Co-generation of data

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

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.

This report was developed by Experts and Specialists involved in the Global Partnership on Artificial Intelligence's project on co-generation of data. The report reflects the personal opinions of the GPAI Experts and External Experts involved and does not necessarily reflect the views of the Experts' organisations, GPAI, or GPAI Members. GPAI is a separate entity from the OECD and accordingly, the opinions expressed and arguments employed therein do not reflect the views of the OECD or its Members.

Acknowledgements

This report was developed in the context of the Co Generation of Data, 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**‡. GPAI is grateful for the meaningful contribution of **Naohiro Furukawa***, **Avik Sarkar**, **Solene Festor****, **Josef Drexl****, **Alexandra Salnikow‡, Stefan Janusz** and **Antoine Glory** (CEIMIA)**.**

GPAI would also like to thank PAG members **Seong Oun Hwang***, **Zümrüt Müftüoğlu**, **Jaco Du Toit****, **Kudakwashe Dandajena***, **Zee Kin Yeong***, **Kim McGrail***, **Mikael Jensen***, **Ulises Cortes*, Toshiya Jitsuzumi*** and **Maja Bogataj***.

GPAI recognizes **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 United Nations Commission on International Trade Law, represented by **Alexander Kunzelmann**† as **ELI's Observer.**

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*** (Truera).

* 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: Principles, Report, November 2024, Global Partnership on AI.

Introduction

Preliminary Remarks

Context. These Principles augment the comparative study published as Copyright and Data Protection Rights in Co-Generated Input and Output of Generative AI, Report, November 2024, Global Partnership on AI.

Scope. These Principles are intended to be used by legislators addressing new Copyright or Data Protection Rights in Co-Generated In- and Output. Thus they are not high-level policy principles, such as the Hiroshima Principles, but rather they serve as general guidelines that can be followed by legislators when drafting specific legal provisions and regulations concerning Gen AI. They aim to balance the interests of various parties involved with, or affected by, the development or deployment of Generative AI ('Gen AI'), while fostering innovation and promoting responsible development of AI technologies.

It should be noted that the Principles are of a preliminary nature; they are neither exhaustive nor final. For the ongoing discussion on the regulation of Copyright and Data Protection Rights with regard to generative AI that goes beyond these Principles, this means that the Principles can not be brought forward as an argument for or against additional Rights.

Definitions

These Principles work with the following definitions:

- **'Co-generation'** is an overarching notion that translates differently, and in more specific legal notions, under the Copyright and Data Privacy Laws of the EU, US and Japan, as well as in other jurisdictions. Under those legal frameworks, the notion of co-generation is discussed under statutory and jurisprudence-based notions such as "reproduction", "display", or "communication to the public" for Copyright Law or "processing", covering among others the "collection", "recording", "storage", "retrieval", of personal data, for Data Protection Law;
- **'Co-Generated Input of Generative AI'** means, inter alia, any observed or provided personal data, non-personal data or content – including text, images and other information – used to train (including to adjust, fine-tune or align[1]) GenAI tools;
- **'Co-Generated Output of Generative AI'** means any result generated by the GenAI tool, including text, images and other information;
- **'Content'** means any structured information that has some resonance or meaning for an individual at the semantic level (for ex in the form of some story or express knowledge), it can be produced by humans or machines, such as GenAI tools.
- **'Data'** (that includes metadata) means any digital representation of acts, facts or information at the syntactic level;

[1] This means that Co-generated Input covers the whole process of developing an AI model, not only its early (pre-)training with input data, but as well the additional data needed to adjust, fine-tune or even align the model (insofar as those phases require the ingestion of data, not solely the calibration of the algorithms). Co-generated Input can also cover the feed data provided by the user of the model (whether through prompts or the ingestion of other content at the initiative of the user).

- **‘Data subject’** means any identified or identifiable person enjoying specific rights within the applicable data protection law;
- **‘Generative AI (or GenAI)’** (tools) means AI models or systems that are trained with a large amount of data and that emulate the structure and characteristics of input data in order to generate derived synthetic content (including images, videos, audio, text, and other digital content);
- **‘Non-Personal Data’** means any data that does not qualify as personal data;
- **‘Personal Data'** means any data that refers to an identified or identifiable natural person within the applicable data protection law;
- **‘Provider’** means any individual or entity that develops an AI model and puts it on the market;
- **‘Publicly available data’** means data which is available to an unlimited number of persons or entities (and usually accessible online);
- **‘Rightholder’** means any individual or entity that enjoys some rights on the data or content used as input for the generation of some output;
- **‘User’** means:

  (a) any individual or entity who uses personal data, non-personal data or content – including text, images and other information – to add relevant input to an already trained Generative AI tool.

  (b) any individual or entity who utilises Generative AI tools to generate a result.

Differing treatments of Copyright and Data Protection Law

The constellation of interests and the public policy considerations supporting the control on data that copyright grants and data protection guarantees are very different. This implies that the reuse of copyrighted content and of personal data raises separate issues and must be treated differently by the law. The scraping and processing of data needed for the functioning of Gen AI tools raises different risks that need to be assessed within the balance of interests and policy compromise supporting those two areas of digital law. For example, while copyright focuses on incentivizing creativity and innovation, data protection prioritizes individual autonomy and the prevention of harm. Therefore, legislators must carefully consider these distinct objectives when crafting legal frameworks for the development or deployment of AI tools depending on whether they make use of copyrighted works or personal data.

