Co-generation of data

Copyright and Data Protection Rights in Co-Generated Input and Output of Generative AI

November 2024

GPAI / THE GLOBAL PARTNERSHIP ON ARTIFICIAL INTELLIGENCE

This report was planned prior to the integration of the Global Partnership on Artificial Intelligence (GPAI) and the Organisation for Economic Co-operation and Development (OECD) mid-2024. Consequently, the report was not subject to approval by GPAI and OECD members and should not be considered to reflect their positions.

Acknowledgements

This report was developed in the context of the 'From Co-generated data to generative AI' project, with the steering of the Project Co-Leads and the guidance of the Project Advisory Group, supported by the GPAI Data Governance Working Group. The GPAI Data Governance Working Group agreed to declassify this report and make it publicly available.

Co-Leads:
Christiane Wendehorst[*], University of Vienna
Kyoko Yoshinaga[*], Keio University

The report was written by **Alain Strowel**[‡] (UCLouvain) and **Sebastian Schwamberger**[‡], (Rostock University), with the engagement of the European Law Institute[‡] (ELI). GPAI is grateful for the meaningful contributions of **Naohiro Furukawa**[*] (ABEJA, In-house Counsel), **Avik Sarkar**[*] (Indian School of Business), **Solene Festor de Suremain**[‡] (Pierstone Brussels), **Josef Drexl**[*] (Max Planck Institute for Innovation and Competition), **Alexandra Salnikow**[‡] (University of Vienna), and **Stefan Janusz** and **Antoine Glory** of CEIMIA.

GPAI would also like to thank Project Advisory Group members **Seong Oun Hwang**[*] (Gachon University), **Zümrüt Müftüoğlu**[*] (Yildiz Technical University), **Jaco Du Toit**[**] (UNESCO), **Kudakwashe Dandajena**[*] (AIMS-II), **Zee Kin Yeong**[*] (IMDA) , **Kim McGrail**[*] (University of British Columbia), **Mikael Jensen**[*] (D-Seal), **Ulises Cortes**[*] (Universitat Politècnica de Catalunya), **Toshiya Jitsuzumi**[*] (Chuo University) and **Maja Bogataj*** (ODIPI).

GPAI recognises ELI's Advisory Committee members **Benoit van Asbroeck**[†], **Neil B Cohen**[†], **Sjef van Erp**[†], **Simon Geiregat**[†], **María Lubomira Kubica**[†], **Ana Keglević Steffek**[†], **Tetsuo Morishita**[†], **Pascal Pichonnaz**[†], **Sam De Silva**[†], **Lord John Thomas**[†], **Christian Twigg-Flesner**[†], **Jos Uitdehaag**[†], **José Antonio Castillo Parrilla**[†], **Jacques de Werra**[†] and Observer to ELI's Advisory Committee for the United Nations Commission on International Trade Law, represented by **Alexander Kunzelmann**[†].

Finally, GPAI would like to acknowledge the efforts of colleagues at the International Centre of Expertise in Montréal on Artificial Intelligence (CEIMIA). We are particularly grateful for the dedication of the Working Group Co-Chairs **Bertrand Monthubert**[*], Conseil National de l'Information Géolocalisée, and **Shameek Kundu**[*], Infocomm Media Development Authority.

* Expert
** Observer
† Invited Specialist
‡ Contracted Parties by the CofEs to contribute to projects

Citation

GPAI 2024. Copyright and Data Protection Rights in Co-Generated Input and Output of Generative AI, Report, November 2024, Global Partnership on AI.

Table of Contents

A. Preliminary Considerations

I. Aim of this study

The aim of this study is first to review the legal frameworks applying to co-generation scenarios (considering, at least, co-generation of data as such, co-generation of models, and co-generation of content). The conclusions envisaged at the end of the study aim to reflect on the current ethical and policy debates around co-generation. This study focuses on the regulatory frameworks of the EU, Japan and the US, and offers a high-level overview on the hard and soft law developments particularly in the U.S. and Japan on copyright, as well as on the policy discussions regarding rights in co-generated input and output of Generative AI.

Before reviewing the legal frameworks, it is important to assess the notion of Generative AI (GenAI) as well as discuss what co-generation, as an overarching concept, means in the context of this study (A.II.). Furthermore, as explained below this study will distinguish **co-generated input** and **co-generated output of GenAI** (A.IV.) rather than rely on the differentiation between co-generated data, co-generated models or co-generated content.

II. Generative AI

The term GenAI often encompasses "foundation models"[1], "large language models"[2], 'large generative models"[3] or "general purpose AI"[4].[5] In simple terms, a system of GenAI is trained with a large amount of training data that is represented as probability distributions that through sampling and mixing , can generate content beyond the training data set.[6] GenAI is thus a category of AI tool that produces content (which could include data protected or non-protected by intellectual property or (personal) data rights).

The explanatory memorandum on the updated OECD definition of an AI system[7] proposes the following:

> "*An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment.*"

This is also in line with the U.S. Executive Order[8] that sets out the following definition of GenAI:

---

[1] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A. and Brunskill, E. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021).
[2] Ganguli, D., Hernandez, D., Lovitt, L., Askell, A., Bai, Y., Chen, A., Conerly, T., Dassarma, N., Drain, D. and Elhage, N. Predictability and surprise in large generative models. ACM Conference on Fairness, Accountability, and Transparency (2022), 1747-1764.
[3] Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J. and Clark, A. Training compute-optimal large language models. arXiv preprint arXiv:2203.15556 (2022).
[4] Art. 3(63) AI Act, Regulation (EU) 2024/1689.
[5] Hacker/Engel/Mauer, Regulating ChatGPT and other Large Generative AI Models, 7 February 2023, 3, available via https://www.europeannewschool.eu/images/chairs/hacker/Hacker_Engel_Mauer_2023_Regulating_ChatGPT_Feb07.pdf.
[6] Idem.
[7] https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/03/explanatory-memorandum-on-the-updated-oecd-definition-of-an-ai-system_3c815e51/623da898-en.pdf
[8] Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, 30 October 2023, Sec. 3(p), available via https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/ .

> *"The term "generative AI" means the class of AI models that emulate the structure and characteristics of input data in order to generate derived synthetic content. This can include images, videos, audio, text, and other digital content."*

The definition of general-purpose AI in the EU AI Act[9] appears broader (but is at the same time more limited as it does not cover the uses (for ex. for R&D) before the placing on the market):

> *"'general-purpose AI model' means an AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications, except AI models that are used for research, development or prototyping activities before they are placed on the market;"*

A shorter definition (focusing on the output) is used by the Japanese Ministry of Internal Affairs and Communications and the Ministry of Economy, Trade and Industry[10]:

> *"Generative AI*
> *A general term representing AI developed from an AI model that can generate texts, images, programs, etc."*

We do not aim at establishing a separate (additional) definition of GenAI, but rather to work within the framework of existing definitions. By doing so, we ensure consistency and alignment with the current understanding and use of the term in the field in these respective legal frameworks.

III. Co-generation

The broad concept of Co-Generation has been developed and coined by the ALI-ELI Principles for a Data Economy.[11] But it has now also been adopted – at least implicitly – by the legislator, for example in the EU Data Act.

1. ALI-ELI Principles

The ALI-ELI Principles for a Data Economy set out factors to be taken into account in determining whether, and to what extent, data is to be treated as co-generated, in the following priority in its Principle 18(1)[12]:

(a) the extent to which that party is the subject of the information coded in the data, or is the owner or operator of an asset that is the subject of that information;

[9] Art. 3(63) AI Act, Regulation (EU) 2024/1689.
[10] Ministry of Internal Affairs and Communications (MIC), Ministry of Economy, Trade and Industry (METI), AI Guidelines for Business Ver1.0, April 19, 2024, p. 10, available via https://www.meti.go.jp/english/press/2024/0419_002.html.
[11] Cohen/Wendehorst, The ALI-ELI Principles for a Data Economy, available via https://europeanlawinstitute.eu/projects-publications/publications/ali-eli-principles-for-a-data-economy-data-transactions-and-data-rights/.
[12] Cohen/Wendehorst, The ALI-ELI Principles for a Data Economy, available via https://europeanlawinstitute.eu/projects-publications/publications/ali-eli-principles-for-a-data-economy-data-transactions-and-data-rights/.

(b) the extent to which the data was produced by an activity of that party, or by use of a product or service owned or operated by that party;

(c) the extent to which the data was collected or assembled by that party in a way that creates something of a new quality; and

(d) the extent to which the data was generated by use of a computer program or other relevant element of a product or service, which that party has produced or developed.

In addition the Principles provide for a definition of co-generated data which means "data to the generation of which a person other than the controller has contributed, such as by being the subject of the information or the owner or operator of that subject, by pursuing a data-generating activity or owning or operating a data-generating device, or by producing or developing a data-generating product or service" (Principle 3(1)(h)).

This concept of 'co-generated data' has been adopted by the European Commission in its European Data Strategy[13] and has influenced some core provisions of the EU Data Act[14], it was also endorsed by the German Data Ethics Commission[15], the default rules for data provision contracts by UNCITRAL[16] and the Framework Paper for GPAI's Work on Data Governance 2.0[17].

2. GPAI report "From co-generated data to generative AI: New rights and governance models in digital ecosystems"

The ODI/Apati Study[18] refers to the ALI-ELI Principles with regard to co-generated data. Besides that, the study defines 'co-generated technology' as technology which has been generated by multiple parties. 'AI co-generated works' is defined as new outputs of GenAI models, in the form of data or content such as text, image, audio or video.

3. Implicit definitions

The **EU Data Act**[19] puts forward several provisions that are based on the broad concept of co-generated data of the ALI-ELI Principles and the underlying notion that a party who had a share in the generation of the data should be afforded certain rights in regard to that data. According to **Articles 4 and 5**, users of Internet of Things (IoT) products or related services shall have the right to access and use data generated by the use of an IoT product or related service and to share the data with third parties. Thus, the Data Act recognises that in certain cases of co-generation (in the IoT context), some right to access the output is justified. Furthermore, the literature on data generated by connected cars (an example of IoT product) suggests that the generation of data resulting from their use could raise issues of data access and data sharing with regard to its various stakeholders: car owners, drivers and passengers, other traffic participants, manufacturers, add-on service providers, as well as car dealers and distributors[20].

[13] COM(2020) 66 final, p. 10.
[14] Regulation (EU) 2023/2854.
[15] Opinion of the German Data Ethics Commission, 2019, p. 133 ff.
[16] https://documents.un.org/doc/undoc/gen/v24/066/85/pdf/v2406685.pdf.
[17] https://gpai.ai/projects/data-governance/Data%20Governance%20-%20A%20Framework%20Paper%20for%20GPAI%E2%80%99s%20Work%20on%20Data%20Governance%202.0%20.pdf.
[18] GPAI 2024. From co-generated data to generative AI: New rights and governance models in digital ecosystems, Report, May 2024, Global Partnership on AI.
[19] Regulation (EU) 2023/2854.
[20] Determann, No One Owns Data, 70 Hastings L.J. 1 (2018) 29 ff.

Like the IoT, GenAI involves some form of co-generation of data or content as output, it also relies for its training on various datasets that are combined and processed in order to develop, validate and fine-tune the AI model.

GenAI is legally defined in the U.S. Executive Order (see above A.II.). The US Notice of Inquiry, 88 Fed. Reg. 59948[21] issued by the U.S. Copyright Office refers to GenAI as an application of AI used to generate outputs in the form of expressive material such as text, images, audio, or video. GenAI systems may take commands or instructions.

The OECD paper "AI, Data Governance and Privacy: Synergies and Areas of International Co-Operation" explores the opportunities and challenges that AI, particularly GenAI, presents in relation to data protection and privacy.[22] The report highlights AI's growing reliance on vast amounts of global data, underscoring the need for international policy synchronisation and cooperation to address privacy risks. Currently, AI and privacy policy communities often work separately, resulting in divergent approaches across jurisdictions. Issues like the use of personal data for AI training, which raise significant privacy concerns, have not been widely addressed collectively. Without cooperation, there could be regulatory conflicts and added complexities in compliance. Despite differences in approach—AI being innovation-driven, and privacy policies being cautious and established—both communities have valuable lessons to share. The paper identifies synergies and encourages cooperation to align policies, improve consistency, and address key issues like terminological differences, emphasising the importance of international collaboration in AI and privacy governance. In this paper, the OECD[23] defines GenAI as covering systems that create new content — including text, image, audio, and video — based on their training data and in response to prompts.

4. Co-generation within this study

In this study, the term "co-generation" relies on the various origins of data used by, or generated by, digital technologies. Data sharing is the *conditio sine qua non* for the generation. For example, in the case of co-generated data, the data would not exist without the interaction between different parties—such as users and digital platforms—whose combined inputs and actions cause the generation of that data. Co-generation requires a form of causation. Some forms of causation are considered as justifying the legal recognition of data access or data sharing rights. Thus, **individuals who have made a sufficient (causal) contribution to the creation or development of digital data are granted a right in relation thereto**. The critical question is under which conditions and circumstances a party is (and should be) granted a right to access, control and/or share the co-generated data.

IV. Relevant differentiation between co-generated input and output

To enquire when, under the existing frameworks for copyright and personal data laws, a party is only causally contributing or is also granted a right, this study focuses on the use of data (or content) at different stages in the life cycle of a GenAI tool and differentiates between **co-generated input** and **co-generated output**.

The co-generation of *models* does not seem to involve similar legal issues, at least if the development of an AI model is (narrowly) defined as covering steps such as structuring the problem (classification task, regression task, etc.), choosing the model type (decision tree, neural network, etc.), tuning hyperparameters, all tasks that do not involve the processing of data. If the development phase of the

[21] https://www.govinfo.gov/content/pkg/FR-2023-08-30/pdf/2023-18624.pdf .
[22] https://www.oecd-ilibrary.org/science-and-technology/ai-data-governance-and-privacy_2476b1a4-en.
[23] OECD Artificial Intelligence Papers , September 2023, No. 1: https://www.oecd.org/en/publications/initial-policy-considerations-for-generative-artificial-intelligence_fae2d1e6-en.html.

model is (broadly) defined as covering also some additional operations on data (such as selection, ingestion, verification, processing, etc. of data), then the legal issues raised by the creation of the model merge with those raised by the co-generation of input. The input phase covers every occurrence where data or content is fed into a model to train it or predict the output, it thus includes the training and validation data, as well as the data provided later by the user of the AI tool (such as the prompts) and the data (available online) the AI tool is using (in real time) for generating the output.

The differentiation between co-generated input and co-generated output also means that the model that is trained with the data or content could also be regarded as "input" serving for the generation of some output. This is also because some models are able to "remember" certain data or content they have been trained with.[24] In these scenarios, it would mean that it is not possible to differentiate between the model and the data or content it was trained with, and thus, considered together with the training data and content, the model is part of the (co-generated) input.

No meaningful distinction between data and content. It is not possible or useful to distinguish between (co-generated) data and content, as content used to train an AI tool (input) or produced by the AI tool (output) could include data (and in reverse order, the data used for training or produced as output could include protected content). Data, in essence, refers to binary code — zeros and ones — and thus operates on a syntactic level. Content, on the other hand, pertains to the semantic level, such as the image or meaning represented by the data. **The differentiation between data and content, therefore, merely reflects the level at which the input or output is being analysed**. The distinction might be relevant in determining whether the input or the output involves content that is protected by an intellectual property (IP) right (as pure data is not protected by IP rights). However, the question of IP protection is precisely (part of) the subject matter of this study, and thus, relying on the distinction data/content would be premature and potentially misleading.

V. Underlying policy debates around human rights, democracy and economic welfare

The World Economic Forum's Global Risk Report 2024 highlights key global risks associated with AI, including misinformation, job displacement, cyberattacks, and decision-making biases. These challenges illustrate the broad impact AI is likely to have across all sectors.

