CEDPO
CONFEDERATION OF EUROPEAN DATA PROTECTION ORGANISATIONS

Bonn, Bucharest, Dublin, Lisbon, Madrid, Milan, Paris, The Hague, Vienna, Warsaw

Generative AI: The Data Protection Implications

CEDPO AI Working Group
16 October 2023

Contact information:
https://cedpo.eu
info@cedpo.eu

About this Guidance

Artificial intelligence is not a new concept for DPOs and data protection professionals. Generative AI, however, is. When OpenAI's ChatGPT launched in November 2022, the majority of data protection professionals had never heard of generative AI, and were certainly not concerned with such technologies in their day-to-day work.

Now, with ChatGPT in the hands of over 100m users globally, and many other providers such as Google Bard and Anthropic's Claude entering the market, it has become an operational reality, and necessity, for data protection professionals to deal with the consequences of generative AI tools being rapidly utilised within organisations. Whether these tools are adopted simpliciter or are fine-tuned by organisations using their own data sets, novel and as-yet unexamined data protection implications exist, all of which data protection professionals must rapidly come to terms with.

The aim of this paper is to guide data protection professionals through the maze of issues that are unfolding as these technologies gain rapid adoption in organisations. Amongst other key issues, this paper looks at data-sharing risks, accuracy of personal data, conducting DPIAs on generative AI tools, implementing data protection by design, selecting a lawful basis for training generative AI systems, optimising organisational structures, applying privacy-enhancing techniques, and handling data subject rights in the context of these technologies.

There will be no future without generative AI, and with data playing such a pivotal role in the training and operating of these systems, DPOs will play a central role in ensuring that both data protection and data governance standards are at the heart of these technologies.

CEDPO
CONFEDERATION OF EUROPEAN DATA PROTECTION ORGANISATIONS

Table of Contents

1. Accuracy of Personal Data

The accuracy of personal data processed by generative artificial intelligence (AI) tools is a fundamental data protection issue with such technologies. Article 5 (1) (d) of the GDPR states that 'Personal data shall be accurate, and where necessary, kept up to date; every reasonable step must be taken to ensure that personal data that are inaccurate (…) are erased or rectified without delay.'

It is obvious, and a matter of common sense that the processing of inaccurate personal data can have very real-world implications for the data subject behind the data, yet generative AI tools, such as OpenAI's widely used text-based chatbot, ChatGPT, inherently have numerous inaccuracies in the data they process. By their nature these tools ingest vast amounts of training data, sourced from massive data scraping exercises across the internet. Necessarily, this data comes with all of its imperfections, and becomes a part of the data bank which users of ChatGPT make queries against. When a user receives an answer that is either wholly or partly inaccurate, this generates what AI providers call 'hallucinations' or, in the vernacular, 'falsehoods'.

Even OpenAI itself, on its website, warns users about the perils involved and that the accuracy of data retrieved cannot be automatically trusted. Under a section entitled 'Limitations' it notes 'ChatGPT sometimes writes plausible-sounding but incorrect or nonsensical answers'[1]. Compounding the issue, OpenAI further notes that the tool will often add to inaccuracies by essentially guessing what an uncertain user means. It states: 'Ideally, the model would ask clarifying questions when the user provided an ambiguous query. Instead, our current models usually guess what the user intended.[2]'

When coupled with the fact that ChatGPT's data processing terms make it clear that the user is the data controller, while OpenAI is merely the data processor, it should be clear, for users, that this is very much a 'buyer beware' market. Why? Because if any party further processes inaccurate personal data, it will become liable for any non-compliance with Article 5 (1) (d) above. In the context of ChatGPT, then, relying upon inaccurate personal data provided by the tool will make the user liable to non-compliance with the GDPR, especially where such re-use impacts the fundamental rights and freedoms of data subjects.

[1] https://openai.com/blog/chatgpt
[2] Ibid.

Organisations should understand that this is not merely a theoretical point and that regulators have already called generative AI companies to account for the accuracy of their data. In March 2023, the Italian Data Protection Authority blocked the deployment of ChatGPT in Italy, noting, amongst other matters, that the data was frequently not accurate. It noted, based on 'tests carried out so far, the information made available by ChatGPT does not always match factual circumstances, so that inaccurate personal data are processed.'[3]

Data protection officers (DPOs) must remain aware, then, of the risks of processing inaccurate data. Users within a DPO's organisation should be given clear guidelines to help them to understand that the outputs of any generative AI tools, such as ChatGPT, come with a health warning, namely, that the human user is still ultimately responsible for verifying the accuracy of any personal data obtained. This is a critical point.

A further related risk comes from the second clause of Article 5 (1) (d), that personal data shall be 'kept up to date'. ChatGPT, and similar tools such as Google's Bard and Anthropic's Claude, rely on data scraping activities up to a certain point in time, meaning that their data bank becomes out-of-date and so, eventually, are necessarily not responding to up–to-date events. This creates the clear risk that users will obtain personal data that is no longer relevant, or perhaps lacks context, or is simply outrightly inaccurate, given how events have changed or how information has moved forward in the intervening period.

DPOs should also remain aware of the ways in which unmitigated bias and discrimination in the training sets could indirectly lead to inaccurate data outputs, again opening up the user to the risks of further processing inaccurate data.

A final, global risk with generative AI chatbots is the tone that they adopt: an oracular level of certainty and authority that might almost be called a dark pattern, so misleading is it in its effect on the evaluation of search results. When generative AI chatbots are palpably wrong or inaccurate, they are often wrong in a very confident and confusingly definitive manner, an attitude that masks the fact that, as OpenAI, for instance admits, the answer may simply be 'nonsense'. In any search results, the tone of the response should be ignored, and again, users should realise that the output of these tools requires human evaluation, certainly when it concerns questions over the accuracy of any personal data involved.

[3] https://www.gpdp.it/web/guest/home/docweb/-/docweb-display/docweb/9870847#english

2. Sharing Personal Data with Generative AI Tools

Artificial Intelligence (AI) has rapidly evolved from a concept of science fiction to a relatively common feature of our life. A rapidly emerging branch of AI is generative AI, which can create new, previously non-existing data that closely mimics the input data. Generative AI models can, under the right conditions, generate high-quality text, images, music, and more. However, the convenience and innovative potential of generative AI comes with a cost. Despite its promising capabilities, the sharing of personal data with these systems presents substantial risks for privacy, confidentiality, and the integrity and security of data. Understanding these risks is essential in order to protect individual data protection rights, and to maintain a secure digital environment.

