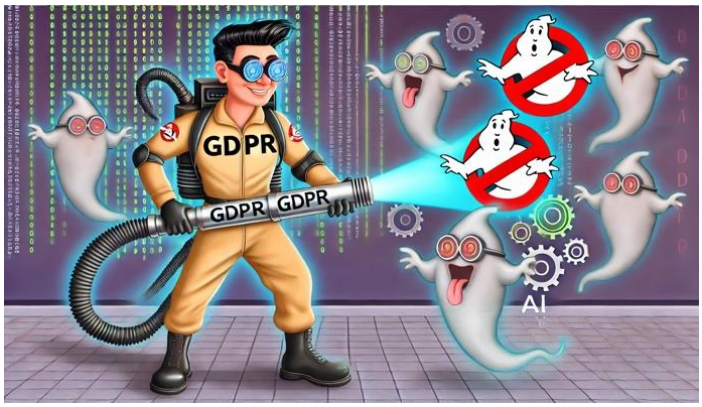
AI Hallucinations and Data Subject Rights under the GDPR: Regulatory Perspectives and Industry Responses

Theodore Christakis

AI Hallucinations and Data Subject Rights under the GDPR: Regulatory Perspectives and Industry Responses

Theodore Christakis

Professor, Chair AI-Regulation, University Grenoble Alpes (France)



The emergence of general-purpose artificial intelligence (GPAI) is significantly impacting various industries, from healthcare to finance, with its ability to produce human-like text, images, videos, and even music. Powered by advanced machine learning algorithms, these AI systems are becoming increasingly sophisticated, offering new possibilities for efficiency, creativity and economic growth. However, alongside these developments lies a persistent challenge: AI hallucinations.

AI hallucinations occur when, for several reasons, GPAI systems produce content that is convincing and seems plausible but is false or nonsensical. Several highly publicized instances of AI hallucinations marked the emergence of GPAI systems in 2022–2023, leading critics to label these models as "stochastic parrots" and even… "bullshit generators". They accuse GPAI systems of disseminating "careless speech [...] poised to cumulatively degrade and homogenize knowledge over time".

Despite significant improvements in AI technologies during 2024 (see part 3), these hallucinations remain a thorny issue. They not only undermine the reliability of AI-generated content but also pose potential risks when such content is disseminated as factual information. This is particularly concerning in contexts where accuracy is paramount, such as legal matters, healthcare, and news reporting.

The implications of AI hallucinations also extend into the realm of personality rights and data protection. Over the past two years, several stories have circulated about individuals who have been erroneously portrayed by GPAI systems as criminals, involved in sexual harassment, or even as deceased, sometimes leading to defamation lawsuits. Despite their exceptional nature and infrequent occurrence, these incidents have attracted significant media attention.

In September 2023 ChatGPT's hallucinations reportedly drew the first GDPR data protection complaint in Poland by a privacy and security researcher who found he was unable to have incorrect information about him corrected by OpenAI. Even more notoriously, in April 2024

the consumer organization Noyb also filed a complaint with the Austrian Data Protection Authority (DPA), claiming that ChatGPT incorrectly stated the date of birth of a public figure, presumed to be its founder Max Schrems.

Noyb claims that ChatGPT's erroneous answers violate article 5(1)(d) GDPR and the principle of accuracy. It argues that, on the basis of this article, the controller has an obligation to erase or rectify inaccurate data without delay, but Open AI failed to do so, despite "made aware of the accuracy issue by the data subject". Noyb explains that OpenAI responded by stating that "the only way to prevent the inaccurate information from appearing would be to block any information concerning the data subject. This would in turn violate the controller's freedom to inform and the general public's right to be informed, as the data subject is a public figure". Noyb also contends that Open AI violated articles 12(3) and 15 GDPR related to the right of access by the data subject because "the data subject has not received any information on what data concerning him is processed by OpenAI, for the purpose of powering ChatGPT…". As a result, Noyb is requesting a series of corrective measures, including "an effective, proportionate, and dissuasive administrative fine."

While these complaints are under examination, two other DPAs have proposed nuanced approaches that seem to contribute positively to the problem of hallucinations.

In July 2024, the Hamburg Data Protection Authority (DPA) published a Discussion Paper that ignited extensive debate. This paper's significance lies in the Hamburg DPA's detailed explanation of the critical distinction between GPAI systems and Large Language Models (LLMs), which constitute only one component of GPAI systems. According to the Hamburg DPA, LLMs themselves do not contain personal data and, as such, fall outside the scope of the GDPR—a stance that has drawn criticism from some commentators.

