EDPS

2004
2024
EDPS

EUROPEAN DATA PROTECTION SUPERVISOR

The EU's independent data protection authority

03 June 2024

Generative AI and the EUDPR.

First EDPS Orientations for ensuring data protection compliance when using Generative AI systems.

edps.europa.eu

These EDPS Orientations on generative Artificial Intelligence (generative AI) and personal data protection intend to provide practical advice and instructions to EU institutions, bodies, offices and agencies (EUIs) on the processing of personal data when using generative AI systems, to facilitate their compliance with their data protection obligations as set out, in particular, in Regulation (EU) 2018/1725. These orientations take a technology-neutral approach and do not prescribe specific technical measures. Instead, they put an emphasis on the general principles of data protection that should help EUIs comply with the data protection requirements according to Regulation (EU) 2018/1725.

These orientations are a first step towards more detailed guidance that will take into account the evolution of Generative AI systems and technologies, their use by EUIs, and the results of the EDPS' monitoring and oversight activities.

The EDPS issues these orientations in its role as a data protection supervisory authority and not in its new role as AI supervisory authority under the AI Act.

These orientations are without prejudice to the Artificial Intelligence Act.

Introduction and scope

1. These orientations are intended to provide some practical advice to the EU institutions, bodies, offices and agencies (EUIs) on the processing of personal data in their use of generative AI systems, to ensure that they comply with their data protection obligations in particular as set out in the Regulation (EU) 2018/1725 ('the Regulation', or EUDPR). Even if the Regulation does not explicitly mention the concept of Artificial Intelligence (AI), the right interpretation and application of the data protection principles is essential to achieve a beneficial use of these systems that does not harm individuals' fundamental rights and freedoms.

2. The EDPS issues these orientations in his role as a data protection supervisory authority and not in his new role as AI supervisory authority under the AI Act.

3. These orientations do not aim to cover in full detail all the relevant questions related to the processing of personal data in the use of generative AI systems that are subject to analysis by data protection authorities. Some of these questions are still open, and additional ones are likely to arise as the use of these systems increases and the technology evolves in a way that allows a better understanding on how generative AI works.

4. Because artificial intelligence technology evolves quickly, the specific tools and means used to provide these types of services are diverse and they may change very quickly. Therefore, these orientations have been drafted as technologically neutral as possible.

5. These orientations are structured as follows: key questions, followed by initial responses along with some preliminary conclusions, and further clarifications or examples.

6. These initial orientations serve as a preliminary step towards the development of more comprehensive guidance. Over time, these orientations will be updated, refined and expanded to address further elements needed to support EUIs in the development and implementation of these systems. Such an update should take place no later than twelve months after the publication of this document.

1. What is generative AI?

Generative AI is a subset of AI that uses specialised machine learning models designed to produce a wide and general variety of outputs, capable of a range of tasks and applications, such as generating text, image or audio. Concretely, it relies on the use of the so-called foundation models, which serve as baseline models for other generative AI systems that will be 'fine-tuned' from them.

A foundation model serves as the core architecture or base upon which other, more specialised models, are built. These models are trained on the basis of diverse and extensive datasets, including those containing publicly available information. They can represent complex structures like images, audio, video or language and can be fine-tuned for specific tasks or applications.

Large language models are a specific type of foundation model trained on massive amounts of text data (from millions to billions of words) that can generate natural language responses to a wide range of inputs based on patterns and relationships between words and phrases. This vast amount of text used to train the model may be taken from the Internet, books, and other available sources. Some applications already in use are code generation systems, virtual assistants, content creation tools, language translation engines, automated speech recognition, medical diagnosis systems, scientific research tools, etc.

The relationship between these concepts is hierarchical. Generative AI is the broad category encompassing models designed to create content. A foundation model, such as a large language model, acts as the foundational architecture upon which more specialized models are built. Specialised models, built upon the foundation model, cater to specific tasks or applications, using the knowledge and capabilities of the foundational architecture.

The life cycle of a generative AI model covers different phases, starting by the definition of the use case and scope of the model. In some cases, it might be possible to identify a suitable foundation model to start with, in other cases a new model may be built from scratch. The following phase involves training the model with relevant datasets for the purpose of the future system, including fine-tuning of the system with specific, custom datasets required to meet the use case of the model. To finalise the training, specific techniques requiring human agency are used to ensure more accurate information and controlled behaviour. The following phase aims at evaluating the model and establishing metrics to regularly assess factors, such as accuracy, and the alignment of the model with the use case. Finally, models are deployed and implemented, including continuous monitoring and regular assessment using the metrics established in previous phases.

