ltz.nomos.de

LegalTech

01 | 2025
04. Jahrgang
Seiten 1 – 96

Zeitschrift für die digitale Rechtsanwendung [LTZ]

Aus dem Inhalt

Nomos

Digitaler Zugriff:
beck-online
DIE DATENBANK

ders verlässt (*permutation feature importance*) sowie kontrafaktische Erklärungen (*counterfactual explanations*), die Auskunft darüber geben, wie eine geringfügige Veränderung der Eingabedaten zu einer anderen Entscheidung hätte führen können.

Diese Informationen werden den Betroffenen jedoch in der Regel nicht in die Lage versetzen, komplexe algorithmische Entscheidungen auf ihre Richtigkeit überprüfen zu können, sofern nicht – wie im Ausgangsfall – offensichtlich widersprüchliche Informationen vorliegen. Inwieweit zwischen der verwendeten Methode und den herangezogenen Kriterien einerseits und dem Ergebnis der automatisierten Entscheidung andererseits eine „objektive nachprüfbare Übereinstimmung" und ein „objektiv nachprüfbarer Kausalzusammenhang" besteht, kann letztlich nur von Experten und Aufsichtsbehörden auf der Grundlage vollständiger und kontextbezogener Informationen beurteilt werden, die dann jedoch nicht nur Zugang zu den Datenbanken und Trainingsmodellen, sondern auch – wie in Art. 74 Abs. 13 KI-VO vorgesehen – Einsicht in Algorithmen und den zugrundeliegenden Quellcode bekommen müssten. Es ist daher schon vom Ansatz her verfehlt, wie GA de la Tour sowohl eine allgemeine Verständlichkeit der Informationen als auch eine Überprüfbarkeit der algorithmischen Entscheidung zu fordern.

Prof. Dr. Martin Ebers

The case of LAION: The first public (German) court decision on text and data mining (TDM) in the context of machine learning

Paulina Jo Pesch

The Regional Court of Hamburg has published the first German and probably also the first European court decision on text and data mining (TDM) in the context of machine learning. Albeit the decision concerns the preparation of training datasets only, the judgment addresses the broader question of whether reproductions of works in the context of training (generative) machine learning models can be permissible under Art. 3 f. DSM Directive[1], Sections 44b, 60d of the German Copyright Act[2] . A shortened machine translation of the judgment is provided below. The subsequent comment briefly introduces the relevant technical basics and discusses the court's statements, identifying both strengths and weaknesses of the judgment. The comment discusses especially whether, and if so, to which extent machine learning training constitutes TDM within the meaning of Art. 3 f. DSM Directive, and Sections 44b, 60d of the German Copyright Act by which the German legislator has transposed the DSM provisions into national law. Furthermore, the comment analyses the requirement of machine-readability for rights reservations concerning online content.

Paulina Jo Pesch is an Assistant Professor of Civil Law, Law of Digitalisation, Data Protection Law and Artificial Intelligence Law at FAU Erlangen-Nuremberg. Her research focusses on technologies that pose specific challenges to privacy. At present, she mainly investigates legal issues of image-generative AI Models and Large Language Models.

Sections 60d, 44b of the German Copyright Act, Art. 3, 4 DSM Directive

Editorial headnotes of the judgment

1. **The automated analysis of the consistency of image files with their textual descriptions aims at obtaining information about correlations and, therefore, constitutes text and data mining within the meaning of Section 44b para 1 of the German Copyright Act.**
2. **Reproductions for the purpose of machine learning training are not excluded from the scope of the TDM limitations laid down in Sections 44b para 2, 60d para 1 of the German Copyright Act.**
3. **Scientific research within the meaning of Section 60d para 1 of the German Copyright Act includes preparatory steps aimed at subsequent knowledge gain (here: the creation of a machine learning training dataset).**
4. **The creation of a dataset for the purpose of making it available to the public free of charge is an act in pursuit of non-commercial purposes pursuant to Section 60d para 2 no. 1 of the German Copyright Act, irrespective of any commercial use of the dataset by third parties.**

Regional Court of Hamburg, judgment of 27 September 2024 – 310 O 227/23 – LAION[3]

1 Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market.
2 Urheberrechtsgesetz (Gesetz über Urheberrecht und verwandte Schutzrechte – UrhG).

Tenor

1 1. The action is dismissed.

2 2. The plaintiff bears the costs of the proceedings.

3 3. The judgment is provisionally enforceable. The plaintiff may avert the defendant's enforcement by providing security in the amount of 110% of the amount enforceable under this judgment, unless the defendant provides security in the amount of 110% of the enforceable amount prior to enforcement.

Facts of the case

4 […]

5 The defendant […, the non-profit organisation LAION] makes a so-called dataset for image-text pairs publicly available free of charge under the name [LAION-5B]. This is a type of table document that contains hyperlinks to publicly accessible images or image files on the internet as well as further information about the corresponding images, including an image description (also known as alternative text) that provides information about the content of the image in text form. The dataset comprises 5.85 billion corresponding image-text pairs. The dataset can be used to train so-called generative artificial intelligence.

6 The creation of the dataset took place after the defendant's establishment in the second half of 2021. The […] defendant had used an existing dataset […], which contained the respective URLs along with textual descriptions of the content of the images, representing a kind of random cross-section of images available on the internet. The defendant then extracted the URLs to the images from this dataset and downloaded the images from their respective storage locations. The defendant then used software to check the images to assess whether the description of the image content […] actually matched the content to be seen in the image. Images for which the text and image content did not match sufficiently were filtered out. For the remaining images, the metadata, in particular the URL of each image's storage location and description, were extracted and transferred to a newly created dataset, [LAION-5B]. Whether the downloaded image files were subsequently deleted is in dispute between the parties – at least with regard to the photograph in dispute.

7 As part of the aforementioned process, the image in dispute was also captured, downloaded, analysed and included with its metadata in [LAION-5B]. Specifically, the image file was downloaded from the website of the stock agency […].

8 On the website of the stock agency […] the following text has been on the subpage https://www…com/de/usage.html since at least 13 January 2021:

9 "RESTRICTIONS YOU MAY NOT: (…)

18. use automated programs, applets, bots or the like to access the …com website or any content thereon for any purpose, including, by way of example only, downloading content, indexing, scraping or caching any content on the website."

10 The plaintiff alleges an infringement of copyright in the photograph at dispute in the form of unauthorised reproduction by the defendant as part of the analysis process.

11 The plaintiff claims that he is the author of the photo [at dispute]. The [stock agency] would have been entitled to offer and also publish on its website the photo in dispute, and to offer licenses for the photo; [the stock agency] would have been in this respect the owner of [non-exlusive], sublicensable rights of use.

12 The – undisputed – reproduction that took place as part of the analysis process would have infringed the plaintiff's rights under Section 16 of the German Copyright Act, in particular it would not have been covered by the limitation provisions of Sections 44a, 44b and 60d of the German Copyright Act:

13 The limitation provision of § 44a of the German Copyright Act (UrhG) would not be applicable, as the independent download of a photograph does not, in particular, constitute a temporary act within the meaning of this provision.

14 The reproduction would also not be covered by Section 44b of the German Copyright Act. The aggregation of data for the purpose of AI training would not be text or data mining within the meaning of Section 44b of the German Copyright Act. Neither the European nor the German legislator would have had such a use "in mind" when creating the limitation provision of Art. 4 DSM Directive or Section 44b of the German Copyright Act. In the case of text and data mining within the meaning of Section 44b of the German Copyright Act, only "information hidden in the data should be made accessible", "but not the content of the intellectual creation should be used". However, the so-called "AI web scraping" at issue here would be precisely about the intellectual content of the works used for training purposes "and ultimately about the creation of identical or similar competing products". […]

15 Additionally, "the mass incorporation of copyright-protected works for training purposes in the context of generative AI" would impair the normal exploitation of copyright-protected works, as it would create the conditions to replace authors in many cases or, at the very least, significantly would hinder the exploitation of the work due to free competing [products]. However, according to Art. 7 para 2 DSM Directive in conjunction with Art. 5 para 5 InfoSoc Directive, this would preclude the application of the limitation rule.