An effective and balanced approach towards AI should aim at taking into account the intertwined nature of human rights, democracy and economic welfare and engage into a holistic approach for AI governance, involving multiple stakeholders like government agencies, private sectors, and non-profits. The EU legislator recognised this interrelation in its commitment to integrating human rights and democratic values into the AI governance model of the AI Act. The potential risks that AI systems pose to health, safety, privacy, and other core rights are directly addressed in the AI Act which could serve as a first template of technological governance aimed at fostering an environment that encourages responsible AI development while promoting economic growth and competitiveness. Other countries and intergovernmental institutions are increasingly recognising the need for cooperative strategies in AI governance. For instance, the Council of Europe developed its *Framework Convention on Artificial Intelligence, Human Rights, Democracy, and the Rule of Law* [25], which emphasises similar principles and emerges as the first international legally binding treaty on the subject. Australia, Canada, Japan and other countries participated in the drafting

[24] Carlini et al., Proceeding 30*th* USENIX Security Symposium 2021, p. 2633; s.a. Carlini et al., Quantifying Memorization Across Neural Language Models, rev. 2023, arXiv:2202.07646.
[25] See https://rm.coe.int/1680afae3c .

process and the Framework Convention was signed in September 2024 by Andorra, Georgia, Iceland, Norway, the Republic of Moldova, San Marino, the United Kingdom as well as Israel, the United States and the European Union.

While this study cannot fully explore the breadth of these developments, it's important to highlight that AI raises a complex set of issues that countries around the world are beginning to tackle, particularly concerning privacy and copyright.

Privacy is clearly an central fundamental right (recognised among others in Article 8 of the European Convention on Human Rights and in Article 7 of the EU Charter on Fundamental Rights). Data protection, while being part of privacy, is sometimes elevated as a separate fundamental human right (as in Article 8 of the EU Charter on Fundamental Rights). While privacy and data protections are considered as strengthening the autonomy of persons and their balanced development within their communities and society, and thus also contribute to a democratic life, considerations of economic welfare remain quite separate from the justification and the implementation of data protection laws.

Copyright issues raised by AI, on the contrary, are discussed taking into account economic welfare and the promotion of innovation. Usually, copyright is presented as a challenge to the development of AI, sometimes even as a potential brake for innovation. But this view must be nuanced. Discussions on AI innovation should fully account for the extent to which copyright protects a category of assets that plays a crucial role in economic growth. Studies tend to show a positive correlation between strong intellectual property rights and economic growth, particularly in creative industries. In addition, beyond the need to balance copyright with other interests so as to guarantee it fosters the progress of knowledge and the dissemination of (entertaining) content, copyright embodies some values (such as the autonomy of the creators) and it is protected under the fundamental right to property, enshrined in the EU in Article 17 of the European Charter of Fundamental Rights. Public policies regarding copyright-protected assets destined to be used for training GenAI tools should take the different facets of copyright and values it protects seriously.

This ultimately raises the need to balance intellectual property protection with innovation. While strong IP safeguards promote competitiveness, promoting an overly strict IP regulation can hinder creativity and collaboration, especially in fast-evolving areas like AI. The fundamental freedom to conduct a business, enshrined in Article 16 of the EU Charter of Fundamental Rights, must be considered and weighed against the need for IP protection, ensuring that regulations do not stifle entrepreneurship or limit access to essential technologies. In AI, the use of copyrighted materials for training algorithms raises questions about both IP infringement and the freedom to innovate. Courts and lawmakers are increasingly adapting IP laws to address these challenges, recognising that fair use and flexible licensing can support innovation while respecting IP rights.

VI. Rights in co-generated input and co-generated output

This Study does not deal with the whole regulatory framework with regard to co-generated input and co-generated output but focuses on certain rights (that were selected for this Study). The focus on rights is already reflected in the ALI-ELI Principles for a Data Economy that aim to delineate data rights.

1. Data protection rights

When it comes to data protection, rights have long been a cornerstone of privacy regulation in the U.S. but also in the EU and worldwide.[26] Some countries soon opted for the introduction of so-called “ARCO” rights.

[26] Solove, The Limitations of Privacy Rights, Notre Dame Law Review 98, 2023, 975 (980 ff.).

The acronym ARCO stands for the right to Access, to Rectification, to Cancellation and of Opposition.[27] In the U.S., however, the early privacy laws only provided for the right to information, access and correction, but several laws provide individuals with the rights to opt out or opt in to the collection and use of their data.[28]
In Europe, the Data Protection Directive of 1995 originally included a set of privacy rights. In 2016, the European legislator enshrined these privacy rights into Articles 12-22 GDPR and added the right to data portability. Thus, the GDPR now contains a right to information, to Access, to Erasure, to Restriction, to Data Portability and to Object.[29]

2. Copyright and data

According to data science, data proceeds "by abstracting the world into categories, measures, and other representational forms – numbers, characters, symbols, images, sounds, electromagnetic waves, bits – that constitute the building blocks from which information and knowledge are created".[30] Data is always framed by the instruments, practices, contexts used to generate, select, represent and analyse them (for example the Celsius or Fahrenheit grading system for temperatures).

Copyright does not aim to protect data as such and thus does not grant rights with regard to data. One could say that the whole enterprise of copyright (and more broadly of IP), has been devoted to making the subject matter of copyright (and of other IP rights except trade secrets) distinct from raw data (or simple information): copyright-protected works must be distinguished from uncopyrightable data (or information and ideas).

That said, the way data has been defined in the recent EU legislation points towards some form of protection by copyright. Indeed, several pieces of EU legislation adopted since 2020 contain a definition of data: "'Data' means any digital representation of acts, facts or information and any compilation of such acts, facts or information, including in the form of sound, visual or audiovisual recording" (Art. 2(1) Data Governance Act; same in Art. 2(1) Data Act). This definition does not only refer to the 'bits' (the digital representation of acts, facts and information), but to their compilation (or combination). As compilations are included in the definition, IP might "kick in" (for example, the database right on a substantial investment in obtaining, verifying or presenting data or on a "substantial part of the contents of a database"). The link between copyright (IP) and data is not clearly articulated in this recent EU definition. The example of "sound, visual or audiovisual recording" is not helpful either as (simple) recordings are not protected by copyright (in the EU) but might be protected under some related IP rights (that do not require originality or authorship). There is at least a continuity between data and copyright (IP) in the sense that at some point when data is compiled and arranged, a potentially IP-protected item might be available within the dataset.

This is also why it is difficult to distinguish data and content for the analysis of the copyright issues: at some point a combination of data becomes a work that can be protected by copyright.
If there is a copyright (IP) protection, a bundle of rights are automatically applicable, including the right of reproduction, of communication to the public (including the performance right), the right of distribution, the right of adaptation, the right of display, etc (as well as the non-economic or moral rights that in the Continental European tradition flow from authorship and the existence of a protected work). There is no common list of rights deriving from the existence of copyright, and, in particular, the bundle of rights making copyright will vary between the U.S., the EU and other countries.

[27] Ibid., 981.
[28] Ibid.
[29] See below B.I.1.b).
[30] Rob Kitchin, *The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences* (Sage 2014), 1.

3. ALI-ELI Principles for a data economy

Central to the ALI-ELI Principles for a Data Economy is the recognition of data rights stemming from a party's contribution to the creation or collection of data. The Principles assert individuals who play a role in generating data, whether through active participation, passive observation, or even unintentional capture, the following Data Rights:

- the right to be provided access to data by means that may, in appropriate circumstances, include porting the data (Principle 20);
- the right requiring the controller to desist from data activities (Principle 21);
- The right requiring the controller to correct (incorrect or incomplete) data (Principle 22);
- The right to receive an economic share in profits derived from the use of data (Principle 23).

Granting a Data Right requires a balancing act, which is dependent on the type of data right. Hence, the Principles spell out specific grounds for the four types of data rights (Principles 20 to 23) that should be taken into account together with the general grounds, such as the share a party had in generating the data, any imbalance of bargaining power or any public interest including the interest to ensure fair and effective competition. For example, for access rights, Principle 20 stipulates that the user of a service or product should *inter alia* have access to the user generated data, if the access is necessary for the normal use, maintenance or re-sale of the product or service and the controller is part of the supply network and can reasonably be expected to have foreseen this necessity (see Principle 20(1)(a)).

The grounds to receive an economic share in the profits derived from the data enshrined in Principle 23 are much narrower than for access rights. As a general rule, a party is not entitled to an economic share in profits derived by another party from the use of co-generated data. However, under exceptional circumstances a right to receive an economic share in the data can be justified. According to the Principles, this is the case when a party's contribution to the generation of the data was particularly unique or based on an extraordinary investment and the denial of participating in the profits would be inconsistent with doctrines such as unjust enrichment, because the profits were exceptionally high and the party seeking the economic share could not efficiently bargain (See Principle 36).

Principle 24 emphasises that granting data rights in the public interest, particularly when the rights holder wasn't involved in generating the data, should be carefully considered and justified. It asserts that such rights should only be granted when absolutely necessary to serve a specific public interest, and only if the impact on the data controller or other involved parties is proportional to the intended benefit. This ensures a balanced approach that respects existing rights and interests while enabling legitimate public interest uses of data. Essentially, it acts as a safeguard against unwarranted encroachment on data control, requiring a rigorous demonstration of necessity and proportionality before granting such public interest data rights.

B. Legal framework

The following section examines the copyright and data protection legal frameworks in Europe.

I. European Union

1. Co-generated input of GenAI

a) Copyright

Preliminary remarks. The key notions of European copyright law such as the right of reproduction and the right of communication to the public are historically shaped by technological developments. The right of reproduction is the exclusive right of an author to authorise the reproduction of a content, hence it is an essential concept in a discussion of copyright and GenAI. Over time, its scope was broadened to include various forms of digital copying, making it ever more challenging for any online operator to ignore the obligation to obtain the copyright holders' consent to use their works in any form of digital activities. In parallel, and as a reaction to this, the European legislator introduced an ever-increasing set of exceptions in the copyright framework to exempt certain activities from copyright infringement based on their technical character (e.g.: cache copies are exempted under the temporary copying exception). Directive 2019/790 ("the CDSM directive, or the "CDSM dir.") provides for a recent example of such enabling exceptions within the digital ecosystem under Articles 3 and 4 for text and data mining ("TDM exceptions"). Under specific conditions, those exceptions permit artificial intelligence operators to mine/extract, use/copy, and generate copyrighted content for research (Art. 3 CDSM dir.) or commercial purposes (Art. 4 CDSM dir.), without having to obtain a licence and without violating copyright laws. As we outline later, the right of reproduction must be considered both at the training and output stages of GenAI. We first examine below the implication of the right of reproduction in GenAI training (i), briefly discuss the inapplicability of the temporary copying exception to GenAI training (ii) and discuss the TDM exception in relation to commercial purposes (iii).

Whether a party who made a causal contribution to an input has a right in the co-generated data under copyright law firstly depends on whether their contribution is itself protected by a copyright. EU Copyright law does not protect mere facts, styles, random data or ideas but the work of an author's own intellectual creation which results from free and creative choices. To the question 'does any party contributing to a co-generated input have a right under copyright rules', the answer is negative. Only content that is eligible for copyright protection in the first place will "carry along" a valid copyright to the benefit of the original author on a co-generated input. If this first condition of copyright protection is met, the second condition relates to whether the act of training with that content is considered a reproduction in the meaning of copyright rules.

(i) The right of reproduction at the training level

Right of reproduction. The reproduction right is defined in Article 2 of the 2001/29 (the "InfoSoc dir.") as the "*exclusive right to authorise or prohibit* (…) *temporary or permanent reproduction by any means and in any form, in whole or in part*" of the works of authors. The right of reproduction encompasses reproductions in both analogue and digital forms hence it does not exclude GenAI. It does not matter by which means or in what kind of medium the reproduction is made. Importantly, partial reproductions are also covered. A reproduction could also occur if the form of the work is substantially changed, i.e. in the case of transformative reproductions. Reproduction (broadly defined) covers in certain copyright laws not only material copies, but also intellectual reproductions such as translations; however other national copyright laws distinguish material reproductions from adaptations and arrangements.

As some form of copying might occur in the training phase of a GenAI the question is whether such (technical) copies fall under the right of reproduction in the copyright sense.

One of the current positions is that Generative Pre-trained AI systems engage in acts of reproduction during the training process, based on several considerations. First, the concept of reproduction is an autonomous principle under EU law, requiring a uniform interpretation. The Court of Justice of the European Union (CJEU) has consistently adopted a broad interpretation of the reproduction right. In the landmark *Infopaq* ruling,[31] the CJEU held that even a short sequence of 11 words extracted from a press article could qualify for protection, thus requiring authorisation from the rightsholder. This case is interesting as it involves various processes, including the technical process of Optical Character Recognition (OCR) that can be compared with some steps in the process of training a GenAI tool.[32] Furthermore, the Court's discussion of temporary copies in *Infopaq* raises important questions about the temporary copying exception, a provision that may still apply to certain technical activities performed by GenAI during training next to the main provision for text and data mining for commercial purposes.

(ii) The available exception prior to the introduction of the text and data mining exceptions: temporary reproduction

Under Article 5(1) of the InfoSoc directive, some temporary copies are considered reproductions but can be allowed without needing permission from the rights holders. The Temporary Reproduction exception (“TR exception”) relies on five conditions: (1) the copies are temporary, (2) transient or incidental, (3) are an essential part of a technical process, (4) their main purpose is to enable transmission between third parties or lawful use of the content, and (5) they have no independent economic significance, i.e. they do not conflict with the normal exploitation of the work. The CJEU also highlights that deletion of these copies should happen automatically, without human involvement. Activities like browsing or caching are examples of such permitted acts.

Some uncertainty results from the interpretation of the conditions for this exception to apply to AI systems. First, in the *Infopaq* case[33], the CJEU determined that the cumulative effect of processing and storing multiple extracts increases the chances of reproductions. It was emphasised that an assessment is needed to determine whether there is a risk that the reproduction could "remain in existence for a longer period, according to users' needs." Consequently, it is a question of whether reproductions during the training phase can be considered transient (criterion 2), meaning limited to the completion of the technological process and then deleted. Copies, though not inherently stored, cannot be permanently deleted as they are often fed back into black boxes and training models.

Second, the use/copies might have some economic significance. The fifth requirement (that the use of the work has no economic significance) means that the implementation of those acts does not enable the generation of an additional profit, going beyond that derived from lawful use of the protected work.[34]

(iii) The text and data mining exception for commercial purposes

The introduction of two exceptions for text and data mining by the legislator gives an indication that the technical acts performed to pursue TDM activities are not likely to be considered temporary copies, or else no new exception should be needed. The EU legislator also explained that: “*users of text and data mining*

[31] Case Infopaq II C-302/10.
[32] Optical character recognition (OCR) is a technology that converts an image of a text into a machine-readable text format.
[33] Idem.
[34] Article 5(1) of the 2001/29 (InfoSoc) directive.

could be faced with legal uncertainty as to whether reproductions and extractions made for the purposes of text and data mining […] do not fulfil all the conditions of the existing exception for temporary acts of reproduction provided for in Article 5(1) of Directive 2001/29/EC." (recital 18 of the CDSM dir.).

The TDM exception introduced in Article 4 of the CDSM directive (the "Article 4 TDM exception") is a statutory authorisation to reproduce lawfully accessed copyrighted materials to perform TDM activities for commercial purposes, unless right holders reserve the use of their works in an appropriate manner.

The main conditions are that a) the content is "*lawfully accessible*", b) the storing/preservation of the input data is time-limited (for the TDM operation), and c) that there is no appropriate opt-out by the rightsholders. This last condition is discussed below.

TDM is permitted if the use of works has not been "*expressly reserved by their rights holders in an appropriate manner, such as machine-readable means in the case of content made publicly available online*." Depending on whether the content was publicly available online or not, rightsholders should either reserve their rights using machine-readable means (including meta-data and terms and conditions of a website or a service) or use contractual agreements such as unilateral declaration (recital 18 of the CDSM dir.).

