From co-generated data to generative AI

New rights and governance models in digital ecosystems

May 2024

GPAI | THE GLOBAL PARTNERSHIP ON ARTIFICIAL INTELLIGENCE

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:
Joe Massey‡ and **Elena Simperl**‡, Open Data Institute
Vinay Narayan‡**, Avani Airan**‡**,** and **Astha Kapoor**‡, Aapti Institute

GPAI would like to acknowledge the tireless efforts of colleagues at the International Centre of Expertise in Montréal on Artificial Intelligence (CEIMIA) and GPAI's Data Governance Working Group. We are grateful, in particular, for the support of **Camille Seguin**, **Stephanie King**, and **Stefan Janusz** from CEIMIA, and for the dedication of the Working Group Co-Chairs, including up until December 2023 **Jeni Tennison*** (Connected by Data) and **Maja Bogataj Jančič*** (Intellectual Property Institute, Slovenia), and from January 2024 **Shameek Kundu*** (Truera) and **Bertrand Monthubert*** (Université Toulouse-III-Paul-Sabatier). The report also benefited from the detailed expert review of Working Group members, including in particular **Avik Sarkar**, **Zee Kin Yeong**, and **Kudakwashe Dandajena**.

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

Citation

GPAI 2024. From co-generated data to generative AI: New rights and governance models in digital ecosystems, Report, May 2024, Global Partnership on AI.

Table of Contents

Executive Summary

Within two months of the launch of ChatGPT in 2022, it had 100 million active users a month, making it the fastest-growing consumer application in history.[1] While the core technology behind such generative artificial intelligence (AI) models has been around for some time, tools such as ChatGPT have made it accessible to the masses. As a result, there is a lot of talk about AI and new questions being asked about the role data plays in making these tools as good as they are.

'Co-generated data' refers to data generated by more people or entities than solely the data holder. However, co-generation is not limited just to data. Co-generation occurs across the data ecosystem; when we scroll through social media, when we speak to our voice assistants and when we prompt conversational AI tools. In each of these examples there can be multiple data co-generators, with different levels of involvement in the co-generation process and awareness of its potential financial, social, legal or ethical implications. As generative AI enters the mainstream, we must update our thinking about co-generated data, technology and AI-generated works.

This report is part of the research project 'From co-generated data to generative AI,' documenting work conducted between June 2023 and May 2024. Commissioned by the Global Partnership on Artificial Intelligence (GPAI) and executed by the Open Data Institute (ODI) and Aapti Institute, with support from Pinsent Masons, the research involved analysing six co-generation scenarios. This concept of co-generation being relatively new, we decided to first observe and analyse the cases which involve co-generations as it would allow us to understand what kind of rights may be involved. This was done using a framework of legal rights developed through a literature review, supplemented by expert interviews and workshops.

The six scenarios cover a wide range of types of co-generation, co-generators and legal contexts: remunerated work (Karya), collaborative crowd-sourced data collection (OpenStreetMap), social media platforms (Instagram), Internet of Things (BMW connected cars owned by Europcar), and conversational generative AI (Whisper and Midjourney).

Our findings indicate that co-generators in these scenarios navigate a complex web of rights, spanning intellectual property, data, and other rights like labour rights. Nonetheless, significant gaps persist, particularly regarding the co-generation of technology, AI-generated works, and rights for communities and collectives. Where rights exist, they are often mediated through statutory exceptions (such as for text and data mining), terms and conditions (T&Cs) of contracts or other mutual agreements, and licences, which do not always function as intended, as evidenced by the documented failures of T&Cs. Determining the existence and extent of rights over various types of co-generated data remains challenging, especially for generative AI models. Further research is essential to map the rights landscape fully, identify gaps, and explore non-legal mechanisms like new technologies, governance models, and licences.

[1] Hu, K. (2023). Reuters. 'ChatGPT sets record for fastest-growing user base'. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

Background

New forms of co-generation in digital ecosystems

Data[2] has become an important part of society. It is now a resource in its own right, when it was previously a by-product of industrial, commercial, consumer and other activities. Data has become the lifeblood of some of the largest companies in the world, which has enabled enormous profits for these businesses. Data enables the development of new and improved technologies that have become important parts of our lives. At the same time, access to data is critical to tackling some of society's biggest challenges, such as the climate crisis and health inequalities.[3]

The rapid expansion of the data economy raises serious questions for governments, businesses and the public about who has access to data and for what reasons, who gets to decide what data is used for, and ultimately who is able to realise the value of data.[4] It also raises questions about how to limit the misuse of data, how to preserve people's privacy and how to hold those causing harm to account.[5]

There is increasing attention being paid to the notion that parties who have contributed to the generation of data should have some rights in the utilisation of such data.[6] This has played out across the data ecosystem. There have recently been debates over the rights of users of Internet of Things (IoT) devices, such as sensors within autonomous vehicles or smart speakers. Substantial amounts of data are collected for the development of this technology, but users of these devices, who are critical to the generation of this data, may not have rights over the data.[7] Following the Cambridge Analytica scandal, there has been significant debate over the data rights of users of social media platforms.

