BRANDI
RECHTSANWÄLTE

USER CONTENT AS TRAINING DATA FOR AI

Information on data protection I September 2025

Introduction

AI models, especially language models such as ChatGPT (OpenAI) or Gemini (Google), rely on training data. The collection and compilation of training data is in itself a form of data processing. The more extensive and diverse this data is, the more powerful the models become. Providers use various means to obtain the data. The use of their own data or data licensed from third parties is unproblematic. However, the use of content that users of online services share online themselves is particularly controversial. The data material may regularly include personal data, in which case the requirements of the General Data Protection Regulation (GDPR) must be observed. For this reason, the use of content published by users of online services raises questions about the legal framework for data protection, in particular the relevant legal basis and the rights of data subjects.

The framework conditions for AI training

When training AI models with content published by users of online services, compliance with the individual principles of Art. 5 GDPR appears problematic at first glance.

The principle of transparency under Art. 5 (1) (a) (3) GDPR requires that data processing must take place in a manner that is comprehensible to the data subject. However, the resulting information obligations under Art. 14 (5) (b) GDPR are dispensable if they are impossible to fulfill or can only be fulfilled with disproportionate effort. AI training is characterized by autonomous processing and self-learning processes, meaning that data processing by AI is often not comprehensible to the providers themselves, making it difficult and disproportionately burdensome for them to fulfill their information obligations. Consequently, the principle of transparency can be complied with when processing personal data for the purpose of AI training.

According to the principle of data minimization under Art. 5 (1) (c) GDPR, data processing must be limited to what is necessary for the purposes. The fact that AI models rely on large amounts of data for their quality and functionality does not preclude compliance with this principle. The principle of data minimization does not imply an obligation to process only a small amount of data. Rather, it is intended to ensure that data processing is carried out for a specific purpose. Providers should therefore not process data that does not serve the purpose of developing and improving the AI model. Here, when considering each piece of personal information individually, it may become apparent that this particular piece of information is not necessary to achieve the purpose. However, an overall assessment must be made, i.e. whether the data as a whole is considered necessary for training a high-quality AI model. This can be affirmed in the development and improvement of AI models with the 'the more, the better' argument.

The principle of data accuracy under Art. 5 (1) (d) GDPR is intended to ensure that data is factually correct and, where necessary, up to date. Inaccurate data should be deleted or corrected without delay using appropriate means. The quality of the output of an AI model depends on the quality of the previous input. Compliance with this principle is also in the provider's own interest in a high-quality AI model, so that the provider will generally endeavor to ensure the accuracy of the data within the scope of technical possibilities.

When considering the individual principles, it should also be noted that the GDPR is based on an interpretation that promotes innovation. Compliance with the individual principles of Art. 5 GDPR should therefore also be possible in the case of innovative developments, such as AI models.

Legal basis for AI training

The processing of personal data is generally subject to a preventive prohibition with reservation of permission pursuant to Art. 6 (1) GDPR. This means that data processing is generally considered unlawful and can only be justified on the basis of Art. 6 (1) (a) to (f) GDPR. For the processing of data published by users of online services, the legal bases of consent pursuant to Art. 6 (1) (a) GDPR and legitimate interest pursuant to Art. 6 (1) (f) GDPR are particularly relevant.

Consent according to Art. 6 (1) (a) GDPR

The lack of contact with users of online services poses a particular obstacle to obtaining consent in accordance with Art. 6 (1) (a) GDPR. The provider would have to identify in advance the individuals whose published data is being used for AI training and obtain their consent. In addition, consent pursuant to Art. 4 No. 11 GDPR must be given in an "informed manner," which is hardly feasible due to the complex and autonomous processing procedures of AI. If individuals later revoke their initial consent, the data would have to be deleted. This is hardly technically feasible once the AI has already been trained. In addition, providers run the risk of violating the prohibition of coupling under Art. 7 (4) GDPR. According to this provision, the voluntary nature required for consent is excluded if the performance of a contract is made dependent on consent and the specific data processing is not necessary for the performance of the contract. The necessity of collecting and using data published by users in the context of AI training is generally to be denied for the operation and provision of an online service. Thus, the provision of the online service may not be made dependent on consent—not even de facto through the restriction of certain functions of the online service. Ultimately, consent would potentially lose value due to a lack of actual choices for the user and inadequate provision of information, and would be degraded to mere "de facto" consent.

Legitimate interest according to Art. 6 (1) (f) GDPR

The legitimate interest pursuant to Art. 6 (1) (f) GDPR is associated with a certain degree of legal uncertainty due to the need for consideration, but it also allows the technological specifics of AI training to be taken into account.

First, there must be a legitimate interest for the provider in processing the data. Here, economic interests come to the fore, for example, in order to offer the AI model on the market. In the medical and healthcare sectors, societal and social interests may exist. In the field of research, an idealistic interest can be derived from freedom of science and information.