Relevant use cases in generative AI are general consumer-oriented applications (such as ChatGPT and similar systems that can be already found in different versions and sizes[1], including those that can be executed in a mobile phone). There are also business applications in specific areas, pre-trained models, applications based on pre-trained models that are tuned for specific use in an area

[1] The size of a Large Language Model is usually measured as the number of parameters (tokens it contains. The size of a LLM model is important since some capabilities only appear when the model grows beyond certain limits.

of activity, and, finally, models in which the entire development, including the training process, is carried out by the responsible entity.

Generative AI, like other new technologies, offers solutions in several fields meant to support and enhance human capabilities. However, it also creates challenges with potential impact on fundamental rights and freedoms that risk being unnoticed, overlooked, not properly considered and assessed.

→ The training of a Large Language Model (LLM) (and generally of any machine-learning model) is an iterative, complex and resource intensive process that involves several stages and techniques aiming at creating a model capable of generating human-like text in reaction to commands (or prompts) provided by users. The process starts with the model being trained on massive datasets, most of it normally unlabeled and obtained from public sources using web-scraping technologies (- data protection authorities already have expressed concern and outline the key privacy and data protection risks associated with the use of publicly accessible personal data). After that, LLMs are - not in all cases - fine-tuned using supervised learning or through techniques involving human agency (such as the Reinforcement Learning with Human Feedback (RLHF) or Adversarial Testing via Domain experts) to help the system better recognize and process information and context, as well as to determine preferred responses, whether to limit output in reply to sensitive questions and to align it with the values of the developers (e.g. avoid producing harmful or toxic output). Once in production, some systems use the input data obtained through the interaction with users as a new training dataset to refine the model.

---

2. Can EUIs use generative AI?

As an EUI, there is no obstacle in principle to develop, deploy and use generative AI systems in the provision of public services, providing that the EUI's rules allow it, and that all applicable legal requirements are met, especially considering the special responsibility of the public sector to ensure full respect for fundamental rights and freedoms of individuals when making use of new technologies.

In any case, if the use of generative AI systems involves the processing of personal data, the Regulation applies in full. The Regulation is technologically neutral, and applies to all personal data processing activities, regardless of the technologies used and without prejudice to other legal frameworks, in particular the AI Act. The principle of accountability requires responsibilities to be clearly identified and respected amongst the various actors involved in the generative AI model supply chain.

EUIs can develop and deploy their own generative AI solutions or can alternatively deploy for their own use solutions available on the market. In both cases, EUIs may use providers to obtain all or some of the elements that are part of the generative AI system. In this context, EUIs must clearly determine the specific roles - controller, processor, joint controllership - for the specific processing operations carried out and their implications in terms of obligations and responsibilities under the Regulation.

As AI technologies advance rapidly, EUIs must consider carefully when and how to use generative AI responsibly and beneficially for public good. All stages of a generative AI solution life cycle should operate in accordance with the applicable legal frameworks, including the Regulation, when the system involves the processing of personal data.

→ The terms trustworthy or responsible AI refer to the need to ensure that AI systems are developed in an ethical and legal way. It entails considering the unintended consequences of the use of AI technology and the need to follow a risk-based approach covering all the stages of the life cycle of the system. It also implies transparency regarding the use of training data and its sources, on how algorithms are designed and implemented, what kind of biases might be present in the system and how are tackled possible impacts on individual's fundamental rights and freedoms. In this context, generative AI systems must be transparent, explainable, consistent, auditable and accessible, as a way to ensure fair processing of personal data.

3. How to know if the use of a generative AI system involves personal data processing?

Personal data processing in a generative AI system can occur on various levels and stages of its lifecycle, without necessarily being obvious at first sight. This includes when creating the training datasets, at the training stage itself, by infering new or additional information once the model is created and in use, or simply through the inputs and outputs of the system once it is running.

When a developer or a provider of a generative AI system claims that their system does not process personal data (for reasons such as the alleged use of anonymised datasets or synthetic data during its design, development and testing), it is crucial to ask about the specific controls that have been put in place to guarantee this. Essentially, EUIs may want to know what steps or procedures the provider uses to ensure that personal data is not being processed by the model.

The EDPS has already cautioned[2] against the use of web scraping techniques to collect personal data, through which individuals may lose control of their personal information when these are collected without their knowledge, against their expectations, and for purposes that are different from those of the original collection. The EDPS has also stressed that the processing of personal data that is publicly available remains subject to EU data protection legislation. In that regard, the use of web scraping techniques to collect data from websites and their use for training purposes might not comply with relevant data protection principles, including data minimisation and the principle of accuracy, insofar as there is no assessment on the reliability of the sources.

Regular monitoring and the implementation of controls at all stages can help verify that there is no personal data processing, in cases where the model is not intended for it.

→ EUI-X, a fictional EU institution, is considering the acquisition of a product for automatic speech recognition and transcription. After studying the available options, it has focused on the possibility of using a generative AI system to facilitate this function. In this particular case, it is a system that offers a pre-trained model for speech recognition and translation. Since this model will be used for the transcription of meetings using recorded voice files, it has been determined that the use of this model requires the processing of personal data and therefore it must ensure compliance with the Regulation.