The "appropriateness" of the expression is the key condition for Article 4 on text and data mining (TDM), but how it's interpreted is still being debated. Mentioning a unilateral declaration could impact the interpretation of the "machine-readable means" requirement for online content. For example, terms and conditions that are not clearly accepted as part of a contract might be seen as a one-sided expression of intent, which could work as a valid opt-out.[35] Similarly, by sending a notice, another type of unilateral declaration could be accepted as a legitimate opt-out. The opt-out process comes with legal and technical challenges that industry players need to consider, especially when it comes to using technical tags like robots.txt to make the opt-out machine-readable, and determining who is authorised to opt-out on behalf of authors.

Thus Article 4 of the CDSM dir. can constitute a legal basis to claim that the authors' right of reproduction is not infringed upon when an author, who did not exercise their right of opt-out, contributed to an input because their copyrighted content was extracted and used.

(iv) The existence of a right of remuneration for the author of a co-generated input

Whether there exists a right to remuneration for the acts of reproduction effectuated by GenAI is a debated question. Under the classic copyright framework, when an author authorises reproduction, i.e.: use of its works, they are entitled to remuneration. The CDSM dir. even introduced the principle that authors are entitled to a "fair and proportionate" remuneration in all their primary contracts. The question whether remuneration, or compensation should be given to authors of copyrighted content used to train GenAI is not solved in the applicable laws nor in the doctrine.

Under the temporary reproduction exception ("TR exception"), no remuneration is provided to authors. Temporary copies, like those used for buffering or caching, are considered transient and serve technical purposes. Since these copies are not intended for permanent use or distribution, legislators have determined that no harm is done to rights holders hence no compensation is necessary.

[35] Any reference ?See recital 18 of the 2019/790 CDSM directive: ""*in the case of content that has been made publicly available online, it should only be considered appropriate to reserve rights by the use of machine-readable means, including metadata and terms and conditions of a website or a service*" ; "*in other cases, it can be appropriate to reserve the rights by other means, such as contractual agreements or a unilateral declaration*".

Regarding Article 4 of the TDM exception, it remains unclear whether authors are entitled to remuneration when their works are used without a valid opt-out (see below the output stage a proposal where authors would be remunerated regardless of whether they opted-out). The CDSM directive explicitly excludes compensation for TDM under Article 3 (research purposes) but is silent on remuneration for Article 4, which covers commercial use. Given that some copyright exceptions, like private copying, include compensation, national legislators may consider similar models for GenAI activities.

C. Geiger and V. Iaia propose a *statutory licence fee* for machine learning processes, a remuneration right that seeks to offer a balancing framework to facilitate the development of AI technology and the rightful remuneration of authors. Their proposal recommends an input-based remuneration system which would consist of a "permitted-but-paid" model, or else called "limitation-based remuneration right" like the private copying exception and levies that already exist under EU copyright law.[36] However the proposal lacks an analysis of the relationship with the Art. 4 TDM exception of the CDSM dir.: it proposes to replace the opt-out mechanism by this statutory remuneration right. The solution is thus not applicable as long as the TDM exceptions are in place – but removing this TDM exception recently adopted (in 2019) is not a realistic option in the short and medium term (see below for a different mechanism based on output).

Outside Article 4, if an author opts out of TDM, they retain full control of their works under the exclusive right of reproduction. Any use of their works would then require permission, allowing authors to oppose unauthorised use and pursue infringement claims. In such cases, GenAI operators would need to secure licensing agreements and pay royalties based on negotiated terms, ensuring authors are compensated for use outside TDM exceptions.

b) Data protection

Preliminary Remarks. In the European Union, the main legal framework for Data Protection, is the GDPR[37]. The Regulation uses a broad notion of personal data, which means any information relating to an identified or identifiable natural person.[38] To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used.[39] That includes, inter alia, texts from books, scientific articles as well as videos and photographs on the internet. If the model trained with the input data is able to "remember" the input data, the model itself may be qualified as personal data if the reproduction of the training data is generally likely.[40]

(i) Rights arising from unlawful use of data

A legal basis is required for the processing of personal data to be lawful. In this context, consent[41] and the fulfilment of a contract[42] are regularly not applicable when the data is scraped on the internet. Instead,

[36] Christophe Geiger, Vincenzo Iaia, The forgotten creator: Towards a statutory remuneration right for machine learning of generative AI, Computer Law & Security Review, Volume 52, 2024, 105925
[37] Regulation (EU) 2016/679.
[38] Art. 4 Nr. 1 GDPR.
[39] Recital 26 GDPR.
[40] Cf. Pesch/Böhme, Verarbeitung personenbezogener Daten und Datenrichtigkeit bei großen Sprachmodellen ChatGPT & Co. unter der DS-GVO, MMR 2023, 917 (920); see further Veale/Binns/Edwards, Certification systems for machine learning: Lessons from sustainability, Phil. Trans. R. Sic. A A:376:20180083;
[41] Art. 6 para. 1 lit. a GDPR.
[42] Art. 6 para. 1 lit. b GDPR.

processors would typically rely on legitimate interests[43] as the legal basis for the data processing.[44] Thus, the lawfulness for data processing with regard to the input of GenAI requires a balancing of interests.[45]

Legitimate interests lie both in the economic interests of the providers and in the public's broader interest in accessing information.[46] Whether a balancing of interests can serve as a valid basis depends, among other factors, (1) on the extent of the data processed and whether this goes beyond the information that a user would typically access on the internet. What may also be taken into account is (2) the purpose of the development of GenAI and (3) if the persons concerned are minors. Another aspect is (4) how and what information about individuals is disclosed to users when interacting with the AI.

While these considerations apply to 'normal' personal data, the processing of special categories of personal data faces additional hurdles: For the training of GenAI, the processing of (special categories of) personal data could also be based on the permission of Art. 9(2)(e). Accordingly, the processing of special categories of personal data is permitted if it concerns personal data "which are manifestly made public by the data subject". This exception appears to apply not only to special categories of personal data, but - *a maiore ad minus* - also to the processing of "normal" personal data. However, the CJEU has recently clarified, that also in this case, a legal ground under Art. 6 GDPR – e.g. the processing based on legitimate interests – is necessary.[47] But even for special categories of data, Art. 9(2)(e) may face practical hurdles, according to literature, because it is difficult for data controllers to assess whether the data has manifestly been made publicly available by the data subject themselves or by a third party.[48]

The consequence of unlawful data processing is that affected individuals may seek cessation of processing and/or deletion of their personal data,[49] and compensation for damages[50]. In addition, the data subject may also have claims for unjust enrichment. Although the GDPR does not provide for these claims itself, they may arise from the applicable national law. According to the widespread opinion, the GDPR does not preclude the use of these national provisions for unjust enrichment, as Recital 146 s. 4 expressly permits competing claims in addition to Art. 82 GDPR.

(ii) Right to erasure

The data subject is entitled to erasure in case of wrongful data processing under Article 17 GDPR. This is *inter alia* the case, where the personal data is no longer necessary in relation to the purposes for which they were collected or otherwise processed or if the data subject has withdrawn consent on which the processing is based according to point (a) of Article 6(1), or point (a) of Article 9(2), and where there is no other legal ground for the processing.

Article 17(3) sets out, under which cases the right to erasure shall not apply. For example, when the processing is necessary for exercising the right of freedom of expression and information, for compliance with a legal obligation, for reasons of public interest in the area of public health or for the establishment, exercise or defence of legal claims.

---

[43] Art. 6 para. 1 lit. f GDPR.
[44] Cf. Novelli/Casolari/Hacker/Spedicato/Floridi, Generative AI in EU Law: Liability, Privacy, Intellectual Property, and Cybersecurity, 15 March 2024, 8 f., available via https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4694565; Ruschemeier, Generative AI and Data Protection, 6 et seq, available via https://ssrn.com/abstract=4814999.
[45] See Ruschemeier, Generative AI and Data Protection, 8 et seq, available via https://ssrn.com/abstract=4814999.
[46] Ibid.
[47] CJEU 21.12.2023 C-667/21 - ECLI:EU:C:2023:1022
[48] EDPB, Report of the work undertaken by the ChatGPT Taskforce, 23 May 2024, mn. 18, available via https://www.edpb.europa.eu/system/files/2024-05/edpb_20240523_report_chatgpt_taskforce_en.pdf.
[49] Art. 17 GDPR.
[50] Art. 82 GDPR.

The request for erasure[51] poses major challenges for processors if the AI model can "remember" trained data or if the trained personal data can be extracted or inferred from the model.[52] Deleting the model raises concerns about balancing individual rights with the broader societal benefits of AI technologies as well as with aspects of environmental sustainability due to the resources used to train these models again.[53] Especially, because it remains unclear how processors should fully comply with the right to erasure in these cases.[54]

(iii) Right to information

Irrespective of the lawfulness of the data processing, the data subjects always have to be informed on the processing of their personal data. Regarding the information on data collection, in accordance with Articles 13 and 14 GDPR, notification must be provided, including when the data is collected from publicly available sources. Informing all affected individuals about the processing of their data from public sources is not feasible for providers of GenAI.[55] For such contexts, Article 14(5)(b), first sentence GDPR applies, which excludes the application of the information duty if providing the information would involve disproportionate effort.[56] Ultimately, a balancing of the resulting effort with the informational interests of the affected individuals must be carried out. This balance again depends on the extent to which, and whether exclusively, public data is processed, such as data that can be found via a search engine. In such cases, the informational interest of the affected individuals would be considerably lower, and the effort outweighing this interest would be justifiable. Even though Art. 13 et seq. GDPR are often to referred to as a right[57], the provisions only set out a duty on the processor, but they are not a right of the data subject.

(iv) Right to access

In contrast to the Information duties in Art. 13 et seq., the regulation sets out a right of access in Art. 15 GDPR, which entitles the data subjects to be informed on the processing and also to obtain a copy of the processed personal data. Art. 15 applies to rightfully and wrongfully processed data. However, this right shall not enable the data subject to further process the data themselves or via a third person. It rather serves the interest of the data subject to be informed about the data processing and to be able to assert any other rights, such as the right to erasure, to compensation or to rectification of the data.

(v) Right to rectification

Data subjects may also assert their right to rectification of inaccurate personal data concerning them under Art. 16 GDPR. This right only applies to input data if the model can remember and therefore reproduce the data. Personal data is initially incorrect if its content is untrue. In principle, the claim only relates to factual information, as only this can be "incorrect". Furthermore, Art. 16 GDPR also applies to incomplete data.

---

[51] Art. 17 GDPR.
[52] Novelli/Casolari/Hacker/Spedicato/Floridi, Generative AI in EU Law: Liability, Privacy, Intellectual Property, and Cybersecurity, 15 March 2024, 12, available via https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4694565.
[53] Hacker, Sustainable AI Regulation, Common Market Law Review (forthcoming), https://arxiv.org/abs/2306.00292.
[54] Novelli/Casolari/Hacker/Spedicato/Floridi, Generative AI in EU Law: Liability, Privacy, Intellectual Property, and Cybersecurity, 15 March 2024, 12, available via https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4694565.
[55] Cf. Hacker/Engel/Mauer, Regulating ChatGPT and other Large Generative AI Models, 7 February 2023, 2 f., available via https://www.europeannewschool.eu/images/chairs/hacker/Hacker_Engel_Mauer_2023_Regulating_ChatGPT_Feb07.pdf.
[56] Cf. EDPB, Report of the work undertaken by the ChatGPT Taskforce, 23 May 2024, mn. 27, available via https://www.edpb.europa.eu/system/files/2024-05/edpb_20240523_report_chatgpt_taskforce_en.pdf.
[57] Solove, The Limitations of Privacy Rights, Notre Dame Law Review 98, 2023, 975 (994).

(vi) Right to data portability

Data subjects may also assert their right to data portability under Art. 20 GDPR concerning their personal data. The provision entitles data subjects to receive the personal data they have provided[58] to a controller based on consent or contract. Provided data is commonly referred to as data that is provided by individuals when they explicitly share information about themselves or others.[59] However, it has been argued, especially by the European Data Protection Board, that the provision also applies to observed data, i.e. data that is created where activities are captured and recorded and where – in contrast to provided data – the role of the data subject in the case of observed data is passive.[60] Taking this interpretation, the right to data portability would also apply to most of the personal data that has been used for the training of GenAI.

Under Art. 20 GDPR, the personal data has to be supplied in a structured, commonly used and machine-readable format[61] and the data subjects have the right to have the data transmitted directly from one controller to another, where technically feasible. The right to data portability in Art. 20 GDPR should strengthen the data subject's control over data[62] and is considered as a means to tackle "lock-in" situations.[63] However, if the personal data is linked with other personal data, the right to data portability shall not adversely affect the rights and freedoms of others.[64] This means that not only the data protection rights of third parties, but also every individual interest protected by European primary law must be taken into account. This means that *a balancing of interests is also necessary at this point and the right to data portability is therefore void in those cases in which the protection of other interests prevails.*

Due to its (potential) limited scope, the right to data portability in Article 20 of GDPR has recently been accompanied by far reaching *data access rights in the Data Act*[65]*.* The Regulation obliges producers of connected devices and providers of related services to design and manufacture/provide their products in services in such a manner, that product data and related service data, including the relevant metadata necessary to interpret and use those data, are, by default, easily, securely, free of charge, in a comprehensive, structured, commonly used and machine-readable format, and, where relevant and technically feasible, directly accessible to the user.[66]

Where data cannot be directly accessed by the user from the connected product or related service, data holders shall make readily available data, as well as the relevant metadata necessary to interpret and use those data, accessible to the user without undue delay, of the same quality as is available to the data holder, easily, securely, free of charge, in a comprehensive, structured, commonly used and

[58] Examples include creating a social network profile and entering credit card information for online purchases" (see OECD (2019) p 20). It has been argued, especially by the European Data Protection Board (see Working Party 29 (2017) Guidelines on the right to data portability, WP 242 rev.01, p 9 ff. https://ec.europa.eu/newsroom/article29/items/611233. Accessed 23 November 2021), that the provision also applies to observed data, i.e. data that is created where activities are captured and recorded and where – in contrast to provided data – the role of the data subject in the case of observed data is passive. However, the EDPB does not apply the provision to derived data which is created by the data controller on the basis of the data "provided by the data subject".

[59] De Hert/Papakonstantinou/Malgieri/Beslay/Sanchez, The Right to Data Portability in the GDPR: Towards User-Centric Interoperability of Digital Services,Computer Law & Security Review (2018) 193 (199 et seq.).

[60] See Working Party 29 (2017) Guidelines on the right to data portability, WP 242 rev.01, p 9 ff, available via https://ec.europa.eu/newsroom/article29/items/611233

[61] The GDPR does not provide for a certain format in this regard, but only encourages data controllers to develop interoperable formats, see Recital 68 GDPR.

[62] See Recital 68 GDPR.

[63] See Crémer J, de Montjoye Y-A, Schweitzer H (2019), Competition policy for the digital era. Publications Office of the European Union, Luxembourg, p 81 ff. https://ec.europa.eu/competition/publications/reports/kd0419345enn.pdf. Accessed 23 November 2021.

[64] Art. 20(4) GDPR.

[65] Regulation (EU) 2023/2854.

[66] Art. 3(1) Data Act.

machine-readable format and, where relevant and technically feasible, continuously and in real-time. This shall be done on the basis of a simple request through electronic means where technically feasible.[67]

In addition, the Data Act also provides for a data portability right in its Art. 5: "Upon request by a user, or by a party acting on behalf of a user, the data holder shall make available readily available data, as well as the relevant metadata necessary to interpret and use those data, to a third party without undue delay, of the same quality as is available to the data holder, easily, securely, free of charge to the user, in a comprehensive, structured, commonly used and machine-readable format and, where relevant and technically feasible, continuously and in real-time."