Cogenerated data, like all data, can come from different activities, and could be used by the multiple actors (data generators and users) in an organisation to improve its functioning, to drive research and innovations in particular domains (for example in academia or medicine), or to provide insights that improve how the organisation interacts with others within a given ecosystem.

Generative AI is one of the many sources of cogenerated data. It refers to a category of AI-based algorithms that generate new outputs based on the data they have been trained on.[8] Unlike

[2] 'Data' means *information recorded in any machine-readable format suitable for automated processing, stored in any medium or as it is being transmitted* – European Law Institute
[3] Clutton-Brock, P. et al. (2021). Global Partnership on AI Report, Climate Change AI, & the Centre for AI & Climate. 'Climate Change & AI: Recommendations for Government'. https://www.gpai.ai/projects/climate-change-and-ai.pdf
[4] Ada Lovelace Institute (2021). 'Participatory data stewardship'. https://www.adalovelaceinstitute.org/report/participatory-data-stewardship/
[5] ODI (2023). 'Responsible data stewardship'. https://www.theodi.org/article/defining-responsible-data-stewardship/
[6] Leffer, L. (2023). Scientific American. 'Your Personal Information Is Probably Being Used to Train Generative AI Models'. https://www.scientificamerican.com/article/your-personal-information-is-probably-being-used-to-train-generative-ai-models/
[7] Janeček, V. (2018). 'Ownership of personal data in the Internet of Things'. Computer Law & Security Review, Volume 34, Issue 5. https://doi.org/10.1016/j.clsr.2018.04.007.
[8] Routley, N. (2023). World Economic Forum. 'What is generative AI? An AI explains'. https://www.weforum.org/agenda/2023/02/generative-ai-explain-algorithms-work/

traditional AI-based systems, generative AI models create *new* content in the form of data, including images, text, audio, and more. Some examples of generative AI models include ChatGPT, which produces text responses to human prompts; DALL-E, which produces images in response to human prompts; and Voicebox, which produces audio responses to human prompts. Generative AI models are trained on huge amounts of data, with this database often being created by scraping the internet (web scraping)[9]. In many instances, users are unaware that this data is being used for these purposes.[10]

The widespread use of generative AI, catalysed by ChatGPT and similar tools, has reignited interest in co-generation in the digital realm—highlighting that digital technologies are refined through data generated by diverse contributors.

The ability of generative AI models to generate new data, including audio and video content, provides vast potential for innovation for societal good. At the same time, these models' ability to mimic styles in the fields of journalism, poetry, painting and photography poses potential threats to the livelihoods of creative professionals, in addition to raising questions around fairness and transparency for society more broadly.

The role of data rights in AI development and deployment

Data is indispensable to AI, underpinning model development and deployment at every stage. High-quality training datasets are crucial for success, requiring extensive wrangling, cleaning, and enrichment. Generative AI models particularly emphasise data specificity; for instance, image-generating models necessitate extensive image datasets.

Data rights are crucial in today's digital landscape, providing individuals with privacy protection, transparency and control over their personal information. While they manifest in a variety of ways in different jurisdictions, these rights broadly cover the collection of data, access to data, the portability of data, desistance from the use of data (the right to resist use of data through to the right to delete data), correction of data and the deletion of data.[11] These rights empower individuals to make informed decisions about data collection and sharing, reducing the risk of privacy breaches, and promoting transparency in data practices. The trend towards considering data as an 'asset', even though there is currently no formal, all-encompassing property right per se to data, has led to a wealth of commentary and developments in relation to the various overlapping and sometimes conflicting frameworks of rights that exist over data.[12]

[9] Web scraping is a data science technique that deploys tools for the extraction of data from websites. See: Thapelo, S. T. et al. (2021). Data Science Journal. 'SASSCAL WebSAPI: A Web Scraping Application Programming Interface to Support Access to SASSCAL's Weather Data'. https://datascience.codata.org/articles/10.5334/dsj-2021-024.

[10] Chakravorti, B. (2020). Harvard Business Review. 'Why It's So Hard for Users to Control Their Data'. https://hbr.org/2020/01/why-companies-make-it-so-hard-for-users-to-control-their-data

[11] Global Partnership on AI (2020). A Framework Paper for GPAI's work on Data Governance, Global Partnership on AI, Data Governance Working Group. https://gpai.ai/projects/data-governance/gpai-data-governance-work-framework-paper.pdf

[12] Wendehorst, C. & Cohen, N. (2023). American Law Institute and European Law Institute. 'ALI-ELI Principles for a Data Economy - Data Transactions and Data Rights'.

The centrality of data to AI models, from conception to operation, makes data rights even more important. Given that data provides the information that a machine learning model is trained on and learns from,[13] the nature of rights over data – including who can access what data, under what conditions, and what uses different types of data can be put to and under what circumstances – has important ramifications for the innovation of generative AI models and larger public good.

In addition to data rights, the other major legal rights framework involved in generative AI models is intellectual property (IP) rights. In training AI models, there is a need to utilise huge quantities of data. Usually this requires organisations to scrape this data and content from the internet. By using this method, they collect a great variety of data and content, all of which is covered by a range of legal protections. Some of the content may be protected by copyright and others may be covered by a specific type of licence, among other protections. However, in the process of scraping, these intricacies are lost, and the legal protections are not adhered to. The ability of generative AI models to also mimic styles of individual creatives reignites the debate around the protection of style (and its link to livelihood).[14] There is a clear need to attune IP rights frameworks to balance protection of creatives with the use of content for the benefit of society.