16 In any case, the reproduction would be inadmissible due to the reservation of rights declared on the website […] pursuant to Section 44b (3) of the German Copyright Act. The corresponding declaration of the stock agency would be attributable to the plaintiff, as it distributes the photo in dispute for him. Contrary to the defendant's view, the reservation would

3 The judgment was machine-translated using both, DeepL translator and ChatGPT 4o. No guarantee is given for the accuracy or completeness of the translation. With regard to the defendant, LAION, the judgment was de-pseudonymised, as the identity of the defendant is not only obvious but has also been clarified by its legal councils, see Heidrich Rechtsanwälte, press release (English version below), https://www.recht-im-internet.de/presseanfragen/pressemeldung-laion (last accessed on 25 November 2024).

also be machine-readable within the meaning of Section 44b (3) sentence 2 of the German Copyright Act. [T]he text would […] have been recognisable as a reservation for a computer program [as] specific tools such as WebOpt-Out would be able to recognise also reservations such as on [the website of the stock agency].

17 Furthermore, the defendant could not invoke the limitation provision of Section 60d of the German Copyright Act. The plaintiff disputes that the defendant fulfils the requirements of Section 60d of the German Copyright Act in fact, namely

[…]

21 – that the defendant is exclusively engaged in research or was engaged in such activities at the time of the reproduction in question; that the defendant conducts scientific research, pursues non-commercial purposes, and reinvests all profits into scientific research or operates within the framework of a government-recognised mandate in the public interest. Furthermore, according to the defendant's submitted statutes, its purpose is solely the 'promotion of research' and not 'research' itself. It would also be unclear what aspect of the collection created by the defendant, which is (undisputedly) made available to other companies, constitutes research;

[…]

23 – that the defendant aimed to provide other researchers and interested parties the opportunity to train their own AI models; […] the dataset at issue was also used to train […] services [that are] operated by (purely) commercial companies […].

24 Furthermore, the defendant could not invoke […] Section 60d para 2 sentence 3 of the German Copyright Act pursuant to Section 60d of the German Copyright Act. The defendant is apparently working intensively with commercial AI providers[.]

[…]

30 The plaintiff now requests

31 that the defendant be ordered, under penalty of a fine of up to €250,000 for each individual case of violation, or alternatively imprisonment of up to 6 months, to refrain from reproducing and/or allowing the reproduction of the [photograph in dispute] for the creation of AI training datasets, as occurred in the context of the production of the dataset [LAION-5B.]

32 The defendant requests

33 that the action be dismissed.

[…]

35 Above all, however, the (one-time) download of the image in question, which took place while creating the dataset, would indeed constitute a reproduction relevant under copyright law. However, this would be covered by the limitation provisions of Sections 44a, 44b, and 60d of the German Copyright Act (of the German Copyright Act) […].

[…]

41 The reservation would also clearly not have been intended as one under Section 44b para 3 of the German Copyright Act. The fact that, according to the plaintiff's submission, the clause was already present on the website as of January 13, 2021, underscores that it could not have been created "with regard to the provision in Section 44b para 3 of the German Copyright Act," as the legal provision had not yet come into force at that time. Furthermore, it is "not credible" that a U.S.-based provider would rely on a reservation of rights according to German law.

[…]

Reasons for the decision

53 I.

54 The admissible action is unsuccessful on the merits. By reproducing the photograph in dispute, the defendant has infringed the plaintiff's exploitation rights. However, this interference is covered by the limitation provision of Section 60d of the German Copyright Act. Whether the defendant can additionally invoke the limitation provision of Section 44b of the German Copyright Act does not need to be conclusively assessed against this background.

55 The photograph in dispute is in any case protected as a photograph pursuant to Section 72 para 1 of the German Copyright Act. After inspecting the raw data on the plaintiff's laptop, the court also has no doubts about the plaintiff's status as the photographer, Section 72 para 2 of the German Copyright Act. The plaintiff is also entitled to assert infringement claims pursuant to Section 97 of the German Copyright Act, including the claim for injunctive relief pursuant to para 1 of the provision; the fact that the plaintiff has granted the stock agency […] more extensive than (sublicensable) [non-exclusive] rights of use has not been demonstrated by the defendant. The stock agency […] applied a watermark to the photo; this constituted a non-free alteration within the meaning of Section 23 para 1 sentence 1 of the German Copyright Act, which means that the plaintiff's consent as the author was fundamentally required for its use. In the course of the download, the defendant reproduced this version within the meaning of Section 16 para of the German Copyright Act without obtaining the plaintiff's consent.

56 However, the defendant was entitled to do so based on statutory permission. The reproduction was not covered by the limitation provision of Section 44a of the German Copyright Act (hereinafter 1.), and whether the defendant can invoke the limitation provision of Section 44b of the German Copyright Act appears doubtful (hereinafter 2.). However, the latter does not require a final decision in the present case, as the act of reproduction was in any case covered by the limitation provision of Section 60d of the German Copyright Act (hereinafter 3.).

57 1.

58 The reproduction […] is not covered by the limitation provision of Section 44a of the German Copyright Act.

59 Accordingly, temporary acts of reproduction are permitted which are transient or incidental and constitute an integral and essential part of a technological process and whose sole purpose is to enable a transmission in a network between third parties by an intermediary or a lawful use of a work or other subject-matter and which have no independent economic significance.

60 The present reproduction was already neither transient or incidental.

61 a)

62 A reproduction is transient within the meaning of Section 44a of the German Copyright Act if its lifetime is limited to what is necessary for the proper functioning of the technical process in question, whereby this process must be automated in such a way that it automatically deletes the act, i.e., without the involvement of a natural person, as soon as its purpose – enabling the execution of such a process – has been fulfilled. (CJEU, judgment of 16 July 2009 – C-5/08 – Infopaq/Danske Dagblades Forening, para 64 (juris) on Art. 5 para 1 DSM Directive).

63 Insofar as the defendant refers in this respect to the fact that the files were deleted "automatically" as part of the analysis process carried out by [the defendant], this does not establish the transience of the reproduction in the aforementioned sense. Apart from the fact that the defendant has not stated anything about the concrete duration of the storage, the deletion was not "user-independent", but rather due to a corresponding deliberate programming of the analysis process by the defendant.

64 b)

65 A reproduction is incidental within the meaning of Section 44a of the German Copyright Act if it is neither independent of the technical process of which it is a part nor serves an independent purpose (CJEU, judgment of 05.06.2014 – C-360/13, para 43 (juris)).

66 In this case, the image files were downloaded in a targeted manner in order to analyse them using specific software. This means that downloading is not merely an accompanying process to the analysis carried out, but a conscious and actively controlled procurement process upstream of the analysis.

67 2.

68 Whether the defendant can invoke the limitation provision of Section 44b of the German Copyright Act appears to be doubtful in the present case. It is true that the download carried out by the defendant is in principle subject to the limitation provision of Section 44b para 2 of the German Copyright Act, in particular it was carried out for the purpose of text and data mining within the meaning of Section 44b para 1 of the German Copyright Act (hereinafter a)). However, without this requiring a final decision in the present case, there is some evidence to suggest that the act of reproduction was not already permissible under Section 44b para 2 of the German Copyright Act due to an effectively declared reservation of use within the meaning of Section 44b para 3 of the German Copyright Act (hereinafter b)).

69 a)

70 The act of reproduction at issue is in principle subject to the limitation provision of Section 44b para 2 of the German Copyright Act.

