guidelines were drafted collaboratively by the Network Artificial Intelligence in Employment and Social Protection Services. The Network AI is a project managed by the Policy Lab Digital, Work & Society at the Federal Ministry of Labour and Social Affairs (BMAS), which provides funding and coordinates the project with the support of the iRights.Lab. The Policy Lab invited all the agencies and authorities[2] within the area of responsibility of the BMAS to take part in spring 2021. Twenty agencies delegated experts and employees to the Network, who were actively involved in drafting the guidelines. Numerous experts contributed their different perspectives and specialist knowledge in workshops, known as AI Labs.

The Network will continue to pursue dialogue and exchange views on the use of AI in public administration. The guidelines will be reviewed regularly and adapted in line with new technological developments, the changing demands of society, and legal requirements. Experience and lessons learned from using AI in administrative practice will also be incorporated. One key milestone in their continued development will be to review the guidelines once the European AI Regulation has been adopted and bring them into line as necessary with the then applicable legal framework. This ongoing development of the guidelines provides an opportunity to update them and adapt them to the respective social, legal, and administrative requirements using a proven participatory process. Feedback on the guidelines is very welcome and may be taken into account in further work.

In this first version of the guidelines, the Network focused initially on the four core topics: human-centric introduction, risk and impact assessment, data quality and bias, and transparency and explainability. It has not yet been possible to incorporate other important aspects in detail. These include, for instance, the necessary skills of administrative staff and their acquisition through vocational and professional training, the fundamental question of staffing at public agencies and authorities, further debate about data privacy aspects, and the relationship between the guidelines and other considerations, such as cost-benefit analysis. Discussion of these topics is to be continued in greater detail in the Network's future work.

If you would like to get involved in the Network's discussions or make suggestions, please write to us at: ***ki-in-der-verwaltung@bmas.bund.de***

[2] This includes the social insurance agencies.

2. Value Foundation

The Network Artificial Intelligence in Employment and Social Protection Services has agreed on common rights, values, and principles for the deployment of AI. These values are based on the Ethics Guidelines for Trustworthy AI from the High-Level Expert Group on Artificial Intelligence (AI HLEG) set up by the European Commission, the report by the German Data Ethics Commission, the final report by the AI Study Commission of the German Bundestag, the German federal government's AI strategy, the Hambach Declaration on Artificial Intelligence by the German Data Protection Conference, and the Recommendation of the OECD Council on Artificial Intelligence. The seven ***value pairs*** *that have been defined are:*

- ***Human-centricity & Common good***
- ***Fairness & Non-discrimination***
- ***Explainability & Transparency***
- ***Privacy & Personality rights***
- ***Security/Safety & Robustness***
- ***Human Intervenability & Responsibility***
- ***Ecological sustainability & Conservation of resources***

Human-centricity & Focus on the common good

Human-centricity means making human beings and their well-being the starting point and the objective whenever AI is used. AI is there for people, and not the other way around. The human-centric development and application of AI means thinking things from the premise of human beings and their needs, in order to establish trust and acceptance and to uphold the rights of members of the public and staff. The use of AI creates an opportunity to redesign and improve interactions between technology, people, and the environment. It is vital, however, to consider AI systems holistically and to see them within the context of their respective use. To this end it is important to include all actors, their use requirements, needs, and values in the development and implementation process. The focus on the common good emphasises the shared bigger picture. It means that AI should ideally benefit everyone in society and that any social consequences of using AI and any conceivable impact on the fundamental values of society, such as democracy and the rule of law, must be taken into account.

What does this mean for the administrative work of lemployment and social protection services?

The planning, development, and use of AI must always be designed such that these values are respected at all times. When defining the objective of an AI use case, it is vital to ask whether the AI really serves the people involved and the common good. The "people involved" includes both the members of the public affected by the administrative actions concerned and the staff of the public authorities. To what extent does the AI application meet their needs and take place in their interest? For instance, AI may relieve staff of monotonous, unpleasant routine tasks, reduce waiting times for citizens, or improve the quality of services and decisions. Inclusion and accessibility for people with disabilities must be also ensured when AI is used.

Fairness & Non-discrimination

AI-based decisions must be fair. The definition of what is fair or just in any individual case varies considerably depending on fundamental positions regarding morals, culture, and world view and is thus subject to a process of negotiation. It is therefore important to involve all stakeholder groups when developing AI. At the legal level, fundamental rights inform key value judgements which are directly binding on public-sector agencies and authorities. These include the imperative of equal treatment, which states that cases which are essentially the same may not be treated differently without proper justification. Furthermore, the Basic Law of Germany stipulates special protection against discriminatory unequal treatment on the basis of certain characteristics. Among other things, this encompasses the disadvantaging or favouring of any person because of their disability, for racist reasons, or based on their gender, ethnicity, language, origin, religious beliefs, or political opinions. Any such discriminatory unequal treatment can only be justified in exceptional cases for particularly serious reasons. These requirements of the Basic Law are formulated more specifically in the German General Equal Treatment Act (AGG) and in the volumes of the German Social Code (SGB).

Anyone who develops or uses an AI system in their organisation must therefore absolutely prevent the use of the AI system from having any discriminatory impact, particularly in terms of those characteristics that are specially protected by the Basic Law. State activities must also do justice to the interests of the persons affected in the sense of procedural justice or fair process: an administrative procedure supported by AI must equally ensure that steps such as hearings and the participation of employee representation are taken, because they are elements of this fairness and are required by law. This makes it possible to obtain better results and thereby increase acceptance of the applications.

What does this mean for the administrative work of employment and social protection services?

Discrimination may occur unintentionally when AI is used; if the data sets with which the AI is trained are distorted, for instance, perhaps because certain groups are over- or underrepresented (i.e. there is systemic bias). There is then a risk that the AI systems will reproduce and subsequently reinforce the existing inequalities from the analogue sphere. Unless something is explicitly done to prevent it, the results of AI systems reflect the discriminatory reality from which they are fed via the training and operating data. By examining data sets compiled from previous administrative practice, the process of introducing AI can also contribute to identifying existing discrimination and finding solutions for it. To avoid discrimination by AI-based systems, the quality of the training data and the AI models are therefore very important. Moreover, it is vital that the developers and users have the requisite competency and awareness for addressing these challenges. In order to identify discrimination by AI systems, the systems themselves must be sufficiently transparent and explainable (see Explainability & Transparency). Diversity at team level can additionally help to avoid discrimination or to identify it early on and eliminate it.

It is furthermore essential that AI systems be free of discrimination because once they have been introduced, they typically influence a large number of official decisions. If the outputs of an AI system were to be discriminatory, this would have an effect on every single one of those decisions. Conversely, if a system is sufficiently explainable AI, errors are easier to spot, and eliminating them improves all the use cases of the AI system. By making calculations repeatable and outputs reproducible, the use of AI can contribute to greater consistency in official decisions.

Explainability & Transparency

In the context of AI, explainability and transparency mean that the users and affected persons can understand, verify, and question the functionality and the outputs of AI systems. Depending on their role (e.g. AI developers, authorities and their staff, members of the public) and prior knowledge, this will require different types and amounts of information and explanations. This is the only way in which other values can be effectively implemented: by making it possible to recognize that data sets are biased, for instance, or that the AI application uses discriminatory parameters. At the same time this forms the basis for the human oversight and correction of the AI system. Explainability and transparency also mean that members of the public can always tell that they are dealing with an AI system (chatbots are identified as such) or that an AI system was involved in the decision-making process (even in a preparatory role).

What does this mean for the administrative work of employment and social protection services?

AI models should be designed in such a way that it is possible to explain, and thus to understand, how their recommendation come about. Especially in the context of employment and social protection services, it is likely that many different people will come into contact with an AI system. Developers and users within a public authority, supervisors and actors from other agencies, but also members of the public should be able to understand how the AI system works if they need to. Citizens should therefore be told every time an AI system is used to prepare a decision. Furthermore, members of the public should always be given easily accessible ways of finding out how the AI system arrived at it's conclusion.

Overall, the aim is to empower people when working with AI systems by giving them the necessary information. This may mean showing the relevant staff in a given agency what the error probabilities for the results are, or giving them the option of checking and correcting the results "manually" by accessing the data and documents behind the user interface.

Privacy & Personality rights

Privacy stands for a personal sphere, where everyone can realise their individual wishes freely and which is protected against public or state scrutiny. This protection is extended by fundamental right to personality rights: everyone is entitled to decide whether, when, and how data referring to them are used. Even if the information value of individual data points may only be slight, how they are handled can have a significant impact on the privacy and individual freedom of the person concerned (known as the "data subject"), depending on the purpose of the data collection and the links that are made between them. There is a particular risk when AI systems process personal data, because AI may create profiles of individuals and user types from data sets, evaluate them, and then make decisions that can have serious consequences for the people concerned. As a rule, state actors are not allowed to create personality or user profiles. If personal data are used for decisions that are likely to put the rights and freedoms of natural persons at serious risk, or if they are included in the typically large volumes of training data, then a risk assessment must be carried out for these data (data privacy impact assessment). Generally speaking, compliance with data privacy legislation must be ensured throughout the life cycle of an AI application, whereby the primary source of legislation is the GDPR.

What does this mean for the administrative work of employment and social protection services?

The employment and social protection services process personal data from members of the public that are often particularly sensitive (e.g. data on sickness, vocational training and careers, or information about their personal, family, social, and financial situation). Such administrative agencies therefore have to pay particular attention to protecting the privacy of data subjects and their right to control information referring to them, and to comply with the data privacy regulations that guarantee these rights. The principles for processing personal data must be applied when AI systems are used as well. Among other things, this means that the data can only be used for the purpose for which they were originally collected and for which they are necessary, and that they may only be stored for as long as is required for this purpose. The data subjects must also be told clearly which of their personal data are processed with the aid of AI systems and for which purposes. The data must be factually correct and up-to-date, to the extent that this is relevant. In addition, data that do not have to refer to a particular person must be anonymised. Employees responsible for processing personal data in the context of AI must be made particularly aware of the data privacy requirements. Functioning data privacy processes build trust and acceptance with data subjects at the same time, meaning that making AI systems privacy-friendly pays off twofold. The administrative staff of social protection services also have to be protected. It is therefore particularly important to reject AI systems being used for the surveillance of an organisation's own employees.

Security/Safety & Robustness

Security in terms of AI and IT more generally means that an AI system has to be appropriately protected against misuse, attacks, and security breaches (e.g. against hacking) and that there must be appropriate contingency plans to deal with security risks. Safety is used to refers to protecting the people interacting with the system.

Robustness means that the results generated by the AI systems are reliable and can be correctly reproduced under all circumstances, and that an AI system evaluates situations correctly (precision). This is extremely important for applications used by the Employment and Social Protection Services to determine whether the conditions for receiving certain benefits are fulfilled.

What does this mean for the administrative work of employment and social protection services?