However, the true significance of the Discussion Paper, in my view, lies in its call to shift regulatory attention toward other components of GPAI systems—particularly their outputs, where the GDPR clearly applies—rather than the internal mechanics of LLMs. While the Hamburg DPA's stance has sparked heated debate, it underscores an important point: LLMs do not store personal data in discrete records or operate as traditional structured databases. Consequently, applying the GDPR's accuracy requirement in its conventional form may be neither feasible nor appropriate.

Similarly, the UK's Information Commissioner's Office (ICO) proposed a risk-based approach to the issue of AI hallucinations, tailoring accuracy requirements to the purpose and context of AI use and emphasizing information and transparency. The combination of these guidances could be very helpful, in my opinion, to mitigate the risks of violating the principle of accuracy and data subject rights under the GDPR when LLMs generate incorrect personal information, without hindering the development of these technologies in Europe.

This article also explores the multifaceted efforts by GPAI system creators to address these issues, highlighting the technical and legal measures implemented to reduce hallucinations and mitigate associated risks. Developers like OpenAI, Google, Meta, Anthropic, and others, have introduced strategies across data training, model architecture, and system outputs to enhance reliability, transparency, and the exercise of data subject rights. While these measures represent significant progress, they might not yet be sufficient, and ongoing refinement is necessary as the technology evolves.

These developments indicate that policymakers should avoid binary decisions and instead foster continued advances in both GPAI systems use and safeguards. This article advocates for the ongoing, nuanced development of regulatory and technical measures, emphasizing that collaboration among regulators, industry stakeholders, civil society, and researchers is

essential. As we gain more experience with GPAI systems, such engagement will help refine strategies to effectively manage AI hallucinations, protect individual rights, and support responsible innovation within the GDPR framework.

Importantly, this article focuses exclusively on situations where the GPAI system and the LLM are deployed by the same data controller (e.g., OpenAI and the consumer version of ChatGPT) without the involvement of any third data controllers/processors. Further research is needed to address scenarios where LLMs are utilized by third parties—for example, via an application programming interface (API)—and where the responsibilities of each party in managing the risk of inaccurate outputs must be clearly defined. This clarification depends on specific factual and legal circumstances, including whether the parties are in a relationship of joint controllership, data processing on behalf of a controller, or independent controllership. In such cases, contractual, legal, and technical measures should help address these issues in a manner that protects data subject rights.

1. Hamburg DPA: Focus on Outputs

The Hamburg Commissioner for Data Protection and Freedom of Information (HmbBfDI) published on July 2024 a very important Discussion Paper on Large Language Models and Personal Data aiming to present a technology-informed perspective on when and how the GDPR applies to GPAI.

The paper is based on the crucial distinction between a GPAI system and any LLM it may incorporate, a distinction also operated by the EU AI Act (cf. Recital 97 and Article 3(1)). As the paper explains, an AI system (e.g., ChatGPT) consists of multiple components and an LLM is only one of such components. Other components include the user interface, input and output filters, and potentially enrichment processes, such as Retrieval Augmented Generation (RAG – see part 3).

The paper argues that LLMs as such do not contain personal data and, therefore, are not covered by the GDPR. As a result, the focus should instead be on other components of the AI system, particularly the outputs, where the GDPR plainly applies.

The reasons for which the Hamburg DPA, following a similar position by the Danish DPA, the Datatilsynet,[1] considers that the GDPR does not apply to LLMs as such can be summarized as following:

- Unlike traditional data systems, LLMs process tokens and vector relationships (embeddings). “Within LLMs texts are no longer stored in their original form, or only as fragments in the form of these numerical tokens, further processed into ‘embeddings’”.
- “When training data contains personal data, it undergoes a transformation during machine learning process, converting it into abstract mathematical representations. This abstraction process results in the loss of concrete characteristics and references to specific individuals. Instead, the model captures general patterns and correlations derived from the training data as a whole”.
- “Everything that LLMs produce is "created" in the sense that it is not a simple reproduction of something stored (such as an entry in a database or a text document), but rather something newly produced. This probabilistic generation capability fundamentally differs from conventional data storage and data retrieval”.
- Privacy attacks and personal data extraction do not mean that LLMs contain personal data. According to the CJEU only lawful means of identification that don't require

> disproportionate effort in practice should be considered. Generally, designing and executing effective privacy attacks on LLMs require substantial technical expertise and time resources that the average user lacks.