71 (1) The download at issue was made for the purpose of text and data mining within the meaning of Section 44b para 1 of the German Copyright Act. Accordingly, text and data mining is the automated analysis of individual or multiple digital or digitised works in order to obtain information, in particular about patterns, trends and correlations. In any case, this is to be affirmed for the act of reproduction at issue in the present case (below (a)); a teleological reduction of the restricted act cannot be deemed appropriate in this respect (below (b)).

72 In the present case, there is therefore no need to decide the further question, which has been discussed in detail in literature, as to whether or not the training of artificial intelligence in its entirety is subject to the limitation provision of Section 44b of the German Copyright Act (in detail on the state of opinion BeckOK UrhR/Bomhard, 42. Ed. 15.2.2024, UrhG § 44b Rn. 11a–11b with further references; see also in detail the study "Urheberrecht & Training generativer KI-technologische und rechtliche Grundlagen", commissioned by the Initiative Urheberrecht and submitted as Annex K11).

73 (a) The defendant carried out the act of reproduction for the purpose of obtaining information on "correlations" in the literal sense of Section 44b para 1 of the German Copyright Act. The defendant downloaded the photograph at issue from its original storage location in order to obtain information about the correlations using software that was already available – apparently the application … from … – to compare the image content with the image description already stored for the text. This analysis of the image file in order to compare it with a pre-existing image description constitutes without further ado an analysis for the purpose of obtaining information about "correlations" (namely the question of the consistency of images and image descriptions). The fact that the defendant analysed the images included in the dataset [LAION-5B] in this way was not disputed as such by the plaintiff.

[…]

75 (b) The act of reproduction at issue is also not to be excluded from the limitation provision of Section 44b of the German Copyright Act by way of teleological reduction.

76 Insofar as an exclusion of the reproduction of data for the purpose of AI training by way of teleological reduction is occasionally advocated in literature on the grounds that Section 44b of the German Copyright Act only covers the extraction of "information hidden in the data", but not the use of "the content of the intellectual creation" (Schack, NJW 2024, 113; in this direction also Dor[n]is/Stober, Urheberrecht und Training generativer KI-Modelle, Annex K11, pp. 67 et seq. with a differentiation between semantics and syntax), there are doubts as to whether this is convincing, as it is not sufficiently clear what the difference is between "information hidden in the data" and "the content of the intellectual creation" in the case of digitized works.

77 Insofar as it is additionally argued that "AI web scraping" is about the intellectual content of the works used for training purposes and "ultimately" about the creation of identical or similar competing products (Schack, ibid.), in the opinion of the Chamber, this argument does not distinguish strictly enough between

78 – firstly, the creation of a dataset (which is the sole subject of dispute here) that can also be used for AI training,

79 – secondly, the subsequent training of the artificial neural network with this dataset, and

80 – thirdly, the subsequent use of the trained AI for the purpose of creating new image content.

81 This latter functionality may already be the aim when the training dataset is created. However, at the time of compiling the training dataset, it is not possible to foresee how successful the second step (training) will be, nor what specific content can be generated by the trained AI in the third step (in the AI application). The specific application possibilities for a rapidly developing technology such as AI are therefore not conclusively foreseeable at the time the training dataset is created and therefore cannot be determined with legal certainty. Due to this legal uncertainty, the mere general intention to obtain future AI-generated content when the training dataset is created is not a suitable criterion for assessing the legal admissibility of the creation of the training dataset as such.

82 Finally, when a teleological reduction of the limitation provision of Section 44b of the German Copyright Act is argued on the grounds that the European legislator "simply did not yet have the AI problem" "on its radar" when the underlying directive provision (Art. 4 DSM Directive) was created in 2019 (Schack, ibid.; likewise for the training of AI models Dor[n]is/Stober, ibid., pp. 71 ff., 87 ff.), this finding alone is clearly not sufficient for a teleological reduction. In particular, it must be taken into account that the technical development in the field of artificial intelligence since 2019 concerns less the type and scope of the (disputed) data mining for the procurement of training data, but rather the performance of the artificial neural networks trained with the data (accordingly, Dor[n]is/Stober, ibid., p. 95, also assume that the mere creation of training datasets "in advance of the actual training" may well fall under the TDM limitation). It should also be noted that the Common Crawl Foundation database retrieved by the defendant has been created since 2008 (!), cf. https://commoncrawl.org/overview.

83 Apart from this, at least the current European legislator of the AI Regulation (Regulation (EU) 2024/1689 of 13.06.2024, OJ L of 12.07.2024 p. 1) has undoubtedly expressed that the creation of datasets intended for the training of artificial neural networks is also subject to the restriction of Art. 4 of the GDPR. According to Art. 53 para 1 lit. c AI Regulation, providers of AI models with a general purpose are obliged to provide a strategy, in particular to identify and comply with a legal reservation asserted in accordance with Art. 4 para 3 DSM Directive.

84 The fact that the creation of datasets intended for the training of artificial neural networks is also subject to the limitation provision of Art. 4 of the DSM Directive also corresponds to the assessment of the German legislator in the context of the implementation of the aforementioned limitation provision in 2021 (Begr. RegE BT-Drucks. 19/27426, p. 60).

85 (c) The so-called three-step test set out in Art. 5 para 5 InfoSoc Directive (in conjunction with Art. 7 para 2 sentence 1 DSM Directive) does not justify a different assessment. Accordingly, the standardised exceptions may only be applied in certain special cases in which the normal exploitation of the work or other protected subject matter is not impaired and the legitimate interests of the rights holder are not unduly infringed. These requirements are met in the present case.

86 The reproduction relevant to copyright law in the present case is limited to the purpose of analysing the image files for their conformity with a pre-existing image description and subsequent entry into a dataset. It is not apparent and is not claimed by the plaintiff that this use would impair the exploitation possibilities of the works concerned.

87 It may be true that the dataset created in this way may subsequently be used to train artificial neural networks and the resulting AI-generated content may compete with the works of (human) authors. However, this alone does not justify considering the creation of the training datasets as an impairment of the exploitation rights to works within the meaning of Art. 5 para 5 of the InfoSoc Directive. This must apply simply because the consideration of merely future technical developments, which are not yet foreseeable in detail, does not allow for a legally certain distinction between permissible and impermissible uses (see similarly (b) above).

88 Since the use of knowledge gained through text and data mining to train artificial neural networks – which could then compete with authors – can never be entirely ruled out given the current state of technological development, the opposing view would ultimately require, in its final consequence, a complete prohibition of text and data mining within the meaning of Section 44b of the German Copyright Act. However, such a complete nullification of the limitation provision would clearly contradict the legislative intent and, therefore, cannot constitute a viable interpretative outcome.

89 (2) The image file downloaded by the defendant was also – which the plaintiff does not dispute – lawfully accessible within the meaning of Section 44b para 2 sentence 1 of the German Copyright Act.

90 A work is "lawfully accessible" in this sense in particular if it is freely accessible on the internet (Begr. RegE BT- Drucks. 19/27426, p. 88).

91 This is to be assumed for the image downloaded by the defendant. Contrary to the plaintiff's initial submission, the defendant did not download the "original image" reproduced in the application for injunctive relief initially formulated in the statement of claim – which would only have been made available by the stock agency […] only if a license had been purchased – but downloaded a version of the image with a watermark from the stock agency. This was obviously the preview image posted on the agency's website for advertising

purposes. However, this watermarked preview image had just been made freely accessible on the Internet by the agency.

92 b)

93 However, there are some indications that the limitation provision of Section 44b para 2 of the German Copyright Act does not apply in the present case – without this requiring a final decision – since there was an effectively declared reservation of use within the meaning of paragraph 3 of the provision; in particular, the reservation of use indisputably declared on the website …com is likely to meet the requirements for machine readability within the meaning of Section 44b para 3 sentence 2 of the German Copyright Act.

94 (1) There is much to suggest that the reservation of use stated on the agency's website was issued by a person authorised to do so and that the plaintiff can also rely on this to protect his own rights.