In practice, security/safety means that a risk assessment is carried out in collaboration with the responsible departments and functions (e.g. IT Security), and that a system of protection is established. AI-specific risks include what are known as adversarial attacks, which are attacks aimed at manipulating training or operating data in order to distort the results.

For areas in which the failure of an AI system would have a serious impact, security becomes correspondingly more important. This applies particularly – but by no means exclusively – to critical infrastructures. Here, the usual precautions for IT systems must therefore be reviewed and taken as necessary, such as keeping back-up systems, using state-of-the-art technology, providing training for system users, and drafting fallback plans.

Intervenability & Responsibility

It must be possible to modify and shut down AI systems while they are running. Areas of responsibility for the planning, development, and deployment of AI must be clearly defined and assigned, so that it is clear at all times who is responsible and so that this person feels responsible and acts accordingly. What is particularly important – also in terms of the core objective of human-centricity – is that the final decision is always taken by a human being. The GDPR already states that data subjects have a binding right to human intervention on the part of the controller. The principle of human oversight has furthermore been incorporated into German administrative law and into the European Commission's draft AI Regulation.

From the perspective of members of the public, the right to human intervention means that their formal legal remedies (objection and litigation) may not be restricted by the AI application. It is therefore vital that the AI system is explainable and that clear responsibilities have been assigned. An additional channel for obtaining information and lodging a complaint about a public authority's use of AI may also be necessary or advisable in order to give members of the public opportunities to intervene and object, to the use of chatbots, for example, without the need for formal proceedings.

What does this mean for the administrative work of employment and social protection services?

In addition to the ability to modify and shut down the AI system at any time, the agencies operating the AI system must have staff with the necessary technical knowledge to ensure that they are in control of it. As soon as the need for intervention arises, a person must be able and authorised to make the necessary modifications and/or provisionally shut down the system. This calls for clear role descriptions, responsibilities, and decision-making authority, not only when the AI system is in operation, but during the development phase as well. Fallback plans, back-up systems, etc. may also be make sense in such situations (see Security/Safety & Robustness). The necessary knowledge should be acquired as broadly as possible by the employment and social protection services, so that their staff are able to interact with the AI systems on an informed basis, identify errors, and report them. At the same time, the "automation bias" must be acknowledged both when AI is being developed and during its use. This refers to the tendency of people when taking a decision based on an AI-generated recommendation or information to place excessive trust in the result provided by the AI. For this reason, human decision-makers must be able to understand and make an informed assessment of the result of the AI application.

Ecological sustainability & Conservation of resources

Ecological sustainability and the conservation of resources refer to a forward-looking, thoughtful use of natural resources and the obligation to safeguard and preserve the conditions for human life on the planet for future generations. Acting in an ecologically sustainable way means avoiding environmental pollution, maintaining biological diversity, and fighting climate change.

The development and use of AI consumes resources and emits greenhouse gases. The more often AI is used, and the greater the computing effort of individual AI applications, the more important it becomes to factor in sustainability aspects and resource and energy efficiency here as well. Sustainable AI covers both the use of AI for greater sustainability and the sustainability of AI itself. This can entail the construction and operation of energy-efficient data centres, for instance, or the development and implementation of machine learning models and AI systems that are less energy-intensive and have as long a useful life as possible.

Research into “green AI” aims to develop methods for reducing the amount of computing performed by an AI system in order to cut its energy consumption and help enable AI to be used sustainably.

What does this mean for the administrative work of employment and social protection services?

With regards to sustainable AI, the agencies’ administrations carry great responsibility for the planning, development, and purchasing of the systems. The most sustainable AI systems should therefore always be chosen from the technical options available at any given time. In this way, public authorities can use the market power of their purchasing programmes to boost demand for sustainable AI and thereby make an active contribution to climate action. Furthermore, AI can also be used for climate action in the public sector itself, as when AI is applied to technical facility management and helps to save energy.

3. Designing human-centric introduction processes & defining objectives

3.1 Introduction

The introduction of IT systems – which includes AI applications – starts by answering a number of key questions: what is supposed to be achieved, improved, or solved? How are these objectives to be achieved? Who will be affected and how? The process of finding answers to these questions should have an open outcome and avoid a predetermined choice of a particular technology. If it turns out that a learning system is a suitable tool, then the structure not only for the project management, but also for the design of the AI itself is defined in the introduction phase. It is therefore very important that AI projects are understood as being human-centric by design right from this initial phase, so that this specification can be implemented in the subsequent design process. The draft AI Regulation additionally stipulates that risk management measures and precautions to ensure human oversight are to be included in the conceptual planning phase.

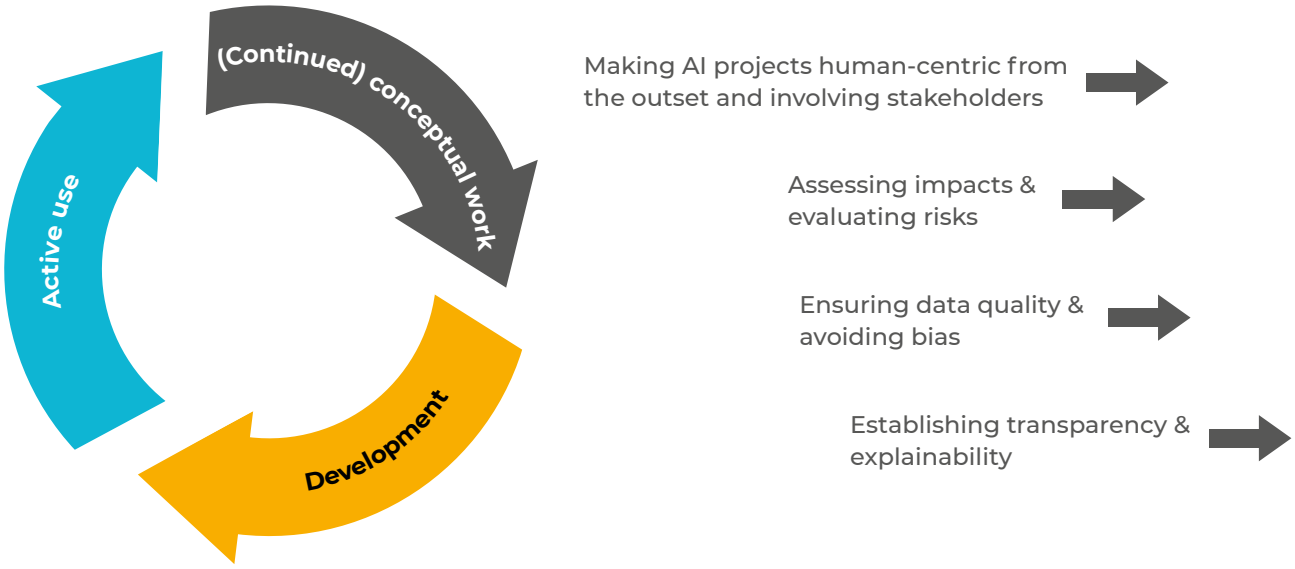
In simplified form, the configuration of AI systems can be broken down into three phases: the conceptual design, when objectives and scope are defined, the technical development of the system, and finally its operational use. Each of these phases can be divided into many smaller steps, in each of which measures must be taken to ensure that the system being developed is human-centric and value-driven.[3] As with the configuration of AI systems in general, the order in which the individual steps are taken depends to a large extent on the individual case. In a needs-based, agile[4] methodology these three phases are not necessarily completed consecutively, but may rather be revisited and entwined in a series of iterations. Furthermore, the design of AI systems does not stop when they are introduced. Their use and the reactions to the system and its impacts lead to further reflection, improvements, and continued conceptual work. AI systems and the contexts of their application may evolve over the course of their use. It must therefore be ensured that such systems are reviewed

[3] Cf. Puntschuh, Fetic (2020): *Algo.Rules: Handreichung für die digitale Verwaltung*, accessed from: https://algorules.org/fileadmin/files/alg/Handreichung_fuer_die_digitale_Verwaltung_Algo.Rules_12_2020.pdf. For information in English: https://algorules.org/en/home.

[4] This does not mean working strictly according to agile methods, but is intended to ensure open-mindedness and a focus on problem-solving. In practice there are many blended forms of agile and “waterfall” methods.

regularly and where necessary modified, depending on the context in which they are used and their risk assessment.

Later chapters of these guidelines look in more detail at key aspects of the value-driven introduction of AI systems: assessing risk and impact, ensuring data quality and avoiding bias, establishing transparency and explainability. All these aspects carry through all three phases of designing AI systems, and they must be addressed and examined right from the earliest conceptual phase.



The following section focuses on the initial phase, which starts with the conceptual work. This is where key success factors for the AI project are determined or affected – even if they can later be adjusted iteratively in the course of agile development. In particular, this is when:

- project goals are defined, especially overarching goals and mission statements as well as measurable and verifiable sub-targets;
- stakeholders are involved, in order to factor in their perspectives and create acceptance, and
- the entire process is outlined by selecting an appropriate process and participation design.

This results in an overview of the stakeholders and perspectives that are to be incorporated into the project, particularly in the initial phase.

3.2 Initial phase: plan AI projects human-centrically

AI projects can have different starting points, because the idea of launching an AI project can come about in a variety of different ways. It may be that the public authority wants to introduce AI on a trial basis, in order to gain experience with the technology. It is also possible that the staff have identified a concrete need for an AI system. Alternatively, a problem may have arisen that has to be solved and the final approach has not yet been decided, but AI is one potential instrument.

There are many questions to be clarified, stakeholders to consult, and basic decisions to be taken in the initial phase. The Network has drawn up general recommendations and questions to ask to aid in this process.[5] How important the individual questions are and in what order they should be answered depends on the starting point for the specific AI project. For example, depending on where the idea comes from or how the existing processes for introducing AI systems are structured, the questions for stakeholders and the core questions will have to be approached differently.

[5] The starting point here as well was the *Algo.Rules: Handreichung für die digitale Verwaltung*, accessed from: https://algorules.org/fileadmin/files/alg/Handreichung_fuer_die_digitale_Verwaltung_Algo.Rules_12_2020.pdf. For information in English: https://algorules.org/en/home.

3.3 General recommendations

- **Make participation as open as possible:** broad involvement makes for more successful AI projects. When relevant stakeholders have been involved appropriately, their perspectives can be factored into the development process. Co-creative introduction processes result in a better product, because errors are spotted sooner and requirements are captured better.[6] This also helps to boost acceptance of the system, increase the speed of uptake, and facilitate its productive use.[7] Involving the future users is particularly important in this context. Generally speaking, it should be ensured that all the stakeholders can participate effectively in accordance with their role. It may be necessary, for instance, to consider specific needs such as availability at particular times or a lack of financial resources, especially for marginalised groups or representatives of civil society. In certain situations it may be expedient or appropriate to provide expense allowances.

- **Analyse, model, and optimise workflows:** precise knowledge of processes is a prerequisite for optimising or (partially) automating them by means of AI systems. The analysis should make use of all the participants' process knowledge, particularly of the employees in the areas concerned. Before automation begins, however, it must be checked that the existing process is already of a reasonable quality, in order to avoid turning a poor process into a poor automated process.