The conclusions of this technical and legal assessment are the following:

- The mere storage of an LLM does not constitute processing within the meaning of article 4 (2) GDPR. This is because no personal data is stored in LLMs.
- Given that no personal data is stored in LLMs, data subject rights as defined in the GDPR **cannot relate to the model itself**.
- However, claims for access, erasure or rectification can certainly relate to the input and output of an AI system of the responsible provider or deployer.

Relevance in the field of AI hallucinations and data subject rights

The consequences of this guidance in the field of AI hallucinations are explained by the Hamburg DPA itself:

- "As LLMs don't store personal data, they **can't be the direct subject of data subject rights** under articles 12 et seq. GDPR".
- However, when an AI system processes personal data, **particularly in its output** … the controller must fulfill data subject rights".

In relation with AI hallucinations, the Hamburg DPA's guidance thus invites all stakeholders to switch the attention away from the LLM itself, and towards the **output** of the AI system containing the LLM.

This permits to address a fundamental challenge created by GPAI technologies. As a matter of fact, traditionally, the accuracy principle in the GDPR is primarily concerned with the correctness of personal data that organizations collect, store, and process about individuals. As the ICO has summarized it, the resulting obligations for data controllers are that they "should take all reasonable steps to ensure the personal data they hold is not incorrect or misleading as to any matter of fact" and "if they discover that personal data is incorrect or misleading, they must take reasonable steps to correct or erase it as soon as possible".

LLMs generate content dynamically, predicting the next word in a sentence based on patterns learned from vast amounts of data. They do not store information about individuals in discrete, retrievable records; instead, any mention of personal data results from statistical correlations in the training data. The inaccuracies (or 'hallucinations') are unintended artifacts of the generative process, not deliberate misrepresentations of stored personal data. Since LLMs lack discrete records and do not function as databases, applying the GDPR's accuracy requirement in the traditional sense may be neither feasible nor appropriate. The Hamburg DPA thus emphasizes a widely accepted view: the outputs of LLMs are probabilistic in nature, and despite the risk of occasional regurgitations, LLMs are "not databases from which outputs are pulled".

The Hamburg DPA discussion paper has already led to a lot of discussions and criticisms (see for instance the discussions and comments here and here and here and here and here and here and here).

A first set of criticisms asserts that the likelihood of personal data extraction is enough to assume data storage, given the inherently informative nature of language generation. Critics emphasize that this is particularly true for specific pieces of information widely present in training data, as a result of being widely present on the internet (for instance, Donald Trump's

birth date). Others apply the binary distinction between anonymization and pseudonymization to argue that it is possible to extract personally identifiable information (PII) through combinations of different parts found on LLMs. Others again claim that what is relevant for the existence of personal data is whether the information processed in an LLM, whatever its form, may result in impacting on individuals. More broadly, many commentators find it difficult to reconcile the seemingly paradoxical situation in which an LLM claimed to be free of personal data can still generate outputs that contain such data. This is the *"If it comes out, it must be in there"* argument. The Hamburg DPA has responded to many of these criticisms in a paper published recently (see Thomas Fuchs, "Hamburger Thesen zum Personenbezug in LLMs. Ein Debattenimpuls zur Anwendbarkeit der DS-GVO auf Large Language Models" in: Künstliche Intelligenz und Recht (KIR), October 2024).

Some other critiques may reflect expansive interpretations of the Hamburg DPA's position. For example, Lokke Moerel and Marijn Storm argue that: "If the Hamburg and Danish DPAs' guidance is followed, LLM providers would not be responsible for any inaccurate outputs relating to public persons. This would lead to a gap in data protection…" leaving data subjects without recourse for inaccurate outputs. However, the Hamburg DPA's stance emphasizes that while LLMs themselves do not store personal data and therefore cannot be subject to data subject rights directly, AI systems that process personal data—particularly when such data appear in outputs—must uphold data subject rights. The discussion paper includes examples of how organizations deploying LLMs should address data subject requests, including the provision of information, rectification, and erasure. Therefore, the concern that "if this guidance were followed, LLM chatbot providers would not be responsible for any inaccurate outputs" may not fully capture[2] the nuances of the Hamburg DPA's position, which explicitly calls for LLM chatbot providers to implement GDPR rights at the level of outputs that might contain personal data.