95 According to the wording of Section 44b para 3 of the German Copyright Act, "the rightholder" can declare the reservation of use. This means that not only declarations of reservation by the author herself, but also by subsequent rights holders, whether they are legal successors or holders of rights derived from the author, must be taken into account. According to the plaintiff's coherent submission […], he had granted the stock agency […] simple rights of use to the original image that could be sublicensed. Consequently, the stock agency itself became the rightholder for the images posted on its platform and was therefore fully entitled to issue a reservation of use under Section 44b para 3 of the German Copyright Act. There is no evidence or claim that there were any agreements with in rem effect within the contractual relationship between the plaintiff and the stock agency that would have precluded this.

96 The plaintiff is likely also entitled to rely on the reservation of rights declared by his licensee. From an economic perspective, the exploitation of the disputed original photograph occurred through the agency. In practice, this meant that the agency made the specific decisions about which third party would be authorised for which type of use; it was under no obligation to contract. In such a situation, the court considers it reasonable that the author, when asserting the prohibition rights retained by him, may rely on a reservation declared by his licensee under Section 44b para 3 of the German Copyright Act.

97 (2) The defendant's objection that the prohibition of use for web crawlers [sic!] stated in the agency's general terms and conditions towards its customers could not be formulated "in relation to Section 44b para 3 of the German Copyright Act" in terms of time alone is also irrelevant. It is not a prerequisite for the legal effects of the declaration that it is consciously declared with regard to a specific version of the law.

98 (3) The wording of the reservation is also sufficiently clear. Art. 4 para 3 of the DSM Directive requires an explicit declaration of the reservation of use. Consequently, this requirement of explicitness must be considered in a directive-compliant interpretation of Section 44b para 3 of the German Copyright Act (see also the explanatory memorandum to the draft bill, BT-Drucks. 19/27426, p. 89). The declared reservation must therefore be made explicitly (not implicitly) and with such precision (concretely and individually) that it unequivocally covers a specific content and a specific use (Hamann, ZGE 16 (2024), p. 134). The reservation of use formulated on the website of the stock agency […] fully meets these requirements.

99 Insofar as it is also argued that a reservation of use declared for all works posted on a website contradicts the expressiveness requirement of Section 44b para 3 of the German Copyright Act (according to Hamann, ibid., p. 148, extending his own abstract derivation), this is not convincing. This is because even the reservation explicitly declared for all works posted on a website can be determined beyond doubt in terms of its scope and content and is therefore expressly declared.

100 (4) Finally, there is considerable evidence to suggest that the reservation of use meets the requirements for machine readability within the meaning of Section 4[4]b para 3 sentence 2 of the German Copyright Act.

101 In view of the underlying legislative intention to enable automated queries by web crawlers [sic!] (see explanatory memorandum to the draft bill, BT-Drucks. 19/27426, p. 89), the term "machine readability" will certainly have to be interpreted in the sense of "machine comprehensibility" (see Hamann, ibid., pp. 113, 128 et seq.).

102 The court tends, however, to consider a reservation of use drafted solely in "natural language" to be "machine-readable" (contrary to what appears to be the prevailing view in the literature; see Hamann, ibid., pp. 131 ff., 146 ff., with further references to the state of opinion, including a reference to an article by the defendant's representatives in this case, namely Akinci/Heidrich, IPRB 2023, 270, 272, who apparently also share the court's perspective. However, the court did not have direct access to the article prior to finalising the judgment). Nevertheless, the question of whether and under what specific conditions a reservation declared in "natural language" can also be considered "machine-readable" will always have to be answered in relation to the state of technological development at the time the relevant work is used.

103 Accordingly, the European legislator has also stipulated within the framework of the AI Act that providers of AI models must have a strategy in place, in particular to identify and comply with a legal reservation asserted in accordance with Art. 4 para 3 of the GDPR "including by means of the most advanced technologies" (Art. 53 para 1 lit. c of the AI Regulation). However, these "state-of-the-art technologies" unambiguously include AI applications that are capable of recognising the content of text written in natural language (according to the defendants' representatives Akinci/Heidrich in particular in the article IPRB 2023, 270, 272, not directly accessible to the Chamber, cited here after Hamann, ibid., p. 148, the latter also affirming this possibility in technical terms). In this respect, there is every indication that the legislator of the AI Act had precisely such AI applications in mind with its reference to "state-of-the-art technologies".

104 An objection to this view is sometimes raised, arguing that it would lead to a circular reasoning: if it was required that the operator of text and data mining must use AI applicati-

ons to verify whether a reservation of use has been declared, then such AI-supported search itself would require pattern analysis, which already would constitute text and data mining as defined in Section 44b para 1 of the German Copyright Act; in other words, the application of the limitation would first determine the permissibility of its own application (see Hamann, ibid., p. 148). The court does not share this assessment: contrary to the above-mentioned view, the copyright-relevant act of use requiring justification is not the execution of a "pattern analysis" as such but rather the reproduction of the copyright-protected work as defined in Section 16 of the German Copyright Act. The argument that the preceding discovery of such works on the internet and their verification to determine whether reservations pursuant to Section 44b para 3 sentence 2 of the German Copyright Act have been declared necessarily requires an additional, quasi-preparatory text and data mining within the meaning of Section 44b para 1 of the German Copyright Act does not appear compelling. This is particularly because one could imagine processing website content using web crawlers [sic!], which result only in transient and incidental reproductions that are already justified under Section 44a of the German Copyright Act.

105 Furthermore, the broader understanding of the term "machine readability" considered by the Chamber is also objected to, as this term is understood more narrowly by the European legislator in a different context. In this context, reference is made to recital 35 of the PSI Directive (Directive (EU) 2019/1024), which requires, among other things, "simple" recognisability for "machine readability" within the meaning of this Directive (according to BeckOK UrhR/Bomhard, 42nd ed. 15.2.2024, UrhG § 44b para 31 with further references); this cannot be assumed for a reservation formulated only in natural language. However, such an argument presupposes that the terms of both directives must be understood in the same way. The Chamber has doubts as to whether such an equation of the terms is convincing, as the directives have different objectives: While the PSI Directive deals with the purely unilateral access of the public to information or the purely unilateral obligation of public authorities to publish certain information, Article 4 para 3 of the DSM Directive deals with a balance between the interests of the users of text and data mining (to be able to operate this as simply and as legally securely as possible) and the interests of the rightholders (to secure their rights as simply and as effectively as possible). In the opinion of the Chamber, this balance of interests cannot be resolved unilaterally in favor of the users of text and data mining by solely considering the simplest conceivable technical solution for them as sufficient for the effectiveness of a declared reservation of use. Such an understanding would also be contradicted by the assessment of the legislator of the DSM Directive, which in recital 18 does not require the declaration of a reservation "in the simplest possible manner", but only "in an appropriate manner". And the German implementing legislator also only requires a declaration in a way that is "appropriate to the automated processes of text and data mining" (explanatory memorandum to the draft bill BT-Drucks. 19/27426, p. 89).

106 In the Chamber's view, it would also be a certain contradiction of values to allow the providers of AI models to develop increasingly powerful text-understanding and text-creating AI models via the barrier in Section 44b para 2 of the German Copyright Act on the one hand, but not to require them to use existing AI models within the framework of the barrier in Section 44b para 3 sentence 2 of the German Copyright Act on the other.

107 Whether and to what extent, at the time of the act of reproduction in dispute in 2021, sufficient technology for the automated content recognition of the disputed reservation of use was already available at the time of the act of reproduction in dispute in 2021 has not yet been demonstrated by the plaintiff; in this respect, the plaintiff has only referred to services available in 2023 […]. However, there are indications that the defendant already had suitable technology. According to the defendant's own submission, the analysis carried out as part of the creation of the dataset [LAION-5B] in the form of a comparison of image content with pre-existing image descriptions obviously also and precisely required the content of these image descriptions to be recorded by the software used. Against this background, there is some evidence that systems were already available in 2021 – especially to the defendant – that were capable of automatically recording a reservation of use formulated in natural language.