- **Make development an open-outcome, needs-based process:** even if many fundamental questions are discussed in the initial phase, they must never be considered to be definitively settled. The first assumptions about requirements may turn out to be wrong, initial approaches to solving a problem with AI may not be very effective, and other opportunities may present themselves. The focus must always be on the problem to be solved, e.g. the defined needs of stakeholders. Conversely, the aim is not to make a specific approach or technology the starting point. These can and should always be adapted to the problem or objective at hand. For example, if the waiting times for an administrative service are too long, the objective should not be to make the staff work faster, or enable them to do so, but rather for members of the public to receive the service faster. To achieve this, the processes from submission of the request through to communication of the decision must be analysed and ways of optimising them devised.

- **Assess impacts and potential risks:** possible social consequences of the AI application for the persons affected and any conceivable impact on fundamental values such as democracy and the rule of law must be taken into account early on and potential risks categorised and measured (see Chapter 4).

- **Create diversity in all roles:** diversity in the development and project teams helps to make AI systems more error-resistant, fairer, and thus better designed. Diverse teams are able to identify potential sources of discrimination and bias earlier, for instance, and take action accordingly.[8] Diversity covers various dimensions, but particularly the professional and personal background of the team members. It is recommended that systems are not designed exclusively by computer scientists, but also by social scientists or organisational psychologists, depending on the project. In addition, the team should be diverse in terms of its members' social origins, immigrant background, and gender identity.[9] The aim is for the team to reflect the make-up of the society, especially that of the users and persons affected. It is particularly important that the public sector insist on diversity, in order to stimulate demand for corresponding teams from private sector partners and contractors, and in order to act as a role model for society as a whole.

---

[6] Krüger, Lischka (2018): *Damit Maschinen den Menschen dienen*, accessed from: https://algorithmenethik.de/wp-content/uploads/sites/10/2018/05/Algorithmenethik_L%C3%B6sungspanorama_final_online.pdf.

[7] Na et al. (2022): *Acceptance Model of Artificial Intelligence (AI)-Based Technologies in Construction Firms*, accessed from: https://www.mdpi.com/2075-5309/12/2/90/html.

[8] Cf. Iyer, Neema & Achieng, Garnett (2022): *Inclusion, Not Just an Add-On*, accessed from: https://pollicy.org/wp-content/uploads/2022/01/Inclusion_Not_Just_an_Addon_guide.pdf; and Livingston, Morgan (2020): *Preventing Racial Bias in Federal AI*, accessed from: https://doi.org/10.38126/JSPG160205.

[9] Other dimensions of diversity can be found in RAA Berlin (2017): *Diversitätsorientierte Organisationsentwicklung*, accessed from: http://raa-berlin.de/wp-content/uploads/2018/12/RAA-BERLIN-DO-GRUNDSAETZE.pdf.

- **Ensure transparent and regular communications:** in order to keep stakeholders and possibly also the general public sufficiently informed about the AI system, communication about the project should be proactive and open. This enables stakeholders to follow the design of the system independently and make contributions themselves.

- **Learn from one another:** there are many agencies and authorities in the employment and social protection services that are currently experimenting with the design and introduction of AI systems. it can be helpful at an early stage of the process to look at other projects that have taken a similar approach or are pursuing similar goals. Comparing notes[10] with the respective project teams can help to transfer knowledge from their projects to yours. In return, others can benefit from your past experience if you share it as well.

[10] In the AI labs run by the Network Artificial Intelligence in Employment and Social Protection Services, the participating representatives present their own AI systems and share their experiences of the development and operating process. Events and formats like this can help to share knowledge between public authorities and network the participants.

3.4 Checklist

1. Define the problem to be solved and the goals

What is the AI system meant to achieve?

2. Identify and involve stakeholders

Which stakeholders have which interests?
How should they be involved?

3. Design project structure

How can agile, open, and human-centric project management be ensured?

For the individual steps:

→ ***3.4.1 Defining the problem to be solved and the goals***

What is the AI system meant to achieve?

Questions to ask for determining and discussing the project's goals:

What is the problem to be solved and which goals should be reached?

- Which problem was the starting point for reflection?
- For whom is the problem to be solved?
- To which overarching goal is this intended to contribute?
- What other goals are to be achieved, e.g. economic, budgetary, or financial goals?
- Which work processes are connected to the problem to be solved?
  Which work processes within the organisation or by users are intended to change?
  In what way is the process to be improved? How can it be rethought?
  How should the process change from the perspective of staff?
  Which tasks do they want to carry out themselves, for instance, and which do they think should be automated?
- What role can an AI system play in the existing workflows?
  Is AI even a suitable means for solving the problem?

What impact could the AI application have on society or the fundamental rights of staff or individuals affected? (See Chapter 4) What consequences does this have for the goals?

How can it be ensured that the project achieves the goals?

- When is the project a success? How can this be measured?
- What are the conditions for this success? How can their fulfilment be ensured?

→ 3.4.2 Identifying and involving stakeholders

Which stakeholders have which interests? How should they be involved?

Questions to ask when identifying the stakeholders and their perspectives:

Who are the relevant stakeholders? (to be determined using the list below)

How are the stakeholders to be involved?

- What knowledge and which perspectives do the stakeholders contribute that the project can benefit from?
- What should their successful participation look like? Which formats are constructive, e.g. because they fit the project, the stakeholders, and the organisation's culture of work and collaboration?
- Who should be involved from the outset?
  Who should only be involved later, in the conceptual or development phase?
- Should they be involved on a permanent basis, for specific events, as needed, or at regular intervals?

What expectations and interests do the stakeholders have with regard to the project?

- How do the stakeholders view the project at its initiation?
  Is it perceived as being problematic per se?
  What does this mean for the participation processes?

Which impacts on stakeholders can be predicted?

- Which performance indicators (e.g. satisfaction of staff and members of the public, shorter processing times, number of applications processed per day) can be used to measure the impacts?
- How can they be used to make the definition of targets more specific?
- How does the project structure have to be designed?

3.4.3 Designing the project structure

How can agile, open, and human-centric project management be ensured?

Questions to ask regarding the involvement of stakeholders in the project:

Which basic project structure should be chosen?

- How are the existing structures in the agency or authority designed? What does this mean for the structure of this project?
- How can agile working be ensured?

How can the performance indicators be tracked over the course of the project?

Which forums and formats need to be created in order to enable participation?

- Do working groups, project advisory groups, etc. need to be created? What tasks do these have and who is involved with which perspective?

How can sufficiently broad participation be ensured?

- What is the target to be achieved with respect to the participation and diversity of stakeholders?
  Which groups must definitely be involved, for instance, and to what extent, and which groups can be involved on an optional basis?

3.5 Overview of perspectives and stakeholders: what do they contribute in the initial phase?

Various stakeholders should have a say in the design of the subsequent AI system and be involved in the development process. They each contribute valuable perspectives that can help to make the system better.

The overview below lists all the perspectives and potential stakeholders that are important in the initial phase. It is based on the guide *Algo.Rules: Handreichung für die digitale Verwaltung*, which the Network has expanded and adapted to the work of the employment and social protection services.[11] A specific perspective may come from a specific person or specific organisation/unit, but a person or organisation can also have several perspectives at the same time. The list can help to identify perspectives and stakeholders for the AI project at hand and ensure that they are involved in the conceptual phase and beyond in a structured way. The idea here is not that all the stakeholders have to be involved all the time or equally, but rather that they participate in relevant phases or topics.

[list in alphabetical order]

Affected persons perspective

- It considers how the use of the AI system impacts affected persons (data subjects), primarily their interests and/or fundamental rights.
- Examples: members of the public, job seekers, employees, member organisations. They may also be represented by intermediaries, e.g. interest groups.
- In initial phases, affected persons can be involved as experts regarding the impact of the AI application on their lives.

Coordination perspective

- It is the leading interface for the planning and development functions and interactions between developers, project owners, and implementers. It translates needs and goals into specific requirements and process steps and is responsible for their technical and practical implementation by the other participants. It is also responsible for communicating the project constructively within the agency/organisation.
- Examples: policy officers, project managers.
- The coordinator steers the reins from the outset and plays a key role in the early conceptual work on the AI system.

Data perspective

- It considers the work with and the management of the data sets that may be used for training and/or operation of the AI system.
- Examples: technical experts, possibly internal data science departments in public authorities, data analysts, data owners.
- In initial phases, these stakeholders can provide an expert overview of the available data and their quality, the time and expense of preparing them for an AI application, etc.

Data privacy perspective

- It ensures that data privacy requirements are met and provides advice on matters of data protection and the right to privacy.
- Example: data protection officer.
- Data protection officers can help in initial phases to determine whether and to what extent personal data are processed by the planned AI system and if this is permitted.

Decision-making perspective

- This ensures that the organisation's senior level is represented. It covers the allocation of resources (e.g. money, time, and personnel), the definition of higher-level requirements, integration with the overarching policy framework, and ownership of overall responsibility.
- Examples: heads of teams, departments, or organisations.
- Their support at the outset can enable or facilitate the necessary steps. In addition to questions of cost-effectiveness, the public perception of the AI project, measures taken to ensure its success, and embedding the AI project in overarching political and administrative strategies may all play a role.

[11] The starting point here was the *Algo.Rules: Handreichung für die digitale Verwaltung*, p. 9–10, accessed from: https://algorules.org/fileadmin/files/alg/Handreichung_fuer_die_digitale_Verwaltung_Algo.Rules_12_2020.pdf. For information in English: https://algorules.org/en/home.

Development perspective

- This considers the development of the AI system, including the model and the underlying software, in technological terms. The specifications made by coordinators are implemented.
- Examples: external private IT service providers, in the case of in-house developments, people from the organisation's IT department, product owners.
- Even if developers only develop the system at a later date, they must be involved from the outset in order to determine the technical feasibility, estimate time and expense, and determine the technical specifications.

IT and information security perspective

- It is responsible for ensuring the security of IT systems, especially against external attacks, and for ensuring the confidentiality, availability, and integrity of technical systems. This includes advising the heads of the agency/organisation and accompanying the AI development. The AI system cannot go into operation without its approval.
- Examples: IT security officers, information security officers.
- In initial phases, these stakeholders can assess aspects of the (technical) feasibility from the perspective of IT security and identify the applicable security standards.

Operational and implementation perspective

- It considers the organisational and technical implementation of the AI systems in existing processes. This includes linking the system with existing data and integrating it into an operational environment. It also monitors the professional use of resources and provides the software for users.
- Examples: IT departments of public authorities, technical administrators/operators, and external IT services providers if appropriate.
- In initial phases, these stakeholders can provide insights into organisational workflows and the technical infrastructure where the AI system is to be installed. They provide an overall view of the processes with regard to the users. They can also provide an assessment of key aspects of the technical and financial feasibility of the project and its operating costs.