[2] *Opinion 41/2023, of 25 September 2023 , on the Proposal for a Regulation on European Union labour market statistics on businesses*

4. What is the role of DPOs in the process of development or deployment of generative AI systems?

Article 45 of the Regulation establishes the tasks of the data protection officer. DPOs inform and advise on the relevant data protection obligations, assist controllers to monitor internal compliance, provide advice where requested regarding DPIAs, and act as the contact point for data subjects and the EDPS.

In the context of the implementation by EUIs of generative AI systems that process personal data it is important to ensure that DPOs, within their role, advise and assist in an independent manner on the application of the Regulation, have a proper understanding of the lifecycle of the generative AI system that the EUI is considering to procure, design or implement and how it works. This means, obtaining information on when and how these systems process personal data, and how the input and output mechanisms work, as well as the decision-making processes implemented through the model. It is important, as the Regulation points out[3], to provide advice to controllers when conducting data protection impact assessments. Controllers must ensure that all processes are properly documented and that transparency is guaranteed, including updating records of processing and, as a best practice, carrying out a specific inventory on generative AI - driven systems and applications. Finally, the DPO should be involved in the review of compliance issues in the context of data sharing agreements signed with model providers.

From the organisational perspective, the implementation of generative AI systems in compliance with the Regulation should not be a one-person effort. There should be a continuous dialogue among all the stakeholders involved across the lifecycle of the product. Therefore, controllers should liaise with all relevant functions within the organisation, notably the DPO, Legal Service, the IT Service and the Local Informatics Security Officer (LISO) in order to ensure that the EUI works within the parameters of trustworthy generative AI, good data governance and complies with the Regulation. The creation of an AI task force, including the DPO, and the preparation of an action plan, including awareness raising actions at all levels of the organisation and the preparation of internal guidance may contribute to the achievement of these objectives.

→ As an example of contractual clauses, the European Commission, through the "Procurement of AI Community" initiative, has brought together relevant stakeholders in procuring AI-solutions to develop wide model contractual clauses for the procurement of Artificial Intelligence by public organizations. It is also relevant to consider the standard contractual clauses between controllers and processors under the Regulation[1].

[3] *Article 39(2) of the Regulation*

5. An EUI wants to develop or implement generative AI systems. When should a DPIA be carried out?

The principles of data protection by design and by default[4] aim to protect personal data throughout the entire life cycle of data processing, starting from the inception stage. By complying with this principle of the Regulation, based on a risk-oriented approach, the threats and risks that generative AI may entail can be considered and be mitigated sufficiently in advance. Developers and deployers may need to carry out their own risk assessments and document any mitigation action taken.

The Regulation requires that a DPIA[5] must be carried out before any processing operation that is likely to implicate a high risk[6] to fundamental rights and freedoms of individuals.. The Regulation points out the importance of carrying out such assessment, where new technologies are to be used or are of a new kind in relation to which no assessment has been carried out before by the controller, in the case of generative AI systems for example.

The controller is obliged to seek the advice of the data protection officer (DPO) when carrying out a DPIA. Because of the assessment, appropriate technical and organisational measures must be taken to mitigate the identified risks given the responsibilities the context and the available state-of-the-art measures.

It may be appropriate, in the context of the use of generative AI to seek the views of those affected by the system, either the data subject themselves or their representatives in the area of intended processing. In addition to the reviews to assess whether the DPIA is rightly implemented, regular monitoring and reviews of the risk assessments need to be carried out, since the functioning of the model may exacerbate identified risks or create new ones. Those risks are related to personal data protection, but are also related to other fundamental rights and freedoms.

All the actors involved in the DPIA must ensure that any decision and action is properly documented, covering the entire generative AI system lifecycle, including, actions taken to manage risks and the subsequent reviews to be carried out.

It is EUI's responsibility to appropriately manage the risks connected to the use of generative AI systems. Data protection risks must be identified and addressed throughout the entire life cycle of the generative AI system. This includes regular and systematic monitoring to determine, as the system evolves, whether risks already identified are worsening or whether new risks are appearing. The understanding of risks linked to the use of generative AI is still ongoing so there is a need to keep a vigilant approach towards

[4] Article 27 of the Regulation
[5] Articles 39 and 89 of the Regulation.

[6] The classification of an AI system as posing "high-risk" due to its impact on fundamental rights according to the AI Act, does trigger a presumption of "high-risk" under the GDPR, the EUDPR and the LED to the extent that personal data is processed.

non-identified, emerging risks. If risks that cannot be mitigated by reasonable means are identified, it is time to consult the EDPS.