108 3.

109 However, the defendant can invoke the limitation provision of Section 60d of the German Copyright Act with regard to the reproduction at issue.

110 Accordingly, reproductions for text and data mining for the purposes of scientific research by research organisations are permitted.

111 a)

112 As explained above, the reproduction was made for the purpose of text and data mining within the meaning of Section 44b para 1 of the German Copyright Act. It was also made for the purposes of scientific research within the meaning of Section 60d para 1 of the German Copyright Act.

113 Scientific research generally refers to the methodical and systematic pursuit of new knowledge (Spindler/Schuster/Anton, 4th ed. 2019, UrhG § 60c para. 3; BeckOK UrhR/Grübler, 42nd ed. 1.5.2024, UrhG § 60c para. 5; Dreier/Schulze/Dreier, 7th ed. 2022, UrhG § 60c para. 1). The term "scientific research", by comprising the methodical and systematic "pursuit" of new knowledge, should not be understood so narrowly as to cover only those steps directly associated with the generation of new insights. Rather, it is sufficient that the step in question is aimed at (future) knowledge generation, as is often the case with various data collections that must first be conducted to later draw empirical conclusions. Notably, the term "scientific research" does not require a subsequent research success.

114 Accordingly, contrary to the plaintiff's opinion, the creation of a dataset of the type at issue, which can form the basis for the training of AI systems, can certainly be regarded as scientific research in the aforementioned sense. Although the creation of the dataset as such may not yet be associated with a gain in knowledge, it is a fundamental work step with the aim of

using the dataset for the purpose of gaining knowledge at a later date. It can be affirmed that such an objective also existed in the present case. It is sufficient that the dataset was – undisputedly – published free of charge and thus made available to researchers (also) in the field of artificial neural networks. Whether the dataset – as the plaintiff claims with regard to the services … and … is also used by commercial companies for the training or further development of their AI systems, is irrelevant because the research of commercial companies is still research – even if not privileged as such under Sections 60c ff. of the German Copyright Act.

115 Against this background, the question in dispute between the parties as to whether the defendant also carries out scientific research in the form of the development of its own AI models in addition to the creation of corresponding datasets is irrelevant.

116 b)

117 The defendant is also not pursuing commercial purposes within the meaning of Section 60d para 2 No. 1 of the German Copyright Act.

118 The question of whether research is non-commercial depends solely on the specific type of scientific activity, while the organisation and financing of the institution in which the research is carried out are irrelevant (Recital 42 InfoSoc Directive).

119 The non-commercial purpose pursued by the defendant in relation to the disputed creation of the dataset [LAION-5B] already results from the fact that the defendant indisputably makes it publicly available free of charge. The fact that the development of the dataset in dispute would also at least serve the development of the defendant's own commercial offer (cf. on this criterion BeckOK IT-Recht/Paul, 14th ed. 1.4.2024, UrhG § 60d para 10) is neither submitted by the plaintiff nor otherwise apparent. The fact that the dataset in dispute may also be used by commercially active companies for training or further development of their AI systems is irrelevant for the classification of the defendant's activity. The mere fact that individual members of the defendant also pursue paid activities with such companies in addition to their work for the association is not sufficient to attribute the activities of these companies to the defendant as its own.

120 c)

121 The defendant is also not barred from invoking the limitation provision of Section 60d of the German Copyright Act pursuant to para 2 sentence 3 of the provision.

122 Accordingly, research organisations that cooperate with a private company that has a decisive influence on the research organiaation and preferential access to the results of scientific research cannot invoke the limitation provision of Section 60d of the German Copyright Act. According to the wording of the provision, the burden of presentation and proof for the actual requirements of the counter-exclusion pursuant to Section 60d para 2 sentence 3 of the German Copyright Act lies with the plaintiff.

123 (1) Insofar as the plaintiff initially referred to the fact that the company … had direct influence on the defendant through the financing of the dataset in question and the filling of "relevant positions" at the defendant by its own employees […], this submission lacks substance.

124 In this respect, the plaintiff merely refers to the fact that one of the co-founders of the defendant, Mr. …, is employed at … as "Head of Machine Learning Operations", and that a member of the defendant, Mr. …, is also employed there as a "Research Scientist" […]. This activity of two members of the association for the company … does not prove that this company has a decisive influence on the defendant's research work.

125 Apart from this, the plaintiff has not even claimed that the defendant granted the company … preferential access to the results of its scientific research, namely the dataset at issue. Rather, it is only submitted in this respect that … its service … with the help of the dataset in dispute […].

126 (2) Insofar as the plaintiff […] refers to a chat that took place in 2021 on the platform … according to which the co-founder of the defendant, Mr. …, is said to have agreed to grant the company … on the basis of a financial contribution of USD 5,000.00 made by the latter to grant early access to the (then smaller) dataset, this submission also does not fulfill the exception in Section 60d para 2 sentence 3 of the German Copyright Act.

127 It can be left open whether this chat record – not disputed by the defendant as such […] – supports the interpretation drawn by the plaintiff at all. It also remains to be seen whether the declaration of such a willingness to grant early access – the plaintiff has not stated whether this was actually granted – can be sufficient for holding preferential access to the research results within the meaning of Section 60d para 2 sentence 2 of the German Copyright Act.

128 In any case, it has neither been shown nor is it otherwise apparent that the company … would have a decisive influence on the defendant. Insofar as there are any personal ties between the defendant and companies in the AI sector, these are the companies … and … […].

[…]

Comment

As the first German – and potentially the also the first European – court, the Regional Court of Hamburg published a judgment on text and data mining (TDM) under copyright law in the context of machine learning technologies in general and image-generators in specific. Although the case does not directly concern such models' training but the preparation of training data, the court clearly opposed attempts to exclude reproductions aimed at machine learning training in general – and the training of generative models in particular – from the scope of the TDM limitations. Unfortunately, the court furthermore decided to add questionable obiter dicta to the ongoing debate on the understanding of machine readability within the meaning of Section 44b para 3 sentence 2 of the German Copyright Act, Art. 4 para 3 DSM Directive. The comment introduces

the most relevant technical basics (I.), briefly explains the TDM limitations under the German Copyright Act (II.), outlines the most important facts of the case (III.), summarises the decision (IV.), and, with references to the judgment, addresses two controversial questions in the context of TDM and machine learning training (V.), namely the question of whether machine learning training constitutes TDM within the meaning of Art. 2 para 2 DSM Directive, Section 44b para 1 of the German Copyright Act (1.), and the scope of the term "machine readability" within the meaning of Art. 4 para 3 sentence 2 DSM Directive, Section 44b para 3 sentence 2 of the German Copyright Act (2.) It then draws a conclusion, pointing out both strengths and weaknesses of the judgment (VI.).

I. Technical basics: Machine learning, memorisation and scraping training data

Assessing the implications of the judgment for machine learning training requires an understanding of some technical basics.[4] Machine learning is currently considered the most significant subfield of Artificial Intelligence (AI). Machine learning models are based on artificial neural networks (ANN), complex structures of mathematical instructions represented in a high number of nodes and connections between them.[5] ANN include parameters that act as placeholders for information from the training data. A model with specified parameters is built through training an ANN on usually large training datasets such as LAION-5B. During training, information from the dataset is not systematically stored in the models' parameters.[6] To take text-to-image generators such as Midjourney[7], DALL-E 3[8] or Stable Diffusion[9] as examples, their training aims at transferring abstract information on the design and composition (e. g. motives, colours, art styles) of training images to the model parameters to enable the model to creatively generate new images. However, existing image generators, for some training images, store almost complete information in their parameters in such a way that they allow for their reproduction.[10] This phenomenon, that is commonly referred to as "memorisation", has not been extensively researched and is therefore still insufficiently understood and not quantifiable. The fact that machine learning models are blackboxes and efficient methods for the targeted extraction of training data do not exist, limits research on memorisation. Existing knowledge of memorisation is based on experiments the results of which cannot be transferred to other models or model versions or generalised across domains.