Planning perspective

- Within individual institutions it determines the need for an AI system and formulates it, e.g. as product specifications, tender documents, or contracts. It also determines the goals of the software and plans how it will later be used. It, too, must consider the needs of the target groups.
- Example: division head. May be the same as the operational perspective.
- In the initial phase, the planners represent the perspective of the fundamental need for AI systems on the one hand, and on the other they formulate the initial product specifications, together with other roles, especially the coordinators.

Representatives for equality/women/diversity

- Focus on promoting and implementing equality and diversity in the agency/organisation. They often represent the interests of disadvantaged groups.
- Examples: equality officer, women's representative, equal opportunities representative.
- These representatives play an especially important role at the outset if the planned AI system particularly affects the interests of the groups whom they represent. Their expertise should be sought to determine if their groups are affected in such a way.

Representatives of people with disabilities

- They represent the concerns of people with disabilities and ensure their inclusion and equal treatment within the agency/organisation. They also work to achieve accessibility.
- Examples: representative of people with severe disabilities, ombudsperson of people with disabilities.
- These representatives play an especially important role at the outset if the planned AI system particularly affects the interests of people with disabilities. Their expertise should be sought to determine if their groups are affected in such a way.

Review and quality assurance perspective

- Ensures independent review and quality assurance. Depending on the context, project, and project status, it may do this in terms of a wide range of factors (e.g. user-friendliness, technology, freedom from bias). It can be performed by separate units and/or be covered by other roles, whereby its independence from the development unit, for instance, must be ensured.

- Examples: testers, quality assurance officers.

- In the initial phase, these stakeholders can pave the way for ensuring that sufficient testing and quality assurance measures are included in the project.

Staff council and staff perspective

- It represents the interests of staff vis-à-vis the heads of the agency or public authority.

- Example: staff council.

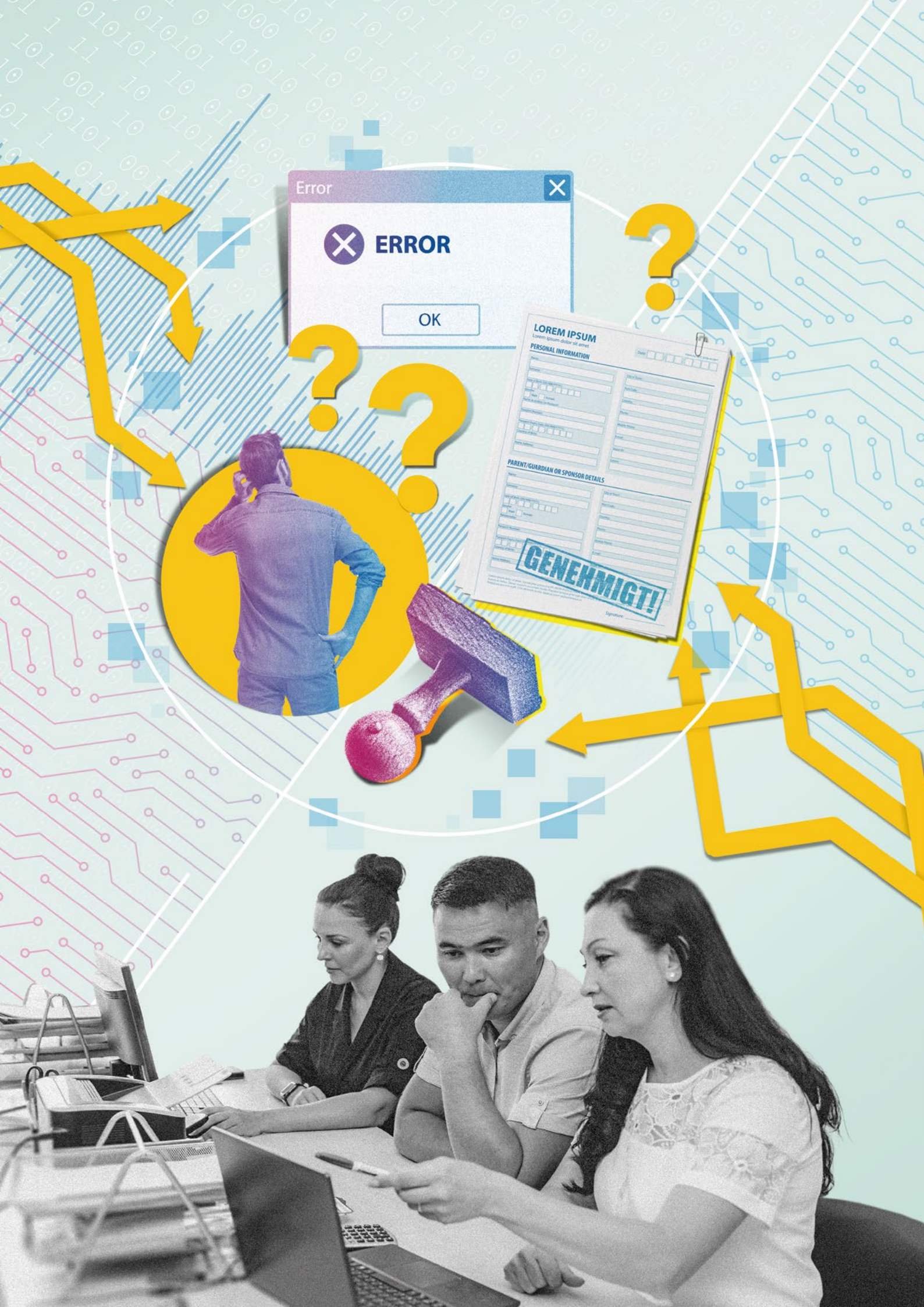
- In initial phases, the staff council can put forward the staff perspective (in some cases the staff council must by law be involved) and facilitate their successful involvement in the project as well as its communication to staff members. This makes it possible to identify any reservations on the part of staff, such as concerns about the monitoring of behaviour and performance, protection of staff data, job losses, or (too fast) changes to the world of work. For instance, it may be possible to agree on parallel training concepts for the subsequent users. Their involvement can make a key contribution to acceptance of the project.

User perspective

- This considers interactions between users and the AI system when it is in operation.

- Examples: staff in the public authority, such as those who operate or use the AI application in their department; members of the public who use e.g. chatbots or assistants for completing forms (who may at the same time also be affected persons).

- At the outset, users can be involved as experts regarding the processes to be optimised and potential changes to their world of work.


Error
ERROR
OK
LOREM IPSUM
Lorem ipsum dolor sit amet
PERSONAL INFORMATION
PARENT/GUARDIAN OR SPONSOR DETAILS
GENEHMIGT!

4. Assessing impacts & evaluating risks

4.1 Introduction

In the public sector it is particularly important when deploying AI applications within the applicable legal framework to assess the expected impacts and evaluate risks at an early stage in the process.

This practical impact and risk assessment, focused on specific applications and areas of deployment, is an essential step when planning the use of AI systems. If applications for social benefits are processed and ultimately decided by an AI system, for example, then a wrong decision can have far-reaching consequences for the persons affected. At the same time, there are also potential applications for AI systems where the adverse effects of a wrong decision are far less serious, such as when a chatbot merely provides non-binding answers and information from a government website in a different form.

It is absolutely vital to identify the potential risks, because they determine the demands made of the implementation process, the technical structure of the system, and its integration with existing processes and workflows in a public-sector agency, and make it possible to take corresponding measures to mitigate them. An initial assessment should therefore take place in the planning phase. Impacts should be reviewed again once the system is in operation, however, e.g. in the event of complaints, imprecise results, or errors, or if any changes are made to the AI system or the context in which it is being used. With AI systems that are deemed to be system-critical, it must also be ensured that reviews take place regularly without any concrete prompt and that the necessary adjustments are made.

To evaluate the potential impact and risks of using AI systems by public authorities, a procedure based on the "criticality matrix" devised by Tobias Krafft and Katharina Zweig,[12] for instance, can be used. In this model, the potential consequences and risks of AI systems are assessed and evaluated in two dimensions.

[12] Krafft, Tobias & Zweig, Katharina (2019): *Transparenz und Nachvollziehbarkeit algorithmenbasierter Entscheidungsprozesse: Ein Regulierungsvorschlag aus sozioinformatischer Perspektive*, accessed from: https://www.vzbv.de/sites/default/files/downloads/2019/05/02/19-01-22_zweig_krafft_transparenz_adm-neu.pdf. Krafft & Zweig later updated the layout of the matrix, so that the most-critical systems are shown in the top right corner, cf. http://aalab.cs.uni-kl.de/resources/img/RM_KI.png.

The dimensions are the potential harm caused by the use of the AI system and the degree to which the persons affected depend on the AI system being used. This assessment methodology therefore evaluates the impact of the AI application and enables the risks of the concrete AI application to be assessed in relation to the context of where it is used. The recommendations of the German federal government's Data Ethics Commission[13] and the draft by the European Commission for an EU Artificial Intelligence (AI) Regulation (COM(2021)206) also include risk-based approaches as essential elements. When it takes effect, the EU regulation will establish binding rules for risk assessment according to general, abstract criteria, which in turn correspond to a defined set of requirements. The concrete, practical risk assessment described below must be carried out within the general framework defined by the EU regulation and in particular may not fall below or result in any dilution of these standards. The risk assessment process devised by the Network on the basis of a criticality matrix was developed in anticipation of the AI Regulation and will be revisited as soon as the EU regulation has been adopted. The procedure described here must comply with the existing legal framework for the use of AI, as defined in Article 22 GDPR and Article 31a German Social Code X for automated processes, for example.

The potential for harm is determined by the question: what damage can the AI system potentially cause to individuals and society? The degree of dependence is determined by the question: how great is the dependence on the AI-based decision and what possibilities for re-evaluation are there? These questions are answered by means of the answers to further specific questions (see checklist on pp. 36 ff.).

[13] The expert opinion (from 2019) is available at: https://www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/it-digitalpolitik/gutachten-datenethikkommission.pdf?__blob=publicationFile&v=6.

Note:

The European Commission's proposal for an EU Artificial Intelligence (AI) Regulation (COM(2021)206) creates a category of prohibited AI systems (for some forms of "social scoring" by public authorities, for instance) as well as the category of "high-risk AI systems". To guarantee the necessary legal certainty, the draft regulation does not provide for a risk assessment for every system using a specific test in each individual case. Rather, the draft regulation classifies (on an abstract level) all AI systems intended to be used in several named application areas as high-risk AI systems. In the field of (social security) administration, for instance, it classifies as high-risk AI systems "intended to be used by public authorities or on behalf of public authorities to evaluate the eligibility of natural persons for public assistance benefits and services, as well as to grant, reduce, revoke, or reclaim such benefits and services". The draft regulation stipulates that once the regulation has come into effect, the European Commission may adopt what are known as "delegated acts" to classify other application areas for AI systems as high-risk; for this purpose it must perform a differentiated assessment of the impact and risks of the AI application for the specific purpose. An assessment of this kind also formed the basis for the original identification of application purposes deemed to be high-risk. According to the draft, there are concrete requirements and obligations for the providers and users of all high-risk AI systems, for example concerning risk management, data quality, and transparency. The risk management system required for systems classified as high-risk further includes a risk assessment of the specific system. In addition to the rules for high-risk AI systems, codes of conduct must be used for systems with lower risks. The draft regulation is currently being discussed by the Council and the European Parliament.