Like most AI systems, generative AI is data-driven. Traditional AI training involves feeding large datasets into AI models which can then learn patterns and features from this data. Once the training is complete, the AI system is equipped to generate outputs based on the patterns and features learned. This means that once personal data is part of the AI's training set, it contributes to the formation of the AI's internal model, and will invariably influence its behaviour and outputs. Effectively, the data becomes "part" of the AI, in the sense that it informs the system's understanding and knowledge. This presents significant data protection concerns where personal data features as training data.

Generative AI models trained on personal data can potentially extract sensitive information like names, addresses, health information, or even financial data, and then republish that data in search results for different users. Additionally, generative AI models can amplify exposure by generating more data similar to the original input. Third parties may then exploit this data for unlawful activities including invasive advertising, phishing scams, or in more serious cases, fraud or identity theft. This highlights the complexities of controlling how personal data is used by generative AI models. Once personal data has been shared with generative AI models, managing and tracking its usage becomes an intricate (if not impossible) task, due to the nature of how AI systems process information as well as store and replicate data across different systems. Therefore, retracting personal data shared with generative AI models may be incredibly difficult or unrealistic. The lesson for DPOs is that users *must* understand precisely what kinds of information *can* and *cannot* be shared with generative AI tools, because once personal data is shared, the Rubicon has been crossed, and it will be very difficult to undo what has been done.

One of the more alarming risks associated with sharing personal data with generative AI is the creation and proliferation of 'deepfakes'. Deepfakes refer to the application of AI to create, alter or manipulate content, such as images, audio, and video, in such a way that it fabricates hyper-realistic but entirely false content. By training on personal data, generative AI can generate synthetic media that convincingly impersonate natural or legal persons. These deepfakes can then be used maliciously, such as in disinformation campaigns, fraud, or harassment. Related to this is the fact that the accuracy of generative AI decisions heavily depends on the quality and diversity of the input training data. If this personal data is biased, the AI's outputs can also become biased, leading to unfair consequences.

Generative AI holds significant promise for numerous applications, but its use of personal data must be carefully managed to mitigate potential risks. By employing strong data protection controls, ethical AI practices, and robust legal protections, it may be possible to harness the potential of generative AI while safeguarding individual data protection rights and fostering a safe and secure digital environment.

3. What is an Appropriate Lawful Basis?

The lawful basis that properly applies to the training of AI systems with personal data is a key consideration. Prima facie, there is no obvious candidate that would both clearly legitimise this processing activity and also uphold the data protection rights of affected individuals. This is a critical consideration because the volume of training data that is used for generative AI applications is enormous, and only growing in size. If such training activities are to continue, and if AI is to deliver on its promise, then it cannot be founded on an uncertain lawful basis as regards personal data. Moreover, the Artificial Intelligence Act (AI Act) is not particularly instructive on this point given that Article 10, (which deals with data governance and the governance of training data for AI systems), does not create a lawful basis specific to the use of personal data for the training of AI systems. It is, then, to the GDPR that we must turn for a suitable lawful basis for this activity.

Firstly, we will briefly look at how data is used to train generative AI systems. This takes place in four broad ways:

1. Based on personal data that has been scraped from the internet;
2. Where the personal data has been provided by the users of the AI system, such as when they submit prompts to Generative AI tools;

3. Where the personal data has been collected from third parties, such as data brokers, or companies that have databases which are relevant to the AI training phase (for instance, a database of court decisions for a predictive AI tool in the legal domain); and
4. When AI developers/operators use the personal data held in their own databases to train the AI system.

In these cases, under Article 6 of the GDPR, three lawful bases are most relevant: contract, legitimate interest and consent.

1. **Contract:**

Article 6(1)(b) of the GDPR notes that contract may form a legal basis for processing personal data where that 'processing is necessary for the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract.'

The application of the first branch of the contract legal basis (*i.e.*, the performance of the contract itself) would require demonstrating that the training of the AI system (and not the use of the AI once trained) is strictly necessary to the performance of a contract with the data subject.

This necessity requirement is interpreted very narrowly by the data protection authorities. According to the European Data Protection Board (EDPB), it should not be possible to perform the main subject-matter of the specific contract with the data subject, if the processing of the personal data in question does not occur. In other words, processing the personal data in this way should be a necessary condition for performing the contract.

Considering this narrow interpretation, there is very little room for the contract basis when training an AI system. This basis could theoretically be applied when the use of the AI system *is* the subject matter of the contract entered into between the AI operator and the user, and when there is no other way to perform this contract than to train the AI with the data of the users.

As for the second branch of this legal basis, *i.e.*, the pre-contractual steps, its application would require demonstrating that a data subject made a request in the context of potentially entering a contract and that there is no other way to meet his/her demands than to train (and not only use once trained) the AI. This is an even more restrictive and limited option than the first part of this legal basis.

On the whole, the circumstances in which the lawful basis of contract might be used to justify training AI systems with personal data are very limited and, in practical terms, this basis will not be a viable option for grounding such processing activities.

In the case of generative AI, contract as a lawful basis is, in any case, particularly unsuitable given that typically no contract exists between the data subjects whose data is used, and the organisations responsible for training such systems with that data.

2. Legitimate Interests

The legitimate interests basis could only apply provided that a legitimate interest assessment is completed by the data controller to ensure that these interests are not overridden by the interests or fundamental rights and freedoms of the data subject.

This may however be challenging especially since, most of the time, the organisation behind training generative AI tools, such as OpenAI, is not in direct contact with the data subjects, nor does it have any form of relationship with those data subjects. In this regard, the recent actions of the Italian Data Protection Supervisory Authority (Garante per la protezione dei dati) against ChatGPT should be noted. In March 2023, the authority blocked ChatGPT in the Italian territory until OpenAI was able to satisfactorily answer certain questions, one of which was that OpenAI needed to specify the lawful basis for training ChatGPT with personal data. In its response to this point, OpenAI identified legitimate interests as the lawful basis. This is a highly significant commitment and statement by OpenAI as it effectively ties the huge task of training generative AI systems to a lawful basis that is inherently uncertain, given data subjects' explicit right under Article 21 of the GDPR to object to such processing.