→ The EDPS has established a template allowing controllers to assess whether they have to carry out a DPIA [annex six to Part I of the accountability toolkit]. In addition, the EDPS has established an open list of processing operations subject to the requirement for a DPIA. Where necessary, the controller shall carry out a review to assess if the data processing is being performed in accordance with the data protection impact assessment, at least when there is a change to the risks represented by processing operations. If following the DPIA, controllers are not sure whether risks are appropriately mitigated, they should proceed to a prior consultation with the EDPS.

---

6. When is the processing of personal data during the design, development and validation of generative AI systems lawful?

The processing of personal data in generative AI systems may cover the entire lifecycle of the system, encompassing all processing activities related to the collection of data, training, interaction with the system and systems' content generation. Collection and training-related processing activities include obtaining data from publicly available sources on the Internet, directly, from third parties, or from the EUIs' own files. Personal data can also be obtained by the generative AI model directly from the users, via the inputs to the system or through inference of new information. In the context of generative AI systems, the training and use of the systems relies normally on systematic and large scale processing of personal data, in many cases without the awareness of the individuals whose data are processed.

The processing of any personal data by EUIs is lawful if at least one of the grounds for lawfulness[7] listed in the Regulation is applicable. In addition, for the processing of special categories of personal data to be lawful, one of the exceptions[8] listed in the Regulation must apply. When the processing is carried out for the performance of task carried out in the public interest[9] or is necessary for the compliance with a legal obligation[10] to which the controller is subject to, the legal ground for the processing must be laid down in EU law. In addition, the referred EU Law should be clear and precise and its application should be foreseeable to individuals subject to it, in accordance with the requirements set out in the Charter of Fundamental Rights of the European Union and the European Convention for the Protection of Human Rights and Fundamental Freedoms.

Moreover, where a legal basis gives rise to a serious interference with fundamental rights to data protection and privacy, there is a greater need for clear and precise rules governing the scope and the application of the measure as well as the accompanying safeguards. Therefore, the greater the interference, the more robust and detailed the rules and safeguards should be. When relying on internal rules, these internal rules should precisely define the scope of the interference with the right to the protection of personal data, through identification of the purpose of processing, categories of data subjects, categories of personal data that would be processed, controller and processors, and storage periods, together with a description of the concrete minimum safeguards and measures for the protection of the rights of individuals.

The use of consent[11] as a legal basis may apply in some circumstances in the context of the use of generative AI systems. Obtaining consent[12] under the Regulation, and for that consent to be valid, it needs to meet all the legal requirements, including the need for a clear affirmative action by the individual, be freely given, specific, informed and unambiguous. Given the way in which generative AI systems are trained, and the sources of training data, including publicly available information, the use of consent as such must be carefully considered, also in the context of its use by public

[7] Article 5 of the Regulation.
[8] Article 10(2) of the Regulation.
[9] Article 5(1)(a) of the Regulation.
[10] Article 5(1)(b) of the Regulation
[11] Articles 5(1)(d) and 7 of the Regulation.
[12] EDPB Guidelines 05/2020 on consent under Regulation 2016/679, available at https://www.edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_202005_consent_en.pdf

bodies, such as EUIs. In addition, if consent is withdrawn, all data processing operations that were based on such consent and took place before the withdrawal - and in accordance with the Regulation - remain lawful. However, in this case, the controller must stop the processing operations concerned. If there is no other lawful basis justifying the processing of data, the relevant data must be deleted by the controller.

Service providers of generative AI models may use legitimate interest under the EU General Data Protection Regulation[13] (GDPR) as a legal basis for data processing, particularly with regard to the collection of data used to develop the system, including the training and validations processes. The Court of Justice of the European Union (CJEU) has held[14] that the use of legitimate interest lays down three cumulative conditions so that the processing of personal data covered by that legal basis is lawful. First, the pursuit of a legitimate interest by the data controller or by a third party; second, the need to process personal data for the purposes of the legitimate interests pursued; and third, that the interests or fundamental freedoms and rights of the person concerned by the data protection do not take precedence over the legitimate interest of the controller or of a third party In the case of data processing by generative AI systems, many circumstances can influence the balancing process inherent in the provision, leading to effects such as unpredictability for the data subjects, as well as legal uncertainty for controllers. In that regard, EUIs have a specific responsibility to verify that providers of generative AI systems have complied with the conditions of application of this legal basis, taking into account the specific conditions of processing carried out by these systems.

As controllers for the processing of personal data, EUIs are accountable for the transfers of personal data that they initiate and for those that are carried out on their behalf within and outside the European Economic Area. These transfers can only occur if the EUI in question has instructed them or allowed them, or if such transfers are required under EU law or under EU Member States' Law. Transfers can occur at different levels in the context of the development or use of generative AI systems, including when EUIs make use of systems based on cloud services or when they have to provide, in certain cases, personal data to be used to train, test or validate a model. In either case, these data transfers must comply with the provisions laid down in Chapter V[15] of the Regulation, while also subject to the other provisions of the Regulation, and be consistent with the original purpose of the data processing.