The jury is still out on whether LLMs contain personal data. Other DPAs, such as the French CNIL or the German DSK, seem to have adopted a more nuanced position than the Hamburg DPA, and to advocate for a case-by-case analysis. Based on my discussions with several data scientists, it appears that the absence of personal data is not an inherent property of *all* LLMs[3]. Tokenization choices[4], anti-memorization and de-duplication measures, along with various other safeguards, can help minimize the possibility of extracting personal data.

Regardless of the resolution to this hotly disputed issue, the Hamburg DPA's guidance—praised by some as a genuine effort to "approach regulation and legal compliance with a deep understanding of how the technology actually works"—is of great interest from the perspective of this paper, as it illustrates how fundamentally the functionality of LLMs differs from conventional data storage methods. It demonstrates how the way LLMs process tokens and vector representations, along with their probabilistic and generative functions, fundamentally differs from traditional data storage and retrieval systems where the GDPR's principle of accuracy has traditionally applied. Models are not databases of information or structured repositories of facts or personal data. They do not operate by retrieving information from a database or by "copying and pasting" portions of existing data. As the Hamburg DPA explained: "The GDPR was conceived for a world where personal data is stored and processed in clearly structured databases. LLMs break this framework and present us with the challenge of applying current law to a new technology".

By shifting the focus from the LLM component to the AI system as a whole in the exercise of data subject rights, this guidance offers a GDPR-compliant, innovation-friendly solution to the issue of AI hallucinations. GPAI system providers can mitigate AI hallucinations and effectively uphold data subject rights by concentrating on training data, inputs and, most importantly, outputs, without resorting to measures that stifle innovation, such as constant re-

identification attacks or continuous retraining of their LLMs. Such measures are not only technically challenging and sometimes inefficient—since inaccuracies are probabilistic in nature—but could also entail significant economic and environmental costs.

2. ICO: Focus on Purpose and Transparency

On April 12, 2024, the Information Commissioner's Office (ICO) published provisional guidance and initiated a public consultation on the "Accuracy of training data and model outputs". While the final guidance, enriched by numerous public responses, is expected soon, the provisional document already contains several noteworthy proposals that could complement the suggestions of the Hamburg DPA regarding AI hallucinations.

Unlike the HmbBfDI, the ICO does not address whether LLMs contain personal data; instead, it emphasizes the purpose of the AI system as a whole and underscores the necessity of adequate information and transparency. The ICO correctly points out that the accuracy requirements of an AI system can vary significantly depending on its specific application. Before determining the need for accuracy in a GPAI system's outputs, organizations deploying such technology must first define the intended purpose of the model and assess its suitability for that purpose in collaboration with developers. Generative AI models created solely for creative purposes do not require strict accuracy standards. For example, the ICO cites a scenario where "a model used to help game designers develop storylines" can produce outputs associating fictional elements with real people without a requirement for factual accuracy.

We concur with the ICO's assessment. Probabilistic outputs may lead to AI hallucinations but can also serve as a powerful tool for enhancing human creativity and generating new ideas and content. This potential extends beyond purely recreational or gaming contexts to a variety of generative AI functions used daily by millions of individuals with minimal risk of infringing on the GDPR's principle of accuracy. These functions include creative writing that eschews factual constraints, brainstorming and idea generation, as well as tasks like translation, grammar correction, and offering alternative phrasing for stylistic clarity. Restricting and over-regulating generative AI in the name of GDPR "accuracy" may be counterproductive and stifle innovation in such low-risk scenarios.

Conversely, as the ICO emphasizes, certain uses of GPAI systems can pose risks to data subject rights, necessitating proactive measures by both developers and deployers to mitigate these risks. The ICO illustrates this point with the example of "a model used to summarize customer complaints," which must generate accurate outputs to effectively fulfill its purpose. This requirement demands both statistical accuracy and adherence to data protection standards to ensure customer information is correctly represented and handled. The ICO concludes: *"The specific purpose for which a generative AI model will be used is what determines whether the outputs need to be accurate. Therefore, it is of paramount importance that there is clear communication between the developers, deployers and end-users of models to ensure that the final application of the model is appropriate for its level of accuracy"*.