Machine learning typically requires large datasets, which are often collected automatically from the internet. Both the creation and the use of such datasets require bots, i.e. automated programs that perform tasks online, concretely web crawlers and web scrapers. Crawlers browse websites to index content, for example to enable search engines to provide links to specific online content or to collect links to specific types of data (e.g. text-image pairs) for building datasets such as LAION-5B. Scrapers automatically extract data online,[11] e.g. image files that datasets such as LAION-5B refer to. LAION, the defendant in the present case, used a scraper to download the images referred to in the pre-existing US dataset. Some website publishers wish to limit the use of crawlers and scrapers on their websites. Robots.txt files have become a de facto standard for preventing search engine indexing, instructing crawlers not to index the content of a website, and are generally respected by search engine operators. A growing number of website publishers use robots.txt files to preclude specific crawlers and scrapers from indexing and extracting the content on their whole website including subsites, in particular those used to collect data for machine learning training sets.[12] Furthermore, the World Wide Web Consortium (W3C) has introduced a standard that is specifically designed to express rights reservations concerning text and data mining for specific web content[13] and that is also used by a growing number of website publishers[14].

II. TDM limitations in the German Copyright Act

On the EU level, Art. 3 f. DSM Directive foresee limitations for text and data mining (TDM). The provisions are intended to provide for legal certainty, and to facilitate innovation and to incentivise the development of new applications and technologies (see only recital 18 sentence 4 DSM Directive). The German legislator transposed Art. 3 f. DSM Directive into German law through the introduction of Sections 44b, 60d into the German Copyright Act.

Sections 44b and 60d of the German Copyright Act allow for reproductions for the purpose of TDM, i.e. the automated analysis of text and data in digital form to generate information such as patterns, trends and correlations, cf. Section 44b para 1 of the German Copyright Act, Art. 2 para 2 DSM Directive. Aiming at generating – more precise: extracting – information, TDM as such falls out of scope of

4 For a more detailed description of image generators and their training, differentiating between GANs and diffusion models, see Pesch/Böhme, GRUR 2023, 997 (998 ff.).
5 Only see Krogh Nature Biotechnology 26/2 (2008) 195.
6 This claim is, however, made by Image Generator Litigation, see amended complaint, p. 27 ff, https://imagegeneratorlitigation.com/pdf/andersen-first-amended-complaint.pdf (last accessed on 25 November 2024).
7 Midjourney, https://www.midjourney.com (last accessed on 25 November 2024).
8 OpenAI, DALL-E 3, https://openai.com/index/dall-e-3/ (last accessed on 25 November 2024).
9 Stability AI, Image Models, https://stability.ai/stable-image (last accessed on 25 November 2024).
10 See only Carlini et al., Extracting Training Data from Diffusion Models (2023), USENIX '23; Image Generator Litigation, amended complaint, exhibits, https://imagegeneratorlitigation.com/pdf/andersen-first-amended-complaint.pdf (last accessed on 25 November 2024).
11 The court seems to have scrapers in mind when referring to crawlers (paras 97, 101, 104 of the judgment).
12 See, for example, the robots.txt file of SPIEGEL magazine that, among others, excludes GPTBot, ClaudeBot and DiffBot, https://www.spiegel.de/robots.txt (last accessed on 25 November 2024).
13 W3C, TDM Reservation Protocol (TDMRep), https://www.w3.org/community/reports/tdmrep/CG-FINAL-tdmrep-20240510/#abstract (last accessed on 25 November 2024).
14 See, for example, the source code of the website of Frankfurter Allgemeine Zeitung, https://www.faz.net/aktuell/ (last accessed on 25 November 2024): "<meta name="tdm-reservation" content="1">".

copyright law, and, per se, never impairs the normal exploitation of a work.[15] However, extracting information by automated means always requires digital copies of the works to analyse. Art. 3, 4 DSM Directive, Sections 44b, 60d of the German Copyright Act permit those reproductions, including technically required modifications[16].

Section 60d of the German Copyright Act applies to research only – without excluding the application of Section 44b to such cases.[17] Other than Section 60d of the German Copyright Act, Section 44b para 3 sentence 1 allows rightholders to expressly reserve their rights, while such a rights reservation, according to sentence 2 of the provision, must be declared in a machine readable manner for any content publicly available online.[18] This corresponds with Art. 4 para 3 DSM Directive.

III. Most important facts of the case

The plaintiff, a stock photographer, alleged that the Hamburg-based non-profit organisation[19] LAION[20], the defendant, infringed his rights when preparing the public LAION-5B[21] dataset.[22] LAION aims to make machine learning datasets publicly available for free to "democratise" machine learning research.[23] LAION-5B refers to a dataset of 5,85 billion image-text pairs. Image-text pairs are especially crucial for the training of text-to-image generators such as Midjourney[24], DALL-E 3[25] or Stable Diffusion[26]. LAION-5B, however, does not include the image files but metadata on the images only, for each image especially its caption, i.e. textual description (ALT text[27]), and its URL, i.e. the web address to access the image online. LAION-5B and other LAION datasets are not exclusively but commonly used for the training of commercial and non-commercial image-generators worldwide.

LAION-5B is based on a pre-existing dataset that contains image URLs, also for the photo in dispute, and their captions. In a first step, LAION downloaded this dataset and the images referenced in it. The photo in dispute was downloaded from the website of a stock agency that, on another subsite, included a general reservation of rights in natural language. In a second step, in an automated manner, LAION checked whether the captions matched the images and filtered out all text-image-pairs that did not match sufficiently. LAION claims it deleted the copies of the images after performing this automated text-image-consistency analysis. In a third step, LAION built its dataset LAION-5B, consisting of the metadata for the remaining images only.

IV. Summary of the decision

The photo in dispute is subject to exclusive rights under Section 72 of the German Copyright Act that grants a related right to any photo that is no original personal creation within the meaning of Section 2 para 2 of the German Copyright Act and therefore not subject to a copyright ("Urheberrecht").[28] Uncontested between the parties, LAION, without obtaining rights before, downloaded the image, and thereby reproduced it within the meaning of Section 16 of the German Copyright Act and Art. 2 InfoSoc Directive[29].[30] The court had to examine whether this reproduction is permissible under the limitations that are laid down in Sections 44a ff. of the German Copyright Act ("Schranken"), cf. Art. 5 InfoSoc Directive.[31]

The court especially considered the limitations for text and data mining (TDM) laid down in Sections 44b and 60d of the German Copyright Act. It leaves open whether the reservation of rights pursuant to Section 44b para 3 of the German Copyright Act had been declared effectively and, accordingly, whether the reproduction could be permissible under Section 44b para 2 of the German Copyright Act.[32] It argues that at least Section 60d of the German Copyright Act applies.[33] The court clarifies that the automated text-image-consistency analysis that LAION carried out is a textbook case of TDM aimed at determining correlations between images and their caption.[34] It points out that, as a prerequisite, the preparation of the dataset already – and not only its subsequent use in machine learning research – constitutes scientific research.[35] It opposes legal