Personal data processing in the context of generative AI systems requires a legal basis in line with the Regulation. If the data processing is based on a legal obligation or in the exercise of public authority, that legal basis must be clearly and precisely set out in EU law. The use of consent as a legal basis requires careful consideration to ensure that it meets the requirements of the Regulation, in order to be valid.

[13] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

[14] Judgment of 4 July 2023, Meta Platforms and Others (General terms of use of a social network), C-252/21, EU:C:2023:537, paragraph 106 and the case-law cited

[15] Articles 46 to 51 of the Regulation

→ For example, the GPA Resolution on Generative Artificial Intelligence Systems states that, where required under relevant legislation, developers, providers and deployers of generative AI systems must identify at the outset the legal basis for the processing of personal data related to: a) collection of data used to develop generative AI systems; b) training, validation and testing datasets used to develop or improve generative AI systems; c) individuals' interactions with generative AI systems; d) content generated by generative AI systems.

7. How can the principle of data minimisation be guaranteed when using generative AI systems?

The principle of data minimisation means that controllers shall ensure that personal data undergoing processing are adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed. There is a misconception that the principle of data minimisation[16] has no place in the context of artificial intelligence. However, data controllers have an obligation to limit the collection and otherwise processing of personal data to what is necessary for the purposes of the processing, avoiding indiscriminate processing of personal data. This obligation covers the entire lifecycle of the system, including testing, acceptance and release into production phases. Personal data should not be collected and processed indiscriminately. EUIs must ensure that staff involved in the development of generative AI models are aware of the different technical procedures available to minimise the use of personal data and that those are duly taken into account in all stages of the development.

EUIs should develop and use models trained with high quality datasets limited to the personal data necessary to fulfil the purpose of the processing. In this way, these datasets should be well labelled and curated, within the framework of appropriate data governance procedures, including periodic and systematic review of the content. Datasets and models must be accompanied by documentation on their structure, maintenance and intended use. When using systems designed or operated by third-party service providers, EUIs should include in their assessments considerations related to the principle of data minimisation.

The use of large amounts of data to train a generative AI system does not necessarily imply greater effectiveness or better results. The careful design of well-structured datasets, to be used in systems that prioritise quality over quantity, following a properly supervised training process, and subject to regular monitoring, is essential to achieve the expected results, not only in terms of data minimisation, but also when it concerns quality of the output and data security.

→ EUI-X intends to train an AI system to be able to assist with tasks related to software development and programming. For this, they would like to use a content generation tool that will be available through the individual IT staff members' accounts. The EUI-X needs to reflect before training the algorithm to make sure they will not be processing personal data that would not be useful for the intended purpose. For example, they may carry out a statistical analysis to demonstrate that a minimum amount of data is necessary to achieve the result. Furthermore, they will need to check and justify whether they will be processing special categories of personal data. Additionally, they will need to examine the typology of data (i.e. synthesised, anonymised or pseudonymised). Finally, they will need to verify all relevant technical and legal elements of the data sources used, including their lawfulness, transparency and accuracy.

[16] In accordance with Article 4(1)(c) of the Regulation, personal data undergoing processing shall be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed.

8. Are generative AI systems respectful of the data accuracy principle?

Generative AI systems may use in all stages of their lifecycle, notably during the training phase, huge amounts of information, including personal data.

The principle of data accuracy[17] requires data to be accurate, up to date, while the data controller is required to update or delete data that is inaccurate. Data controllers must ensure data accuracy at all stages of the development and use of an artificial intelligence system. Indeed, they must implement the necessary measures to integrate data protection by design that will help to increase data accuracy in all the stages.

This implies verifying the structure and content of the datasets used for training models, including those sourced or obtained from third parties. It is equally important to have control over the output data, including the inferences made by the model, which requires regular monitoring of that information, including human oversight. Developers should use validation sets[18] during training and separate testing sets for final evaluation to obtain an estimation on how the system will perform. Although generally not data protection oriented, metrics on statistical accuracy (the ability of models to produce correct outputs or predictions based on the data they have been trained on), when available, can offer an indicator for the accuracy of the data the model uses as well as on the expected performance.

When EUIs use a generative AI system or training, testing or validation datasets provided by a third party, contractual assurances and documentation must be obtained on the procedures used to ensure the accuracy of the data used for the development of the system. This includes data collection procedures, preparation procedures, such as annotation, labelling, cleaning, enrichment and aggregation, as well as the identification of possible gaps and issues that can affect accuracy. The technical and user documentation of the system, including model cards, should enable the controller of the system to carry out appropriate checks and actions regularly to ensure the accuracy principle. This is even more important since models, even when trained with representative high quality data, may generate output containing inaccurate or false information, including personal data, the so-called "hallucinations."

Despite the efforts to ensure data accuracy, generative AI systems are still prone to inaccurate results that can have an impact on individuals´ fundamental rights and freedoms.