While some of the ICO's specific proposals for addressing such risks have elicited criticism—such as the suggestion to "provide clear information about the statistical accuracy of the application," which some argue can be counterproductive due to the "accuracy paradox"—its central recommendation is widely regarded as both uncontroversial and valuable. The ICO states, "Developers need to set out clear expectations for users, whether individuals or organisations, on the accuracy of the output". This is crucial, particularly when considering the human tendency to attribute greater capabilities to technology, a phenomenon known as automation or technology bias.

As we will explore in the following section, GPAI developers have already implemented measures to not only filter potentially inaccurate responses but also to foster user skepticism and critical assessment of GPAI outputs. These efforts include enabling users to verify information through links to credible sources, thereby promoting a more thoughtful and cautious engagement with generative AI systems.

3. Companies' Responses to Hallucination Challenges

Creators of GPAI systems have taken multiple steps to address criticisms about previous releases, aiming to reduce the number of hallucinations and mitigate the potential harms they may cause. While the measures adopted are not yet perfect and more work remains on several fronts, these efforts represent steps in the right direction within a dynamic and constantly evolving field. Further progress in managing AI hallucinations in a GDPR-compliant manner could result from ongoing scientific research and practical experience, helping to better address these complex issues.

To address AI hallucinations and comply with the GDPR's accuracy principle, major GPAI developers—including OpenAI (ChatGPT), Google (Gemini), Meta (Llama), Anthropic (Claude), Mistral, and others—have implemented measures across three key areas: input and training data, the large language model (LLM) itself, and, most importantly, system outputs. While the analysis below primarily focuses on the earliest systems available in the EU—namely those developed by OpenAI and Google—these efforts are indicative of broader industry initiatives.

At the level of **training data**, GPAI developers claim to have introduced **data quality control** measures. These involve not collecting data from untrusted sources and implementing rigorous filtering and cleaning processes during data curation to minimize biased, outdated, or incorrect information that could lead to inaccurate model outputs. They also state that they work on **bias reduction** by analyzing training datasets for patterns of bias and applying techniques such as reweighting, debiasing algorithms, and diverse data sampling to reduce the impact of skewed data distributions.

Regarding the **LLM itself**, GPAI developers assert that they have adopted a series of measures, starting with **model architecture improvements**. These consist of optimizing model architectures and training protocols to enhance interpretability, mitigate issues like overfitting to incorrect patterns, and promote accurate generative behavior. Other measures include **developing hallucination guardrails**—designing specific mechanisms within the LLM to maintain adherence to factual accuracy—and employing **reinforcement learning from human feedback (RLHF)** to refine responses based on human reviewers' input, focusing on accuracy, relevance, and reducing the potential for hallucinations. RLHF is indeed an important post-training technique to teach a model to follow instructions, to decrease the likelihood of it returning inaccurate content, and to add safety features. It does this by having people write sample answers and rate answers provided by the model, and provide those samples and ratings back to the model in follow-up training processes. For example, human reviewers might be asked to pick between a range of different outputs before ranking them in terms of various criteria, like factuality of math responses or responsiveness to the question. Researchers would then use the sample output and ratings to try to teach the model to produce output that is closer to the output that was ranked highly. To **measure the factuality** of language models, GPAI developers use different means such as the recently published new OpenAI benchmark called SimpleQA. **Regularly updating LLMs** with new data and retraining them to reflect more accurate, current, and balanced knowledge is also a major measure used to prevent outdated information from persisting in responses.

However, most of the measures adopted, including several related to data subject rights, concern the level of **GPAI system outputs**. As Google explains:

> *"in general, the focus of privacy controls should be at the application level, where there may be both greater potential for harm (such as greater risk of personal data disclosure), but also greater opportunity for safeguards. Leakages of personal data, or hallucinations misrepresenting facts about a non-public living person, often happen through interaction with the product, not through the development and training of the AI model".* (p.22)

The output safeguards introduced by LLM providers can be summarized in the following categories:

a) Technical Measures to Avoid Inaccurate Outputs

GPAI developers are introducing several methods and tools to limit the number of inaccurate outputs. These include **prompt engineering techniques**—fine-tuning the way the model interprets and responds to prompts to increase factual reliability and minimize potential misinterpretation—and methods that detect when a prompt is likely to produce a confabulation.

Many GPAI systems are capable of **Named Entity Recognition (NER) systems** to identify references to individuals, such as names, job titles, or personal attributes within generated outputs. This recognition capability allows the model to apply additional scrutiny to outputs involving identifiable persons to verify accuracy and relevance. LLM developers then employ sophisticated **output filters** that analyze and moderate content generated by the AI in real-time.