To effectively be able to rely on the legitimate interests basis would require in particular:

- a study on a case-by-case basis of the context of training and of use of the AI, as well of the collection of the personal data used to verify that the data processing will meet the reasonable expectations of the data subjects;
- a demonstration of the strict necessity of this processing and of the fact that the AI cannot work efficiently without being trained with the personal data in question;
- an enhanced transparency of the data processing towards the data subjects. The provision of all the required information under GDPR would need to be provided to the data subjects in an appropriate way;

- an effective opt-out system brought to the knowledge of the data subjects within a reasonable period before their data are provided to the AI system[4];
- more generally an efficient system for ensuring the respect of the data subjects' rights which would be difficult to implement given the particularities of generative AI functioning

3. Consent:

The consent basis could also apply, but only in very clearly circumscribed circumstances. While in the extreme cases, it may be the only legal basis possible (for instance when processing special categories of data or data concerning minors) as a general rule it has very little place in the training of generative AI systems, as currently conceived. The entire apparatus used for the training of AI systems makes it almost impossible to obtain consent. This is because, in the first instance, the majority of the data used to train such systems is purchased from data brokers that have obtained this data by scraping the internet, an activity which necessarily does not involve the obtaining of consent from underlying data subjects. Indeed, the very lawfulness of data scraping as a commercial activity is far from certain and the recent joint communication by twelve global data protection authorities, including the UK's Information Commissioner's Office, underlines this point.[5]

To use consent as a lawful basis, would require meeting all the requirements for valid consent under the GDPR, meaning that it would need to result from a clear affirmative action, be freely given, specific, informed and unambiguous. This is indeed a very high bar to reach in the world of training AI systems.

If the AI provider is not in contact with the data subjects, as is generally the case, this consent would have to be collected by the user of the AI system, the organisation with whom the data subject *does* have a relationship. However, this will usually be after the fact, when the AI system has already been trained, so to object to the processing would most of the time be irrelevant since the data processing would have already occurred. Moreover, it would also be very difficult to

[4] If the opt-out system is brought to their knowledge after the training of the AI, to object to the processing would most of the time be, in one hand, irrelevant since the data processing would have already occurred and, on the other hand, very difficult to stop when vast amounts of personal data regarding numerous data subjects are being fed to an AI (*see part XXX regarding the exercise of the data subjects' rights*).

[5] https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2023/08/joint-statement-on-data-scraping-and-data-protection/

reverse when vast amounts of personal data regarding numerous data subjects will already have been ingested by the AI system.

In conclusion, legitimate interest is most likely the most suitable basis for training AI systems with personal data, however, as stated above, it does not provide a certain foundation given the need for a legitimate interests assessment to be carried out, as well as the fact that data subjects can object to such processing at any point.

4. Risks of Jailbreaking and Data Protection Safeguards

Soon after ChatGPT was released, hackers began attempting to "jailbreak" the AI chatbot, trying to bypass its safeguards and make it say inappropriate or irrational things. These intricately phrased prompts that aim to bypass the restrictions imposed on AI programmes have come to be known as 'Jailbreaks'. This term was originally used in the context of digital technology to refer to the act of gaining access to the operating system of a smartphone or tablet, especially one manufactured by Apple, in order to run modified or unauthorised software.

In the context of Generative AI models, the term now refers to the design of prompts that make the chatbots bypass rules around producing hateful content or writing about illegal acts. These attacks involve manipulating the generative AI systems to produce content that goes against their intended rules, such as generating hateful or illegal material. Another use of these attacks could be slander and a personal attack upon an individual once personal data has been leaked.

A security firm that specialises in AI, was able to break GPT-4, OpenAI's latest text-generating chatbot, in just a few hours after the initial release of the system. Using carefully crafted prompts, the CEO of the security firm bypassed OpenAI's safety systems and quickly had GPT-4 generating homophobic statements, creating phishing emails, and endorsing violence. This deviant behaviour poses a serious risk as it has the potential to expose personal data that has been inadvertently, or perhaps even intentionally, input into the system and, thus, has the potential to be manipulated by bad actors.

A closely-related attack is the prompt injection attack that can quietly insert malicious data or instructions into AI models. A prompt injection attack aims to elicit an unintended response from LLM-based tools. And then achieve unauthorised access, manipulate responses, or bypass security

measures. The specific techniques and consequences of prompt injection attacks vary depending on the system.

Jailbreaks and prompt injection attacks are a form of unconventional hacking, using well-crafted sentences instead of code to exploit weaknesses in AI systems. While these attacks are currently focused on bypassing content filters, security researchers warn of the potential for data theft and widespread cybercriminal activities as generative AI systems become more prevalent.

Numerous popular online services and products heavily rely on large datasets to train and improve their AI algorithms. Data streams from networks, social media platforms, mobile devices, and various other sources contribute to the vast amount of information that businesses utilise to train their machine learning systems. It is, hence, important to note that some of the data contained within these datasets could probably be considered personal data, even by users who are less concerned about data protection. Unfortunately, due to the misuse and mishandling of personal data by certain companies, data protection has consequently become a pressing global policy issue.

In a similar vein, much of our sensitive data is also gathered to enhance AI-enabled processes. This data plays a crucial role in driving the adoption of machine learning, as sophisticated algorithms rely on such data for real-time decision-making. Search algorithms, voice assistants, recommendation engines, and other AI solutions leverage extensive datasets of real-world user data to provide personalised and relevant outputs.

Early in 2023, a website called Jailbreak Chat was launched where prompts for AI chatbots like ChatGPT from online forums are collected and shared. Visitors to the site can contribute their own jailbreaks, try out prompts submitted by others, and vote on their effectiveness. Malicious users could leverage these jailbreaks to gather personal data contained within the systems to carry out crimes like identity theft and to create deepfakes to impersonate living individuals.

The implications of jailbreaks and prompt injection attacks become more significant when these systems gain access to personal and sensitive data. For example, if a successful prompt injection attack instructs a personal assistant AI to ignore previous instructions and send an email to all contacts, it could lead not only to embarrassment on the part of the individual but to widespread issues for the affected individuals and the rapid spread of harmful content across the individual's personal and working networks.

Ensuring the safety of foundation models like ChatGPT is paramount as their use becomes more widespread. The hackers, however, will not give up easily. As AI systems have evolved, the jailbreaks have become more complex. Some involve multiple characters, intricate backstories, translation, and even elements of coding to generate specific outputs.