While providers are implementing advanced training systems to ensure that models use and generate accurate data, EUIs should carefully assess data accuracy throughout the whole lifecycle of the generative AI systems and consider the use of such systems if the accuracy cannot be maintained.

[17] Article 4(1)(d) of the Regulation.
[18] Validation sets are used to fine-tune the parameters of a model and to assess its performance.

→ EUI-X, following the advice of the DPO, has decided that the results of the ASR model, when used for the transcription of official meetings and hearings, will be subject to validation by qualified staff of the EUI. In cases where the model is used for other less sensitive meetings, the transcription will always be accompanied by a clear indication that it is a document generated by an AI system. EUI-X has prepared and approved at top-management level a policy for the use of the model as well as data protection notices compliant with the Regulation requesting the consent of individuals, both for the recording of their voice during meetings and for its processing by the transcription system. A DPIA has also been carried prior to the deployment of the AI system by the EUI.

9. How to inform individuals about the processing of personal data when EUIs use generative AI systems?

Appropriate information and transparency policies can help mitigate risks to individuals and ensure compliance with the requirements of the Regulation, in particular, by providing detailed information on how, when and why EUIs process personal data in generative AI systems. This implies having comprehensive information - that must be provided by developers or suppliers as the case may be - about the processing activities carried out at different stages of development, including the origin of the datasets, the curation/tagging procedure, as well as any associated processing. In particular, EUIs should ensure that they obtain adequate and relevant information on those datasets used by their providers or suppliers and that such information is reliable and regularly updated. Certain systems (i.e. chatbots) may require specific transparency requirements, including informing individuals that they are interacting with an AI system without human intervention.

As the right to information[19] includes the obligation to provide individuals, in cases of profiling and automated decisions, meaningful information about the logic of such decisions, as well as their meaning and possible consequences on the individuals, it is important for the EUI to maintain updated information, not only about the functioning of the algorithms used, but also about the processing datasets. This obligation should generally be extended to cases where, although the decision procedure is not entirely automated, it includes preparatory acts based on automated processing.

EUIs must provide to individuals all the information required in the Regulation when using generative AI systems that process personal data. The information provided to individuals must be updated when necessary to keep them properly informed and in control of their own data.

→ EU-X is preparing a chatbot that will assist individuals when accessing certain areas of its website. The controllers affected, with the advice of the DPO, have prepared a data protection notice, available in the EU-X website. The notice includes information on the purpose of the processing, the legal basis, the identification of the controller and the contact details of the DPO, the recipients of the data, the categories of personal data collected, the retention of the data as well on how to exercise individual rights. The notice also includes information on how the system works and on the possible use of the user´s input to refine the chat function. EU-X uses consent as a legal basis, but users can withdraw their consent at any moment. The notice also clarifies that minors are not permitted to use the chatbot. Before using the EUI's chatbot, individuals can provide consent after reading the data protection notice.

[19] Article 14 of the Regulation.

10. What about automated decisions within the meaning of Article 24 of the Regulation?

The use of a generative AI system does not necessarily imply automated decision-making[20] within the meaning of the Regulation. However, there are generative AI systems that provide decision-making information obtained by automated means involving profiling and /or individual assessments. Depending on the use of such information in making the final decision by a public service, EUIs may fall within the scope of application of Article 24 of the Regulation, so they need to ensure that individual safeguards are guaranteed, including at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.

In managing AI decision-making tools, EUIs must consider carefully how to ensure that the right to obtain human intervention is properly implemented. This is of paramount importance in case EUIs deploy autonomous AI agents that can perform tasks and make decisions without human intervention or guidance.

EUIs must be very attentive to the weight that the information provided by the system has in the final steps of the decision-making procedure, and whether it has a decisive influence on the final decision taken by the controller. It is important to recognise the unique risks and potential harms of generative AI systems in the context of automated decision-making, particularly on vulnerable populations and children[21].

Where generative AI systems are planned to support decision-making procedures, EUIs must consider carefully whether to put them into operation if their use raises questions about their lawfulness or their potential of being unfair, unethical or discriminatory decisions.

[20] Article 24 of the Regulation.
[21] Global Privacy Assembly (GPA) (2023). Resolution on Generative Artificial Intelligence Systems.

→ EUI-X is considering using an AI system for the initial screening and filtering of job applications. Service provider C has offered a generative AI system that performs an analysis of the formal requirements and an automated assessment of the applications, providing scores and suggestions on which candidates to interview in the next phase. Having consulted the documentation on the model, including the available measures on statistical accuracy (measures on precision and sensitivity of the model) and in view of the possible presence of bias in the model, EUI-X has decided that it will not use the system at least until there are clear indications that the risk of bias has been eliminated and the measures on precision improve, to the analysis of formal requirements.