Challenges with co-generated works

Debates around co-generation in the digital context are not new. Free and open source software is a prime example of the development and refinement of digital technologies through the contributions of various individuals and entities.[15] The domain of gig work has also seen debates around co-generation, with differing views[16] on gig workers having control of, and obtaining value from, the data they generate on a gig work platform, given that this data is used by the platform to improve their algorithms. There has also been significant debate about the use of health data by private companies to train and develop AI models for medical research.[17]

The rapidly increasing efficiency and use of generative AI-based models has re-focused the debate around how sufficient existing legal frameworks are to deal with generative AI and other new co-generation scenarios. There is a degree of uncertainty around the rights of co-generators, especially with generative AI models. This uncertainty extends to whether content with IP protection can be used to train AI models without prior permission from the rights holder, whether co-generators hold any rights in the outputs created by generative AI models that have been trained

https://www.europeanlawinstitute.eu/fileadmin/user_upload/p_eli/Publications/ALI-ELI_Principles_for_a_Data_Economy.pdf

[13] ODI (2023). 'What do we mean by "without data, there is no AI"?'. https://theodi.org/news-and-events/blog/what-do-we-mean-by-without-data-there-is-no-ai/

[14] Brownlee, M. (1993). Columbia Law Review. 'Safeguarding Style: What Protection Is Afforded to Visual Artists by the Copyright and Trademark Laws?'. https://www.jstor.org/stable/1122961

[15] Mizushima, K. & Ikawa, Y. (2011). IEEE. 'A structure of co-creation in an open source software ecosystem: A case study of the eclipse community'. https://ieeexplore.ieee.org/document/6017787

[16] van Doorn, N. & Badger, A. (2020). Antipode: A Radical Journal of Geography. 'Platform Capitalism's Hidden Abode: Producing Data Assets in the Gig Economy'. https://doi.org/10.1111/anti.12641

[17] Ipsos (2022). NHS AI Lab Public Dialogue on Data Stewardship. https://www.ipsos.com/sites/default/files/ct/news/documents/2022-11/22-033229-01%20NHS%20AI%20Lab%20Data%20Stewardship%20Dialogue%20-%20Report_0.pdf

on the data of the co-generators, and how the training of generative AI models interfaces with data rights[18] of these co-generators (where a 'data right' means a right against a holder of data that is specific to the nature of data and that arises from the way the data is generated, or from the law for reasons of public interest). This is made even more difficult as those developing generative AI models often do not disclose which data and content has been used to train the models.[19] This complexity is exemplified by the plethora of lawsuits and legal challenges against the organisations developing generative AI models.

Many of the most high profile cases lean on copyright law, and many are currently taking place in the US. For example, a group of artists have filed a lawsuit against Stability AI for the use of their artwork in the training of the Stable Diffusion image-generating AI model;[20] and a group of authors have filed a lawsuit against OpenAI and Meta for 'infringing the authors' copyrights by using copies of their books to train their AI models'.[21] The *New York Times* is in the process of suing OpenAI for breaching copyright, alleging that ChatGPT is able to regurgitate whole sections of *New York Times* articles, among other accusations, which violates copyright law and impacts its business model.[22] However, data rights have also been invoked, as the Italian Government has twice banned ChatGPT for breaching data protection regulation under the General Data Protection Regulation (GDPR).[23]

There is much debate around what counts as 'fair use' of data and content collected from the internet.[24] Jurisdictions around the world have included exceptions for text and data mining (TDM) in their copyright laws for decades. However, in recent years certain jurisdictions have begun reviewing their laws in this regard. Japan, for example, has developed and revised AI-related regulations with the goal of maximising AI's positive impact on society, rather than suppressing it due to overestimated risks.[25]

The continued growth in the number and variety of lawsuits around generative AI, and the increased

---

18 https://www.europeanlawinstitute.eu/fileadmin/user_upload/p_eli/Publications/ALI-ELI_Principles_for_a_Data_Economy_Final_Council_Draft.pdf

19 The Economist (2024). 'Generative AI is a marvel. Is it also built on theft?'. https://www.economist.com/business/2024/04/14/generative-ai-is-a-marvel-is-it-also-built-on-theft

20 Bearne, S. (2023). BBC. 'New AI systems collide with copyright law'. https://www.bbc.co.uk/news/business-66231268

21 Richard Kadrey, Sarah Silverman & Christopher Golden v. Meta Platforms, Inc. (2023). United States District Court Northern District of California. Complaint. https://llmlitigation.com/pdf/03417/kadrey-meta-complaint.pdf; Sarah Silverman, Christopher Golden & Richard Kadrey v. OpenAI, Inc. & Others (2023). United States District Court Northern District of California. Complaint. https://llmlitigation.com/pdf/03416/silverman-openai-complaint.pdf