Some authorised "red teams" prompt attacks on AI models to uncover vulnerabilities. A red team in cybersecurity represents the offensive security team, which is responsible for discovering security vulnerabilities through penetration testing. With GAI, these teams look for exploits that include actual vulnerabilities, influencing the system's behaviour, or deceiving users to get around the system's security. Other attempts come from hobbyists who like to showcase humorous or disturbing outputs on social media. This approach to security is suboptimal as it is fragmented and relies on viral exposure and influential individuals to prompt fixes.

While companies like OpenAI, Google, and Microsoft have taken steps to address jailbreaking and prompt injection attacks, the researchers behind these attacks continue to find new ways to exploit vulnerabilities. The development of generative AI systems requires approaches beyond traditional red-teaming methods, such as using a second AI model to analyse prompts or clearly separating system prompts from user prompts.

Automation and advanced techniques are necessary to identify and mitigate jailbreaks and injection attacks at scale. By automating the process of identifying vulnerabilities and unintended behaviours, researchers aim to discover and address a greater number of these security risks.

These types of automated techniques can be seen as the starting point for a deeper commitment from AI developers to assess and evaluate the safety of their systems. By involving a diverse range of participants and prioritising transparency and accountability, the goal is to enhance the safety, reliability, and ethical use of generative AI technology. Third-party assessments, automated mitigation of jailbreaks and using red-teaming, will play a pivotal role in achieving this goal and improving the practices surrounding AI development in order to meet the requirements of both the GDPR and the forthcoming AI Act.

5. How are Data Subject Rights implemented with Generative AI Tools?

Generative AI, or GenAI, are AI systems capable of generating text, images, or other media in response to prompts. Generative models learn the patterns and structure of input data, subsequently generating new content similar to the training data, but with a degree of novelty, as opposed to merely classifying or predicting data. These AI systems are often based on Generative Pretrained Transformers (GPT), artificial neural networks built on the transformer architecture, pretrained on large sets of unlabelled text data, and capable of generating human-like text. They employ large language models (LLMs) to produce data based on the training dataset that was used to create them.

Understanding the technology behind generative AI is vital to realising that these tools encompass various phases, and personal data can be processed at each phase. However, the processing of personal data at one phase does not necessarily imply data processing at another.

The stages under data protection law where data subject rights pertaining to personal data might apply in the generative AI context include:

1. The training data phase, when personal data is incorporated.
2. The deployment phase, where personal data is used to generate content and the content result itself.
3. The model itself, which might contain personal data.

It is also essential to point out that generative AI software can indirectly process data, particularly related to the user of the solution, such as account data or metadata related to the use of the solution.

In common machine learning models, identifying the individuals that the training data is about is a potential challenge to ensuring their rights. Usually, this data includes only the information pertinent to predictions, without unique data subject identifiers. It undergoes various pre-processing measures to make it suitable for machine learning algorithms, often transforming personal data into a form that's harder (but not impossible) to link back to specific individuals. Data protection laws might, therefore, still apply to this transformed data, as it could still be used to

identify individuals. This process necessitates consideration when responding to individuals' rights requests.

This process is different for generative AI models than it is for common machine learning models as explained in the previous paragraph. Generative AI models are often trained with data accessible on the web, and their value also often lies in generating results related to physical persons, implying a significant amount of personal data in the training data for these models. As a result, these datasets could be the target of data subject requests.

In generative AI models, 'continuous learning' also poses unique challenges for GDPR compliance. These models are regularly updated based on user interactions, meaning personal data is continuously processed. This data mostly originates from the interactions and prompts of the tool's users and it should be noted that the data subjects and the data providers are not necessarily the same entity in the context of continuously learning AI models.

Given these considerations, navigating data rights under the General Data Protection Regulation (GDPR) in the context of generative AI models presents unique challenges, particularly for the rights of Erasure, Rectification, Access, and Objection.

The first shared issue is the non-retrievability of data in generative AI models. As previously mentioned, these models source data from a wide array of origins, like web scraping and user interactions. This multifaceted approach to data collection makes it more difficult to trace individual contributions. Furthermore, in contrast to traditional data storage systems, in GenAI systems, personal data are also deeply embedded within complex algorithms, complicating the isolation of specific data. This makes it challenging to fulfil GDPR rights since identifying whether and where personal data are processed within the system.

Adding another layer of complexity is the issue of "inferred personal data." These are conclusions that the model may draw based on its training. For example, a generative AI model could deduce a user's political affiliations based on past data interactions. The prevailing opinion leans towards including these inferences when responding to rights requests, as they could indirectly reveal personal information. The concept of "inferred group data" also deserves attention. This type of data is generated based on broader patterns recognised during training. Whether this group data is considered personal depends on its subsequent processing and utilisation.

Beside common challenges, there's also specific ones related to individual rights that require data modification or erasure. Notably altering or removing data from the training set after a Data Subject Request (DSR) could impact the model's validation and correctness. The original data often serving as a foundation for such validation processes. Moreover, the erasure or modification of data that is already embedded in the model would often imply to remove or modify this data to retrain the model, a task that is both costly and time-consuming.

In summary, the intersection of GDPR rights and generative AI models presents a labyrinth of challenges, each with its own intricacies and complications. The very nature of these models, from the way they embed and process data to the difficulties in tracking individual contributions, adds layers of complexity to GDPR compliance. While no silver bullet exists to seamlessly navigate these challenges, the evolving landscape does offer some emerging solutions that could serve as starting points for compliance.

To begin with, despite the absence of a one-size-fits-all solution, proactive steps can be taken. Implementing the principle of 'privacy by design and by default' during the GenAI model's creation and deployment phases provides a foundational layer of data protection that is integrated right from the beginning.

In navigating the complex terrain of data protection, one could consider a pre-emptive strategy that narrows down the scope of data and its identifying features. By doing so, it could be possible to potentially alleviate many of the complexities that might arise later in the data processing cycle. Data minimisation could serve as an essential part of this early-stage planning, guiding data controller to collect only what is truly necessary. Building on this, anonymisation techniques of personal data or the use of Privacy Enhanced Technologies (PETs), such as synthetic data, could allow a further reduction in the scope potentially affected by DSR.

Moreover, investing in proactive measures like data mapping and data labelling is crucial. Such measures offer clarity on the origins and characteristics of training data, making it easier to handle rights requests in subsequent phases.