(vii) Right to object

The GDPR grants the data subject the right to object at any time to processing of personal data concerning them. However, data subjects are only entitled to this right if the processing is based on legitimate interests under Art. 6(1)(f) GDPR. Since the processing of personal data for Co-Generated Input of GenAI is mainly based on legitimate interest, the right to object will have high practical relevance.

Under Art. 21 GDPR, the controller is obliged to no longer process the personal data unless they demonstrate compelling legitimate grounds for the processing which override the interests, rights and freedoms of the data subject or for the establishment, exercise or defence of legal claims.

2. Co-generated output of GenAI

a) Copyright

Right holders may exercise the same rights with regard to the output as the one they can claim with regard to the training (and the input), with however one additional right linked to their right of communication to the public (as the output of a GenAI tool is made accessible to a public, contrary than the processing taking place during the training phase).

(i) The exclusivity granted by the right of reproduction

The right of reproduction can give rise to various types of claims when it is infringed i.e.: when a work is used without authorisation.

An AI output can be deemed infringing if the output is *substantially* like a protected work and if the copyright work was used unlawfully. This would give rise to the cease and desist claim against an AI operator to stop infringing a copyrighted work. The standard of substantial similarity for copyright infringement seems to apply in both the EU and the US, with greater codification in the U.S. legal system. In both systems, whether a GenAI output is infringing a copyright will depend on a case-by-case appreciation by the judges.

In the EU, Article 4 TDM exception does not cover reproduction in the output generation phase of a GenAI. The text is only a derogation to the right of reproduction (and of database extraction) for acts of "reproductions and extractions of lawfully accessible works and other subject matter for the purposes of text and data mining", but it does not mention the distribution or communication to the public of the output. This could constitute an infringement to the exclusive rights of reproduction and of communication to the public, but also to the moral right to be credited as an author (see below). Under an unlawful reproduction claim, the copyright holder may seek financial compensation for the losses suffered due to the infringement.

[67] Art. 4(1) Data Act.

Aside, the imposed transparency obligations of the AI Act that will be supervised by the new AI Office could also lead to a regulatory dialog whose outcome might include some commitment on, or determination of, a fair remuneration. The AI Act requires operators of general-purpose AI (GPAI) models to comply with EU copyright laws, including respecting opt-outs from copyright holders, especially under the TDM exception. A key part of this is the obligation to document and publicly share information about the data used to train their models. This transparency will allow rightholders to better enforce their rights, such as the right of reproduction and to receive fair remuneration for the use of their content. However, the ability to claim trade secret protection might weaken the effectiveness of this transparency obligation.

Additionally, in some cases, the AI might not have been trained on the original work, yet still produces a similar output, akin to the concept of "independent (double) creation" seen with human creators. In such instances, courts often examine whether the alleged infringer had access to the original work or had prior exposure to the first work before determining liability. This may pose issues of burden of proof and liability since GenAI providers would need to prove that the model was not trained on the original work. Alternatively, responsibility could rely on the users who could have inputted works during the generation stage in a manner that could lead to copyright infringement.

Finally, GenAI can easily produce hundreds of works in the style of known painters like Matisse or Monet without reproducing (even partially) any of their paintings. Under the current infringement copyright test, no ruling has ever considered that recognisable styles of prior works could qualify as copyright infringements but the extent to which a style could be copied with GenAI may challenge this approach.

(ii) The exclusivity granted by the right of communication to the public

The right of communication to the public is defined in Art. 3(1) of the InfoSoc dir. as: “the exclusive right to authorise or prohibit any communication to the public of their works, by wire or wireless means, including the making available to the public of their works in such a way that members of the public may access them from a place and at a time individually chosen by them”.

The Art. 4 TDM exception is not a derogation to the right of communication to the public. In this sense, if a copyrighted work is communicated to the public via a generated output, the same kind of reasoning as under an infringement to the right of reproduction at the output stage would apply: a GenAI operator would face liability but could also enter into a licensing agreement with an author for the communication of its works.

Some legal difficulties will arise in relation to the definition of an act of communication to the public under Article 3 of InfoSoc dir. The Jurisprudence of the CJEU considered that such act occurs when a link communicates protected subject matter to a “new public”, meaning a public not taken into account when the copyright holder authorised the initial publication of their work or protected subject matter.

(iii) The rights falling within the author’s moral rights (where they exist)

Moral rights, particularly the rights to attribution (paternity) and the protection of the integrity of a work, safeguard the personal and reputational interests of human creators. They are not harmonised through European Union law. In most jurisdictions, moral rights cannot be waived by contract. They are a foundational element of copyright law, designed to protect the personal and non-economic interests of authors and creators.

The two most relevant moral rights are:

a) the right to claim authorship of the work (i.e. to be identified as the author of the work), which implies that the source, including the author's name, be indicated unless this proves impossible,

(b) the right to preserve the integrity of the work (i.e. to request that the work be used without modification or alteration). This right protects the author against material modifications, whether the modification is made by addition, removal or in any other way, but also against those of a non-material nature insofar as they affect the spirit of the work.

The Article 4 TDM exception does not override moral rights. Therefore, even if an author permits the use of their work during the training phase—either through a contract or under the Article 4 TDM exception—their moral rights may still be violated if the resulting output harms the integrity of the original work. For example, if the output mocks, transforms, or dismantles the copyrighted work, this could constitute an infringement of the author's moral rights. Additionally, the right to attribution poses significant challenges when Generative Pre-trained AI models fail to properly credit the original authors whose works have clearly been used to generate the output.

(iv) Output based remuneration

A different type of remuneration scheme based on output is discussed amongst some commentators who believe that copyright holders should benefit from a statutory remuneration right, defined and calculated by the legislator. It would consist of a levy on Gen AI systems rather than be based on their training.[68] The generation of an output would suffice to serve as a reference point and to trigger the obligation to pay remuneration. The output-based remuneration model is based on the argument that remuneration measures introduced in copyright are proposed to ensure that authors receive some income and are incentivised to create more content. It is irrelevant whether the output resembles or contains parts of the protected inputs, i.e.: the right to remuneration would not be triggered by the reproduction right. It would suffice that the tool, to be subject to the remuneration, "*cannot produce results that resemble human literary and artistic works unless it has had the opportunity to analyse human creations.*"[69] The levy would be uniformly applied to all providers of GenAI tools. The lump sum would be calculated on a "*general, abstract assessment of whether an AI system is capable of serving as a substitute for human literary and artistic productions*"[70]. The evidence of this substitution effect may be sufficient to require the payment of a levy. The levy would be administered and enforced through a collective management society. While this idea is gaining traction in discussions about compensating creators whose works are used in training AI models, it has not yet been fully developed or grounded in legal texts, treaties, or binding regulations.

(v) The copyrightability of an AI-generated output ?

A lot of creators take advantage of creating with AI which has the potential to stimulate creativity. Whether generated outputs could be protected by copyright was a debated topic when GenAI were first being deployed on the market. Most EU commentators consider that AI generated outputs cannot be eligible for copyright protection essentially because no moral rights could be attributed to a machine. As explained above, one of the foundational justifications of copyright protection is to protect the expression of free and creative choices stemming from *personal* creation. Copyright law is inherently a personal right which embodies the individual's ability to create.

[68] Senftleben, Martin. "Generative AI and author remuneration." *IIC-International Review of Intellectual Property and Competition Law* 54.10 (2023): 1535-1560.
[69] ibid.
[70] ibid.

However, this does not exclude that AI Assisted human works could be eligible to copyright protection and that authors using AI tools may be entitled to the abovementioned rights based on their AI created works. The U.S. and Japanese Copyright Offices issued guidance on this matter which is outlined below.

b) Data protection

Preliminary Remarks. The generated texts often contain real names. In the context of the text prompts or other information from the generated text, it may be possible to establish a clear reference to certain real persons. This may initially be the case if the GPAI reproduces personal data. This can be "remembered" personal training data or personal data from text prompts. In such cases, the text output is also considered personal data. There is no difference to the retrieval of personal data retrieved from a database. Thus, with regard to "remembered" personal data, reference can be made to the outlines above (B.I.1.b)).

Data subjects have the same rights with regard to the output as with regard to the data input: rights arising from unlawful procession, the right to erasure, to information, to access, to rectification, to data portability, to object and not to automated decision-making. In this respect, reference can generally be made to the above explanations (B.I.1.b)). In the following, it will only be discussed if additional aspects deviating from the above explanations need to be considered.

In addition, the right not to be subject to automated decision-making, which only comes into question with regard to co-generated output, is discussed below.

(i) Rights arising from unlawful use of data

New data generated by GenAI can be related to individuals, for example because a name from the training data or a text prompt is linked to specific information for the first time. Against the background of how GenAI works, the question arises as to whether such text outputs can be objectively understood as relating to real persons.

If data issued by a model can be assigned to a real person, this is regularly recognisably based on information about this person in the training data. In many of these cases, identification is also probable according to general judgement. These data may be qualified as personal data and thus the processing of this data, e.g. generating the output, is only lawful if a legal ground in Art. 6 GDPR and, concerning special categories of personal data, Art. 9 is given. Consent[71] may be applicable when it comes to the legal basis of prompts containing personal data. But this is only the case, when the users include personal information about themselves in prompts. If the prompts contain personal information about third parties, the user cannot validly consent for this other person.

Data subjects have the right to withdraw their consent at any time.[72] This withdrawal must be as easy as it was to give consent initially.[73] Once consent is withdrawn, the data controller must cease processing the data that was based on that consent, unless there is another legal basis for the processing. Importantly, the withdrawal of consent does not affect the lawfulness of processing that took place prior to the withdrawal.[74]

Besides consent and the fulfilment of a contract, the processor may only rely on the legitimate interest for the processing[75]. The consequence of unlawful data processing is that affected individuals may seek cessation of processing and/or erasure of their personal data, and compensation for damages. Reference can be made to the outlines above (B.I.1.b)(i)).

[71] Art. 6 para. 1 lit. a GDPR.
[72] Art. 7(3) Sentence 1 GDPR.
[73] Art. 7(3) Sentence 2 GDPR.
[74] Art. 7(3) Sentence 3 GDPR.
[75] Art. 6 para. 1 lit. f GDPR.

(ii) Right to access

The processing of personal data submitted by users on themselves in a chat interface (prompts) is subject to the information requirements in Art. 13 GDPR.[76] On other personal data, Art. 14 and its exception in Art. 14(5)(b) GDPR applies (See further above B.I.1.b)(iv)).

Additionally, data subjects can assert their right to information under Art. 15 GDPR. Where the data processing included automated decision-making, the information has to include meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject (Art. 15(1)(h) GDPR). Since the scope of this provision has been extended by the rulings in the SCHUFA as well as the Uber and Ola cases, Art. 15(1)(h) GDPR was given a significant role alongside Art. 86 AI Act and it was even doubted that the latter provision will now bring any added value at all compared to Art. 15(1)(h) GDPR.

(iii) Right to not be subject to automated decision-making

Data subjects have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning them or similarly significantly affects them under Art. 22 GDPR. The provision shall not apply if it: (a) is necessary for entering into, or performing, a contract between the data subject and a data controller; (b) is authorised by Union or Member State law to which the controller is subject, provided that such law also establishes appropriate measures to protect the rights, freedoms, and legitimate interests of the data subject; or (c) is based on the explicit consent of the data subject.[77]

On 7 December 2023, the CJEU issued a landmark judgement on Art. 22 GDPR in the *Schufa* Cases.[78] Accordingly the provision also applies, when a third party draws strongly on the probability value or score to establish, implement or terminate a contractual relationship with the data subject. The Court noted a risk of circumventing Article 22 of the GDPR and a lacuna in protections if a narrow approach was taken and the Schufa score was only regarded as preparatory.

3. Interim conclusion

Copyright. In the context of GenAI, *at the level of training*, a co-generated input could require an authorisation from the copyright owner if a protected work has been used (in the form of reproductions) to generate it. Considering that styles, mere facts or information are not protected by copyright, their use in the co-generation process would not fall under copyright law. There is thus uncertainty as to whether a right can be claimed with regard to the co-generation of input.

For temporary reproduction –the author of the co-generated input cannot claim anything in the case of a reproduction that is temporary: a copyright holder does not have the right to authorise or prohibit the use of its works as the exception is a strict exception to the right of reproduction (TR exception), and no right of remuneration for the temporary use of copies exists.

For TDM – a copyright holder still retains some form of control over his/her work when TDM is made for commercial purpose. The author retains a right to oppose and reserve his/her rights, but the way the right must be manifested is still subject to varying interpretations. Concerning remuneration, it still appears possible for national legislators to introduce a remuneration right when works subject to an opt-out were used by GPAIs during the co-generation process.

---

[76] EDPB, Report of the work undertaken by the ChatGPT Taskforce, 23 May 2024, mn. 28, available via https://www.edpb.europa.eu/system/files/2024-05/edpb_20240523_report_chatgpt_taskforce_en.pdf.
[77] Art. 22(2) GDPR.
[78] CJEU 7.12.2023 C-634/21 – SCHUFA Holding u.a. (Scoring).

At the *output stage*, an AI-generated work could infringe the right of reproduction if it is substantially similar to a protected work and was unlawfully used. However, styles and mere ideas, not being protected by copyright, would not trigger such claims. Concerning the exclusive right of communication to the public, a copyright holder retains the right to authorise or prohibit communication of the work via AI-generated outputs, especially if it reaches a "new public" that was not originally intended to access the work.

With regard to *moral rights* – even if lawful use occurred in training –, moral rights may be violated if the output affects the integrity of the original work or fails to credit the author appropriately. These non-economic rights remain protected.

Regarding *remuneration*, there is growing discussion about establishing a new output-based remuneration scheme. Some commentators propose that copyright holders should benefit from a statutory right to remuneration for AI-generated outputs, which would be triggered not by the reproduction right, but simply by the generation of outputs by AI systems. This scheme would function as a levy imposed on all providers of GenAI tools. The levy would be justified by the argument that AI systems, by analysing human-created works during training, become capable of generating outputs that resemble human literary or artistic creations. Under this model, the obligation to pay remuneration would arise regardless of whether the output directly incorporates protected elements from the training data. Instead, it would focus on the general capacity of the AI to substitute human creativity, acknowledging the indirect impact on human creators' ability to benefit from their work. This levy would be uniformly applied and managed through a collective management society, ensuring that creators receive compensation based on an abstract assessment of the AI system's potential to replace human-made works

Data Protection. In the context of GenAI, processing personal data requires a valid legal basis, with legitimate interests often being the most relevant. Consent or contract fulfilment typically does not apply to data scraped from online sources. Determining whether legitimate interests are valid depends on a balance of factors, including the type and extent of data used, the AI's purpose, and whether the data involves minors. Special categories of data, such as sensitive information, require that the data was manifestly made publicly available by the data subject. Unlawful data processing can lead to claims for cessation, erasure or damages. The right to erasure can be difficult when AI models have already been trained on personal data. The right to information may also face limitations due to the significant effort required to notify all affected individuals. Additional rights, such as access, rectification, and data portability, enable data subjects to maintain control over their personal information. While the Data Act introduces new portability rights for data generated by connected devices, its relevance to GenAI remains limited. Finally, the right to object to processing based on legitimate interests is expected to play a key role, given that most AI data processing relies on this legal basis.

The same applies where the GenAI model is able to "remember" the input data. But the use of data by GenAI can potentially generate new information related to individuals, such as linking a name to specific information for the first time. This raises questions about whether such AI-generated outputs can be understood as relating to real persons. If an AI model's output can be tied to a real person, this likely stems from personal data in the training set, making it subject to GDPR requirements. For processing to be lawful, a legal basis under Art. 6 or Art. 9 GDPR must apply, especially in cases involving special categories of personal data. Consent may be a valid legal basis when users input their own personal data into prompts, but not when they include data about third parties. Individuals can withdraw their consent at any time, and processing based on consent must stop unless another legal basis exists. Unlawful data processing allows individuals to request the cessation of processing, erasure of their data, and compensation for damages. Individuals also have the right to access personal data under GDPR, with specific rights for automated decision-making outlined in Art. 15 and Art. 22. Recent rulings, like the Schufa case, have emphasised that

decisions based heavily on automated processes or scoring systems may fall under the protections of Art. 22, even if such scores are only used in preparatory stages of decision-making.