22 Reed, R. (2024). Harvard Law Today. ChatNYT: Harvard Law expert in technology and the law says the New York Times lawsuit against ChatGPT parent OpenAI is the first big test for AI in the copyright space. https://hls.harvard.edu/today/does-chatgpt-violate-new-york-times-copyrights/

23 Reuters (2024). 'OpenAI's ChatGPT breaches privacy rules, says Italian watchdog'. https://www.reuters.com/technology/cybersecurity/italy-regulator-notifies-openai-privacy-breaches-chatgpt-2024-01-29/

24 The Economist (2024). 'Generative AI is a marvel. Is it also built on theft?'. https://www.economist.com/business/2024/04/14/generative-ai-is-a-marvel-is-it-also-built-on-theft

25 The AI Patent Blog (2023). 'Legal protection from generative AI in Japan'. https://www.theaipatentblog.com/legal-protection-from-generative-ai-in-japan

efforts of policymakers to keep up with regulation, indicates the size of the issue at hand. There are two broad key challenges to solve around co-generation:

First, many of these lawsuits argue for a fairer technology ecosystem, in which members of the public, creators and communities have strong legal rights to protect them and their livelihoods. This is borne out of the fact that generative AI models are trained on the content of the creators and communities, to the point of being able to mimic individual styles. The content is accessed for free and the models developed can accrue commercial gains for the organisations creating them.

Second, while these models can represent commercial gains, they also have the potential to support society to tackle some of its greatest challenges, such as achieving sustainable development goals and tackling the climate crisis or solving health issues. Access to data is a crucial part of innovation for AI and beyond.

These lawsuits have created a sense of unease around the sharing of data and content, which has been described as a 'data winter'.[26] There needs to be a balance between the priorities stated above, and regulation that restricts access in the interest of livelihoods but is not detrimental to larger societal good. Finding this balance is complicated by a number of factors, including who contributes data, the nature of their contribution and the role of technology in facilitating the contributions. These factors can lead to different types of co-generation scenarios, and different scenarios are likely to necessitate different combinations of rights to the individuals and organisations involved in the co-generation.

Co-generation is a broad concept, and one that is challenging to regulate in legal terms. This research explores how the concept of co-generation is complicated by the unique aspects of generative AI and seeks to better understand the issues around co-generation in this context. It analyses how current legal frameworks apply to different examples of co-generation, and explores how concepts that are not yet legally binding, such as collective rights, could be asserted by communities that have contributed to co-generation in some meaningful way.

[26] Verhulst, S. (2024). 'Are we entering a "Data Winter"?'. https://sverhulst.medium.com/are-we-entering-a-data-winter-f654eb8e8663

Scope and approach

Defining co-generation

The concept of 'co-generation' is central to this research. Co-generation in the realm of data and technology can occur in various forms, involve different participants, and be governed by different laws depending on the jurisdiction. There are several dimensions of co-generation to consider:
Firstly, the nature of the co-generators will differ. Each co-generator could be a private business, government entity, member of the public, community, or even an AI model.
Secondly, the level of involvement of the co-generators can vary. Co-generation is not binary; thus, the involvement of co-generators spans a spectrum from active to passive participation.
Finally, the type of co-generation, in terms of the final output that is generated, can differ significantly. In this research we consider three different types of co-generation:

'**Co-generated data**' refers to data generated by more people or entities than solely the data holder. The term was first coined by the American Law Institute and the European Law Institute's 'Principles for a Data Economy (ALI-ELI Principles)'.[27] The European Union (EU), as part of its European Data Strategy, adopted the term co-generated data based on these principles. The ALI-ELI Principles define co-generated data as: 'data to the generation of which a person other than the controller[28] 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'. The fact that a party has contributed to the generation of certain data may, together with other factors identified in the ALI-ELI Principles, give rise to certain rights that that party may have vis-à-vis a controller (holder) of the data. Rights in co-generated data may include various rights of access to the data, up to and including, where appropriate, the right to real-time data portability.

This report considers the dimension of data as a representation of information.[29] The concept of co-generated data acknowledges that data is usually generated by different contributions from various parties, eg, by being the subject of the information, by performing an activity through which the data was generated, or by having rights in a product or service that has contributed to the generation of data.[30]

'**Co-generated technology**' refers to technology which has been generated by multiple parties. Debates around co-generation in the digital context have taken place around open source

[27] Wendehorst, C. & Cohen, N. (2023). American Law Institute and European Law Institute. 'ALI-ELI Principles for a Data Economy - Data Transactions and Data Rights'. https://www.europeanlawinstitute.eu/fileadmin/user_upload/p_eli/Publications/ALI-ELI_Principles_for_a_Data_Economy.pdf
[28] Controller here has the same meaning as in Article 4(7) of the EU General Data Protection Regulation, but not restricted to personal data. It is equivalent to the term 'data holder' used elsewhere.
[29] GPAI (2020). Global Partnership on AI. 'A Framework Paper for GPAI's work on Data Governance'. https://gpai.ai/projects/data-governance/gpai-data-governance-work-framework-paper.pdf
[30] Wendehorst, C. & Cohen, N. (2023). American Law Institute and European Law Institute. 'ALI-ELI Principles for a Data Economy - Data Transactions and Data Rights'. https://www.europeanlawinstitute.eu/fileadmin/user_upload/p_eli/Publications/ALI-ELI_Principles_for_a_Data_Economy.pdf

technologies, which have been built through the separate contributions of various individuals and entities.