Data processing must then be necessary to achieve the legitimate interest. The collection and gathering of data is essential for training AI during the development phase. However, if the data is no longer used for the development of AI but for the (minimal) improvement of the quality of AI that has already been developed, the necessity appears questionable. However, especially in the context of economic interests, AI models are subject to competition, and even a minimal improvement in quality is of great importance. In addition, the necessity of achieving the interest must be measured by equally effective means. As long as an improvement in quality depends on a larger data set, other means are not equally effective and data processing remains necessary.

Finally, the interests must be weighed comprehensively against the conflicting interests, fundamental freedoms, and fundamental rights of users. On the user side, it should be noted that users are generally unaware of the processing of their published data and therefore cannot assert any rights as data subjects. In addition, there are certain reverse engineering techniques that can be used to reconstruct and disclose the training data at the output level of the AI. On the part of the providers, it should be noted that they can make use of various organizational and technical security measures. This can prevent or at least sufficiently mitigate risks arising from reverse engineering techniques in particular. If sensitive data categories are processed in accordance with Art. 9 (1) GDPR, data processing can be based on Art. 9 (2) (e) GDPR. In the case of sensitive categories of data, it is important that the data must have been published independently by the users and that the mere public nature of the data is not sufficient. Another argument in favor of the providers is that users may have to expect the use of the data they have published. Whether and when individuals can reasonably expect their public personal data to be used depends largely on the nature of the relationship between the individual and the provider, the type of online service, the context in which the personal data is collected, and the individual's actual knowledge that this personal data is online.

In summary, legitimate interest pursuant to Art. 6 (1) (f) GDPR constitutes a practical legal basis.

OLG Cologne on the processing of personal data for AI systems

The courts have also accepted the processing of user data for the purpose of AI training on the basis of legitimate interest pursuant to Art. 6 (1) (f) GDPR. In summary proceedings before the Higher Regional Court of Cologne, the Consumer Advice Center of North Rhine-Westphalia and a subsidiary of Meta Platforms Inc. faced off (OLG Cologne, decision dated 23.05.2025 - 15 UKl 2/25). The subsidiary is Meta Platforms Ireland Limited, which operates the online services Instagram and Facebook. The Consumer Advice Center sued Meta Platforms Ireland Limited for an injunction against data processing due to violations of Art. 6 and 9 GDPR after Meta announced that it would use public content from its adult users on Facebook and Instagram to train its AI.

The court considered the intended data processing to be lawful. In the court's opinion, the processing could be based on Article 6 (1) (f) GDPR. According to this, data processing is permissible if it is necessary to safeguard the legitimate interests of the controller and does not outweigh the interests of the data subject. The training of the AI model constitutes such a legitimate interest. The court also found that the processing of the data was necessary for the purpose of AI training.

Rights of data subjects

Persons affected by the processing of personal data are entitled to comprehensive rights under the GDPR. When processing data published by users of online services, the right to object under Art. 21 GDPR is of particular importance. Users can object to data processing by using the opt-out functions of online services. However, the objection only applies to content published on the respective user account. If personal data about the objecting user is published on other user accounts, data processing remains possible. An exception exists exclusively for sensitive data categories pursuant to Art. 9 (2) (e) GDPR. A delayed objection does not reverse the data processing that has already been carried out, and deletion is not feasible with the current technical capabilities due to autonomous processing procedures and self-learning processes. As a result, effective protection through the right to object requires that all users in an environment object and do so in a timely manner. Another way to protect your published personal data is to lodge a complaint with the competent data protection authority. However, this may result in longer processing times during which data processing continues.

Conclusion

Access to training data, as well as the quality of this data, is crucial to the success of AI models. The processing of content published by users of online services, which always includes personal data, is therefore of great importance for the training of AI models. It is to be expected that more and more providers will make use of this content in the future.

First of all, it can be said that, upon closer examination, the apparent conflicts with the principles of the GDPR do not preclude the use of public information for training AI models. Rather, the aim is to strike a balance between the protection of personal data and the development and improvement of AI models.

Data processing can be based on the legitimate interest in developing and improving the AI model in accordance with Art. 6 (1) (f) GDPR. The publication of content by the user themselves or by other users significantly reduces the need for protection of personal data, so that the interests of users in protecting their data do not generally outweigh and do not prevent data processing. In contrast to consent pursuant to Art. 6 (1) (a) GDPR, legitimate interest does not require the cooperation of users of online services. Rather, users must act on their own initiative, for example by making use of the opt-out functions provided by online services and, if necessary, encouraging those around them to do the same. As a result, awareness should be raised that one's own behavior and use of online services can form the basis for the extensive processing of one's own personal data.

Marc-Levin Joppek / Mira Husemann

Contact:

BRANDI Rechtsanwälte
Partnerschaft mbB
Adenauerplatz 1
33602 Bielefeld

Marc-Levin Joppek
Research Associate
T +49 521 96535 - 890
F +49 521 96535 - 113
E levin.joppek@brandi.net

Mira Husemann
Research Associate
T +49 521 96535 - 890
F +49 521 96535 - 113
E mira.husemann@brandi.net