The rights in the GDPR, especially the right not to be subject to automated decision-making and the right to information, have similarities with the new duties in the AI Act. According to the right to explanation of *individual decision-making in Art. 86 AI Act*: "Any affected person subject to a decision which is taken by the deployer on the basis of the output from a high-risk AI system listed in Annex III, with the exception of systems listed under point 2 thereof, and which produces legal effects or similarly significantly affects that person in a way that they consider to have an adverse impact on their health, safety or fundamental rights shall have the right to obtain from the deployer clear and meaningful explanations of the role of the AI system in the decision-making procedure and the main elements of the decision taken." Since the scope of Art. 15 para. 1 lit. h GDPR has been extended by the rulings in the SCHUFA as well as the Uber and Ola cases, the GDPR-provision was given a significant role alongside Art. 86 AI Act and it was even doubted that the latter provision will now bring any added value at all compared to Art. 15 para. 1 lit. h GDPR.[79] In addition, Art. 50(2) AI Act contains information duties regarding the Output of GPAI: "Providers of AI systems, including general-purpose AI systems, generating synthetic audio, image, video or text content, shall ensure that the outputs of the AI system are marked in a machine-readable format and detectable as artificially generated or manipulated. Providers shall ensure their technical solutions are effective, interoperable, robust and reliable as far as this is technically feasible, taking into account the specificities and limitations of various types of content, the costs of implementation and the generally acknowledged state of the art, as may be reflected in relevant technical standards. This obligation shall not apply to the extent the AI systems perform an assistive function for standard editing or do not substantially alter the input data provided by the deployer or the semantics thereof, or where authorised by law to detect, prevent, investigate or prosecute criminal offences."

II. USA

1. Co-generated input of GenAI

A. Copyright

Like in the European Union's framework, the protection of the share of an input integrated into a cogenerated data at the level of training will first depend on whether that specific input data is protected by copyright.

Preliminary remarks: copyright protection. Like in the European Union, U.S. copyright laws do not protect styles, facts or ideas. The Copyright Act expressly protects "original works of authorship fixed in any tangible medium of expression, now known or later developed, from which they can be perceived, reproduced or otherwise communicated, either directly or with the aid of a machine or device".[80] Historically, copyright protection was subject to a registration requirement: works were to be registered and given a 'notice' to be considered copyrighted. Although this formality is no longer required for copyright protection, several incentives exist to register copyrighted works in the market. For this reason, the U.S. Copyright Office engages in a more meaningful discussion on the copyright protection of GenAI outputs compared to the EU, as will be discussed.

[79] Hacker, Comments on the Final Trilogue Version of the AI Act, 13 April 2024, 12, available via https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4757603.
[80] 17 U.S. Code § 102 – Subject matter of copyright

(i) The right of reproduction

The right of reproduction constitutes one of the exclusive rights of an author under the Copyright Act which enumerates five fundamental rights that the Act gives to copyright owners: the exclusive right of reproduction, adaptation, publication, performance, and display. Historical notes from the House of Congress explain very simply: the right "to reproduce the copyrighted work in copies or phonorecords" means the right to produce a material object in which the work is duplicated, transcribed, imitated, or simulated in a fixed form from which it can be "perceived, reproduced, or otherwise communicated, either directly or with the aid of a machine or device."[81] This provision applies to training acts of GenAI systems.

Very recently, the U.S. Patent and Trademark Office described the process of AI training as one that will "almost by definition involve the reproduction of entire works or substantial portions thereof"[82]. The Office further estimates that "whether this constitutes copyright infringement will generally be determined by considering the applicability of the fair use doctrine"[83] which is a defence for use of copyright works that promotes freedom of expression by permitting the unlicensed use of copyright-protected works in certain circumstances, for example news reporting, teaching etc.

Fair use is being largely discussed in the USA in recent GenAI litigation cases as courts are grappling with whether AI companies' use of copyrighted materials to train their models without permission qualifies as a lawful fair use or constitutes copyright infringement. We examine this defence below.

(ii) Fair use exception

Unlike the EU, the U.S. copyright framework has not introduced specific exceptions for text and data mining to promote innovation. However, the U.S. system may be seen as less restrictive for innovation due to the fair use doctrine, which allows certain uses of copyrighted material that would otherwise be infringing. Fair use protects these uses from liability, recognising their value in promoting creativity, research, and innovation. [84]

Four factors determine whether fair use may be applicable:

(i) **The purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes**. Without going into much details, as the exception is already largely discussed in the literature, it is worth highlighting that like under the EU exception for text and data mining under Article 4 of the CDSM Directive, the purpose of the use can be commercial. Using copyrighted materials may more easily be considered fair use in the absence of commercial benefits, but the doctrine applies to commercial uses. At the moment, large companies like Open AI seek to rely on fair use for the scraping of their data. A largely debated aspect of this criterion is that if a use is "transformative", a use is more likely to be considered fair. According to the Copyright Office, transformative uses are "those that add something new, with a further purpose or different character, and do not substitute for the general use of the work"[85]. In relation to GenAI, it cannot be excluded that use is transformative by nature at the level of training.

[81] Historical and Revision Notes, House report no. 94–1476.
[82] USPTO, Public Views on Artificial Intelligence and Intellectual Property Policy, October 2020.
[83] ibid.
[84] Connor Moran, "How Much Is T How Much Is Too Much? Copyright Protection of Short Portions of Text in the United States and European Union after Infopaq International A/S v. Danske Dagblades", Washington Journal of Law, Technology & Arts.
[85] U.S. Copyright Office, Fair Use Index.

(ii) **The nature of the copyrighted works.** This factor evaluates how the work being used serves the underlying purpose of copyright, which is to encourage creative expression. Consequently, the use of a more creative or imaginative work (for example, a novel, film, or song) tends to weaken a fair use argument, whereas the use of a factual work (such as a technical article or news report) is more likely to be viewed favourably in a fair use analysis.

(iii) **The amount and substantiality of the portion used in relation to the copyrighted work as a whole.** Courts look at the quantity and quality of the copyrighted material that was used. If the use includes a large portion of the copyrighted work, fair use is less likely to be found; if the use employs only a small amount of copyrighted material, fair use is more likely. It is debatable whether the quantity of the work at the input stage would play a significant role compared to the output phase where it could be less likely that fair use could apply if large quantities of recognisable works appear in generated outputs.

(iv) **The effect of the use upon the potential market for or value of the copyrighted work.** In this context, courts examine whether, and to what degree, the unlicensed use negatively impacts the existing or potential market for the copyright holder's original work. When evaluating this factor, courts consider whether the use is damaging the current market for the original (such as by reducing its sales) and/or whether the use could cause significant harm if it became widely adopted.[86]

Overall, numerous states' courts in the USA differ in the application of fair uses and many legal arguments can be found to support opposing conditions. In the absence of a clear ruling regarding the applicability of the fair use defence, it cannot be concluded whether this presents an adequate venue for text and data mining in the U.S. A judicial precedent should play in favour of AI companies: in the Google Books case (digitisation of books from libraries), the court found that "Google's unauthorised digitisation of copyrighted works, creation of a search functionality and display of excerpts from these works are fair uses that do not infringe copyright. The purpose of the copying is highly transformative, the public display of the text is limited, and the disclosures do not provide a meaningful commercial substitute for the protected elements of the originals. Google's commercial nature and profit motive did not justify denying fair use. But the use of protected content to train GenAItools confers a greater (economic) advantage; moreover, it conflicts more with the interest of rights holders (Google Books excerpts or snippets support book sales); on the other hand, outputs are fairly good substitutes for the creative content used for training, at least for graphic works.

(iii) Right to remuneration

In addition to the ongoing litigation in the judicial landscape, many of the same large defendants and GenAI operators are also entering into a series of agreements with major content providers to gain access to their content and data to train AI systems. OpenAI has signed licensing agreements with Associated Press, the Springer group, Le Monde, News Corp (The Wall Street Journal, The Sun, etc.), The Atlantic, Vox Media, Time and Condé Nast, among others - but when these negotiations fail, the rights holders take legal action, as in the case of The New York Times. Other start-ups include a quid pro quo for content creators in their service offering. Tess AI, for example, is a synthetic image generator that pays artists when their visual style is used (and the tool claims that artists reappropriate their style). Intermediaries are appearing, such as Fairly Trained, who aim to guarantee that fair training has been carried out by the labelling tools. Perplexity also plays the card of good collaboration with the content industries (but after being sued for plagiarism). However, a priori less valued (and monetised) content, such as posts on online networks, are also useful for training: for example, the Reddit platform is said to have signed a content licence (for a sum of 60 million USD per year) with an AI developer (as yet unknown) authorising the use of data exchanged

[86] U.S. Copyright Office Fair Use Index.

by users to train models (this licence on often personal data may pose a problem from a privacy point of view).[87]

While the U.S. copyright system does not offer an automatic right of remuneration or mandatory compensation like those found in the European Union, U.S.-based companies are more engaged in negotiating licensing agreements with copyright holders. This difference is largely due to the dominance of American AI companies, which need access to vast amounts of input data for training their models. As a result, U.S. firms tend to prioritise active negotiations with copyright holders to ensure they have the necessary licences, contrasting with the EU's more regulated approach to remuneration rights and text and data mining.

b) Data Protection

Preliminary Remarks. Unlike the European Union's GDPR, the U.S. lacks a unified data protection regime, leading to a patchwork of regulations that vary depending on the type of data, sector, and jurisdiction. Thus, Data protection law in the United States is characterised by a sectoral approach, where different industries and types of data are regulated by various laws at the federal and state levels, rather than through a comprehensive national framework. Key federal laws include the Federal Trade Commission (FTC) Act, the Gramm-Leach-Bliley Act, the Children's Online Privacy Protection Act (COPPA), the Family Educational Rights and Privacy Act (FERPA), and the Health Insurance Portability and Accountability Act (HIPAA).[88] However, the California Consumer Privacy Act (CCPA), and its successor, the California Privacy Rights Act (CPRA), have established broader rights for consumers, such as data access and deletion rights, serving as a model for other state-level privacy laws, for example the Virginia Consumer Data Protection Act (VCDPA).

The CCPA uses the term "personal information" that is defined as "information that identifies, relates to, describes, is reasonably capable of being associated with, or could reasonably be linked, directly or indirectly, with a particular consumer or household." Similarly, the Virginia Consumer Data Protection Act defines "personal data" as "any information that is linked or reasonably associated to an identified or identifiable natural person".

(i) Rights arising from unlawful use of data

The CCPA and the VCDPA both do not govern publicly available data.[89] The former excludes "publicly available information," which includes "information made available by a person to whom the consumer has disclosed the information if the consumer has not restricted the information to a specific audience."[90] In the same vein, the VCDPA excludes "publicly available information" from its scope. Publicly available information is defined as "information that is lawfully made available through federal, state, or local government records, or information that a business has a reasonable basis to believe is lawfully made available to the general public through widely distributed media, by the consumer, or by a person to whom

[87] Alain Strowel, François Wery, L'intelligence artificielle générative pour les juristes, Larcier, 2025, *forthcoming.*
[88] See Federal Trade Commission Act § 5, 15 U.S.C. § 45 (consumer data generally); Gramm-Leach-Bliley Act tit. V, 15 U.S.C. §§ 6801–6809, §§ 6821–6827 (financial services consumer data); Children's Online Privacy Protection Act §§ 1301–1308, 15 U.S.C. §§ 6501–6505 (children's online data); Family Educational Rights and Privacy Act § 438, 20 U.S.C. § 1232g (codifying FERPA) (educational data); Health Insurance Portability and Accountability Act of 1996, Pub. L. No. 104–191, 110 Stat. 1936 (codified as amended in scattered sections of 18, 26, 29, 42 U.S.C.) (healthcare data). Duffourc/Gerke/Kollnig, Privacy of Personal Data in the Generative AI Lifecycle, N.Y.U. Journal of Intell. Prop. & Ent. Law Vol 13(2) 2024, 219 (227 et seq).
[89] Duffourc/Gerke/Kollnig, Privacy of Personal Data in the Generative AI Lifecycle, N.Y.U. Journal of Intell. Prop. & Ent. Law Vol 13(2) 2024, 219 (237).
[90] Cal. Civ. Code § 1798.140(b) (West 2024).

the consumer has disclosed the information, unless the consumer has restricted the information to a specific audience:"[91]

Personal Information does not include publicly available information or lawfully obtained, truthful information that is a matter of public concern.[92] Thus, the CCPA excludes personal information from its scope of application that a business has a reasonable basis to believe is lawfully made available to the general public by the consumer or from widely distributed media; or information made available by a person to whom the consumer has disclosed the information if the consumer has not restricted the information to a specific audience.[93] Data scraping that concerns such data would be within the scope of both the CCPA and VCDPA, even though it has been stated that it can be difficult to determine whether personal data published to the general public by a third party was originally restricted to a specific audience.[94]

When processing personal data, the CCPA and the VCDPA primarily rely on information obligations and the consumer's ability to opt out of data sharing (or disclosure to third parties) to protect personal data.[95] They require businesses to inform consumers of plans to collect and use personal information for GenAI purposes prior to data collection. When it comes to sensitive personal data, the VCDPA additionally requires consent by the consumers, otherwise the collection of the data is unlawful. In contrast, the CCPA merely requires notification, also in case of procession of sensitive personal data.[96]

For individual consumers, violations of the CCPA provide a right to take legal action, especially in cases involving unauthorised access or disclosure of personal information due to inadequate security measures.[97] Consumers can file private lawsuits seeking statutory damages of $100 to $750 per incident or actual damages, whichever is greater. This legal recourse empowers consumers to claim compensation without needing to demonstrate specific harm, making it easier to hold businesses accountable. Additionally, if businesses fail to comply with consumer requests, such as not disclosing collected personal data or refusing deletion requests, consumers can report the violation to the California Attorney General. While consumers themselves cannot seek fines, the Attorney General can enforce compliance by imposing civil penalties of up to $2,500 for each violation and $7,500 for intentional violations, further ensuring businesses adhere to CCPA requirements.

Violations of the CCPA can only lead to an individual lawsuit if the affected parties have been harmed. Otherwise, they do not have standing to sue.[98] The Supreme Court denied standing to plaintiffs in *TransUnion LLC v. Ramirez* because they could not allege a later injury.[99] According to recent literature, consumers should have the right to sue in the event of data protection violations due to unjust enrichment of the company.[100] It would be possible to enforce data rights even in the absence of damage to the consumer, but the company would have to be enriched.[101]

---

[91] Va. Code Ann. § 59.1-575 (West 2023).
[92] Cal. Civ. Code § 1798.140(v)(2) (West 2024).
[93] Cal. Civ. Code § 1798.140(v)(2) (West 2024).
[94] Duffourc/Gerke/Kollnig, Privacy of Personal Data in the Generative AI Lifecycle, N.Y.U. Journal of Intell. Prop. & Ent. Law Vol 13(2) 2024, 219 (252).
[95] Cal. Civ. Code § 1798.100(a)(1), (2) (West 2023); Va. Code Ann. § 59.1-578 (C) (West 2023).
[96] Cal. Civ. Code § 1798.100(2) (West 2024) (requiring notification), *with* Va. Code Ann. § 59.1-578 (A)(5) (West 2023) (requiring consumer consent).
[97] See Loyola Consumer Law Review Vol 32(2) 2020, 246 (258 et seq).
[98] Chao, Unjust Enrichment: Standing Up For Privacy Rights, Iowa Law Review Online 108(49) 2023, 50 (61).
[99] TransUnion LLC v. Ramirez, 141 S. Ct. 2190, 2201 (2021). See the critique of Solove/Citron, Standing and Privacy Harms: A Critique of TransUnion v. Ramirez, 101 B.U. L. REV. ONLINE 62, 2021, 62.
[100] Chao, Unjust Enrichment: Standing Up For Privacy Rights, Iowa Law Review Online 108(49) 2023, 50 (64); Scholz, *Privacy Remedies*, IND. L.J. 94, 2019, 653 (658).
[101] Ibid.