In this report, we focus specifically on one subset of technology: the co-generation of AI models. As explored in the above section, AI models are trained upon huge amounts of data and content, which is called training data. This training data has been collected from the internet and could take the form of text data from social media websites, code contributed to code-hosting platforms (e.g Github), images from an artist's webpage or videos from online video-sharing platforms (e.g. Youtube). In some cases, these inputs may have been co-generated themselves, such as a web page on Wikipedia, however the main focus is on the co-generation of the technology itself. This data and content forms the basis of these AI models, and thus each input has some (extremely small) involvement in the final output. Ultimately, these models could not be created without the data and content collected in huge quantities for training datasets, and therefore generative AI models should be considered as 'co-generated technology'.

'**AI co-generated works**' here refers to the new outputs of generative AI models, in the form of data or content such as text, image, audio or video. Generative AI models often function via prompts, where a user inputs a natural language input which the model uses to generate an output. In this case there are two clear co-generators: the prompter and the AI model, with both combining to create the desired output. However, as explored in the previous type of co-generation, the AI model itself has been developed with inputs from numerous sources, and different co-generators. Therefore AI co-generated works could be said to be generated by a subset of those involved in the training dataset as well. In the example of the lawsuit being filed against Stability AI by a group of artists - it is the artists' artwork which has enabled the model to generate artworks in the style of the artists.[31]

Methodology

Different examples of co-generation scenarios simultaneously occur all over the world, whether in the generation of new data points, training datasets, algorithms or AI-generated works. This report is concerned with the different legal mechanisms which apply to co-generators in different jurisdictions, covering data rights, IP rights and more. Each legal jurisdiction has its own legal regime, albeit with some similarities across regions. This research takes a case study approach, focusing on six different examples of co-generation to explore the legal ramifications of co-generation in practice. These case studies involve different legal landscapes across jurisdictions and a variety of co-generation activities. A number of aspects can impact cogenerated data:

- The nature of the co-generators.
- The level of involvement of the co-generators.
- The type of co-generation.
- The legal jurisdiction of the data holder and the co-generators.

[31] Bearne, S. (2023). BBC. 'New AI systems collide with copyright law'. https://www.bbc.co.uk/news/business-66231268

This research began with a literature review, examining sources from legal, technical, and data perspectives across both scientific and grey literature, as well as company documents and national legislation pertinent to each case study. The team also conducted 13 interviews with experts from around the world and from a variety of disciplines to explore and confirm our understanding of the co-generation case studies and the complications for co-generators created by AI. We sought to balance legal expertise—those with an understanding of specific areas of law or rights in a given jurisdiction—with industry expertise, encompassing those with knowledge of the realities of co-generation and the impact of generative AI, alongside experts from civil society. Additionally, we held two workshops with experts to discuss the ideas covered in this research.

Selecting the co-generation case studies

Building on the literature review, and in collaboration with the GPAI project working group, we researched different co-generation scenarios to compile a longlist of potential case studies. We then narrowed this list down to six case studies that would allow us to explore the breadth of different types of co-generation. The rights frameworks that apply in each scenario differ and we selected a set of diverse scenarios across the following criteria:

- **The nature of the co-generators**. This criterion looks at who the co-generators are; private businesses, governments, members of the public, communities, or even AI models themselves.

- **The level of involvement of the co-generators**. This criterion assesses whether the generation of data was a result of the active or passive involvement of the co-generator. As noted in the 'Comment to Principle 18 of the ALI-ELI Principles',[32] the share/role the co-generator had in co-generation has a bearing on the co-generator's claim or justification for a right over co-generated data.

- **The type of co-generation**. This criterion refers to the three types of co-generation considered in this report: co-generated data, co-generated technology and AI-generated works.

- **The legal jurisdiction of the data holder and the co-generators**. This criterion is important to understanding the nuances of different legal jurisdictions around the world, and how they apply to co-generators.

The case studies were chosen to demonstrate a variety of co-generation scenarios and cover some of the topical debates at the time of writing. For example, IoT devices are of particular interest with the advent of the Data Act in the EU, which attempts to balance access to data for innovation with rights over data for users of technology. Given the interest in generative AI for this research, we selected two case studies covering generative AI models.

Analysis of the case studies by applicable legal frameworks

[32] Wendehorst, C. & Cohen, N. (2023). American Law Institute and European Law Institute. 'ALI-ELI Principles for a Data Economy - Data Transactions and Data Rights'. https://www.europeanlawinstitute.eu/fileadmin/user_upload/p_eli/Publications/ALI-ELI_Principles_for_a_Data_Economy.pdf

We understand rights to mean the freedoms and entitlements of citizens in a particular political and legal system. Each of the six case studies was analysed using the following rights framework, looking at how these rights are reflected through contracts, licences and T&Cs:

- **Context** for each case study, including the different co-generators involved, their degree of involvement and their awareness of involvement in co-generation.