4.2 Checklist

1. Determine potential for harm

What damage can the AI system potentially cause to individuals and society?

2. Determine dependence

How great is the dependence on the AI-based decision and what possibilities for re-evaluation are there?

3. Determine position on the criticality matrix

What is the best estimate of the AI system's potential impact?

For the individual steps:

→ *4.2.1 Determining potential for harm*

What damage can the AI system potentially cause to individuals and society?

To determine the AI system's potential for harm, the possible consequences of a wrong AI-based decision are considered. The first step is to define a plausible worst-case scenario, whereby the probability of its occurrence is not important at this stage. It may be helpful to think of scenarios for different possible impacts and to work these through. This is particularly useful if the errors and their consequences can vary significantly and it is not possible to define one single appropriate case for the risk assessment. To determine the potential for harm there are two questions to ask, each with more specific sub-questions:

Impact on individuals:
Which individuals are impacted how and with what intensity?

- **Who is impacted?**
  The estimate here should cover the potential **types of persons affected** (e.g. applicants for benefits or users in the public authority) and the **number of persons affected**. Not only individuals or natural persons can be affected, but also legal persons (primarily associations and enterprises).

- **How are these persons affected?**
  To what extent are legitimate interests affected?
  This particularly refers to **fundamental rights and human rights, but other legal rights** (e.g. rights to social benefits) must also be considered. Key fundamental rights in the context of employment and social protection services include the right to choose vocational training and an occupation, the right to physical integrity, personality rights, especially the right to data privacy, the principle of equal treatment, and various non-discrimination rights (including rights related to gender, origin, age, and religious beliefs).

- **To what extent are individuals impacted?**
  The impact on individuals must be measured qualitatively. This means objective aspects (such as the amount of financial damage or the importance of a non-monetary benefit, such as participation in a physical rehabilitation or professional training event) and concrete individual impacts (e.g. the individual's dependence on the financial benefit, special personal circumstances, and the social consequences of withholding the benefit) must be taken into account.

Impact on society and public goods or basic principles:
To what extent does the system entail the direct or indirect, short or long-term risk of harming society as a whole or public goods?

- **To what extent is society affected "as a whole", above and beyond the level of impact on individuals?**
  This may be the case, for instance, if fundamental trust in the accuracy of official information is shaken, or if the AI system impacts larger societal processes, such as elections, employee representative bodies, public debate, or the fundamental relationship between employees and employers.

- **To what extent are public goods such as the rule of law, democracy, the welfare state, or the environment affected?**
  Digital technologies have both a direct and an indirect impact on society and thus may also present challenges for public goods or basic principles. When AI systems are used, it is therefore important to consider the impact they may have on the exercise of democracy or social justice.

4.2.2 Determining dependence

How great is the dependence on the AI-based decision and what possibilities for re-evaluation are there?

Dependence on an AI system is measured along the axes of switchability, (human) oversight, and redress. There are three questions to ask, each with more-specific sub-questions:

Switchability:
How easy is it to avoid the AI system or its decision?

- **Does switchability exist from the perspective of the public authority?**
  Can the process be carried out without the support of the AI system? How easy is it for users to make a decision without the support of the AI system? Is it possible to replace the AI system with another one? If no decision can be taken without the support of the AI system being used, then the dependence on this system is high. The absence of alternatives with which to replace the AI system also increases the degree of dependence.

- **Does switchability exist from the perspective of the members of the public?**
  Can members of the public avoid a public authority's AI system by changing to another public authority? Or can members of the public avoid the AI system within the process of a given public authority? Are there alternative channels for obtaining a benefit, for instance? Are there channels that do not rely on AI? How easy is it to access these channels?

- **Is it ensured that the review that takes place if an objection is raised (e.g. an appeal against a decision) is performed without using the AI system?**
  For an appeal in particular, it is important that a new evaluation actually takes place and that the same process is not simply repeated unchanged. A review of the substance of the individual case should be performed by a human being.

Human oversight:
To what extent are the decisions and actions taken by an AI system regularly checked by means of sensible human interactions?

- **To what extent is the output generated by the AI system verified in the course of the decision? What role does the AI system play in the decision-making process in which it is embedded?**
  The less the results of an AI system are checked by humans in the course of a decision-making process, i.e. the more autonomously the system decides, the more critical the evaluation of the system per the matrix. In this perspective, fully automated decisions would have maximum criticality.[14] If the system is structured to provide support for decisions, it must be ensured that the human oversight and decisions are actually effective. Here it is relevant, for example, that administrative officers acquire or have the information, time, and competences they need to check the outputs. It must also be examined whether the administrative officers still actually go through a decision-making process. To ensure that the results produced by the AI system are not just rubber-stamped, there are technical options (e.g. if the system supplies multiple indicators, which the staff have to use actively) and ways of structuring the social process. It is also important to ask in this context what happens in a public authority if an administrative officer goes against the suggestion made by an AI application.

Correctability:
Is it possible (and how easily) to challenge or correct an AI-based decision?

- **What possibilities do the persons affected have to challenge a decision? Are there legal remedies, for instance, or other options? How accessible are these options?**
  Is the person affected in a specific case actually able to make use of the legal instrument? This is not, or only partially the case, for instance, if the persons affected do not know about the possibility, are not aware that they are interacting with an AI, and/or the procedure is too complicated or time-consuming.

- **How effective are the options?**
  Can entire decisions be challenged, for instance, and is a full re-evaluation performed if an objection is made, or only a cursory review? **How long does the public authority need to process an objection properly? What is the situation for the affected persons while the review takes place?** During the review, the situation of the affected persons may deteriorate and further harm may be caused, perhaps because benefits are not paid on which individuals depend.

[14] For the overall assessment of such a system, the application context is also relevant, however. If a decision is barely relevant, an error can cause virtually no harm, and if the possibilities (for the affected persons and the users) to correct the results afterwards are very good, then such a system may be considered to be non-critical overall.

Spotlight:
How should effective oversight be judged when AI makes a preselection?

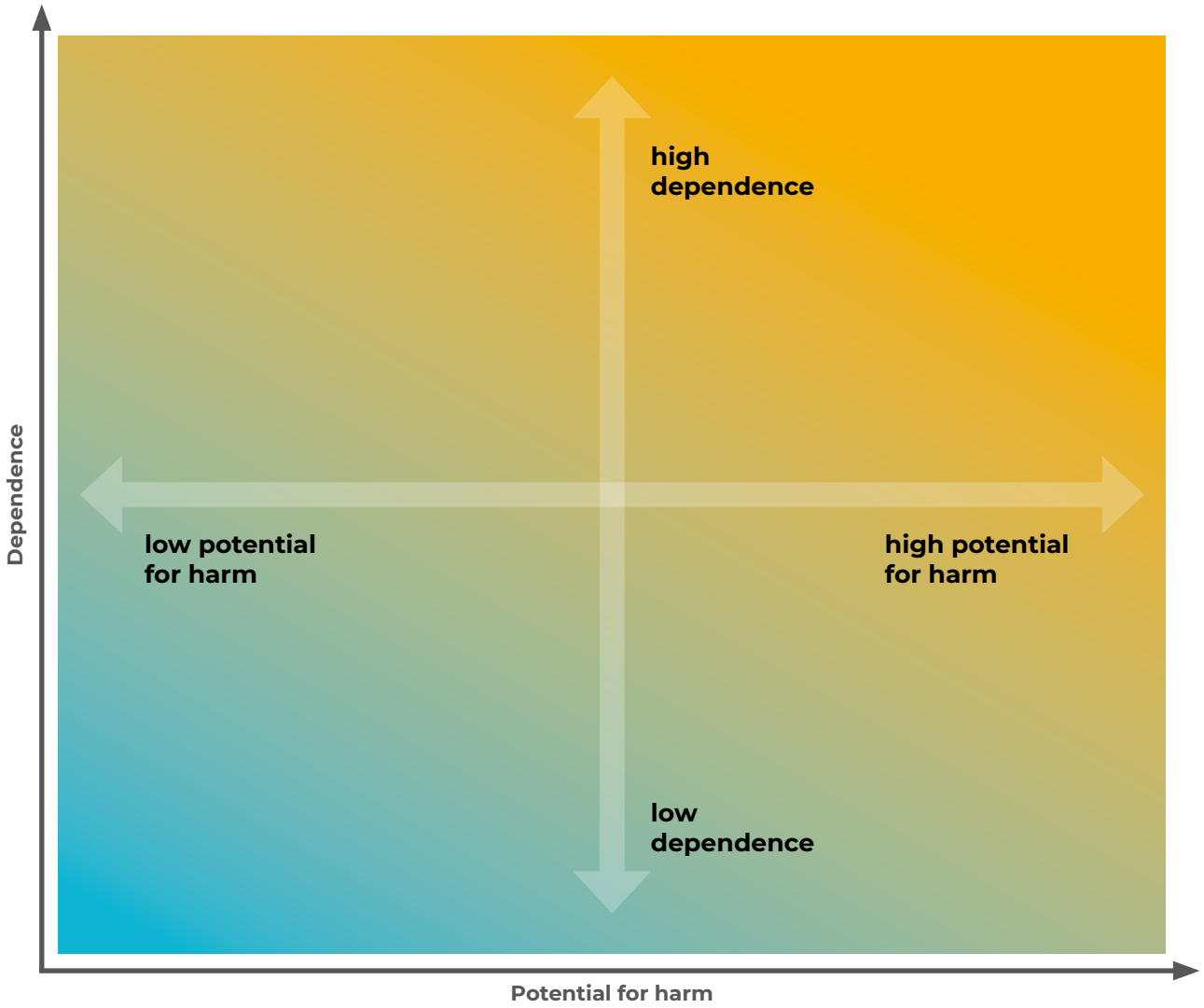
If AI systems prepare or presort decision, but each case is examined in full by a human being, then human oversight is maximised. By contrast, if the person making the decision has virtually no time and/or information or they do not have the competences to judge the preparatory work and therefore just pass the results through as a final decision (not just occasionally), then the human oversight is considered to be minimal. Between these poles, there are situations in which the human being does not check every case, but does process the preselected cases and checks them in full before taking a decision as well as making random checks on an informed basis of the other cases that were not chosen.

When AI systems presort cases, e.g. by drawing up a ranking, care must be taken to examine the need for human oversight of two separate AI outputs. On the one hand, it must be considered whether and to what extent the selected cases are reviewed and decided. On the other, it must be examined what impact the preselection has and whether it is also subject to human oversight, because the preselection determines the group of cases that are ever processed at all.

→ *4.2.3 Determining position on the criticality matrix*

What is the best estimate of the AI system's potential impact?

The position along the two axes is plotted using the answers to the questions above. The worst case is assumed for each answer.