However, selected state law grants standing in the event of unlawful data processing. The Illinois Biometric Information Privacy Act (BIPA) prohibits private entities from collecting a person's biometric identifier or information without the person's written consent.[102] In case of breach, BIPA allows for private lawsuits with statutory damages.[103]

(ii) Right to delete personal information

Consumers always have the right to request the deletion of any personal information a business collects under the CCPA.[104] Businesses also have to inform the consumers that deletion is available to them. Although there are exceptions to the right to delete,[105] these are only relevant in exceptional cases, particularly in connection with GenAI. The right of deletion can be found in different variations in other state law.[106] However, most state laws do not provide a private cause of action.[107] Violations of the right to erasure can only lead to an individual lawsuit if the affected parties have been harmed. However, claims may be viable due to the unjust enrichment of the business (see above B.I.1.b)(ii)).

(iii) Right to information

The CCPA and VCDPA require businesses that collect or use consumers' personal data to inform consumers about the kinds of personal data being collected, the reasons for collecting it, and details about any sharing of that data with third parties.[108]

Additionally, the CCPA contains a right to information:[109] 'A consumer has the right to request that a business collecting personal information disclose the categories of personal information collected, the sources from which the information is gathered, the business or commercial purpose for collecting, selling, or sharing the information, the categories of third parties to whom the business discloses personal information, and the specific pieces of personal information it has collected about that consumer.' The business is obliged to provide the specific pieces of personal information obtained from the consumer in a format that is easily understandable to the average consumer, and to the extent technically feasible, in a structured, commonly used, machine-readable format that may also be transmitted to another entity at the consumer's request without hindrance.[110]

(iv) Right to correct inaccurate personal information

The CCPA, as well as other state law[111], also contains a right to correct.[112] A consumer has the right to request a business that maintains inaccurate personal information about the consumer to correct that inaccurate personal information, taking into account the nature of the personal information and the purposes of the processing of the personal information.[113] A business is obliged to correct inaccurate

[102] 740 ILL. COMP. STAT. ANN. 14/15 (West 2022).
[103] 740 ILL. COMP. STAT. ANN. 14/20 (West 2022).
[104] Cal. Civ. Code § 1798.105. See hereto also Li, The California Consumer Privacy Act of 2018: Toughest U.S. Data Privacy Law with Teeth?, Loy. Consumer L. Rev. 32 (2019-2020), 177 (187).
[105] See Cal. Civ. Code § 1798.105 (d).
[106] https://iapp.org/media/pdf/resource_center/State_Comp_Privacy_Law_Chart.pdf.
[107] Chao, Unjust Enrichment: Standing Up For Privacy Rights, Iowa Law Review Online 108(49) 2023, 50 (61).
[108] Cal. Civ. Code § 1798.100(a)(1)–(2) (West 2024); Va. Code Ann. § 59.1-578(C) (West 2023).
[109] Cal. Civ. Code § 1798.110. Consumers' right to know what personal information is being collected. Right to access personal information.
[110] Cal. Civ. Code § 1798.130(a)(3)(B)(iii).
[111] VA. CODE ANN. § 59.1-577(A)(4) (2023) ("To obtain a copy of the consumer's personal data that the consumer previously provided to the controller in a portable and, to the extent technically feasible, readily usable format that allows the consumer to transmit the data to another controller without hindrance, where the processing is carried out by automated means […]")
[112] Cal. Civ. Code § 1798.107.
[113] Cal. Civ. Code § 1798.107(a).

personal information with the use of commercially reasonable efforts.[114] According to *TransUnion LLC v. Ramirez*, violations of the right to correct may only lead to a claim if the consumer suffered a damage (see above).

(v) Right to data portability

In the United States, privacy laws traditionally did not include a right to data portability.[115] However, more recent state laws are beginning to incorporate this right. For instance, the CCPA mandates that businesses provide personal information 'in a structured, commonly used, machine-readable format that may also be transmitted to another entity at the consumer's request without hindrance.'[116] But the right to data portability in the CCPA only applies to information that has been "obtained from the consumer". Similarly, the VCDPA contains a right to data portability regarding the data that the consumer has provided to the controller.[117]

(vi) Right to limit the use/disclosure of sensitive personal information

The California Privacy Rights Act (CPRA) led to a completely new data right in California, namely the Consumers' right to limit use and disclosure of sensitive personal information.[118] A consumer shall have the right, at any time, to direct a business that collects sensitive personal information about the consumer to limit its use of the consumer's sensitive personal information to that use which is necessary to perform the services or provide the goods reasonably expected by an average consumer who requests those goods or services, to perform the services and as authorised by regulation.[119] Similar opt-out rights are provided by VCDPA and the Colorado Privacy Act.[120]

A business that has received direction from a consumer not to use or disclose the consumer's sensitive personal information from using or disclosing the consumer's sensitive personal information for any other purpose after its receipt of the consumer's direction unless the consumer subsequently provides consent for the use or disclosure of the consumer's sensitive personal information for additional purposes.[121]

2. Co-generated output of GenAI

a) Copyright

(i) Right of reproduction

At the output level, Rightholders retain their right of reproduction in the same way as under the training phase. As explained above, according to the U.S. Patent and Trademark Office, AI typically involves reproducing entire works or substantial portions of them. Whether this reproduction constitutes copyright infringement depends on the fair use doctrine, which allows unlicensed use of copyrighted works under certain conditions, such as for teaching or news reporting. If a fair use defence succeeds (which can only be played out in the judiciary setting), a rightsholder whose content was reproduced in a co-generated output cloud successfully claims an infringement to its right of reproduction.

---

[114] Cal. Civ. Code § 1798.107(c).
[115] Solove, The Limitations of Privacy Rights, Notre Dame Law Review 98, 2023, 975 (1006).
[116] Cal. Civ. Code § 1798.130(a)(3)(B)(iii) (West 2023).
[117] VA. CODE ANN. § 59.1-577(A)(4) (2023): "To obtain a copy of the consumer's personal data that the consumer previously provided to the controller in a portable and, to the extent technically feasible, readily usable format that allows the consumer to transmit the data to another controller without hindrance, where the processing is carried out by automated means."
[118] Cal. Civ. Code § 1798.121.
[119] Cal. Civ. Code § 1798.107(a).
[120] VA. CODE ANN. § 59.1-577(A)(5) (2023); COLO. REV. STAT. § 6-1-1306(1)(A) (2023) (effective July 1, 2023).
[121] Cal. Civ. Code § 1798.107(b).

(ii) Right of public display

As opposed to the enforcement of the right of communication to the public in the EU, few cases involve the right of public display in the U.S. In the ongoing litigation between Open AI and the New York Times, the Times submitted (only) twice in its court arguments that part of the generated output provided by Open AI's product display Times content by showing memorised copies or derivatives of Times Work retrieved from the models, and by showing synthetic search results that are substantially similar to Times works.[122] The ongoing debate in Gen AI more generally involves discussions around the right of reproduction but the right of public display could still be relevant.

(iii) Authorship and copyright protection of AI generated outputs

Particularly because it registers copyrighted works, the U.S. Copyright Office engages in a more meaningful discussion on the copyright protection of GenAI outputs compared to the EU.

"A Recent Entrance to Paradise", an AI-generated artwork created by Stephen Thaler using his "Creativity Machine" algorithm set a precedent on the issue of copyrightability of AI generated works. Thaler sought copyright protection for the piece, but the U.S. Copyright Office denied his application, citing the lack of "human authorship," a requirement for copyright eligibility under U.S. law. This ruling was upheld by U.S. District Court Judge Beryl A. Howell, emphasising that works created solely by AI do not meet the necessary criteria for copyright protection

Other productions can result from more complex interactions and entanglements between GenAIand human input, and in practice it can be difficult to identify whether the threshold of human creativity, or free and creative choice, necessary for originality in the copyright sense has been met. We can expect a series of refinements and clarifications by the U.S. Copyright Office, which cannot avoid these questions when registering protected works, as the first public authority to address these issues on a case-by-case basis. The Office started undertaking a consultation and policy work to offer increasing guidance. For instance, current Guidance instructs applicants seeking to register works containing more than *de minimis* AI aggregated material to disclose that the work contains such material and provide an explanation of the human author's contributions.[123]

(iv) Moral rights

In the U.S., moral rights are limited because the copyright system emphasises economic rights over the personal or reputational interests of creators. While the Visual Artists Rights Act (VARA Act) provides limited moral rights, specifically the rights to attribution and integrity, these protections apply only to certain visual arts and do not extend to other creative fields like literature, music, or film. Additionally, the "work-for-hire" doctrine in the U.S. further limits moral rights because works created within employment are considered the property of the employer, not the individual creator, effectively eliminating moral rights for many creative works produced in professional contexts. In the absence of broad moral rights, creators in the U.S. often rely on contracts to protect their personal interests, such as securing attribution rights, remuneration rights, or limiting alterations to their work. However, these contractual agreements are private and not the same as legally recognised moral rights hence unlike in the EU, successful claims of infringements to authors' moral rights of a GenAI output are not likely to occur unless in the context of the protection afforded by the VARA Act.

[122] United States District Court Southern District of New York, Case 1:23-cv-11195 Document 1 Filed 12/27/23.
[123] USA gov. Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence, March 2023

b) Data protection

Preliminary Remarks. Data subjects have the same rights with regard to the output as with regard to the data input when the model is able to "remember" the personal data: the right to erasure, to information, to access, to rectification, to data portability, to object and not to automated decision-making. In this respect, reference can generally be made to the above explanations (B.II.1.b)). In the following, it will only be discussed if additional aspects deviating from the above explanations need to be considered.

(i) Right to not be subject to automated decisions

The CCPA provides consumers with rights to opt out of automated decision-making, to learn about the algorithmic logic involved, and to know about the likely outcome. The Act mandates the Attorney General to issue regulations that businesses disclose "meaningful information about the logic involved in those decision-making processes, as well as a description of the likely outcome of the process with respect to the consumer".[124]

[124] Cal. Civ. Code § 1798.185(a)(16) (West 2023).

3. Interim conclusion

Copyright.

Input data used in AI training in the U.S. might be protected if the data or content is original. The scope of the right of reproduction under the U.S. Copyright Act is also similar to that of the EU. Notably, the U.S. Patent and Trademark Office, which issued a public communication on AI and IP challenges, has noted that AI training involves reproducing significant portions of copyrighted works. Whether this constitutes infringement depends on the "fair use" doctrine, which allows some unlicensed uses of copyrighted works for purposes like research and education.

Unlike the EU, the U.S. has no specific exceptions for text and data mining but uses the fair use doctrine to balance innovation and copyright protection. The successful application of fair use relies on factors such as the purpose of the use (commercial vs. non-commercial), the nature of the work, the amount used, and the impact on the market. The pertinence of fair use is disputed in the context of GenAI, and court decisions are still evolving. Past court cases, like the Google Books case, found transformative uses of copyrighted works to be fair use, but the use of copyrighted content for training AI tools may present more challenges due to the economic value of generated outputs and the far-reaching substitution effect (the output appearing as a good substitute for the works used for training).

With regard to remuneration, the U.S. copyright law doesn't mandate automatic compensation for copyright holders. Many U.S.-based AI companies negotiate licensing agreements to access content for training. The agreements signed by major players like OpenAI with publishers such as The Associated Press and Le Monde illustrate how the market is rapidly adapting and evolving in response to the growing demand for data to train AI models and the requests from rightholders for fair compensation for the exploitation of their copyrighted content. In contrast, the EU system includes more and stronger remuneration rights.

During the output stage, rights holders retain their reproduction rights as they do during the training phase. Given that AI often reproduces entire works or substantial parts, whether such reproduction constitutes infringement relies on the fair use doctrine. This doctrine, applicable only in judicial settings, allows for unlicensed use of copyrighted works under specific circumstances, like teaching or news reporting. If a fair use defence fails, a rightsholder may successfully claim reproduction rights infringement for content included in a co-generated output.

While in the EU the right of communication to the public is often relied on for digital forms of exploitation, fewer cases address the scope of the right of public display in the U.S. In litigation involving OpenAI and *The New York Times*, the latter has argued that OpenAI's outputs display Times content by reproducing memorised copies or derivatives of its works. While discussions relating to GenAI often focus on the reproduction rights, the public display right may become increasingly relevant in relation to the output (when substantially similar).

On authorship and copyright protection, the U.S. Copyright Office has clarified that works must show some human creativity to qualify for copyright protection, as affirmed in the *Recent Entrance to Paradise* case. However, AI-generated works involving complex human and AI interactions challenge the traditional concept of originality in copyright. As such, the Copyright Office has begun providing guidance requiring applicants to disclose AI contributions, indicating that further clarification will likely follow. The report of the U.S. Copyright Office on the copyright issues linked to the training of GenAI models is expected for the end of 2024 and hopefully will clarify some issues.

Data Protection

The U.S. lacks a unified data protection framework, unlike the EU's GDPR, and instead follows a sectoral approach, with regulations varying by industry, data type, and jurisdiction. Key federal laws include the FTC Act, HIPAA, COPPA, and others. States like California have led with broader consumer rights through the CCPA and CPRA, which have inspired similar laws in other states like Virginia.

Under these laws, rights concerning unlawful data use are addressed by both the CCPA and the VCDPA, which do not govern publicly available data. Consumers can opt out of data sharing, and businesses must inform them of data collection plans. In Virginia, the collection of sensitive personal data requires consumer consent, while the CCPA only mandates notification for processing such data.

Consumers always have the right to request the deletion of personal data under the CCPA, though there are exceptions. Additionally, consumers can request details on the types of personal data being collected, the purposes for collecting it, and information on sharing with third parties. This right to information under the CCPA allows consumers to know how their data is being used.

Consumers also have the right to correct any inaccurate personal information. If businesses maintain incorrect data, they must take commercially reasonable efforts to correct it upon request. Moreover, the right to data portability is included in newer privacy laws, such as the CCPA and VCDPA, allowing consumers to request their personal data in a structured, machine-readable format that can be transferred to another entity.

The California Privacy Rights Act (CPRA) introduces a right to limit the use of sensitive personal information. Consumers can direct businesses to restrict its use for necessary services or as permitted by regulation. This right is mirrored in other state laws like the VCDPA and the Colorado Privacy Act.

These rights also apply to the co-generated output of GenAI Models if the Model is able to remember the input. In addition, the CCPA provides consumers with rights to opt out of automated decision-making, to learn about the algorithmic logic involved, and to know about the likely outcome.

III. Japan

1. Co-generated input of GenAI

a) Copyright

On the global stage, Japan ascended to the rank of a “machine learning paradise” after its 2018 amendment to the Copyright Act. The legislative reform introduced an exception for computer data analysis intended to cover text and data mining activities, which, at first glance, appears to be one of the most permissive provisions compared to the EU and U.S. frameworks. The Japanese 2018 reform was first initiated by an IP strategy in 2016 which, quite early on, aimed at promoting Intellectual Property Innovation for the 4th Industrial Revolution by constructing a *Next Generation IP System adapted to Digitization and Networking*.[125] The intervention significantly emphasises the fair balance between intellectual property and competitiveness, as well as the need to redefine an intellectual property system in line with innovation goals. It also mentioned that numerous mechanisms such as “uncompensated rights limitations, rights limitations with attached remuneration rights, copyright collective licensing, and compulsory licensing of orphaned works” could be used to promote a "graded approach" towards finding flexible solutions to contribute to the creation of appealing content from Japan. Whilst the TDM exception introduced into Japanese law was considered almost unrestrictive to achieve the mentioned goals, the Agency for Cultural Affairs in charge of handling the Copyright Office, issued a Report on AI and Copyright in March 2024

[125] Intellectual Property Strategic Program 2016, Overview of the "Intellectual Property Strategic Program 2016" (Approved on May 9, 2016 by the Intellectual Property Strategy Headquarters).

clarifying numerous aspects regarding the scope of the exception.[126] Essentially, not all uses of copyrighted works in machine learning would fall under the exception, introducing (non-binding) limitations by means of use cases which raise ambiguities and somewhat temper the pro-innovation approach promoted in 2016. The Japanese TDM exception could still be considered broader than the EU's, as it does not explicitly provide for an opt-out mechanism. However, it does not prohibit contractual overrides, meaning in principle, copyright holders may still impose terms to exclude their works from the exception. In contrast, under the EU's TDM regime, copyright holders can directly contract out of the exception, allowing them to limit its applicability through specific contractual terms.