15 See only Hofmann ZUM 2024, 166 (172).
16 Arg. ex Sections 23 para 3, 69d para 4 of the German Copyright Act, Recital 8 sentence 5 DSM Directive ("normalised"), Spindler/Schuster/Kaesling/Pesch, Recht der elektronischen Medien, 5th ed. 2025 (in press), § 44b para 25, cf. BGH 16 May 2013 – I ZR 28/12 NJW 2013, 3789 (3791) – Beuys-Aktion, para 36. Other opinion Novelli et al. CLSR 2024, 106066, p. 10 (without explanation).
17 BT-Drucks. 19/27426, 95. Also cf. Art. 4 para 4 DSM Directive.
18 As TDM aims at information generation that per se falls out of the scope of copyright law, this legislative decision is questionable at best, see only Hofmann WRP 2024, 11 (14). In fact, it enables publishers of online content to prevent the download and analysis even of content that is not subject to any rights under copyright law as scrapers are configured to comply with machine-readable rights reservations, Spindler/Schuster/Kaesling/Pesch, Recht der elektronischen Medien, 5th ed. 2025 (in press), § 44b para 44.
19 In German: Gemeinnütziger Verein.
20 LAION (Large-scale Artificial Intelligence Open Network), https://laion.ai/ (last accessed on 25 November 2024).
21 Beaumont, LAION-5B: A New Era of Open Large-Scale Multi-Modal Datasets, 31 March 2022, https://laion.ai/blog/laion-5b/ (last accessed on 25 November 2024).
22 See Heidrich Rechtsanwälte, press release, English version below, https://www.recht-im-internet.de/presseanfragen/pressemeldung-laion (last accessed on 25 November 2024).
23 LAION, About, https://laion.ai/about/ (last accessed on 25 November 2024).
24 Midjourney, https://www.midjourney.com (last accessed on 25 November 2024).
25 OpenAI, DALL-E 3, https://openai.com/index/dall-e-3/ (last accessed on 25 November 2024).
26 Stability AI, Image Models, https://stability.ai/stable-image (last accessed on 25 November 2024).
27 ALT text, or alternative texts, refers to descriptions embedded in the source code of websites to ensure accessibility for users with disabilities and to improve the visibility of images in search engines. In some web browsers, ALT texts are displayed when the cursor is hovered over an image.
28 See para 55 of the judgment.
29 Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society.
30 See para 55 of the judgment.
31 See paras 56–128 of the judgment.
32 See paras 68–107 of the judgment.
33 See paras 108 ff. of the judgment.
34 See para 73 of the judgment.

scholars[36] who argue that reproductions for the purpose of (generative) AI training were out of scope of the TDM limitations.[37] As LAION published the dataset for free, it would pursue non-commercial purposes pursuant to Section 60d para 2 no. 1 of the German Copyright Act and consequently fall under the personal scope of the limitation.[38] The court deemed irrelevant that the dataset was also used by third parties to train commercial machine learning models.[39] The question of how long LAION could retain the copy of the image under Section 60d V (respective Art. 3 para 2 DSM Directive) was not discussed by the court although it was disputed between the parties whether LAION deleted the copies of the images after the automated text-image-consistency analysis.[40]

IV. Text and data mining (TDM) in the context of machine learning training

The Regional Court of Hamburg's judgment touches upon two controversial questions. It addresses the applicability of the TDM limitations laid down in Sections 44b, 60d of the German Copyright Act to reproductions in the context of machine learning (1.). And it attempts to define the concept of machine readability in relation to rights reservations for online content pursuant to Sections 44b para 3 sentence 2 of the German Copyright Act, Art. 4 para 3 DSM Directive (2.).

1. Training data preparation and training as TDM

Whether and, if so, to which extent the TDM limitations apply in the context of machine learning is controversially discussed. For a constructive debate it is crucial to thoroughly distinguish the different acts of reproduction that occur in the context of machine learning. With respect to LAION-5B and image generators, it is necessary to differentiate between[41]

- the scraping of the data from the pre-existing US dataset by LAION (the reproduction at dispute),
- the scraping of the data from LAION-5B by developers of text-to-image generators and certain modifications[42] of images for different iterations of the training,
- the "memorisation" of training data during training,
- reproductions of training data in the model's output,
- the reproduction of text or images as input for text-to-image or image-to-image generation when using the model.

It is commendable that the Regional Court of Hamburg has clearly rejected the plaintiff's and some legal scholars' attempts to conflate the preparation of training data, the training, and the subsequent use of the trained model and its outputs,[43] also with respect to the potential creation and use of competing outputs with a trained model. The court convincingly concludes that reproductions for the purpose of carrying out TDM techniques (here: automated consistency-analysis for text-image pairs) to prepare AI training datasets (here: LAION-5B) are not precluded from the scope of the TDM limitations. However, when arguing that the difference 'between information hidden in the data' and 'the content of the intellectual creation'[44] was not sufficiently clear,[45] the court fails to recognise the intent and purpose of the TDM limitations[46] to allow for digitally extracting mere information. The court is right to point out that it can be difficult to draw the line between protected personal intellectual creations and unprotected information. However, a clear distinction between copyrightable creations from mere information is essential to the correct application of copyright law in general and of the TDM limitations in specific. In that context, the court could have easily opposed the alleged dichotomy between protected syntax and unprotected semantics[47] in the case of linguistic works. This is because the protection of linguistic works is not limited to their specific arrangement of words, but also other elements of the work such as characters and plots.[48]

The applicability of the TDM limitations to generative AI training is primarily doubtful with regard to the "memorisation" of training data[49]. Under EU and German copyright law, the (almost) complete "memorisation" of training images constitutes a reproduction within the meaning of Section 16 of the German Copyright Act and Art. 2 InfoSoc Directive, i.e. image generators contain copies of some training images.[50] However, this does not justify to preclude the training of generative AI models from the scope of the TDM limitations. Instead, the correct application of the provisions requires not more or less than a clear distinction of reproduction acts. Not only preparatory acts like the text-image-consistency analysis performed by LAION, but also the subsequent training of generative AI models constitute TDM.[51] The specification of model parameters[52] can be considered a generation of information within the meaning of Art. 2 para 2 DSM Directive, Section 44b para 1 of the Ger-

35 See para 114 of the judgment.
36 Primarily referring to Schack, NJW 2024, 113, and also to Dornis/Stober, Urheberrecht und Training generativer KI-Modelle, 2024, Open Access, p. 67 ff. Cf. Dornis, The Training of Generative AI is not Text and Data Mining (2024), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4993782 (last accessed on 25 November 2024).
37 See paras 75–88 of the judgment.
38 See para 119 of the judgment.
39 See para 119 of the judgment.
40 Cf. para 6 of the judgment.
41 Cf. paras 77 ff. of the judgment.
42 Especially the generation of images with (moderate) noise for the training of diffusion models, see Pesch/Böhme, GRUR 2023, 997 (1004).
43 Also see Spindler/Schuster/Kaesling/Pesch, Recht der elektronischen Medien, 5 Ed. 2025 (in press), UrhG § 44b, para 22.
44 With reference to Schack NJW 2024, 113 (114); Dornis/Stober, Urheberrecht und Training generativer KI-Modelle, 2024, Open Access, p. 120 f., cf. Dornis, The Training of Generative AI is not Text and Data Mining (2024), p. 6 ff.
45 See para 76 of the judgment.
46 See II. above.
47 Dornis/Stober, Urheberrecht und Training generativer KI-Modelle, 2024, Open Access, p. 96 ff., 109 f., cf. Dornis, The Training of Generative AI is not Text and Data Mining, 2024, p. 4 ff., 10 ff.
48 See only BGH, 17 July 2013 – I ZR 52/12, GRUR 2014, 258 (260) – Pippi Langstrumpf, para 25 with further references.
49 See I. above.
50 Extensively Pesch/Böhme GRUR 2023, 997 (1005). Also see Novelli et al. CLRI (in press); Mezei EIPR 2024, 461 ff.

man Copyright Act. Albeit the information represented in model parameters is not directly perceptible, it is nonetheless extracted from the training data and utilised.[53] By contrast, the "memorisation" of training data goes beyond the mere information extraction that the TDM provisions allow for. Nevertheless, the conclusion that generative AI training never falls under the TDM limitations is based on the unfounded assumption that generative models per se store copies of their training data. This claim is commonly made by rightholders – some of whom even falsely state[54] that reproductions of all training data samples were stored in the parameters of certain models. It is true that, in experiments with existing generative text-to-image generators and large language models (LLMs), researchers could extract significant amounts of training data samples.[55] However, as the results do not generalise across models, model versions or domains, they do not provide a sufficient basis to draw general conclusions about the applicability of the TDM limitations to the training of generative AI models – and even less about their applicability to the preparation of training datasets such as LAION-5B.