GPAI system developers have declined to block all outputs concerning identifiable persons, as such a move could limit the usefulness of an LLM.[5] However, they can use "a public name detector or another classifier to determine whether a person's name is likely in the query, and if so, whether that person is a public or private figure. If the classifier determines the question likely contains the name of a non-public figure, it can take action to not respond".

Similarly, before generating output about a person, a GPAI system can check whether there is a removal request for that person. Where a removal request has been approved, the system can suppress corresponding outputs. Following the Google v. Spain case law on search engines, which applies here *mutatis mutandis*, a GPAI system provider could, in some cases, refuse to approve such a request if it considers that this is not desirable due to "the role played by the data subject in public life" and "the preponderant interest of the general public in having access to the information" about this person. However, in such cases, every reasonable effort should be undertaken, "within the framework of [the a GPAI system providers] responsibilities, powers and capabilities",[6] as the CJEU famously said, to remove any inaccurate information about such a public person. Contrary to what happened in the Noyb case described above, output filters might become able in the future to remove specific parts of inaccurate information about a public person without removing all outputs concerning that person.

More generally, GPAI systems already use rule-based approaches and advanced algorithms to detect and flag potentially inaccurate or harmful statements about identifiable persons. By incorporating automated detection, the filters can block or alter content that appears false, defamatory, or otherwise harmful, preventing the generation of content that could lead to privacy breaches or spread misinformation.

Output filters may employ real-time **post-processing and fact-checking tools** to validate statements involving identifiable individuals and prevent the dissemination of hallucinated content. Retrieval-augmented generation (RAG) and dynamic real-time information injection are AI frameworks for improving the quality of LLM-generated responses by grounding the model on external sources of knowledge to supplement the LLM's internal representation of information. Grounding answers with factual information and fetched data (using methods like

web browser plug-ins) is a technique increasingly used to connect model outputs to verifiable sources of information. LLM output filters are also generally configured to avoid unnecessary references to identifiable persons unless explicitly relevant and requested by the user.

b) Transparency, Information, and User Empowerment

LLM deployers have introduced warnings within the LLM interfaces to alert users that the AI systems "can make mistakes" (ChatGPT) or "display inaccurate information, including information about individuals" (Gemini), inviting them to "check the answers". We view these not as "disclaimers" but rather as **"explainers"** that inform users about how GPAI systems work.

The **terms of service and privacy policies** of GPAI system deployers provide much more detailed information about how these systems function and the risks associated with AI hallucinations.

Open AI, for instance, notes that:

> *"We design our AI models to be learning machines, not databases. AI models learn from relationships in information to create something new; they don't store data like a database. When we train language models, we take trillions of words, and ask a computer to come up with an equation that best describes the relationship among the words and the underlying process that produced them. After the training process is complete, the AI model does not retain access to data analyzed in training. ChatGPT is like a teacher who has learned from lots of prior study and can explain things because she has learned the relationships between concepts, but doesn't store the materials in her head".*

It goes on to explain that:

> *"Services like ChatGPT generate responses by reading a user's request and, in response, predicting the words most likely to appear next. In some cases, the words most likely to appear next may not be the most factually accurate. For this reason, you should not rely on the factual accuracy of output from our models. If you notice that ChatGPT output contains factually inaccurate information about you and you would like us to correct the inaccuracy, you may submit a correction request to dsar@openai.com. Given the technical complexity of how our models work, we may not be able to correct the inaccuracy in every instance. In that case, you may request that we remove your Personal Data from ChatGPT's output by filling out this form".*

In a similar way Google explains that:

> *"LLM experiences (Gemini Apps included) can hallucinate and present inaccurate information as factual. Under [the GDPR], you may have the right to: object to the processing of your personal data; or ask for inaccurate personal data in Gemini Apps' responses to be corrected. To exercise these rights, you can create a request in our Help Center".*

Beyond these explainers and terms of use/privacy policies, some interfaces display messages in response to prompts, reminding users that the GPAI system might not always be accurate or introducing doubt about the factual nature of a response. While keeping in mind the technical difficulties of doing so systematically and the "within the framework of its powers and capabilities" limitation set by the CJEU, we believe that more progress could certainly be made in this field by avoiding overconfident but incorrect outputs about individuals.