As generative AI models transition from development to deployment, the focus shifts towards optimising adaptability and traceability. In this stage, maintaining meticulous data processing records is not just good practice but becomes indispensable for facilitating responding right requests. This is all the more important given the increased malleability of data at this stage. In addition, the challenges of continuous learning in deployed models can be effectively addressed

through versioning techniques. This enables efficient rollback to a previous model state without the laborious need to retrain from the ground up. This linkage ensures that both adaptability and traceability are addressed, providing a robust framework for compliance.

6. Data Protection by Design: How to Build Generative AI Tools in Compliance with the GDPR

Data protection by design plays a pivotal role in ensuring compliance with the General Data Protection Regulation (GDPR). It entails safeguarding personal data from the very early stages of design throughout the entire lifecycle of the system. The idea of data protection by design came from a more general set of privacy principles entitled Privacy by Design first developed in Canada in the early 2000s. Privacy by Design is an approach to systems engineering that was initially developed by Ann Cavoukian and formalised in a report on privacy-enhancing technologies by a joint team of the Information and Privacy Commissioner of Ontario (Canada), the Dutch Data Protection Authority, and the Netherlands Organisation for Applied Scientific Research in 1995. The Privacy by Design framework was published in 2009 and adopted by the International Assembly of Privacy Commissioners and Data Protection Authorities in 2010. In the same year, the International Conference of Data Protection Authorities and Privacy Commissioners unanimously passed a resolution recognising Privacy by Design as an essential component of fundamental privacy protection. This was followed by the U.S. Federal Trade Commission's inclusion of Privacy by Design as one of three recommended practices for protecting online privacy.

Shortly after 2010, Europe began working on revising its data protection laws. Inspired by Privacy by Design and its principles, Europe put together data protection by design principles which were introduced into law via Article 25 of the General Data Protection Regulation (GDPR) in 2018.

In recent years, the swift development of generative AI has given rise to an increased awareness of potential risks and ethical considerations when designing systems which process personal data. These concerns encompass not only complex data protection risks like the leakage of sensitive information and chat histories but also a range of threats to the data subject rights of EU citizens, including the "right to be forgotten." This right allows individuals to request the deletion of their personal data by a company. While deleting data from databases is relatively straightforward, removing data from machine learning models is a more complex task. Anonymisation techniques

and data minimisation practices can help strike a balance between upholding individuals' rights and preserving the overall usefulness of the generative AI model.

Something to consider from a human perspective is that due to the complexity of modern AI systems, the people involved in building and deploying AI systems are often likely to have a wider range of skills and backgrounds than the usual systems developers, including traditional software engineering, systems administration, data scientists, statisticians, as well as domain experts.

Because of this wide range of expertise, there may be less understanding of broader security compliance requirements, as well as those of data protection law more specifically. For these individuals, security of personal data may not always have been a key priority, especially if someone was previously building AI applications with non-personal data or in a research capacity where personal data was protected in sandboxes.

Biased algorithms are another significant data protection concern. Generative AI systems learn from vast amounts of data, and if that data is biased, the algorithms can perpetuate and amplify these biases in their outputs. This raises ethical questions about fairness, discrimination, and the potential harm caused by biased AI-generated content when used to make important, life-changing decisions about data subjects.

AI hallucinations refer to instances where generative AI systems produce outputs that are not based on real or accurate information. These hallucinations can mislead users and have potential implications for the safety of data subjects. Generative AI systems must provide reliable and trustworthy outputs, especially about European citizens whose personal data and its accuracy is protected under the GDPR.

The rise of deepfakes, which are realistic but manipulated audio or video content, has also been associated with generative AI technology. Deepfakes have the potential to manipulate public opinion, spread misinformation, and pose risks to public safety. The ethical implications of deepfakes highlight the need for robust measures to prevent their creation and to detect and combat their dissemination.

A fundamental aspect of data protection by design is transparency. It plays a crucial role in data protection by design and ensures accountability within AI systems. Organisations must be transparent about their data practices, providing clear explanations of how AI systems work and the decisions they make. However, achieving transparency in AI systems can be challenging due

to their complexity. It is essential to develop methods and tools that enable the explanation of algorithmic predictions to end-users in a meaningful and understandable manner.

Further complications arise because common practices about how to process personal data securely in data science and AI engineering are still under development. As part of an organisation's compliance with the security principle of GDPR, they should ensure that they actively monitor and take into account the state-of-the-art security practices when developing AI systems and when using personal data in an AI context.

It is not possible to list all known security risks that might be exacerbated by the use of AI to process personal data. Whatever the risk, however, companies should ensure that staff have appropriate skills and knowledge to address not only security risks but also data protection risks. This is where the importance of GDPR training comes in.

The effectiveness of AI models heavily relies on the quality of the data they receive, making data protection an integral aspect of their design. The utilisation of sensitive data during the training of generative AI algorithms can result in the emergence of personal information in chatbot outputs or compromise data security during cyberattacks.

Thus, when designing AI products, it is paramount to decouple personal data from individual users through the use of synthetic datasets with full anonymisation and non-reversible identifiers for algorithmic training, auditing, and quality assurance, among other practices. Implementing strict controls on data access within the company and conducting regular audits can help prevent data breaches.

It is also important to acknowledge that more data does not necessarily equate to better solutions. Testing algorithms using data minimisation can help determine the least amount of data required for a viable use case. Additionally, providing a streamlined process for users to request the removal of their personal data is critical.

Adopting adversarial learning techniques, which involve combining conflicting datasets during the machine learning process, can help identify flaws and biases in AI algorithm outputs. Additionally, exploring the use of synthetic datasets that do not contain actual personal data is a potential approach, although further research is required to assess their effectiveness.

Organisations must align responsible use of AI with existing data protection principles outlined in the GDPR. These guidelines should encompass various aspects such as accountability, human intervention, accuracy, security, bias prevention, and explainability of automated decision-making. Continuous investments in privacy measures, upskilling in algorithmic auditing, and the adoption of ethics, security, and data protection by design methodologies are necessary to effectively navigate the opportunities and risks associated with generative AI. Technologies such as differential privacy offer privacy-preserving techniques that can be incorporated into generative AI systems. Scalable methods for cleaning datasets, including deduplication and training data disclosure requirements, contribute to addressing privacy-related challenges.