4.3 Sample impact assessment

To provide a better idea of the different categories and borderline areas in the matrix, there are two realistic but fictitious examples below. They are intended to illustrate how the criticality level depends to a large extent on the context. On the dependence axis, the criticality depends largely on the extent to which the AI is embedded in the decision-making process.

Example 1 for dependence:
examining whether applications are complete

An AI system is used to check whether applications for benefits received online are complete. If an application is incomplete, it sends a warning. If this system is used while the application is first made and the warning is sent to the individual filling in the application, then both the potential for harm and the dependence are low.

Applicants can see immediately that this automated check is being used, can correct their mistakes, and then submit the application. If the AI system not only sends a warning, but also has to approve the submission, then the dependence increases slightly, because members of the public are then no longer able to circumvent the system. If the examination system is not used to support the application process but as part of the public authority's review, the level of dependence increases again, because the system now plays a role (albeit a minor one) in the decision-making process. If a warning is provided that the authority must review before sending a request to the applicant to submit missing information or documents, then the dependence is lower than if the system automatically sorts out incomplete applications and they are then effectively not reviewed before an automatic letter is produced requesting the missing documents. Dependence is higher again if the administrative staff do not have the resources or competences to verify the results of the AI system properly before the application is rejected. It is additionally problematic if proper verification entails serious obstacles for the processing officer (e.g. need to provide lengthy justification to superiors). Dependence increases considerably when the system decides fully automatically whether applications are complete, sends members of the public an automatic rejection, and it is not straightforward for them to submit a new application.

Example 2 for potential harm: payment of (social) benefits

This example only deals with the potential harm dimension and only considers the harm to individuals.

AI systems used for the payment of benefits are all the more critical the higher the amounts of money involved, the greater the recipients' dependence on the benefit, and the more people are affected. The system is more critical if the decision is about €1,000 rather than €100, if the recipients tend to be poorer, and if 500 people are affected rather than 50. If the benefits concerned ensure subsistence income (also for family members) then the system is more critical than for other benefits.

The practical challenge here consists of making a reliable estimate of who is affected by a wrong decision and how severely.

4.4 Measures to deal with high criticality: what are the consequences of the criticality assessment?

Higher risks generally mean that the demands made of the AI system are higher, too. There is not currently any definitive and complete description of the relationships between the AI system, its application context, and the measures that have to be taken. The draft AI Regulation from the European Commission, for instance, includes wide-ranging minimum requirements for high-risk AI systems and prohibits certain particularly critical applications. Within the framework defined by the AI Regulation, the use of the criticality matrix should be seen as a practical tool to assist public authorities with their risk assessment, which can be used to decide whether to use an AI system and develop safeguards. This does not preclude other requirements governing the use of AI systems, however, which must be met additionally and in full (either internal administrative rules or those based on other legislation). Measures and further assessment questions on topics such as data quality, explainability, and transparency can be found in the chapters which follow. Once the AI Regulation takes effect, these self-committing practical guidelines will also be reviewed and amended with a view to impact assessment.

5. Ensuring data quality & avoiding bias

5.1 Introduction

High data quality is essential for all data-based administrative activities. In terms of AI this means that all AI applications must have a sound data basis, i.e. a sufficient quantity of up-to-date, meaningful, representative, and accurate data. The concrete requirements for the data set must be determined on a case-by-case basis, depending on the context and the AI model. Conversely, poor data or insufficient data quality often means that no training can take place for the desired application, or that the outputs of the trained AI system are less precise and less reliable, and that more test runs are necessary. Moreover, poor data increases the probability of distorting the presentation of reality, i.e. of systemic bias. For this reason, the Commission's draft AI Regulation additionally includes binding quality criteria for the training, validation, and testing data sets that must be used when a high-risk AI system is being developed.

Ensuring high data quality when developing AI systems not only improves the specific AI application, but can raise the overall quality of data within an administrative unit, too, and thereby enable further data-based applications (e.g. business intelligence applications, dashboards, etc.).[15] As soon as personal data is involved, the requirements of data protection law, which are also data quality requirements, apply. They include lawfulness, purpose, data minimisation, and confidentiality as well as the principle that the data must be accurate, up-to-date, and complete. This chapter focuses on the relationship between data quality and bias. Unless action is actively taken to prevent it, AI systems reproduce any socially constructed biases inherent in the training data. If a decision, e.g. on social benefits, depends substantively on the output of an AI system, it must be ensured that the underlying data are not biased in terms of gender, ethnicity, religion, or age, for instance. High data quality lowers the chance of bias and therefore reduces one cause of discriminatory AI decisions, against which the Basic Law of Germany requires particular protection and which is defined more specifically in the German General Equal Treatment Act (AGG) and in the German Social Code (see value "non-discrimination").

[15] This generally entails feeding the cleaned data back into the source of the original data, e.g. the specialist processes.

What is bias?
Why is it important in the context of discrimination?

"In computer science, the term bias refers to an error that is the result of a systematic distortion. Since AI systems work on the basis of learned correlations, the characteristics of the data used to train them are generally responsible for creating bias in AI systems."[16] *This is the definition used by the AI Study Commission of the German Bundestag in its final report. Bias is a significant challenge when working with data, especially for machine learning systems.*[17] *An AI-based system is only as good as the data on which it was trained, true to the general rule of "garbage in – garbage out".*

Bias occurs, for example, when a data set presents a distorted view of reality. For instance, a data set may include significantly more data from men than from women, although the total statistical population is balanced. It is therefore important for review purposes to know the structure of the population to which the AI application relates (e.g. the proportion of men and women in the entire workforce). A bias within the data results in discrimination "when the data selection causes a systematic error by the AI system, so that some people are treated more favourably or less favourably without justification due to their external and internal personal characteristics".[18] *Bias also occurs when an AI system reproduces existing discrimination.*[19] *This means that in these cases even if the data is an accurate reflection of reality, it results in discrimination. Generally speaking, an AI system will reproduce existing discrimination unless something is actively done to stop it.*

[16] Final Report of the German Bundestag's Study Commission on Artificial Intelligence. Printed Paper 19/23700 (in German), p. 60. An English translation of the Executive Summary can be accessed here: https://www.bundestag.de/resource/blob/804184/f31eb697deef36fc271c0587e85e5b19/Kurzfassung-des-Gesamtberichts-englische-Uebersetzung-data.pdf.

[17] Bias may also play a key role in AI learning processes for generalising patterns, cf. Final Report of the German Bundestag's Study Commission on Artificial Intelligence. Printed Paper 19/23700 (in German), p. 60 with further references. An English translation of the Executive Summary can be accessed here: https://www.bundestag.de/resource/blob/804184/f31eb697deef36fc271c0587e85e5b19/Kurzfassung-des-Gesamtberichts-englische-Uebersetzung-data.pdf.

[18] Final Report of the German Bundestag's Study Commission on Artificial Intelligence. Printed Paper 19/23700 (in German), p. 61. An English translation of the Executive Summary can be accessed here: https://www.bundestag.de/resource/blob/804184/f31eb697deef36fc271c0587e85e5b19/Kurzfassung-des-Gesamtberichts-englische-Uebersetzung-data.pdf.

[19] Final Report of the German Bundestag's Study Commission on Artificial Intelligence. Printed Paper 19/23700 (in German), p. 61. An English translation of the Executive Summary can be accessed here: https://www.bundestag.de/resource/blob/804184/f31eb697deef36fc271c0587e85e5b19/Kurzfassung-des-Gesamtberichts-englische-Uebersetzung-data.pdf.

There are many causes of bias.[20] They may stem from "biased" data collection processes, which do not take a representative sample of the population, perhaps because certain groups are overrepresented in the sampled cases or disproportionally take part in surveys[21] or because sensors for gathering data are not distributed in a representative way. In the case of AI, bias may also occur via the selection of training data as well as during its operation via the selection of operating data.

If the sample chosen is not representative, then even a data set that was originally representative will be distorted or (for operating data) generate a distorted output.

In addition to bias in the data, there are many other kinds of bias, e.g. cognitive, statistical, and inductive bias[22], which may also have an effect on the design of AI systems.

In order to identify and avoid or correct bias and discrimination, it is necessary to look at the data, the learning, and correction processes, and the socio-technical environment in which the AI system is embedded.

[20] Cf. other causes of discrimination: Innovation Office of the Federal Ministry for Family Affairs, Senior Citizens, Women and Youth (no year of issue): *Wegweiser digitale Debatten*. *Teil 2: Algorithmenvermittelte Diskriminierung*, p. 7, accessed from: https://www.bmfsfj.de/resource/blob/186300/961021829a491933cf24e8f06ff8018f/wegweiser-digitale-debatten-teil-2-data.pdf.

[21] A striking acronym is sometimes used: "WEIRD sample" (western, educated, industrialized, rich, and democratic society).

[22] Cf. overview in *Bias in algorithmischen Systemen – Erläuterungen, Beispiele und Thesen der Initiative D21*, accessed from: https://initiatived21.de/app/uploads/2019/03/algomon_denkimpuls_bias_190318.pdf.

Examples of data bias

There are many well-known recent examples of biased data, like the AI application at Amazon that assigned a score to job applicants and substantially favoured men because the system had learned from past data.[23] The gender of the applicants was not explicitly used in the training data, but there were other characteristics in the applicants' CVs that were strongly correlated with gender, from which the AI was able to infer their gender. If the people responsible identify discrimination of this kind towards a particular gender, they must ensure that the bias is corrected when further data are obtained and analysed.

Example of bias due to different organisational structures for (company) medical care in two groups: the chemical industry and the leather industry

The German Occupational Accident Insurance Fund analysed how many (and which) occupational illnesses occurred in each industrial sector. They collected the cases of occupational illnesses that had been identified and compared them with the number of employees. The result was that the chemical industry had significantly more occupational illnesses (per employee) than the leather industry. However, this result is due to an overrepresentation of illnesses in the chemical industry, which in turn stems from the different ways the case numbers were collected. The healthcare network in the chemical industry was much more tightly organised, with regular examinations by company doctors, for example, which did not exist in the leather industry. In many cases, this made it possible to identify the occupational illnesses in the first place. Unless they are corrected, the data suggest erroneously that the chemical industry is more "dangerous" than the leather industry.

[23] Cf. *The Guardian* (2018): "Amazon ditched AI recruiting tool that favored men for technical jobs", accessed from: https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine.

5.2 Checklist

What steps are needed to ensure good data quality?

The steps below provide some initial guidance for obtaining high-quality data. They do not necessarily have to be done in the order shown below; rather, they are a starting point for accompanying AI processes from the perspective of data quality.

1. ***Define objectives of the AI application, (data) requirements, and application-related data quality criteria***

   Which problem is to be solved and what (quantity of) data is needed in what quality?

2. ***Identify available data and assess data quality***

   Do the available data meet the defined quality requirements?

3. ***Prepare and clean the data***

   How can the data be prepared to ensure they have the necessary quality?