- **Data rights** are legally protected interests that 'arise from the very nature of data as a non-rivalrous resource, which may be used by many different parties at the same time'.[33]

    - **Data protection rights** grant individuals ('data subjects') the right, to a certain degree, to control the collection and use of their personal data in relation to, among other things, processing activities of organisations; the right to be 'forgotten'; the right to have incorrect data rectified; and the right to object to unlawful processing.

    - **Data access and portability rights** allow individuals and organisations the right to access data within defined parameters, ranging from merely being able to read the data on a device, to a fully-fledged data portability right.

    - Rights over **desistance** from the use of data (the right to resist use of data through to the right to delete data), **correction** of data and **economic share in profits derived** from data.

- **Intellectual property** (IP) rights protect creations of the mind and provide people and organisations with recognition and/or financial benefit from what they invent or create.[34,35]

    - **Copyright** typically protects original literary, dramatic, musical and artistic works; sound recordings, films and broadcasts; and the typographical arrangements of published editions.[36]

    - ***Sui generis* database right** is a unique right that protects a compilation of data or information in a database and gives the owner the right to prevent unauthorised extraction of either the whole or part of that database.[37]

    - **Laws of confidentiality and trade secrets** protect any form of information that is secret, including personal data and commercial information or data.

- **Other legal frameworks** which might be involved. These include but are not limited to:

[33] GPAI (2020). Global Partnership on AI. 'A Framework Paper for GPAI's work on Data Governance'. https://gpai.ai/projects/data-governance/gpai-data-governance-work-framework-paper.pdf
[34] WIPO (n.d.). 'What is Intellectual Property?'. https://www.wipo.int/about-ip/en/
[35] GPAI (2022). Global Partnership on AI. ' Protecting AI innovation, Intellectual Property (IP): GPAI IP Primer, Report'. https://gpai.ai/projects/innovation-and-commercialization/gpai-intellectual-property-primer-2022.pdf
[36] WIPO (n.d.). 'Copyright'. https://www.wipo.int/copyright/en/
[37] Directive 96/9/EC Of The European Parliament and of The Council of 11 March 1996 on the legal protection of databases, *Official Journal* L77, p. 20. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A31996L0009

- **Fundamental rights** are basic rights[38] in the nature of natural rights[39] and human rights[40] that are guaranteed to all citizens of a particular country (or region), such as privacy.

- **Labour rights** are rights relating to labour relations between workers and employers.

- **Collective and community rights** are rights held by a group as a whole, rather than individually by its members. They include things such as Indigenous data sovereignty rights.

[38] European Union Agency for Fundamental Rights (n.d.). 'What are fundamental rights?'. http://fra.europa.eu/en/about-fundamental-rights
[39] His Holiness Kesavananda Bharati v. State Of Kerala (1973). Supreme Court of India. Judgement. https://judgments.ecourts.gov.in/KBJ/
[40] Universal Declaration of Human Rights. https://www.un.org/en/about-us/universal-declaration-of-human-rights

Six case studies that demonstrate different aspects of co-generation

In order to assess the effectiveness of different legal rights for co-generators, this section analyses what legal rights possibly apply in the six co-generation scenarios.

Given the cross-border nature of digital services, the applicable legal jurisdiction can differ based on various factors including the location and citizenship of co-generators involved. In the interest of brevity and clarity, the analysis was restricted to the provision of a particular service in the jurisdiction identified.

Scenario	Case study	The nature of the co-generators	The level of involvement of the co-generator	The type of co-generation	Legal jurisdiction of the data holder and the co-generators
Remunerated work	Karya	Data labellers	Very active	Co-generated data, co-generated technology	India, global
Crowdsourced data collection	OpenStreetMap	Contributors to OpenStreetMap	Very active	Co-generated data	UK, global
Social media platforms	Instagram	Social media users	Active and passive	Co-generated data, co-generated technology	US, Brazil
Internet of Things	BMW connected cars owned by Europcar	Occupants of a car	Passive	Co-generated data	EU
Conversational generative AI	Whisper	Users and contributors (the creators of the scraped data)	Active and passive	Co-generated technology, AI-generated works	US, global
Conversational generative AI	Midjourney	Users and contributors (the creators of the scraped data)	Active and passive	Co-generated technology, AI-generated works	US, global

Remunerated work, Karya

Scenario	Case study	Nature of the co-generators	Level of involvement of the co-generators	Type of co-generation	Legal jurisdiction of the data holder and the co-generators
Remunerated work	Karya	Data labellers	Very active	Co-generated technology	India, global

Key takeaways:

- The legal frameworks for IP and data rights in connection with the rights of data labellers to the datasets they contribute towards are largely adequate, creating clarity for those utilising the datasets for training AI models.

- Many workers in the data labelling industry worldwide are subject to poor working conditions and low pay. While not directly related to co-generation rights, given the importance of accurately labelled datasets to AI technologies, there is a need to consider the wellbeing of workers in the creation of co-generated technology such as AI.