In any case, if such system is considered as 'fit for purpose' (i.e. candidates' screening) and compliant with all regulations applicable to the EUI, the EUI should be able to demonstrate that it can validly rely on one of the exceptions under Article 24(2) of the Regulation; that the EUI has implemented suitable measures to safeguard individuals' rights, including the right to obtain human intervention by the EUI, to express her or his point of view and to contest the decision (e.g., non-eligibility).

Information must be provided by the EUI, in accordance with Article 15(2)(f) of the Regulation, if the data is collected from the individual, about the logic involved by the AI system, as well as on the envisaged consequences of such processing for the individual. A DPIA must also be carried out prior to the deployment of the AI system by the EUI.

The EUI-X may decide to use, instead of a generative AI system, a 'simpler' online automated tool for the screening of job applications (for instance, an IT tool checking automatically the number of years of professional experience or of education).

11. How can fair processing be ensured and avoid bias when using generative AI systems?

In general, artificial intelligence solutions tend to magnify existing human biases and possibly incorporate new ones, which can create new ethical challenges and legal compliance risks. Biases can arise at any stage of the development of a generative AI system through the training of datasets, the algorithms or through the people who develop or use the system. Biases in generative AI systems can lead to significant adverse consequences for individuals' fundamental rights and freedoms, including unfair processing and discrimination, particularly in areas such as human resource management, public health medical care and provision of social services, scientific and engineering practices, political and cultural processes, the financial sector, environment and ecosystems as well as public administration.

Main sources of bias can come, among others, from existing patterns in the training data, lack of information (total or partial) on the affected population, inclusion or omission of variables and data that should not or should be part of the datasets, methodological errors or even bias that are introduced through monitoring.

It is essential that the datasets used to create and train models ensure an adequate and fair representation of the real world - without bias that can increase the potential harm for individuals or collectives not well represented in the training datasets - while also implementing accountability and oversight mechanisms that allow for continuous monitoring to prevent the occurrence of biases that have an effect on individuals, as well as to correct those behaviours. This includes ensuring that processing activities are traceable and auditable[22] and that EUIs keep supportive documentation. In that regard, it is important that EUIs adopt and implement technical documentation models, which can be of particular importance when the models use several datasets and / or combine different data sources.

Generative AI systems providers try to detect and mitigate bias in their systems. However, EUIs know best their business case and should test and regularly monitor if the system output is biased by using input data tailored to their business needs.

EUIs, as public authorities, should put in place safeguards to avoid overreliance on the results provided by the systems that can lead to automation and confirmation biases.

The application of procedures and best practices for bias minimisation and mitigation should be a priority in all stages of the lifecycle of generative AI systems, to ensure fair processing and to avoid discriminatory practices. For this, there is a need for oversight and understanding of how the algorithms work and the data used for training the model.

[22] The audit of training data can help to detect bias and other problematic issues by studying how the training data is collected, labelled, curated and annotated. The quality of the audit and its results depends on the access to the relevant information, including the training datasets, documentation and implementation details.

→ EU-X is assessing the existence of sampling bias on the automated speech recognition system. Translation services have reported significantly higher word error rates for some speakers than for others. It seems that the system has difficulties to cope with some English accents. After consulting with the developer, it has concluded that there is a deficit in the training data for certain accents, notably when the speakers are not native. Because itis systematic, EU-X is considering refining the model using its own-generated datasets.

12. What about the exercise of individual rights?

The particular characteristics of the generative AI systems mean that the exercise of individual rights[23] can present particular challenges, not only in the area of the right of access, but also in relation to the rights of rectification, erasure and objection to data processing. For example, one of the most relevant elements is the difficulty in identifying and gaining access to the personal data stored by the system. In large language models, for example, individual words like "cat" or "dog" are not stored as strings of text. Instead, they are represented as numerical vectors through a process called word embedding. These vectors derive from the model's training on vast amounts of text data. The consequence is that accessing, updating or deleting the data stored in these models, if possible, is very difficult. In this sense, proper management of the datasets can facilitate access to information, which is difficult in the case of unsupervised training based on publicly available sources incorporating personal data. It is equally complex to manage the production of personal data obtained through inference. Finally, the exercise of certain rights, such as the right to erasure, may have an impact on the effectiveness of the model.

Keeping a traceable record of the processing of personal data, as well as managing datasets in a way that allows traceability of their use, may support the exercise of individual rights. Data minimisation techniques can also help to mitigate the risks related to not being able to ensure the proper exercise of individual rights in accordance with the Regulation.

EUIs, as data controllers, are responsible for and accountable for implementing appropriate technical, organisational and procedural measures to ensure the effective exercise of individual rights. Those measures should be designed and implemented from the early stages of the lifecycle of the system, allowing for detailed recording and traceability of processing activities.