4. ***Find and eliminate bias***

   How can the staff (principally data controllers, but also users) be made aware of bias? What (technical) means of avoiding bias are there? How can the discriminatory impact of bias be prevented?

For the individual steps:

→ ***5.2.1 Defining objectives of the AI application, (data) requirements, and application-related data quality criteria***

Which problem is to be solved and what data is needed in what quality?

Defining objectives of AI application and data requirements
The objectives of the AI application and its planned application context have a decisive impact on the type, scope, volume and quality of the necessary data. The AI model used and the way in which it learns from the data also determine the requirements for the data set. Many machine learning models currently need large amounts of data, with very different characteristics depending on the use case. With reinforcement learning, it is possible to start with a relatively small data set, but greater input is required during operation.

Determining data quality requirements
There are many quality parameters for data and data sets. They range from accuracy and completeness to currentness and consistency. The diagram below provides an initial overview of some standard requirements, but is not intended to be exhaustive.[24]

Different AI applications may make very different demands of the data quality. One may need a large volume of historical data, for instance, to use for training purposes, in order to trace developments over many years and identify patterns. The criteria "up-to-date" and "complete" would therefore be weighted differently. In another case, the aim is for files (consisting largely of text) to be scanned directly and with as few errors as possible, so the criteria of "machine-readability" is particularly important here.

up-to-date	**error-free**	**exact**	**consistent**	**standardised & structured**	**machine-readable**
representative	**extensive & granular**	**trustworthy**	**reliable**	**comprehensible**	**complete**

→ *5.2.2 Identifying available data and assessing data quality*

Do the available data meet the defined quality requirements?

Identifying available data and verifying the quality
The necessary data and quality requirements defined in the first stage are now compared with the available data sources and data:

- Which data sources with which data are available for the specific development? Many different data sources may be relevant, whereby the authorities' internal software programs are a typical source. In addition to their own sources, other data may be provided by partner organisations, public sources (e.g. official statistics or open data portals), or purchased from data providers. In this case it will be necessary to clarify the conditions for access to and use of the data.

- Are the data of the necessary quality? This assessment is vital and must be all the more thorough the less is known about the data or their source. It is particularly important with external data sources, when it must be ensured that the data provider is sufficiently trustworthy. The available data must then be evaluated using the catalogue of requirements that has been developed. This is followed by the preparation and cleaning of the data.

[24] A definition of terms can be found in the glossary. Cf. Fraunhofer FOKUS: *Leitfaden für qualitativ hochwertige Daten und Metadaten*, p. 14 ff., accessed from: https://cdn0.scrvt.com/fokus/e472f1bf447f370f/32c99a36d8b3/NQDM_Leitfaden-f-r-qualitativ-hochwertige-Daten-und-Metadaten_2019.pdf.
The FAIR principle attempts to bring together the main criteria in the context of making the data widely accessible (the criteria are not an aspect of data quality in the narrow sense): findable, accessible, interoperable, and re-usable.

→ 5.2.3 Preparing and cleaning the data

How can the data be prepared to ensure they have the necessary quality?

Good data preparation is an essential and often time-consuming[25] process of cleaning and qualifying the data for their further use.

> ***What does data preparation entail?***
> ***How is it different from data cleaning?***
>
> *Data preparation includes data cleaning. The data are put into a standardised, machine-readable format, faulty data are eliminated, e.g. by removing duplicates, correcting errors, or adding missing data (using data augmentation techniques). They are then prepared for the use for which they are intended (e.g. for analysis or as training data for machine learning). This preparatory work can take place in many different ways, e.g. by reduction (aggregation or generalisation) or by structuring the data and storing it in high-performance database systems.*

Stages of data preparation
The following steps are vital for data preparation:[26]

1. **Formulate standards:**
   Define the relevant quality requirements for the use case, identify any conflicting goals, and translate these into operable standards for evaluating the available data.

2. **Integrate data:**
   The data from the different sources are transformed and merged.

3. **Evaluate and validate data:**
   Errors and discrepancies in the data compared with the standard are identified here.

4. **Plausibilise and impute data:**
   If the data are wrong, missing, unreliable, out-of-date, or similar, they are replaced whenever possible by correct values or the false data are removed. The plausibility of the corrected or imputed data should then be evaluated again.

Depending on the application context, it may be necessary to carry out the steps in a different order or in several iterations.

---

[25] Cf. Press, Gil (2016): "Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says", *Forbes*, accessed from: http://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says. This is described in Prof. Naumann's presentation in AI Lab #5, cf. video, accessed from: https://www.denkfabrik-bmas.de/projekte/ki-in-der-verwaltung/ki-labs-zu-datenqualitaet-und-datenreinigung-fuer-ki-anwendungen, and slide 5 of the corresponding slide show.

[26] Cf. *Qualitätshandbuch der Statistischen Ämter des Bundes und der Länder*, 2021, p. 102 ff., accessed from: https://www.destatis.de/DE/Methoden/Qualitaet/qualitaetshandbuch.html); and Naumann (2021): "Datenqualität und Datenreinigung für KI-Anwendungen", presentation in AI Lab #5 by the Network Artificial Intelligence in Employment and Social Protection Services in the Policy Lab Digital, Work & Society, BMAS, accessed from: https://www.denkfabrik-bmas.de/projekte/ki-in-der-verwaltung/ki-labs-zu-datenqualitaet-und-datenreinigung-fuer-ki-anwendungen.

Causes of low data quality
There can be many reasons for errors, for example:

1. *Faulty data collection or entry*
2. *Different formats from different data sources*
3. *Incomplete data sets (empty cells or entire columns/rows)*
4. *Inconsistent data types (numbers vs. letters)*
5. *Different scales/units*
6. *Different category names*
7. *Several tables in one file*
8. *Errors in data preparation, e.g. in aggregation*
9. *Diverging structures of CSV files (e.g. differing numbers of columns or rows)*
10. *False attribution of values to variables (e.g. when a foreign document is translated)*[27]

Data preparation methods
There are various methods of preparing data. The chosen method must fit the data types and formats, the AI model, the causes of errors, and the aim of the preparation.

One option is to aggregate the data. This does not rectify errors, but the aggregation process generates the appropriate level of granularity for the respective application context. Another option is data augmentation, i.e. a method that fills in missing data with synthetic data. If data are missing for a group or object class (women, people from ethnic minorities, etc.), this method can create synthetic, but "authentic" data to ensure that the total statistical population is balanced or sufficiently large, e.g. for training purposes. Data cleaning processes are also carried out (e.g. ensuring a harmonised machine-readable format, eliminating duplicates).

The limits of the respective method must be taken into account when preparing data. If synthetic data are created by means of data augmentation, for example, then it must be ensured that the data set remains representative overall. This requires in-depth knowledge of the initial data and an understanding of the method and corresponding evaluation processes.

[27] Cf. the many examples in Prof. Neumann's presentation in AI Lab #5, video accessed from: https://www.denkfabrik-bmas.de/projekte/ki-in-der-verwaltung/ki-labs-zu-datenqualitaet-und-datenreinigung-fuer-ki-anwendungen, and slides 7 ff. of the corresponding slide show.

5.2.4 Finding and eliminating bias

How can staff be made aware of potential bias?
What (technical) means of avoiding bias are there?

Bias can occur throughout the AI life cycle (planning, development, introduction, and operation),[28] whereby for machine learning systems it is in the training phase that any data bias is transposed into the decision-making rules for the AI system. Care must be taken in the operating phase as well that the data is representative; partly in order to prevent the trained algorithm from generating discriminatory outputs, but also and especially if the operating data are used to continue training the AI.[29] Only by evaluating the AI system on an ongoing basis can it be ensured that subsequent risks are identified and the necessary adjustments can be made to the AI system to eliminate bias. In administrative applications, it is vital to avoid bias, because depending on the context in which the AI system is being used, bias in the data may result in discrimination.

Encourage staff awareness and build in feedback loops

In the context of AI, the aim must be to examine the underlying data critically and avoid or eliminate bias. Staff must be trained accordingly, depending on their roles. At the same time, they should additionally be made aware that AI systems are fundamentally fallible, so that they can adopt a critical approach and avoid what is known as automation bias.[30] Having a diverse team can also help to spot, avoid, or correct bias. Involving different stakeholders at an early stage of planning and development can provide important inputs in this area as well. For example, bodies responsible for equal treatment and anti-discrimination can be involved in the context of identifying and avoiding discrimination.

Furthermore, depending on the risks of bias and discrimination, testing and feedback loops should also be taken into account and included in the development, introduction, and operation of the AI system. Particularly at the start of training, the total statistical population, and thus the relevant basis for identifying any bias in the data, is not or not precisely known. At the same time, these feedback loops make it possible to record and report the system's impact, especially on the affected persons, which can be taken as the starting point for making changes.

Methods of identifying and dealing with bias

Automated methods for identifying bias in the data are currently still in a trial phase.[31] They will certainly be an important detection tool in future. In certain application cases, however, recognising bias means knowing the structure of the total statistical population, which is not always the case in practice.

One way of preventing bias is to limit the training data for the AI to what is actually needed. This is because characteristics that are not relevant to the aim in question may cause bias in machine learning systems, since the AI system (also) recognises patterns based on these irrelevant features.

---

[28] Cf. chapter 3.

[29] For instance, if the data are to be used for subsequent changes or the AI is self-learning.

[30] This refers to people's tendency to prefer proposals from automated decision-making systems and to take any contradictory information less seriously.

[31] Developers at IBM have created an open source toolbox that can be downloaded from https://aif360.mybluemix.net/ or via Github, https://github.com/Trusted-AI/AIF360.

For instance, if gender is not important for an application use case, then this characteristic should not be included in the training data set so as to avoid any reference being made to it. At the same time, care must be taken that gender is not captured indirectly via other characteristics. Deciding which characteristics are necessary requires knowledge of AI learning processes as well as in-depth understanding of the application context (particularly the tasks and processes) and the available data.

If bias is identified during the development process, it can be eliminated by subsequently deleting the data characteristics that produced the bias or discrimination. However, this is not guaranteed to be successful if these characteristics are correlated with others. Here, too, good knowledge of the data set and interdependencies between data is called for. The consequences of bias in the data and/or discriminating outputs from AI systems applied to past decisions, such as administrative decisions and orders, have to be examined on a case-by-case basis. Do they make the administrative order unlawful or invalid, and if it is unlawful, can it be rectified or does the order have to be withdrawn or revoked?


Hauptantrag
Antrag auf Leistungen zur Sicherung
nach dem Zweiten Buch Sozialgesetzbuch
©BCE ECB

6. Establishing transparency & explainability

6.1 Introduction

How an AI system works and how a given output is generated must be understandable and comprehensible enough that different target groups (e.g. users in the public authorities, members of the public as affected persons, employee representatives, or interest groups from civil society), depending on their role, can operate the system correctly, understand and make further use of the results, or challenge and review the system.[32] This is particularly important for public authorities because administrative activities must always be transparent, explainable, and justifiable for members of the public. AI systems that are transparent and explainable create trust and acceptance of public administration activities among members of the public and among staff members with regard to the use of AI in the public authorities.