Remunerated work is where individuals are paid for their work contributing to digital systems and services. Availability of accurately labelled data is one of the biggest obstacles to AI adoption in industry.[41] Remunerated work is a common approach used for data labelling, speech recognition and content moderation.[42] Remunerated work tends to be carried out by third party organisations, who are contracted by technology companies to deliver these services. It is often poorly paid and extractive, with a history of problematic work cultures.[43] Remunerated work employers are often (but not always) located in the Global South, allowing companies to obtain cheaper labour with less stringent local employment obligations (or enforcements).

The process of data or image labelling, or contributing your voice to a training dataset for machine learning, is an important part of training AI models. These datasets can be used to train models to translate underrepresented languages, or to develop image classifiers. It is of particular interest for this research due to the power dynamics between the workers and the companies which buy their labour. The question of rights over data for data labellers is an interesting dynamic to explore, especially with the ever-growing need for new data to train AI-based models.[44]

[41] Chui, M. et al. (2018). McKinsey. 'What AI can and can't do (yet) for your business'. https://www.mckinsey.com/capabilities/quantumblack/our-insights/what-ai-can-and-cant-do-yet-for-your-business

[42] Dzieza, J. (2023). The Verge. 'AI Is a Lot of Work'. https://www.theverge.com/features/23764584/ai-artificial-intelligence-data-notation-labor-scale-surge-remotasks-openai-chatbots

[43] Perrigo, B. (2022). Time. 'Inside Facebook's African Sweatshop'. https://time.com/6147458/facebook-africa-content-moderation-employee-treatment/

[44] Gomer, R. & Simperl, E. (2020). Cambridge University Press. 'Trusts, co-ops, and crowd workers: Could we include crowd data workers as stakeholders in data trust design?'. doi:10.1017/dap.2020.21

Karya, India

Karya is a data labelling and annotation service for machine learning models set up in 2018. Karya enables people from rural India to earn money by annotating images and videos and reading out passages of text in native languages to create audio datasets. These activities are mediated through a mobile application where workers are provided with tasks. Karya provides clients with 'high-quality data annotation services for AI/ML models across various types of media',[45] but also has created a data catalogue for datasets which are owned by the community. If datasets for annotation are not provided by the client, Karya engages workers to build the dataset (such as a speech dataset of a local dialect of a native language).

Karya seeks to be an ethical alternative to other, more extractive, data labelling services:[46]

- It pays its contributors significantly better than many other services,[47] at a minimum of $5 (US dollars) per hour, over 20 times the Indian minimum wage.

- It has developed a Public Data License which allows workers to profit from repeated sale of a dataset they have contributed to. This is explored further in the rights section of this scenario.

- It supports workers by upskilling them, and is seeking to develop a data collective model to reap the benefits of collective ownership and rights.

Co-generator stakeholders and involvement

The following entities are involved in the data annotation and labelling, and dataset creation:

- **Karya**. Karya collects the requirements from clients, breaks down the requirements into digital tasks, provides these tasks to workers, collects the data from them, validates the data and synthesises the data into AI/machine learning (ML) training datasets.

- **The client**. Clients actively contract Karya, and its workers, to tag or annotate a specific dataset. Clients may provide datasets for the Karya workforce to work on or purchase labelled datasets from Karya.

- **Karya workers**. The workers are paid for tagging content on their mobile apps or for providing speech data by reading out sentences provided through the app. To date, over 30,000 workers have completed 30 million paid digital tasks through Karya.[48]

[45] Karya (n.d.). https://www.karya.in/.
[46] Dzieza, J. (2023). The Verge. 'AI Is a Lot of Work'. https://www.theverge.com/features/23764584/ai-artificial-intelligence-data-notation-labor-scale-surge-remotasks-openai-chatbots
[47] Karya (2022). 'The Future is Data (Cooperatives)'. https://karya.in/resources/blog/the-future-is-data/
[48] Karya (n.d.). 'Impact'. https://karya.in/impact

Rights

Karya is incorporated as a private limited company under Indian law and, as such, Indian law applies to the work carried out by Karya and the data labellers Karya employs.

Data rights

Karya collects basic personal information, including voice data, from workers in order to establish the contribution of each worker. Such data is protected under the Digital Personal Data Protection Act 2023 (DPDPA) and can only be collected by Karya after obtaining the specific consent of the workers, which must include the purpose for which this data is being obtained. In line with the DPDPA, the workers have the right to obtain from Karya a summary of the personal data that Karya has collected and any third party that Karya may have shared this data with. Further, the workers have the right to request the erasure of this data.

Through the completion of tasks on the app, workers create two types of data for Karya: annotated data and new speech data. Karya notes that it takes steps to ensure that this data is non-personal in nature and, as such, it will not fall under the ambit of the DPDPA. Given that Karya also engages with clients that are based outside of India, relevant data protection legislations for those jurisdictions might also apply. Datasets provided to Karya by clients and datasets provided by Karya to its clients do not contain any personal data. Karya has not yet engaged with any clients based in the EU.

Intellectual property rights

The databases created by Karya involve contributions from the workers. However, individual workers do not create an entire database; rather they engage with tasks on the app, the results of which are then collected, aggregated and verified by Karya. Based on this, Karya has a copyright over the database that is created by operation of the Copyright Act 1957,[49] displaying creativity and originality in the creation of the database. The rights in this database are licensed to the client. Depending on the nature of the agreement between the client and Karya, this licence is either exclusive or not. For databases that are non-exclusive, Karya uses a licence that allows for the resale and use of the database. In some cases, Karya also makes the databases openly available.