The collective efforts of the data protection and engineering community, coupled with the commitment of individual organisations and privacy professionals, play an indispensable role in addressing the data protection concerns surrounding generative AI. By adhering to the principles of data protection by design and integrating comprehensive data protection and fundamental rights assessments, organisations can strive towards the trustworthy implementation of generative AI while maintaining GDPR compliance. It is essential to continue investing in data protection training, upskilling in algorithmic auditing, and integrating ethics, security, and data protection by design methodologies to ensure the responsible and ethical use of generative AI.

7. Privacy-Enhancing Techniques and Synthetic Data

Generative AI tools are complex tools, and like all such technologies, they present many significant legal challenges. Generative AI is hungry for data, but such data, (especially quality data), may be hard to come by, or may be legally protected, either from an intellectual property or a data protection legislation' s standpoint.

From the data protection perspective, privacy enhancing technologies (PETs) may represent a valid solution to tackle data protection concerns, in terms of data minimisation, integrity, confidentiality, and data protection by design. The European Union Agency for Cybersecurity (ENISA) defines PETs as "*software and hardware solutions (e.g., systems encompassing technical processes, methods or knowledge) to achieve specific privacy or data protection functionality or to protect against risks of privacy of an individual or a group of natural persons*",

Among the various PETs that could be deployed in the context of generative AI, data synthesis algorithms which generate "artificial" data, better known as synthetic data, can play a pivotal role.

According to the European Data Protection Supervisor (EDPS) "*Synthetic data is artificial data that is generated from original data and a model that is trained to reproduce the characteristics and structure of the original data (...). The generation process, also called synthesis, can be performed using different techniques, such as decision trees, or deep learning algorithms. Synthetic data can be classified with respect to the type of the original data: the first type employs real datasets, the second employs knowledge gathered by the analysts instead, and the third type is a combination of these two.*"

In essence, synthetic data is computer-generated data which is derived from existing real data, or from algorithms and models which replicate, fully or partially, features, patterns and properties of real-world data.

The use of synthetic data may therefore bring many advantages when it comes to the training of generative AI tools, particularly as it:

a) reduces the need for harvesting large amounts of real personal data. In the AI model-training phase this is especially important as it allows engineers to generate much larger data sets from relatively small amounts of personal data;
b) allows near-perfect labelling (e.g., exactly defined for the developing of a specific AI model) and higher quality data, thereby supplementing or substituting real world datasets. A study from Gartner has predicted that "*by 2024, 60% of the data used for the development of AI and Analytics projects will be synthetically generated*";
c) if properly detected and corrected, potentially reduces the bias or statistical imbalance of the original datasets, thereby increasing the fairness of decision making that relies on the data;
d) strengthens privacy and reduces the cybersecurity attack surface by limiting the risk of loss of confidentiality, integrity or availability of real personal information;
e) reduces the costs involved at all stages of the data value chain by limiting the need for excessive data collection, cleaning, preparation, and data storage.

However, this does not mean that synthetic data is the complete solution for all data protection issues. There are still some legal concerns that must be taken into consideration by DPOs.

Firstly, synthetic data does not necessarily correspond to anonymous data, which means that re-identification risk, to one degree or another, will remain. In practice, synthetic data aims at replicating real world data and the more it is an accurate proxy, keeping all the features and patterns of the original data, the more efficient it will be for the generative AI model trained on

such data; but, on the other hand, the downside is that such efficiency will, in direct proportion, increase the risk of **re-identification**. This means that the risk of inferring data related to a specific individual from the synthetic dataset, or from the AI model itself, will not be extinguished.

As noted by the UK's Information Commissioner's Office (ICO). "*You should focus on the extent to which people are identified or identifiable in the synthetic data, and what information about them would be revealed if identification is successful. Some synthetic data generation methods have been shown to be vulnerable to model inversion attacks, membership inference attacks and attribute disclosure risk. These can increase the risk of inferring a person's identity….*

The use of other PET's (such as differential privacy) or the suppression of outliers (data points with some uniquely identifying features), can serve to reduce the risk of re-identification of personal data, but not entirely eliminate it.

Furthermore, synthetic data's generation phase may involve the processing of personal data, especially upon collection and analysis of real datasets, which entails the need to abide by the GDPR and related obligations.

Specific mention should also be made of the duty to provide full information under Art. 13 of GDPR to data subjects whose data is being collected and then used for AI training purposes, as well as to identify a lawful basis of processing under Art. 6 of GDPR.

Finally, the obligation to strictly respect the principles under Art. 5 of GDPR always stands where personal data is concerned. In particular, some of the following principles from Art. 5 are worth mentioning in the case of generative AI:

a) transparency: this is not limited to the information to be provided to data subjects under Art. 13 GDPR as mentioned above, but also towards users, with reference to synthetic outputs generated by AI models, in order to avoid the risk of deep fakes and/or social manipulation;
b) purpose limitation: as synthetic data may be derived from real data, which may contain personal information, there is the need to outline that such data has been collected for specified, explicit and legitimate purposes and that the further processing (e.g., for data synthetisation and subsequent AI model training) is not incompatible with the initial purposes.

A similar principle has been established in relation to the anonymisation process by WP Art. 29 (opinion 5/2014) according to which: "t*he anonymisation process, meaning the processing of (…) personal data to achieve their anonymisation, is an instance of "further processing". As such, this processing must comply with the test of compatibility in accordance with the guidelines provided by the Working Party in its Opinion 03/2013 on purpose limitation*".

Especially with regard to the training phase of AI models, the reference to the "statistical purposes" as not in principle incompatible with the initial purposes under lett. b) of art. 5, ss.1, might serve this purpose[6].

c) accuracy and fairness: attention must be given here to avoiding the risk of "hallucination", or of duplicating bias, errors or inaccuracies contained in the original dataset. This is particularly important if the AI model trained by the synthetic data will then be used to adopt decisions which might affect people's rights or interests.

   Of paramount importance for this specific purpose will be the development of techniques that enable the explainability of the outputs generated by AI systems trained by making use of synthetic data.

8. Issues specific to Image- and Audio-Based Generative AI

In the case of non-text-based generative AI applications, such as image, audio and video generating tools, clear data protection implications exist. Popular applications, such as Midjourney and Stable Diffusion, which allow users to rapidly generate images and videos by inputting text prompts, are built on large volumes of image and video content. This underlying data includes numerous categories of personal data sufficient to identify data subjects, the central one being the very image and likeness of a data subject that will often be represented in the outputs.