Depending on the application, various measures are necessary to make the operating methods and decisions of AI systems understandable and comprehensible.[33] If AI systems use rule-based models with fixed criteria, such as simple decision trees with a small number of branches and few levels, and these are known or made public, then decisions taken by such systems are relatively easy for users, affected persons, and other target groups to understand and interpret. This means that the data used and the methodology must be transparent and presented in an appropriate way for the user group. Models like these, based on comprehensible inputs, are known as white-box models. They are contrasted with black-box models, such as neural networks, which are not intuitively understandable for human beings, even if the data and how the model functions are transparent. With AI systems based on black-box models it is necessary to take further steps to explain the decisions taken by the system. Various methods exist to do this, such as explaining in a way that humans can understand the factors that have a significant influence on the outputs of the AI system. These approaches are often known collectively as "explainable AI".

[32] Cf. *Algo.Rules: Handreichung für die digitale Verwaltung* (2020). *Algorithmische Assistenzsysteme gemeinwohlorientiert gestalten*.

[33] Cf. iit-Institut für Innovation und Technik in VDI/VDE Innovation + Technik GmbH (2021): *Erklärbare KI*: *Anforderungen, Anwendungsfälle und Lösungen*, p. 20–21.

A distinction is made between explanations of how the AI system works at a generalised level and explanations of individual decisions. Explanations of general functionality (known as model explainability or global explainability) help users, the affected persons, and other target groups to understand how an AI model functions overall, for instance by describing the interactions or connections between the data used in a manner appropriate for the groups concerned. As a rule, however, such general explanations do not make it possible to reconstruct how the individual outputs of an AI system are reached, which is why additional explanations of individual results (known as local explainability or data explainability) have to be provided. At the same time it is important to protect the personal data of other data subjects (e.g. other job applicants to a given position, other people requesting public benefits).

There is a statutory requirement to make AI systems explainable, defined particularly in the GDPR rules on information obligations and access rights, which call for "meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject" when personal data are processed (Art. 13 [2] point f, Art. 15 [1] point h). Even if the GDPR only applies to the processing of personal data, in this context and with respect to the explainability of AI systems it means that the main factors affecting individual decisions must be shown.

Moreover, the provisions of the GDPR quoted above on the need for explainability only apply to fully automated decisions. AI-based decision-support tools are not covered by these provisions of the GDPR, which is a significant gap. Due to the potential for automation bias and the fact that recommendations and decision-making support from an AI system may have a similar impact on the members of the public affected as a fully automated decision, transparency and explainability should also be ensured in cases like these. Citizens and staff members should be fully informed about the use of AI so that they can recognise when an AI system is being used or was involved in a decision-making process (see Explainability & Transparency in the fundamental values section on p. 19).

Depending on the type of AI application, this can be achieved by identifying the AI application directly (as with chatbots, AI-based assistants for completing forms, or other AI applications with which members of the public interact directly), via a note in official correspondence (if AI applications were used to prepare or support processing), or even centrally on a website (if AI applications are used in other areas of the same authority that do not affect the processing of benefits applications, such as for advance warning of server overloads). As a supplement to markings and notes in official correspondence, the website, too, can provide more detailed information about the AI applications used.

For members of the public, explanations of the specific decision concerning them are often sufficient. But more detailed explanations of the AI system should also be prepared so that developers and users of AI systems can check and ensure that it is functioning correctly. It should be taken into account here that people have differing levels of technical knowledge. Staff members should receive training to ensure that they have the competences needed to deal with explanations of AI systems. No matter how intuitive an AI system is and how easily understandable the explanations, they can never replace proper training. Training courses help staff to identify any operating errors, for example, and give them confidence when using the application.

The following four-stage checklist is intended to help draft suitable explanations for core target groups (e.g. developers, users, and affected persons). For each stage, the checklist contains several questions that should be asked. It can therefore help to make concrete and sensible application of any binding rules stipulated in the AI Regulation regarding transparency and explainability towards users and affected persons.

6.2 General recommendations

- If in a given use case it is possible to use an easily understandable AI model (white-box model, e.g. decision tree based on clear input variables), then this should be preferred over an equally suitable but less easily understandable AI model (black-box model, such as a neural network).

- If AI models are used that are not understandable, a variety of methods of explanation can be applied. When possible, prototypes or counterfactual explanations[34] should be used, because these are intuitively understandable, particularly for users.[35]

- Tests should be carried out before introduction and regularly thereafter to determine whether the explanations provided are suitable and sufficient, and changes made as necessary.

[34] Counterfactual explanations show what the smallest necessary changes in the input variables are in order to reach a different result.

[35] A guide to the most common explanation strategies can be found in a study by the iit-Institut entitled “Erklärbare KI”.

6.3 Checklist

1. ***Define target groups and the requirements they have of the explanations***
   Who are the core target groups and what has to be explained to them?
2. ***Explanation of general functionality***
   How can the general mode of operation of an AI system be explained to the respective target group?
3. ***Explanation of the concrete decision in a specific case***
   How can the individual decision taken by an AI system be explained to the respective target group?
4. ***Determine the explanation strategy***
   Which aids can be used to facilitate explanations for various target groups?

For the individual steps:

→ ***6.3.1 Defining target groups and the requirements they have of the explanations***
Who are the core target groups and what has to be explained to them?

It should be remembered that the target groups make different demands of the explanations and also differ in terms of their knowledge of AI systems. Users, for example, must be able to detect any errors in the results that could potentially result from a false data entry. However, users (often) have limited technical background knowledge about AI systems, so explanations for them have to be designed for this target group and easy to understand. By contrast, explanations for developers can assume that they have mathematical, statistical, and/or technical knowledge. In both cases, users and developers should receive relevant training. This is necessary in order for them to be in a position to ensure and verify that the AI system is functioning correctly.

Users, developers, and affected members of the public should be the main focus of the explanations, but other key target groups should be taken into account as well. It may be challenging to draw up easily understandable explanations for target groups whose background knowledge of AI systems can vary widely, as is the case with members of the public. It is nonetheless important for ensuring acceptance. It is also helpful to enable users to provide members of the public with additional explanations on request and if they have any queries.

Questions to identify the stakeholders and their needs:

Who are the core target groups?
The core target groups depend on the context in which the respective AI system is used. Stakeholders involved in the introduction process must be included (see Chapter 3.5) as well as target groups taking part in the subsequent development process or during the system's operation. The starting point for determining the key target groups are the following groups:

- **Developers who**
  - design, conceptualise, and implement the system
  - service the system at the public authority after it has been implemented
  - review and test the system's design and functionality
- **Users in the public authority**
- **Members of the public affected by an AI decision or as users of an AI**
  - possibly sub-groups of affected persons, e.g. employees, families, immigrants
- **Internal functions**
  - decision-makers/senior leaders in the public authority
  - staff representative
  - equal opportunities officers
  - disability rights officers
  - controlling/internal audit
  - data protection officers
  - information security officers
- **External functions**
  - supervisory bodies
  - courts of law
  - civil society actors (trade unions, trade associations, NGOs, etc.)

What information required for the explanations has to be communicated?
Information that has to be communicated may include:

- **System**
  - aims of the system
  - known limitations
  - design decisions
  - assumptions
  - models
  - algorithms
  - training methods
  - quality assurance processes
  - information security measures
  - data protection measures

- **Data used**
  - time and place of data collection
  - reason for data collection
  - scope of data collection
  - method of data collection
  - composition of data set, representativity

- **Application**
  - application
  - processing
  - degree of automation and how the system is embedded in the decision-making process

What demands do the target groups make of the explanation?
Determine what has to be explained for each target group:

- decisive factors for an individual decision, in compliance with data privacy rules, particularly as relating to other data subjects
- models used and how they work
- data used in the model
- data used for the individual decision
- underlying methodology of the decision-making system (how is the decision made and what is the role of the AI system?)

What knowledge, competences, and how much time do the target groups have for dealing with the AI system? What training courses or professional development measures might be sensible or necessary?

→ *6.3.2 Explaining general functionality*

How can the general mode of operation of an AI system be explained to the respective target group?

Each target group has its own individual requirements. Questions to help prepare an explanation of general functionality are:

What are the goals of and context in which the AI system is used?

What are the main criteria that the AI system uses to make decisions? How are these individual criteria weighted?

What are the limits of the AI system? What can it do, what not?

What trends have been noted in the results during test or beta phases or in its operation to date? What have the error rates been to date (false positives and false negatives)?

→ 6.3.3 Explaining the concrete decision in a specific case

How can the individual decision taken by an AI system be explained to the respective target group?

Questions to help prepare an explanation of concrete decisions by an AI system are:

- ***How is it possible to communicate to users and affected persons which factors were relevant for a concrete output (e.g. the decision that concerns them personally)?***
- ***How are the decisions taken by means of interactions between the AI system and humans documented?***
- ***Explainable is not the same as understandable: how can information be presented in the form of text and graphics to make it easily understandable and interpretable?***
- ***How can this information be made easily accessible to users and affected persons, e.g. as part of the notification that AI is being used, as part of the output, etc.?***

→ 6.3.4 Determining explanation strategy

Which aids can be used to facilitate explanations for various target groups?

Questions to ask when determining the method of explanation:

- ***What technical measures can be used to retrospectively determine the factors relevant to the decision?***
  - Additional tools that explain the results of the software and, for example, present the factors behind the output in an understandable way
  - The tools depend on the target group, AI model, and data used[36]
- ***What levels of explainability are there? In which cases does which level have to be attained, i.e. how precisely do the system and its outputs have to be understood?***
  - What does the required level of explainability depend on? What role does the application context play, and in particular the risk assessment of the application?
  - Which level is required for which target groups?
- ***How can it be ensured that the right explanatory model is used for each purpose? How can this be verified? Can the explanations be tested with users and affected persons or their representatives beforehand?***
- ***Were tests carried out regularly or following specific events (e.g. when significant changes have been made to the system) in order to ensure that the explanations are understandable and that external parties are capable of auditing the system?***

[36] A guide to the most common tools can be found in a study by the iit-Institut entitled “Erklärbare KI”.

Notes

If you would like to get involved in the Network's discussions or make suggestions, please write to us at: ***ki-in-der-verwaltung@bmas.bund.de***

Acknowledgements and legal notice

Published by
Policy Lab Digital, Work & Society within the
Federal Ministry of Labour and Social Affairs
Wilhelmstraße 49
10117 Berlin
Germany
Internet: denkfabrik-bmas.de
Email: denkfabrik@bmas.bund.de

Information as of: October 2022

Editorial team
Network Artificial Intelligence in
Employment and Social Protection Services

Linda Wichman
Stephan Frühwirt
Policy Lab Digital, Work & Society

Tim Vallée
Patricia Scheiber
iRights.Lab

Design
Scholz & Friends Berlin GmbH
ressourcenmangel GmbH
365 Sherpas GmbH

10/2022

www.denkfabrik-bmas.de