(i) Right of reproduction

As explained above under the EU and USA frameworks, the right of reproduction can only be triggered if a work is protected by a copyright in the first place - the same applies under Japanese rules. Works of authorship which enjoy protection under the Japanese Copyright Law are defined in the law as "production in which thoughts or sentiments are expressed in a creative way and which falls within the literary, scientific, artistic or musical domain." Japan is party to the Berne convention and to the main international treaties and on a general level, hence the protection afforded under Japanese law is similar to the one under EU and U.S. laws. A work is eligible to copyright protection when it includes "thoughts or sentiments", that are "expressed", "creative" and in the literary, scientific, artistic or musical domain. Should these four conditions be met, a work can be protected.

Under Japanese law, the right of reproduction is defined as an exclusive right of the author.[127] Individuals have a right to oppose the reproduction of their works unless the TDM exception applies.

(ii) The Japanese TDM exception

The Japanese TDM exception is located under Article 30-4 (ii) of the Japanese Copyright Law. It dispenses the need to obtain consent from copyright holders for "exploitation for using the work in a data analysis (meaning the extraction, comparison, classification, or other statistical analysis of language, sound, or image data, or other elements of which a large number of works or a large volume of data is composed, provided that:

1. the exploitation is aimed neither at enjoying nor causing another person to enjoy the work,
2. It does not prejudice the interest of the rights holder.

Because of this last criterion, the provision is said to belong to the 'flexible copyright exception for "non-enjoyment" purposes' (see below on the notion of enjoyment).

As summarised by Japanese scholar Tatsuhiro Ueno, Article 30-4 outlines three specific examples of uses that are not intended for "enjoyment" purposes, including text and data mining (the other two pertain to experiments and the processing of works that do not render them perceivable by the public). Since these three scenarios, including text and data mining, are merely examples in a non-exhaustive list, Article 30-4 serves as a broad provision. It covers additional unspecified exploitation of works, as long as it would not result in public enjoyment of the copyrighted work and unless the exploitation unreasonably prejudices the interests of the copyright holder. This means that the exception is designed to be adaptable and aimed at covering many scenarios of data scraping and acts of exploitation.

[126] AI と著作権に関する考え方について, freely translated as AI and Copyright Concepts March 15, 2024 Subcommittee on Legal System, Copyright Subcommittee, Council for Cultural Affairs.
[127] Article 21: "The author of a work has the exclusive right to reproduce the work"

The Agency for Cultural Affairs emphasises that the Copyright Act seeks to balance the interests of rights holders while facilitating the exploitation of works, particularly in the context of AI and copyright.[128] This approach contrasts with the EU's copyright framework, which focuses primarily on ensuring a high level of protection for rights holders.[129]

Use of work as works for (non)enjoyment. The Japanese exception distinguishes itself from the U.S. and EU mechanisms with this criterion : the exploitation of a copyrighted work for purposes other than the enjoyment of its artistic or literary expressions—such as in AI development or data analysis—may, in principle, be permissible without the copyright holder's permission.

According to a Japanese Copyright Office's interpreting document titled "General understanding on AI and Copyright in Japan Overview" ("the Overview"), published in May 29 2024, "enjoyment" under the Article 30-4 refers to "the act of obtaining the benefit of having the viewer's intellectual and emotional needs satisfied through using the copyrighted work."[130] Therefore when a work is used for text and data mining, permission from the copyright holder is required if the use also involves a "purpose of enjoyment" of the work." This requirement is controversial, and some uncertainty remains as to how Japanese Courts will enforce it in practice, especially in relation to cases where the use of works could be multi-purpose : some acts of training could be for enjoyment and some could not, which could lead to a selective application of the exception for certain fragments of the act of exploitation. For now, multi-purpose uses will not benefit from the exception according to the Overview.

Further, according to the Report on AI and Copyright, acts that are not intended to provide enjoyment of ideas or emotions expressed in a work do not, in principle, impair the opportunity for rights holders to obtain utility from their copyrights. In other words, the economic value of a work is realised when an individual views or listens to it and provides compensation to experience the ideas or emotions conveyed. This "enjoyment" centric approach is closely aligned with the concept that copyright law has an internal limit, aligned with the idea that the ability to enjoy or enable others to enjoy a work is central to the concept of infringement. This approach reflects the broader notion that copyright law operates within an internal limit, where reproduction rights are framed not just in terms of ownership but in relation to the public's ability to experience the work.The Overview published in May 2024 exemplifies this doctrine by providing use cases. For instance, the collection of works as input data to GenAIfor implementation of retrieval augmented generation aims to output the creative expression as a copyrighted inputted work, hence, it cannot benefit from the non-enjoyment/TDM exception. These views may evolve as the landscape and stakeholders' discussions could significantly impact how those issues are addressed moving forward.

Importantly, the Overview also clarifies that the exception may not extend to AI training that 'picks off' a specific creator's style. This is significant because, as mentioned earlier, styles are not protected by copyright, and the EU has yet to take a position on this sensitive issue. If Japanese judges were to adhere to the Overview and prohibit AI-generated works that mimic a particular style, it could lead to a broader scope of copyright protection for those specific creations compared to the EU, assuming the doctrines regarding styles and ideas remain unchanged. This could mean that AI-generated content in Japan would face stricter limitations, enhancing the protection afforded to original works.

[128] See General Understanding on AI and Copyright in Japan" -Overview- (published by the Legal Subcommittee under the Copyright Subdivision of the Cultural Council) in May 2024.
[129] Although some examples of balancing copyright with other objectives exist in the case law of the European Court of Justice, it cannot be excluded that the CJEU would apply proportionality and fair balance in potential TDM cases.
[130] ibid 108.

Another uncertain aspect of the Overview is that it suggests the collection[131] of works used for AI training to produce outputs similar to the copyrighted works—such as fine-tuning or overfitting—would not satisfy the "non-enjoyment purpose" requirement. In practice, this may pose issues regarding the actual scope of the exception.

Use of work without prejudicing the interests of the rights holders. Another limitation of the exception is that it cannot apply if it unreasonably prejudices the interests of the copyright owner, i.e.: when the training is intended to compete in the market of the copyrighted work. The Overview gives as an example the reproduction of a database formatted for information analysis without compensation, despite being available online for a fee. Within the Subcommittee in charge of copyright at the Agency for Cultural Affairs, some have raised concerns that training a generative AI exclusively on a specific creator's works may result in outputs that, while only similar in style, could nonetheless affect the same market as the original works and, as a result, harm the creator's interests. They emphasise that any copyright exception must align with the three-step test, noting that Japan's TDM exception (Article 30-4) includes a provision specifying that this exception does not apply if the use "unreasonably prejudices" the copyright owner's interests. An alternative view is that Article 30-4 may still apply, even when the output would be competing with the original works in the same market, because it would not necessarily "unreasonably prejudice" the creator's interests that are protected by copyright which *stricto sensu* only protects original expressions but not style. In sum, the views on this issue are diverse, reflecting the need to cautiously strike, over time and on a case-by-case basis, the complex balance between protecting creators' interests and accommodating advances in AI.

Additional remarks on the source of the training content. In contrast to the EU's TDM exception, the Japanese exception does not clarify whether the content used for testing, analysing, or processing must be licensed or legally obtained. This omission could potentially allow for the use of pirated content to train AI systems. The Agency for Cultural Affairs clarified in its Report that knowingly collecting data from a site hosting pirated content increases the risk that AI developers or service providers could be held liable for infringement, as it would breach their general duty of care. This may suggest that Japanese courts are likely to interpret the TDM exception similarly to the EU, where it is well-established that content used for TDM must be legally obtained.

b) Data protection

Preliminary Remarks. Japan's legal framework for data protection is primarily governed by the Act on the Protection of Personal Information (APPI). Enacted in 2003 and significantly amended in 2017 and 2020, the APPI sets out rules for the collection, use, and management of personal data by businesses and organisations.

The APPI applies to personal information which is defined as information containing a name, date of birth, or other identifier or the equivalent (meaning all items (excluding individual identification codes) made by writing, recording, sound or motion, or other means, in a document, drawing, or electronic or magnetic record (this includes a record created in electronic or magnetic form (meaning electronic form, magnetic form, or any other form that cannot be perceived with the human senses; the same applies in item (ii) of the following paragraph); hereinafter the same); hereinafter the same) which can be used to identify a specific

[131] The three-step test in the Berne Convention limits exceptions to copyright by ensuring they apply only in specific cases, do not conflict with the normal exploitation of the work, and do not harm the legitimate interests of the copyright holder. This ensures a balance between protecting the rights of authors and allowing for certain uses of copyrighted works under strict conditions.

individual (this includes any information that can be easily collated with other information and thereby used to identify that specific individual); and information containing an individual identification code.[132]

(i) Rights arising from the wrongfulness of data processing

Under the APPI, personal information may only be used for the purposes stated at the time of collection, and any change in those purposes requires the individual's consent.[133] However, there is a procedure to change the purpose without consent. When it comes to sensitive data (like medical records) the collection is only lawful, when the data subjects declared their consent to the processing.

When business organisations acting as data controllers use GenAI, any personal data input into the AI must align with the purposes previously communicated or disclosed to the data subjects. It is currently disputed, whether using personal information for AI training also requires the notice, because it could be qualified as a statistical analysis which is excepted from the notice requirements. However, pseudonymised personal information (data processed in such a way that an individual cannot be identified unless combined with other information) allows for the purposes of use of collected personal information to be altered without requiring the individual's consent, facilitating the use of such data in AI machine learning.

If a business processes data without the consent of the identifiable person[134], in a way that there is a possibility of fomenting or inducing unlawful or unjust act,[135] or if the business acquired the personal information by deception or other wrongful means,[136] the use of personal information is prohibited by Article 19 APPI. Further, the identifiable person may request the business to cease to use or delete the personal information.[137]

Currently, enforcement of the APPI primarily relies on administrative guidance and recommendations from the PPC, with formal orders being extremely rare. As a general rule, criminal penalties are only applied in cases of violation of PPC orders.

According to the most recent review of the APPI, the Personal Information Protection Commission published an Interim Summary, outlining its current thinking based on discussions and examinations to date,[138] a revision of the APPI is on the horizon. The Commission inter alia considered strengthening enforcement by establishing a new system of injunctive relief and restoration of damages by organisations. The Commission also highlights an increasing societal need for businesses and services to utilise personal information without the consent of data subjects in specific contexts. Examples include the use of large data sets for GenAI, applications in healthcare, education, disaster prevention, child protection, and fraud prevention. It notes that the current exceptions under the APPI are insufficient to meet these evolving demands. Therefore, the establishment of new exceptions should be explored, balancing the level of societal need and public interest with the protection of individual rights and interests. Further discussions with businesses and relevant government authorities are anticipated to shape these new exceptions and guidelines.

(ii) Right to correct

Individuals have the right to rectify or delete incorrect personal data under Article 34 APPI: “If the content of personal data a business holds that can be used to identify the identifiable person is not factual, the person

[132] Art. 2(1) APPI.
[133] Art. 17(1) and Art. 18(1) APPI.
[134] Art. 18 APPI.
[135] Art. 19 APPI.
[136] Art. 20 APPI.
[137] Art. 35(1) APPI.
[138] https://www.ppc.go.jp/files/pdf/240626_shiryou-1syuuseigo.pdf.

may request that the business handling personal information make a correction, addition, or deletion (hereinafter referred to as a "correction" in this Article) on the content of the personal data the business holds." If a business has made the correction, it is under the duty to notify the identifiable person to that effect without delay.[139]

2. Co-generated output of GenAI

a) Copyright

(i) Right of reproduction and of communication to the public

Controversies arose with regard to the scope of the TDM exception concerning the generation phase.

The Report published in March 2024 clarified that the non-enjoyment exception does not cover the generation and utilisation stage of GenAI. In GenAI cases, the Report suggests examining whether there exists substantial similarity and reliance, like all copyright infringement cases, on a case-by-case basis.

Under Japanese law, to determine copyright infringement to the right of reproduction, two elements must be proven: "similarity" and "reliance" on an existing copyrighted work. "Similarity" can exist even if only certain key aspects of the copyrighted work appear in the new output. "Reliance" occurs when the creator of the new work is aware of the original copyrighted material.

The Report clarifies that if an AI user is unaware of an existing copyrighted work (and its expressive content), and the GenAI has not been trained on that copyrighted work during its development and learning phase, then even if the AI generates something similar to that copyrighted work, this would be considered a coincidental match.[140] Therefore, reliance is not recognised, and copyright infringement does not occur. The May 2024 Overview provides that when an AI user explicitly asks the AI to generate a work based on a specific copyrighted piece, such as by referencing it in a prompt, reliance is established, and the AI user could be liable for infringement.[141] However, if the AI user unknowingly generates infringing content, responsibility may shift to the AI developer or service provider, particularly if the tool frequently produces infringing outputs. To mitigate this risk, AI developers and service providers are encouraged to implement technical safeguards that prevent copyrighted content from being reproduced. They are also advised to introduce measures to stop users from requesting infringing material.

Both the Report and the Overview constitute a soft-law framework, meaning they are not legally binding but serve as a set of recommendations for interpreting copyright issues related to AI-generated works. Despite their non-binding nature, both documents are notably detailed and foresee a series of scenarios that will realistically be the object of litigation, offering judges of future cases with at least a first line of reasoning.

(ii) Authorship of AI-generated outputs

The 2016 IP Policy Strategy mentioned above was maybe one of the first to have directly recognised that the aggregation of "individually worthless data can also create new value". The Japanese copyright experts already highlighted in the 2016 Strategy that AI was already being used to create original works, such as music, logos, and stories, and predicted that AI would play a major role in driving cultural and technological innovation. Unlike in the EU, where the focus has traditionally been on safeguarding the rights of human

[139] Art. 34(3) APPI.
[140] AI と著作権に関する考え方について freely translated as AI and Copyright Concepts March 15、 2024.
[141] General Understanding on AI and Copyright in Japan" -Overview- (published by the Legal Subcommittee under the Copyright Subdivision of the Cultural Council) in May 2024.

creators, Japan recognised very early on the need to consider how AI-generated works could be accommodated within the existing IP system, noting that a blanket protection of all AI generated works would be excessive, but that the IP system should support this development.

In 2024, the Report issued by the Copyright Office provides for a more nuanced and detailed approach. It provides that the copyrightability of AI generated works will be determined on a case-by-case basis, on the basis of two elements.