Another tool used to encourage and enable users to check the accuracy of results is the introduction of **"double-check"** features. For instance, Gemini's "double-check" feature or ChatGPT Search help users verify its answers by evaluating whether there is content across the

web that substantiates the response. When a statement can be evaluated, the user can click the highlighted phrases to learn more about supporting or contradicting information found by Google Search.

Developers have also introduced **user feedback and reporting systems** to encourage users to report outputs involving identifiable individuals that may be inaccurate or misleading. Such reports trigger internal investigations and allow continuous improvement of filtering mechanisms based on real-world feedback. They also lead, at a subsequent stage, to updates in the model's training or fine-tuning procedures.

c) Towards an era of reason?

User empowerment can become significantly more impactful with the new generation of GPAI systems that generate internal chains of thought before answering questions and claim to be capable of reasoning. Utilizing inference-time computation to allow the model to "reason" might reduce hallucinations and enable models to say “I don’t know” or otherwise avoid overconfident responses. This approach also empowers users to inspect the reasoning steps behind responses, helping them assess the model’s logical consistency and thus question the answers provided. More generally, as Rob Van Eijk from the Future of Privacy Forum (FPF) explained to me, progress in **neuro-symbolic AI**—a type of artificial intelligence that integrates neural and symbolic AI architectures to address the weaknesses of each—could be extremely useful in reducing AI hallucinations and increasing user empowerment in this field. Neuro-symbolic AI combines the pattern recognition capabilities of neural networks with the reasoning and knowledge representation strengths of symbolic AI, resulting in robust AI capable of reasoning, learning, and cognitive modeling. This integration could significantly enhance the ability of AI systems to provide accurate and reliable responses, thereby mitigating the occurrence of hallucinations.

d) Enabling the Exercise of Data Subject Rights

As previously shown, LLM developers have introduced mechanisms that allow data subjects to exercise their rights, including the rights to erasure or rectification in cases of inaccurate outputs. For instance, OpenAI states in its Privacy Policy:

> *“You have the following statutory rights in relation to your Personal Data: Access your Personal Data and information relating to how it is processed; Delete your Personal Data from our records; Rectify or update your Personal Data; […] Restrict how we process your Personal Data; […] Object to how we process your Personal Data when our processing is based on legitimate interests. You can exercise some of these rights through your OpenAI account. If you are unable to exercise your rights through your account, please submit your request through https://privacy.openai.com or send it to dsar@openai.com.”*

Conclusion

Addressing AI hallucinations in GPAI systems presents a complex challenge: how can we uphold the GDPR's accuracy principle without inadvertently stifling technological innovation? Overly stringent interpretations of the GDPR—such as demanding absolute accuracy, imposing complete removal of personal data from outputs, mandating immediate elimination or rectification of personal data from models themselves, or imposing fines for any instance of inaccuracy (even if no harm results, such as when a GPAI system guesses a celebrity’s birth

date incorrectly)—could lead to costly and technically challenging measures. While aiming to protect individuals, such approaches might also limit the adoption and utility of AI applications in the EU, risking the EU's ability to use its computer science expertise for innovation and effective competition with the rest of the world.

The guidance provided by the Hamburg DPA and the ICO offers a more pragmatic and flexible approach to this issue. By focusing on the outputs of AI systems rather than the internal workings of LLMs (which are, in any way, covered by the EU AI Act[7]), the Hamburg DPA acknowledges that even if LLMs themselves may not store personal data in a traditional sense, the outputs generated can still impact data subject rights. This shift in focus allows for the protection of individuals without necessitating restrictive measures that could impede responsible technological progress in Europe. This also seems to align with the methodology followed by the CJEU in relation with search engines, which focused on *ex post-facto* remedies, on the basis of a de-referencing request by the data subject.

Similarly, the ICO emphasizes the importance of the purpose and context in which AI systems are used. By advocating for a **risk-based approach** that considers the intended application of the AI system, the ICO encourages developers and deployers to tailor accuracy requirements appropriately. This perspective recognizes that not all AI outputs need to meet the same levels of accuracy—creative applications may tolerate more flexibility, whereas systems used in critical contexts like healthcare or legal services require stricter accuracy standards.