→ EU-X has included in the data protection notice for the chatbot a reference to the exercise of individual rights, including access, rectification, erasure, objection and restriction of processing in accordance with the EUDPR. The notice includes contact details of the controller and EU-X DPO, as well as a reference to the possibility of lodging a complaint with the EDPS. Following a request of access from an individual concerning the content of his conversations with the chatbot, EU-X replied, after carrying out the relevant checks, that no content is preserved from the said conversations beyond the established retention period, 30 days. The conversations, as indicated to the individual, has not been used to train the chatbot model.

[23] Chapter III of the Regulation.

13. What about data security?

The use of generative AI systems can amplify existing security risks or create new ones, including bringing about new sources and transmission channels of systemic risks in the case of widely used models. Compared to traditional systems, generative AI specific security risks may derive from unreliable training data, the complexity of the systems, opacity, problems to carry out proper testing, vulnerabilities in the system safeguards etc. The limited offer of models in critical sectors for the provision of public services such as health can amplify the impact of vulnerabilities in these systems. The Regulation requires EUIs to implement appropriate technical and organisational measures to ensure a level of security[24] appropriate to the risk for the rights and freedoms of natural persons.

Controllers should, in addition to the traditional security controls for IT systems, integrate specific controls tailored to the already known vulnerabilities of these systems - model inversion attacks[25], prompt injection[26], jailbreaks[27] - in a way that facilitates continuous monitoring and assessment of their effectiveness. Controllers are advised to only use datasets provided by trusted sources and carry out regularly verification and validation procedures, including for in-house datasets. EUIs should train their staff on how to identify and deal with security risks linked to the use of generative AI systems. As risks evolve quickly, regular monitoring and updates of the risk assessment are needed. In the same way, as the modalities of attacks can change, proper access to advanced knowledge and expertise must be ensured. A possible way to deal with unknown risks is to use "red teaming[28]" techniques to try to find and expose vulnerabilities.

When using Retrieval Augmented Generation[29] with generative AI systems, it is necessary to test that the generative AI system is not leaking personal data that might be present in the system's knowledge base.

The lack of information on the security risks linked to the use of generative AI systems and how they may evolve requires EUIs to exercise extreme caution and carry out detailed planning of all aspects related to IT security, including continuous monitoring and specialised technical support. EUIs must be aware of the risks derived from attacks by malicious third parties and the available tools to mitigate them.

[24] Article 33 of the Regulation.
[25] A Model inversion attacks takes place when an attacker extracts information from it through reverse-engineering.
[26] Malicious actors use prompt injection attacks to introduce malicious instructions as if they were harmless.
[27] Malicious actors use jailbreaking techniques to disregard the model safeguards.
[28] A red team uses attacking techniques aiming at finding vulnerabilities in the system.
[29] AI systems in which a Large Language Model bases its answers in a knowledge base prepared by the generative AI system owner (e.g. an EUI) with internal sources and not in the knowledge stored by the LLM itself.

→ EU-X, following a security assessment, has decided to implement the ASR system on premises, instead of using the API services provided for the developer of the model. EU-X will train its IT staff on the use and further development of the system, in close cooperation with the provider. This may include training on how to refine the model. In addition, EU-X will get the services of an external auditor to verify the proper implementation of the system, including on security.

14. Do you want to know more?

- **EDPS work on AI**

  - 45th Closed Session of the Global Privacy Assembly - Resolution on Generative Artificial Intelligence Systems - 20 October 2023
  - EDPS TechDispatch #2/2023 - Explainable Artificial Intelligence
  - EDPS at work: data protection and AI (includes links to several documents published by the EDPS alone or in cooperation with other authorities)
  - EDPB-EDPS Joint Opinion 5/2021 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)
  - EDPS Opinion 44/2023 on the Proposal for Artificial Intelligence Act in the light of legislative developments

    Large Language Models (EDPS website, part of the EDPS "TechSonar" report 2023-2024)

- **Other relevant documents**

  - Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679 (wp251rev.01)
  - CNIL: AI how-to-sheets
  - Spanish Data Protection Authority: Artificial Intelligence: accuracy principle in the processing activity
  - Italian Data Protection Authority: Decalogo per la realizzazione di servizi sanitari nazionali attraverso sistemi di Intelligenza Artificiale – September 2023 (Italian)
  - The Hamburg Commissioner for Data Protection and Freedom of Information - Checklist for the use of LLM-based chatbots - 15/11/2023
  - AI Security Concerns in a nutshell (DE Federal Office for Information Security, March 2023)
  - Multilayer Framework for Good Cybersecurity Practices for AI (ENISA, June 2023)
  - Ethics Guidelines for Trustworthy AI (EC High-Level Expert Group on AI, 2019)
  - Living Guidelines on the responsible use of Generative AI in research (ERA Forum Stakeholders' document, March 2024)
  - OECD AI Incidents Monitor (AIM)
  - OECD Catalogue or tools and metrics for trustworthy AI