First, creative contribution which may be determined with factors such as the amount of instructions, the number of generation attempts of an output, or the selection from multiple output materials which may involve creative choices could help determine whether AI generated outputs could be considered copyrightable. In sum, this approach appears to still give significant weight to the role of human creativity and the emerging role of the 'prompt engineer', recognising that human input and influence over a final product will still influence whether an AI generated work is considered copyrightable.This echoes the way photographers were recognised authorship of copyrighted photos when cameras first emerged as a creative medium. Although the question was hotly debated during the early days of photography, over time the creative choices behind the medium came to be widely recognised as deserving of copyright protection.[142]

Second, the Overview clarifies that if a "creative intention" can be identified in a person using AI as a tool to "creatively express thoughts or sentiments", such material can be considered a "work" and the use of the AI the "author" in the meaning of copyright law.[143]

(iii) Right of remuneration

Contrary to the position suggested in the 2016 Intellectual Property Strategy, the Japanese government explains in the Report that even commercial uses of text and data mining (TDM) should not automatically be subject to exclusive rights or remuneration rights. This is because, from the viewpoint of compensating copyright holders, the use of copyrighted works for information analysis in AI development does not normally harm the interests of copyright holders protected under the Copyright Law. The government finds it difficult to justify the introduction of a compensation system under the Copyright Act, as such uses do not typically undermine the copyright holders' opportunities for compensation. However, in light of the broader goal of promoting a virtuous circle of content creation, the Report acknowledges that discussions around compensation may extend beyond the Copyright Law framework, potentially involving market-based mechanisms and technical solutions to ensure fair compensation.[144]

b) Data protection

In Japan, the major discussion around data protection and GenAI was on the use of GenAI by businesses. This is also reflected by the PPC's "Cautionary Notes on the Use of Generative AI Services" (June 2023)[145] which outline the following points of caution for businesses: The transfer of personal data generally requires consent, but the APPI allows for the transfer of personal data without consent when a personal data holder outsources the utilisation of personal data to a third-party company, such as a GenAI provider. Therefore, transferring personal data to GenAI providers without consent is legal under APIP. However, the APPI also regulates the transfer of personal data to foreign entities. This means that to transfer data abroad, a

[142] ibid.
[143] ibid.
[144] Ueno, Tatsuhiro. "The Flexible Copyright Exception for 'Non-Enjoyment." Purposes—Recent Amendment in Japan and Its Implication" (GRUR International 2021) at 147 (2020).
[145] https://www.ppc.go.jp/files/pdf/generativeAI_notice_leaflet2023.pdf.

business must either obtain explicit consent from the data subject, establish a contract with the GenAI provider that includes provisions for the protection of personal data, or transfer the data to a country within the European Union or the United Kingdom. It's important to note that these regulations on transferring data to foreign entities may not apply if the personal information is used for purposes other than responding to the prompt, as this may not be considered a "transfer" under APPI.

Regarding the rights arising from wrongful processing of personal information as well as the right to correct see above (B.III.1.b)).

3. Interim conclusion

Copyright. Japan's 2018 amendment to the Copyright Act introduced an exception for text and data mining (TDM), facilitating the use of copyrighted works for data analysis without requiring consent from rights holders, provided the use does not aim for enjoyment and does not unreasonably prejudice the interests of the copyright owner. The TDM exception is outlined in Article 30-4 of the Copyright Law and applies to various non-enjoyment purposes, including text and data mining, experimentation, and certain forms of processing works.

The TDM exception in Japanese copyright law does not extend to the generation and utilisation stages of generative AI, meaning that outputs produced by AI systems are not covered by this exception—similar to the European Union's exception, which applies only at the training level. As a result, outputs generated by AI systems are considered works under the traditional copyright regime and may therefore be liable to infringe existing copyrights. The determination of copyright infringement for AI-generated outputs in Japan is based on two key elements: "similarity" to existing copyrighted works and "reliance" on those works. If an AI user is unaware of a copyrighted work, and the AI has not been trained on it, any generated similarity would be considered coincidental and not infringing.

Regarding the copyrightability of AI-generated works, a 2024 Report from the Copyright Office indicates that a work may be considered copyrightable if it involves creative contributions from a human user, such as the level of instruction provided or the selection of generated outputs that reflect creative choices. If a user's creative intention is evident in their use of AI to express thoughts or sentiments, the resulting material can be classified as a "work" under copyright law. This nuanced approach acknowledges the emerging role of the "prompt engineer" while ensuring that the traditional concept of authorship remains tied to human input and influence.

Regarding remuneration, the current Japanese view provides that commercial uses of text and data mining may in principle not be subject to exclusive rights or remuneration rights because uses outside of enjoyment purposes do not undermine copyright holders' opportunities for compensation. This stance represents an initial indication of the government's approach to remuneration in the context of AI and copyright, though it may evolve over time and does not necessarily reflect the views of all stakeholders in the field.

Data Protection. Under the APPI, the use of personal information - also sensitive personal information - generally requires the consent of the data subjects. If the data subjects have to be informed of the use of the data for training GenAI, is still disputed. Breaches of the APPI lead to administrative guidelines and recommendations. However, the Personal Information Protection Commission is currently considering the introduction of a new system of injunctive relief and restoration of damages by organisations in its recent review of the APPI.

C. Conclusions

Looking at the legal framework for rights in co-generated input and co-generated output, the aim of the conclusions is to identify common features of both the copyright and data protection frameworks in the EU, US and Japan. Due to the completely different objectives and policy considerations behind data protection law and copyright law, these two areas will be treated separately here. Copyright primarily aims to incentivise creativity and innovation by granting creators exclusive rights to their works, fostering a rich cultural and intellectual environment. It focuses on protecting the expression of ideas, not the ideas themselves. Copyright, at least in the Continental European tradition of author's rights, also includes moral rights (such as the right to claim authorship and the right to integrity). Those non-pecuniary rights aim at enhancing the autonomy and personal development of creative individuals. Data protection, on the other hand, centres on safeguarding individual autonomy and privacy by regulating the processing of personal data. It aims to prevent harm and empower individuals with control over their information, recognising the potential for misuse and exploitation of personal data in the digital age. These differing objectives necessitate distinct legal frameworks with varying mechanisms for enforcement, exceptions, and limitations.

I. Rights with regard to the co-generated input of GenAI

1. Right to an economic share

Copyright. Rights to have an economic share in the Co-Generated Input of GenAI are still under discussion with regard to copyright law. The right to oppose the training does not exist when there is an exception to copyright (and thus a legally authorised use of the protected data or content) whether it is fair use (as in the U.S.) or exempted text and data mining (in the EU and Japan). The exclusive right of reproduction (which appears somewhat similar to the right of opposition to the processing under data protection) can be assigned by contract.

There is an ongoing lively discussion in certain European countries (not at EU level and, to our knowledge, not in the U.S. or in Japan) on whether a compulsory licence with a corresponding remuneration right or another system should be introduced when protected data is scraped on the internet for training AI models.[146]

Data protection. When looking at data protection law, there is no similar discussion on a right to have an economic share as a result of the co-generation during the training phase. However, it is still possible for the data subjects to consent to the data processing and have a "counter-performance" as a counterpart to their consent. But this will not be feasible in practice when the data is scraped from the internet and the processing is based on the legitimate interests of the controller in the EU or is rightful since the data is publicly available in the U.S. (especially under the CCPA and the VCDPA). Even where processing of publicly available data can only be based on consent, e.g. in Japan, there is no discussion around remuneration; this is probably because the consent is given in advance and there is no room for any negotiations of remuneration.

2. Rights against unlawful use

[146] See for ex. Senftleben, Martin and Izyumenko, Elena, Author Remuneration in the Streaming Age - Exploitation Rights and Fair Remuneration Rules in the EU (October 9, 2024). Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4981352 .

Secondly, rights may arise from the unlawful use of the protected data. Unlawfulness can stem from breaches of contractual but also statutory provisions. Under the latter, the legal treatment of publicly available data is a key commonality between both copyright law and data protection law.

Copyright. Under copyright law, the fact that the protected data or content is publicly available does not grant any right to use it. There is no general principle under copyright law according to which the protected content would be freely reusable because it is public: indeed, no implied licence can be automatically derived from the mere fact that the content is accessible online for crawling or scraping. However, the will of the right holder to disseminate some content online might under certain circumstances be interpreted as a waiver. Under certain exceptions, such as the TDM exception (EU) or the fair use rule (U.S.), protected data can be reproduced and adapted for training an AI tool. Both legal frameworks recognise that publicly accessible data holds a different legal status, reflecting the balance between public interest, innovation, and individual rights.

Data protection. Under data protection law, publicly available data often falls outside the scope of strict privacy protections or serves as a justification for lawful processing, as its accessibility reduces the potential privacy risks typically associated with personal data.

While the right to object is a separate right of data subjects, also within the GDPR, it shall not be seen as a separate right within this study. This is because the outcome of the right to object is that the processor may lack a ground for lawfulness for the processing of personal data, either because the data subject withdrew its consent or it objected to the processing based on legitimate interest and the processor could not put forward an overriding legitimate interest. If the processor, nevertheless, processes the data without a legal ground, rights from unlawful use arise.

The rights arising from unlawful use are the erasure of the relevant content and compensation in the form of damages or unjust enrichment. When it comes to the right to erasure it is still under discussion whether this right should lead to the erasure of the input used for training the AI model, with the consequence that the AI model would have to be (partly) retrained. When looking at the current legal framework (in the EU) the processor has to erase the personal data used by the AI model or its ongoing use of the personal data is wrongful. A strict application of the right of erasure might thus have a considerable (and some would say disproportionate) impact on the use and the deployment of GenAI tools.

3. Rights to information, to rectification and to data portability

Data protection. The rights to information, to rectification and to data portability only exist as such in data protection laws (the right to portability is however embedded in the copyright principle according to which the economic rights on the protected item, and thus the protected item itself, can be assigned or licensed). The main difference between the right to information and the right to data portability is that the right to information only aims at informing the data subject on the processing of personal data while the right to data portability may also enable the data subject to receive or share the personal data in a machine-readable format.

However, the practical relevance of the right to data portability could be undermined if the interpretation of the term ‘provided data’ leads to the conclusion that it only covers data that has been actively transmitted, and thus observed data would be excluded from its scope. In addition, some balancing of interests is required if the interests of other persons are affected by the exercise of the right to data portability.

The right to rectification offers contributors the opportunity to rectify errors or inaccuracies in the co-generated model, content, or data, ensuring that their contributions are accurately reflected and remain

valid. The right to information grants users the possibility to know if data is processed and helps data subjects to assert other rights. However, the right to information does not apply if the provision of such information proves impossible or would involve a disproportionate effort.

II. Rights with regard to the co-generated output of GenAI

1. Rights to Have an Economic Share, to Information, to Rectification and to Data Portability

The Rights in Co-Generated Input of GenAI should apply *mutatis mutandis* to Co-Generated Output when the Co-Generated Output has manifestly been generated with the Co-Generated Input and the Co-Generated Output presents substantial similarities with the Co-Generated Input.

Copyright. With regard to copyright, the fact that the protected item might appear in the output (and might thus be made available to the user of the GenAI tool) has some consequences as it triggers the application of another right within the copyright bundle, the right to communication to the public (under EU law) or the right of public display (under U.S. law). Also, the exception of TDM (in the EU) that exempts some reproductions is not applicable to the output phase.

Data Protection. Data protection law will likely apply if a GenAI system can “remember” and subsequently process input data in a manner that allows for the identification of individuals. Where the GenAI system retains input data and uses it to generate outputs that reveal information about individuals, or if the input data itself can be linked back to specific individuals, then such processing falls under the purview of data protection law.

2. Right to explanation

Data protection. The right to explanation becomes crucial, especially in complex data-driven models, where contributors may demand a clear understanding of how their input was used and how the co-generated asset functions, which is particularly relevant in contexts such as AI or machine learning models.

III. Summary Table

The following table aims at summarising key differences and characteristics of the three jurisdictions under scrutiny in this report.

Jurisdiction	Co-generated input	Co-generated output	Key legal frameworks & considerations
European Union	**Copyright**		
	Subject to reproduction rights. TDM is allowed under the exception for text and data mining (CDSM directive), unless rightsholders opted-out. Licensing with remuneration is an option in case opt-out.	Moral rights persist (e.g.: integrity, attribution). The TDM exception does not cover outputs, creating potential infringement risks due to similarity with existing works. Currently, there is no output-based remuneration for now.	The TDM exception requires lawful data acquisition prior to the performance of TDM activities. The valid exercise of opt-out, on a technical level, remains a debated issue.
	Data protection		
	GDPR requires a legal basis for the processing of the personal data (e.g., consent, legitimate interest). Rights include access, erasure, rectification, objection and data portability. Some key issues, such as the erasure of the whole GenAI model are still disputed.	GDPR rights (e.g., access, erasure) apply to personal data in outputs. Additionally, the right not to be subject to automated decision making applies. Outputs tied to training data must meet lawful processing standards.	GDPR always requires a legal basis for the processing of the personal data. The GDPR grants rights in co-generated in- and output. The question of erasure of the model when the data has wrongfully processed is still disputed..
	Copyright		

United States	Subject to reproduction rights. There is no statutory exception for TDM; the fair use doctrine applies on a case-by-case basis. Many litigation cases are currently pending.	Moral rights persist (e.g., integrity, attribution). TDM exceptions do not extend to outputs, creating risks of infringement based on similarity. No framework for output-based remuneration exists.	Ongoing debates about the copyrightability of AI-generated works are taking place: the U.S. Copyright Office requires some level of human input for works to qualify for protection.
	Data protection		
	The U.S. lacks a unified data protection framework. The sector specific regulations do not govern publicly available data (CCPA, VCDPA). Consumers have the right to request the deletion, correct, limit the use of personal information.	The rights also apply to the co-generated output of GenAI Models if the Model is able to remember the input. In addition : right to opt out of automated decision-making, to learn about the algorithmic logic involved, and to know about the likely outcome (CCPA)	Rights in co-generated input also apply to the output if the model is able to remember. Specific rights for the output.
	Copyright		

Japan	Subject to reproduction rights. TDM is permitted under the "non-enjoyment" exception. While opt-outs are not explicitly provided in the law, they can be contractually enforced.	Infringement may occur if specific conditions are met, such as competing in the same market in a way that harms the rights holder's interests.	The Japanese Copyright Office is actively engaged in discussions on AI-assisted creations and dual-purpose uses (e.g., enjoyment and non-enjoyment). Public guidance is available for specific scenarios where double purpose usage of inputs occurs (i.e.: technical data analysis and possible enjoyment purposes).
	Data protection		
	Under the APPI, the use of personal information - also sensitive personal information - generally requires the consent of the data subjects. The individuals have rights arising from the wrongfulness of data processing and the right to correct.	The rights in co-generated input also apply to the co-generated output.	In Japan, the major discussion around data protection and GenAI was on the use of GenAI by businesses.

IV. Concluding remarks

Legal landscapes for dealing with cogenerated data and technologies, especially generative AI models, vary across the EU, US, and Japan. Copyright and data rights shape this evolving field, as each jurisdiction struggles to balance regulation with innovation in the rapidly advancing field of generative artificial intelligence.

Our analysis of the legal aspects of using data and content in generative AI tools reveals the complex interplay between copyright, personal data laws, and the broader context of human rights, democracy, and economic welfare. By dissecting the generative AI lifecycle and differentiating between co-generated input and output, we have aimed to illuminate the multifaceted challenges of balancing intellectual property protection with the need to foster innovation. We have also dealt with the two main regulatory frameworks that apply to cogenerated data and technologies:

- **Copyright:** While copyright laws don't protect data, they become relevant when data is compiled into potentially protectable works. The use of copyrighted works for AI training raises complex questions around reproduction rights, exceptions for text and data mining, and fair use. The copyrightability of AI-generated outputs is also debated, with varying approaches to human authorship and creative contribution.

- **Data protection:** Rights to information and rights arising from the unlawful use of data (such as the right to erasure) are crucial, but their implementation faces challenges when AI models "remember" data. Publicly available data often enjoys less protection, but that data could also include personal data, raising concerns about its use in AI training.

Our overview of regulations in jurisdictions likely to shape AI development globally demonstrates that a holistic approach to AI governance is crucial. Striking a balance between safeguarding rights and promoting technological advancement requires careful consideration of the ethical, legal, and economic implications of generative AI development.

Further legal and policy development is crucial to address the evolving challenges of AI co-generation. This includes clarifying the scope of exceptions, addressing the copyrightability of AI outputs, and ensuring effective enforcement of data protection rights.

This necessitates ongoing dialogue and collaboration among stakeholders to ensure responsible and beneficial AI integration in society.