Specifically, DPOs can expect the following personal data categories to be involved in such tools:

[6] See on this topic, Study at the request of the Panel for the Future of Science and Technology (European Parliamentary Research Service) "*The Impact of the GDPR on artificial intelligence*", June 2020

- photo images of data subjects;
- artistic representations of data subjects;
- video footage of data subjects; and
- Audio, voice-based data

Organisations will have to understand that the further processing of such data brings the GDPR into scope. For instance, if a marketing department wants to create promotional material, and uses images of data subjects garnered from generative AI, it will have to process those images in line with data protection laws, and respect fundamental principles such as transparency, lawfulness and fairness.

Furthermore, the issue of combining the data from generative AI sources, with data from other sources, should be considered. While the data received from the generative AI tool may not identify the data subject, the act of combining it with alternative data may do so, and once again, bring GDPR requirements into view. This could be particularly relevant where, for example, the pasting together of images from different sources leads to the identification of individuals.

In the more creative use-cases, where organisations may wish to modify, alter or significantly change the presentation of images, videos or audio content, this should be carried out with respect for data subjects' fundamental rights and freedoms. Risks, for example, of defaming or damaging data subjects, should always be taken into account, and where it is considered that the processing may be high risk, a DPIA should be conducted.

Finally, where organisations wish to create legitimate 'deepfake' content, such as, perhaps, official corporate videos, issues of data subject consent and transparency of processing should be key considerations.

9. Managing Data Protection Risk

Carrying out a data protection impact assessment (DPIA) when implementing or using a generative AI system becomes even more crucial when, as is often the case, these tools have not yet been properly understood, both from the perspective of business strategy and risk management. The understanding of the risks to personal data from generative AI processing is still evolving and all DPOs must try to be alive to as-yet unanticipated threats and challenges. To manage these emerging risks, the following factors should be taken into account.

a) Risks to Data Subjects

The relationship between the user and AI, as well as the impacts that the processing will have on individuals should be at the heart of the analysis. Potential risks to data subjects include:

- Impacts from a partially or fully automated decision produced by generative AI. The consequences of such decisions may consist of financial opportunity losses or even restrictions on fundamental rights.

- Risks of reinforcing discrimination and bias against certain users.

- Risks arising from the processing of special category data as outlined in Art. 9 GDPR. For instance, a generative AI tool could infer from certain personal data of the person concerned, (from their expression modalities or the use of certain words), their ethnic origin, political or philosophical positions, or even the sexual orientation of the person concerned, and apply differential treatment on this basis. In order to identify such risks, the company deploying the generative AI tool should conduct a regular review of the quality of the results generated.

- In terms of IT security, information available to the attacker in the AI system can be a threat vector. A so-called "white box" scenario, where the attacker can deduce/find a lot of technical information to prepare his attack creates more exposure compared to a "black box" system where the attacker can only access the information produced by the system as an output. More particularly, the following attacks are specific to defined AI project steps:

Learning phase	attack type	infection	backdooring attacks
			poisoning attacks
		exfiltration	membership inference attacks
			model inversion attacks
			model extraction attacks

Production phase	attack type	manipulation	evasion attacks
			reprogramming attacks
			denial of Services
		exfiltration	membership inference attacks
			model inversion attacks
			model extraction attacks

b) Identifying mitigation measures

The DPIA, as always, should be conducted before project initiation and should then, via data protection by design, inform and guide the design stage for any generative AI tool. In the case of generative AI, the following mitigants should be taken to account to manage the identified risks:

- Supervised fine-tuning with exemplary conversations where an LLM is trained to reproduce a corpus of conversations that illustrate what is deemed to be a desired behaviour.

- Fine-tuning with a human value model where human operators will reward the most satisfactory results.

- In addition, organisational measures should aim to ensure a constant evaluation of the results provided by the generative AI tool, both at the level of the human operator who uses it and an organisational entity that analyses the results on a large scale in order to ensure a high level of result quality over time.

- Similarly, we should strive as much as possible for a situation of explainability of the decisions taken by the generative AI model to allow genuine human control. In this regard, human control ultimately remains the best method of mitigating risks raised by generative AI systems. By this means, excessive confidence in the results produced by generative AI

tools can be avoided. Such overconfidence would lead, in the absence of effective human controls, to the production of entirely automated decisions.

An additional consideration for DPOs is the emerging AI governance requirement to conduct Fundamental Rights Impact Assessments (FRIAs). In the draft text of the AI Act, which, at the date of publication of this paper, is still in the trialogue stage of discussions within the EU Legislature, a requirement to carry out FRIAs is included. The intention is that such an assessment would need to be completed by either a provider or user of an AI system, where there are risks to the fundamental rights and freedoms of individuals who are affected by the output.

Given that FRIAs are, in effect, akin to DPIAs for the world of AI, with particular overlaps in understanding how processing activities impact fundamental rights, DPOs should expect that this work will be assigned to them once the AI Act comes into effect. Although, in some respects DPOs are uniquely placed, and qualified, to do this work, they are not necessarily naturally conversant in the novel technological risks that are rapidly being created by AI technologies. For this reason, DPOs should already be researching and understanding AI-specific risks to personal data.

From the practical perspective, it may be possible to conduct FRIAs and DPIAs as one exercise, but whatever method is ultimately chosen, DPOs must start developing knowledge of AI risk now, in anticipation of the AI Act.

10. Transparency and Generative AI

When gathering and feeding data including personal data to an AI for the purposes of its training and when this data processing is governed by GDPR, the entity operating this training (the AI operator) must ensure the transparency of said data processing pursuant to Article 5 § 1 a) and 12 et seqq. of said regulation.

Three different sources of data can be identified:

- The scraping of data from websites with the help of robots or AI systems (Use Case 1);
- The provision of data by users of the system or data suppliers concerning other individuals (Use Case 2);
- The provision of data concerning themselves by users of the AI (Use Case 3).

For each of these Use Cases, the ways to ensure data processing transparency vary according to the type of training AI required.

Use Case 1

Transparency is a delicate and perhaps challenging issue when considering online data scraping, mainly due to the fact that any personal data gathered in this manner is not gathered directly from the data subject. As a result, Article 14 of the GDPR should apply to such data, i.e., personal data that has not been gathered from the data subject directly, entitles the data subject to the right to obtain from the controller confirmation as to whether their personal data is being processed, and, if so, access to their personal data should be provided along with other vital information such as the purpose for processing and the categories of data that is being processed and so on.