If the court had thoroughly considered the intent and purpose of the TDM limitations, it also could have easily opposed the flawed argument[56] that the legislator, when drafting the DSM Directive, has not foreseen generative AI models and, therefore, their training would fall out of the scope of the TDM limitations. The application of legal provisions to specific technologies never requires their anticipation by the historic legislator. For Art. 3 f. DSM Directive, the legislator has even explicitly made that clear that the provisions shall create legal certainty especially with regard to rapidly developing technologies.[57]

2. Machine-readability of rights reservations

The court's apparent reluctance to take into account the intent and purpose of the provisions in Sections 44b, 60d of the German Copyright Act, Art. 3 f. DSM Directive is also evident in the obiter dicta on the machine-readability of rights reservations. Neither the DSM Directive nor the German Copyright Act define machine-readability within the meaning of Section 44b para 3 of the German Copyright Act, Art. 4 para 3 DSM Directive. Other than the court argues, rights reservations in natural language do not meet the machine readability requirement[58]. This follows from the intent and purpose of the TDM limitations to create legal certainty in the field of TDM. Such legal certainty is only achieved if it is possible to download large amounts of data in compliance with rights reservations in an automated manner without manual legal assessments.[59] The court argues that all technologies available must be considered and claims that AI models (concretely: large language models (LLMs) could be used to interpret rights reservation statements in natural language, adding that LAION clearly had such a tool for the image-text-consistency analysis. In fact, no LLM or other application can reliably find and correctly interpret all rights reservations in natural language,[60] even less so in 2021 when the reproduction in dispute took place. Anyone who follows the court's reasoning, could consider even an image of handwritten text in any language machine-readable, as AI models are also used to recognise handwritten text in images and translate text. The court points out that Section 44b para 3 of the German Copyright Act, just as Art. 4 para 3 DSM Directive, does require rights to be reserved not in the easiest way but only in an appropriate manner.[61] The court – that cannot be blamed too much as the defendant's representatives seem to share its view[62] – misses the point that Section 44b para 3 of the German Copyright Act, just as Art. 4 para 3 DSM Directive only achieve their goal to incentivise innovation if rights reservations can be identified with certainty in an automated manner. To enable web scrapers to reliably filter content for which rights have been reserved, at this time, the reservation of rights must be declared in the robots.txt[63] file or using the W3C standard[64].

The court, in this context, misunderstands the objection[65] against the use of TDM techniques to interpret TDM rights reservations.[66] This objection has nothing to do with circular reasoning. The interpretation of website content by LLMs constitutes TDM. Identifying a rights reservation in natural language would require a reproduction of all texts included in the website. To the extent the texts are protected under copyright law, is only permissible under Section 44b para 2 if there is no effective rights reservation. When the LLM then identifies a rights reservation, however, this rights reservation would already have been violated.[67] The court's argument that the reproduction of website content could be permissible under Section 44a of the German Copyright Act if they are only transient or incidental, igno-

51 See only Bomhard DRITB 2023, 255 (260); de la Durantaye ZUM 2023, 645 (651); Hofmann WRP 2024, 11 (13); Maamar ZUM 2023, 481 (483); on Art. 2 Nr. 2 DSM Directive Margoni/Kretschmer GRUR Int. 2022, 685 (687 ff.); Pesch/Böhme GRUR 2023, 997 (1006); Spindler/Schuster/Kaesling/Pesch, Recht der elektronischen Medien, 5th ed. 2025 (in press), UrhG § 44b para 19 ff. with further references.

52 See I. above.

53 Spindler/Schuster/Kaesling/Pesch, Recht der elektronischen Medien, 5th ed. 2025 (in press), UrhG § 44b para 19.

54 See only Image Generator Litigation, amended complaint, p. 27 ff., https://imagegeneratorlitigation.com/pdf/andersen-first-amended-complaint.pdf (last accessed on 25 November 2024).

55 See I. above.

56 Dornis/Stober, Urheberrecht und Training generativer KI-Modelle, 2024, Open Access, p. 121 ff.; Schack NJW 2024, 113 (114); von Welser GRUR-Prax 2023, 516 (518), cf. Dornis, The Training of Generative AI is not Text and Data Mining, 2024, p. 22 f.

57 Cf. recitals 8, 18 xoxo

58 Maamar ZUM 2023, 481 (484); also cf. Bomhard DRITB 2023, 255 (266).

59 Jacobsen/Hartmann MMR-Aktuell 2021, 441332.

60 Spindler/Schuster/Kaesling/Pesch, Recht der elektronischen Medien, 5th ed. 2025 (in press), UrhG § 44b para 40.

61 See para 105 of the judgment.

62 Akinci/Heidrich IPRB 2023, 270 (272).

63 See only Baumann NJW 2023, 3673 (3675); Bomhard DRITB 2023, 255 (267); Spindler/Schuster/Kaesling/Pesch, Recht der elektronischen Medien, 5th ed. 2025 (in press), UrhG § 44b para 41 with further references. On robots.txt files see I. above.

64 Maamar ZUM 2023, 481 (484); Spindler/Schuster/Kaesling/Pesch, Recht der elektronischen Medien, 5th ed. 2025 (in press), UrhG § 44b para 42; Schippan ZUM 2024, 670 (676). On the standard see I. above.

65 Hamann ZGE 2024, 134 (148).

66 Cf. para 104 of the judgment.

67 Cf. Hamann ZGE 2024, 134 (148).

res the legislator's goal to provide for legal certainty. As the German legislator has made clear,[68] the provisions would miss that point if text and data miners had to draw the line between reproductions within the meaning of Section 44a of the German Copyright Act and other reproductions.

V. Conclusion

The judgment of the Regional Court of Hamburg in LAION case has both strengths and weaknesses. The biggest strengths of the judgment lie in the thorough differentiation of reproduction acts and the court's clear rejection of attempts to categorically exclude reproductions in the context of generative AI models from the scope of the TDM limitations. What weakens the line of argumentation of the court, however, is the lack of consideration of the intent and purpose of the TDM provisions. This shortcoming has led the court to an, albeit hesitant, unfoundedly wide interpretation of the machine-readability requirement for TDM reservations according to Section 44b para 3 sentence 2 of the German Copyright Act, Art. 4 para 3 DSM Directive. It remains to be seen how other courts, especially the CJEU, interpret the TDM provisions and to which extent they apply the TDM limitations to reproductions in the context of machine learning models in general and generative AI models in specific. As the case of LAION touches not only upon national and EU secondary law but also fundamental rights, in particular the freedom of information guaranteed by Art. 5 para 1 sentence 1 alternative 2 of the German Basic Law (Grundgesetz), it could occupy civil courts, the federal constitutional court and the CJEU for many years. Hopefully, the LAION case does not turn into a second "Metall auf Metall". The Regional Court of Hamburg could have mitigated this risk by referring the questions of (1.) whether reproductions for the purpose of (generative) AI training fall under the scope of the Art. 3 f. DSM Directive and (2.) whether rights reservations in natural language meet the machine-readability requirement of Art. 4 para 3 sentence 2 DSM Directive to the CJEU for a preliminary ruling.[69]

Paulina Jo Pesch

68 BT-Drucks. 19/27426, 88, also cf. recital 18 sentences 3 f.
69 Cf. LG Hamburg MMR 2024, 973 (978), comment by Hoeren.