Principles of copyright and data protection rights in co-generated input and output of generative AI

Principle 1: Use of Publicly Available Data

A. Publicly available data that is not protected by contract (such as a subscription agreement) or statutory law is freely available for reuse to train or to feed Generative AI (models).

B. To assess whether publicly available personal data, non-personal data or content – including text, images and other information – should be freely available for reuse, the following factors, in particular, should be taken into account:

  1. Whether the use was reasonably anticipated by the Rightholders, e.g. because the Input was made publicly available on behalf of the Rightholder;
  2. Whether the AI tool has socially beneficial applications;
  3. Whether the User is using technology to maximize the protection of the Rightholders;
  4. Whether the output is a good substitute for the input;
  5. The nature and scope of the use;
  6. The effect of the use on the Rightholder position;
  7. The level of transparency of the use;
  8. The control granted to the Rightholders.

Principle 2: Right to an Economic Share

A. A Right to an economic share in the profits or other economic advantages derived from the Co-Generated Input of Generative AI should only exist when it stems from:

  1. A contractual agreement between the Rightholder and the Provider of the Co-Generated Input.
  2. A statutory provision.

B. The Right to an economic share in profits or other economic advantages derived from the Co-Generated Input of Generative AI may only apply to personal data when its nature as personality right does not speak against benefiting from such an economic share.

C. This Principle is without prejudice to the Rights arising from unlawful use of Co-Generated Input under Principle 3.

Principle 3: Rights arising from unlawful use of co-generated input

A. The use of Co-Generated Input is unlawful if it violates any applicable contractual obligation or statutory provision.

B. If the use of Co-Generated Input is unlawful, the available remedies may include:

1. Damages;
2. Injunctions aimed at prohibiting the use of the Co-Generated Input;
3. Erasure of the Co-Generated Input.

C. The erasure of, or the injunction aimed at prohibiting the use of, unlawfully obtained Co-Generated Input of Generative AI should only lead to the erasure of inextricably linked and lawfully obtained Co-Generated input or to an injunction where the legitimate interests of the Rightholders or the general public significantly override the interests of the Provider. When determining whether the inextricably linked and lawfully obtained Co-Generated Input shall be erased or whether injunction should be granted, the following aspects should be considered:

1. The share the Rightholder had in the Co-Generated Input;
2. The purpose of the use of the Co-Generated Input of Generative AI
3. Whether the use may lead to significant harm for the Rightholder
4. Whether the user had notice of its unlawfulness
5. Whether the erasure leads to substantial environmental harms, for example because the generative AI tool has to be retrained.

Principle 4: Right to Information and Transparency

A. Providers of Generative AI tools shall inform the data subjects on the use of their personal data to understand which personal data is used and should put in place mechanisms to assert the Right to Data Portability, to Rectification and to Authentic Attribution or Rights stemming from unlawful use of the personal data, only where this does not prove impossible or would involve a disproportionate effort.

B. Providers of Generative AI tools should prepare and make publicly available a sufficiently detailed summary about the copyright-protected content used for training the AI model.

Principle 5: Right to Rectification and to Authentic Attribution

A. Each data subject shall have the right to request the Provider to use its best efforts to ensure accurate personal data is used when training Gen AI tools and to avoid that incorrect personal data is generated by those tools. The use and generation of incorrect personal data should be considered under Principle 3.

B. Authors of protected content shall - in exceptional cases to be defined by statutory law - have the right to be adequately credited if an extensive collection of their creations has been used to train a generative AI tool that is able to imitate a creative style.

Principle 6: Rights to an Economic Share, to Information, to Rectification and Authentic Attribution in Co-Generated Output

The Rights to an Economic Share (Principle 2), the Rights arising from Unlawful Use of Co-Generated Input (Principle 3), the Right to Information and Transparency (Principle 4), and the Right to Rectification and to Authentic Attribution (Principle 5) should apply *by analogy* 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.

Principle 7: Right to Explanation and Transparency

A. Whether a person should have the right to explanation, i.e. the right to obtain clear and meaningful information on the role of the Co-Generated Input and the main elements of the Co-Generated Output, should depend inter alia on the following factors:
    1. Whether the Person is subject to a decision which is mainly based on Co-Generated Output of Generative AI;
    2. Whether the Co-Generated Output of Generative AI has an adverse impact on that person;
    3. Whether it is possible to retrieve the information from another source.

B. Providers of Generative AI shall ensure that the source or provenance of the output of Gen AI is identified in a machine-readable format and that the output is made detectable as artificially generated or manipulated. Otherwise the use should be considered unlawful under Principle 3.

Concluding remarks and future work

These principles lay the foundation for addressing the intersection of copyright and data protection in the co-generation of data by generative AI. By emphasizing transparency, accountability, and fairness, these principles seek to protect the rights and interests of individuals and stakeholders while fostering innovation and ethical development. They recognize the diverse legal and cultural contexts in which generative AI operates, offering guidance that balances the promotion of technological progress with the need to safeguard fundamental rights. This balanced approach positions these principles as a critical tool for legislators and stakeholders navigating the challenges posed by rapidly advancing AI technologies.

However, as the field evolves, refining accountability mechanisms, and addressing ethical concerns like bias and misuse are key priorities. Sustainability must also be considered, given the environmental impacts of AI development; and must be balanced with the practical application of these principles. Periodic reevaluation of definitions and exploring the implications of emerging applications like synthetic media will ensure adaptability. These principles are a vital starting point, but ongoing collaboration and refinement will be crucial for navigating future challenges responsibly.