GPAI systems providers have taken steps to align with these regulatory insights by implementing technical safeguards, enhancing transparency, and enabling the exercise of data subject rights. As developers themselves acknowledge, these solutions are yet far from perfect, and more work needs to be done on several fronts, including introducing more clarity about the purposes of GPAI systems, as the ICO invites them to do. Still, these measures seem like steps in the right direction in a dynamic and constantly evolving field. Further progress in managing AI hallucinations in a GDPR-compliant way could result from ongoing scientific research and practical experience, helping to better address these complex issues. Furthermore, when LLMs are used by third parties—for example, via an application programming interface (API)—the responsibilities of each party in managing the risk of inaccurate outputs must be clearly defined.

Importantly, the scope of this article has focused solely on AI hallucinations, the GDPR principle of accuracy, and related data subject rights. It has not addressed other significant issues, such as the conditions and legal basis under which LLMs can be trained with publicly available personal data. The interpretation of the GDPR on these matters could significantly and in very different ways affect the challenges and solutions related to AI hallucinations and the accuracy principle. Decisions by the European Data Protection Board (EDPB) and Data Protection Authorities (DPAs) on these issues, particularly the forthcoming Article 64(2) GDPR Opinion expected on December 23, 2024, could undoubtedly influence the ongoing debate on accuracy and related rights.

Continuous collaboration and dialogue among regulators, industry stakeholders, civil society, and researchers remain essential to further develop effective measures. This engagement will help refine existing strategies, address emerging challenges, and support the development of AI systems that are both innovative and aligned with fundamental data protection principles. With a balanced and adaptive regulatory framework, it may be possible to harness the transformative potential of generative AI while safeguarding the rights and interests of individuals.

[1] As stated in p. 7: "The Danish Data Protection Agency assumes that an AI model as a clear starting point does not in itself contain personal data, but is only the result of the processing of personal data. This results from the fact that a statistical report is also not be considered personal data if the report only contains conclusions and aggregated data that are the results of the statistical analysis".

[2] The criticism by Moerel and Storm seems to concern however only situations where LLM providers make their models available via an API to third-party deployers – an issue not discussed in this paper as mentioned in my introduction. Such situations definitely require more research and adequate legal and technical measures in order to protect data subject rights.

[3] For example, an overparameterized function (like an LLM) could be fit on a small dataset such that the dataset is preserved in its entirety (similar to a compressed file).

[4] As Rob Van Eijk from the Future of Privacy Forum (FPF) summarized it to me: "Tokenization affects data through a connected series of stages in language models. During initial processing, text gets divided into tokens (parts of words, whole words, or punctuation marks) and converted to numerical IDs (tokens) in the model's vocabulary. The tokens are stored in the model's vocabulary section. Memorization refers to when a language model reproduces content from its training data. Memorized information is retained within the attention layers of the transformer architecture. The retention of information in LLMs occurs at the parameter level, where repeated patterns in the vocabulary become encoded across the model's weights. During output generation, the model can reconstruct personal information through token prediction, potentially revealing generalizable patterns from training data or even verbatim reproduction of training sequences, or misrepresented in hallucinations"

[5] A concrete example—and a common use case for LLM products—where blocking all personal data in outputs could harm user experience is revising a resume or CV (e.g., when a user asks the model to critique or improve a CV). This also highlights an important distinction between personal data in the training data and data present in the model's context during inference.

[6] The GDPR introduces this idea in Article 5(1)(d), which requires that "every reasonable step must be taken" to correct inaccurate data. This phrasing suggests that measures deemed excessively costly or technically unfeasible might not meet the "reasonable" threshold. Similarly, Article 25(1) of the GDPR, addressing "data protection by design and by default," allows for consideration of factors such as "the state of the art, the cost of implementation, and the nature, scope, context, and purposes of processing," as well as the risks of varying likelihood and severity to individuals' rights and freedoms, when implementing measures to uphold data protection principles.

[7] Indeed, as the Hamburg DPA explains, "feared regulatory gaps will be closed by the AI Act, which came into force in August 2024, according to which LLMs can be regulated as AI models and removed from the market in case of legal violations (cf. Art. 93(1)(c) AIA)".

Acknowledgments, Funding and Disclaimers

The author expresses gratitude to the numerous individuals who provided valuable feedback on earlier versions of this study. Special thanks are due to Chrysi Chrysochou, Rob van Eijk, Lokke Moerel, Peter Swire, and Yann Padova for their insightful contributions. Any remaining errors or omissions are solely the responsibility of the author.

This independent study was supported by funding from the Computer and Communications Industry Association (CCIA Europe). The views and opinions expressed in this work are those of the author alone and do not necessarily reflect the positions of CCIA Europe or its members.