Additionally, Article 15 of the GDPR regarding the right of access by the data subject to their personal information should apply.

In such a scenario, however, several difficulties present themselves to the AI operator. Especially the following:

- Identifying personal data among the data automatically retrieved by the AI, which usually consists of vast amounts of data;
- Directly identifying each individual data subject;
- Obtaining sufficient contact information to inform each data subject of the processing of their data.

In light of these difficulties, Article 14.5 (b) of the GDPR could be applied. This section of the article stipulates that a data controller would not have to provide the specified information to each data subject when "*the provision of such information proves impossible or would involve a disproportionate effort*". Case law from various data protection authorities shows that this exception should be interpreted very strictly. This being said, given the difficulties identified above regarding generative AI models, it could be applied here. If so, the AI operator would, however, still be bound under the transparency requirements to the data subject.

Pursuant to said Article 14.5 (b), the data controller should take appropriate measures to protect the data subject's rights and freedoms and legitimate interests. Such measures include the publication of the controller's privacy policy on its website, but also, possibly more stringent

measures like the example given by the Italian Data Protection Authority when regulating ChatGPT earlier in 2023. Ultimately, OpenAI agreed to carry out an information campaign, of a non-promotional nature, across all the main Italian mass media (radio, television, newspapers and the Internet) to inform people of the probable collection of their personal data for the purpose of training ChatGPT. They also agreed to make a tool available on the data controller's website, through which all interested parties could exercise their right to access their personal data.

On the other hand, regarding such a right to access, Article 11 of the GDPR may also apply, which stipulates that:

"*1. If the purposes for which a controller processes personal data do not or do no longer require the identification of a data subject by the controller, the controller shall not be obliged to maintain, acquire or process additional information in order to identify the data subject for the sole purpose of complying with this Regulation.*

2. Where, in cases referred to in paragraph 1 of this Article, the controller is able to demonstrate that it is not in a position to identify the data subject, the controller shall inform the data subject accordingly, if possible. In such cases, Articles 15 to 20 shall not apply except where the data subject, for the purpose of exercising his or her rights under those articles, provides additional information enabling his or her identification".

Additionally, we are reminded in Recital 4 of the GDPR that, "*the right to the protection of personal data is not an absolute right; it must be considered in relation to its function in society and be balanced against other fundamental rights, in accordance with the principle of proportionality*". As a result, it could be argued that disproportionate efforts cannot be imposed on the AI operator to identify the applicant and detect their personal data in the training data of the AI.

In light of the above, the AI operator facing an access request should:

1. Verify if the personal data concerning the applicant can be identified;
2. Provide the applicant will all personal data identified;
3. Inform the data subject that there may be personal data concerning them that the AI operator is not in a position to detect/provide given the characteristics of the data processing being carried out.

Also, to comply with Article 25 GDPR and the data protection by design principle, the AI operator may also be obliged to demonstrate that they can anticipate such access requests and that they have reviewed all the technical possibilities that they could reasonably deploy to detect the personal data concerning each applicant (and that it reassesses regularly these possibilities).

Use Case 2

Since data is usually supplied the AI operators along the supply chain by other third parties further up the supply chain (a user or a data supplier). These third parties could assist the AI operator in ensuring transparency in the processing of data by providing tools and guidance on how best to extract personal data from the data set, given that it is these third parties that supply the data sets in the first place. These third parties could also help the AI operator when dealing with access requests from data subjects for the same reasons.

Use Case 3

When personal data is collected directly from the users, Article 13 of the GDPR applies. The data controller must provide specific information to the data subject at the time of collection, for instance the identity and the contact details of the data controller; the contact details of their data protection officer; the purposes of the processing for which the personal data is intended as well as the legal basis for the processing; along with other specific information.

11. Optimising Organisational Structures

Within any organisation, from a management structure-perspective, the topic of Generative AI will have to be addressed in a multidimensional way, as a reflection of the complexity of the technology and its impacts. It will not be viable for companies to have each function working alone and not interacting with each other.

The impact of AI is an enterprise issue; therefore, it requires a joined-up enterprise-wide approach. Such an integrated approach is essential in order to avoid duplication of efforts, but more importantly to ensure that key decisions receive multi-disciplinary input.

To achieve this, organisations should put in place an AI taskforce, focusing on responsible AI and its governance. The creation of such a task force could be an initiative driven by the DPO, as one of the functions that will have the biggest exposure to this topic due to the fact that he has to

manage some AI considerations in a context where personal data is involved. Alternatively, it could be initiated and led by an IT function, such as a Chief Data Officer, or Chief Technology Officer.

This task force will significantly involve the legal department, compliance functions and, specifically, data protection. For the technical aspects, the IT Security department should be represented. The task force may involve communications and PR staff, as it will be necessary to communicate internally, and potentially externally, on the decisions taken by the task force. The leader of the task force may establish focus groups in which selected members of the taskforce focus on specific questions and report back their results to the taskforce. The above diagram gives an indicative idea of the composition of these focus groups and how they would relate to the Responsible AI Governance Taskforce.

The mission of the task force is to respond to the immediate need for Responsible AI governance within the organisation and to examine and manage the risks in the use of generative AI, specifically, with regard to personal data, bias, ethical concerns, emerging AI regulation and numerous legal issues such as intellectual property rights and liability exposure.

The main goal of this task force will be to define an action plan. A critical aspect of this action plan will be to conduct an inventory of the AI systems used in the company, which includes generative AI. Another critical aspect is to define roles and responsibilities for all the functions in the group.

The role of this AI taskforce is also to raise the awareness of AI issues at all levels of the company. This point is important, as the risk will naturally come from the employees that are the day-to-day users of the technology, but it has to be linked with the highest level of decision-making, because deciding on the way to use (or not use) generative AI is an enterprise strategy.

As an initial task, the task force should prepare preliminary guidance for the organisation regarding the responsible use of generative AI which would, for example, include the recommendation not to enter personal data in prompts of relevant tools like ChatGPT, nor to upload images with identifiable persons.

Regardless of the complexity of the technology, and its implementation, the DPO's role in this taskforce is ultimately to ensure that any personal data processed via AI technologies is compliant with the GDPR.