It is unlikely that the labelling contributions of the individual workers give rise to copyright as they do not constitute an original work. Under Indian law, copyright in speech recordings typically vests with the 'producer' of the sound recording, who is defined as 'the person who takes the initiative and responsibility for making the work'.[50] Karya is more likely to be recognised as the producer in this case, although there is scope to argue that the speech recordings are a work of joint authorship. The rights of the workers, vis-a-vis the speech recordings, are likely to be governed by contracts between Karya and the workers. None of these contracts are in the public domain, and we are unable to access them.

[49] Copyright Act, 1957, Section 13. https://www.indiacode.nic.in/bitstream/123456789/1367/1/A1957-14.pdf
[50] Copyright Act, 1957, Section 2(uu). https://www.indiacode.nic.in/bitstream/123456789/1367/1/A1957-14.pdf

Workers could possibly assert moral rights over the speech recordings, which are non-assignable and inalienable under Indian law.[51] The workers could claim damages if distortion, mutilation, modification or such other acts would be prejudicial to their honour or reputation.[52]

Contract law

Without seeing the contracts in place between Karya and its workers, and Karya and its clients, it is unclear how any existing IP and other rights are addressed.

Karya maintains it pays its workers significantly more than other employers in the industry. Its website states that it aims to operate as a data cooperative and that it creates worker-owned datasets so that its workers earn royalties when their data is sold. This suggests that their rights in such datasets are recognised but it is unclear whether this is on the basis of any acknowledged IP or data rights.

Contracts between Karya and its clients presumably licence access to and use of the data in the database, as well as terms providing or excluding warranties (and liability) as regards data quality and IP rights.

Uncertainties and gaps

Extrapolating from the above analysis, workers in the data labelling industry are likely to have personal data rights in relation to the data collected about them by platforms or companies that engage them. Subject to individual circumstances, they are unlikely to have any copyrights in the databases they contribute to, based on how the tasks are divided between workers and how the databases are created. However, the focus on rights of workers does not lie in data or IP rights, but in other frameworks such as labour rights.

Data labellers are a very important component of the supply chain of development of generative AI models. The datasets that the labellers create are used to train AI models. The licensing which Karya uses on its datasets ensures that data is accessible and available to use (either freely or at the right price). While the labellers are typically unlikely to have any rights over the datasets themselves, there is a need to consider their general wellbeing.

The Karya case study is not typical of the data labelling industry.[53] Research suggests that many workers in the data labelling and content moderation industry worldwide are engaged on considerably less favourable terms by their digital labour platforms.[54] Typical concerns are around low pay, opaque and difficult to challenge algorithmic labour management, and insufficient work or rejection of work with limited powers of redress. In most data labelling scenarios, there is a need to

[51] K&S Partners and J. Sagar Associates (2021). Thomson Reuters. 'Intellectual property right assignments Q&A: India'.
https://www.jsalaw.com/wp-content/uploads/2021/07/Intellectual-property-right-assignments-QAndA-India.pdf
[52] Copyright Act, 1957, Section 57. https://www.indiacode.nic.in/bitstream/123456789/1367/1/A1957-14.pdf
[53] Perrigo, B. (2022). Time. 'Inside Facebook's African Sweatshop'.
https://time.com/6147458/facebook-africa-content-moderation-employee-treatment/
[54] International Labour Office (2018). 'Digital labour platforms and the future of work: Towards decent work in the online world'.
https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/---publ/documents/publication/wcms_645337.pdf

re-evaluate the rights frameworks that govern these workers, and push for rights frameworks that contend with the unfair power dynamics, and provide workers with rights, including a living wage and basic labour and human rights protections. While efforts to address these issues are already underway, with codes of ethics[55] and principles[56] for data labelling emerging, there is a need for stronger regulatory frameworks that protect the rights of workers. Given the importance of accurately labelled datasets for AI technologies, ensuring the welfare of these workers can also have a bearing on innovation and development.[57]

Crowdsourced data, OpenStreetMap

Scenario	Case study	Nature of the co-generators	Level of involvement of the co-generators	Type of co-generation	Legal jurisdiction of the data holder and the co-generators
Crowdsourced data collection	OpenStreet Map	Contributors to OpenStreetMap	Very active	Co-generated data	UK, global

Key takeaways:

- OpenStreetMap (OSM) clearly provides the mechanism for how contributors can exercise their data rights in their Privacy Policy.
- With regard to their IP rights, given that OSM is an open source, free, collaborative database, the current legal framework that governs the rights of most contributors is largely adequate.
- Contributions from the OSM community help refine OSM for both the community and for society at large – improving OSM as a free, open source alternative that can be used for a variety of purposes for public good.

Crowdsourced data collection[58] is a participatory method of building a dataset with the help of a large group of people. This approach has been used, for instance, to map biodiversity, create the largest body of knowledge in the world in Wikipedia, and even to find potholes on London roads.

Crowdsourced data is an example of co-generation because individuals and communities are directly involved in the generation of new data. Involvement of co-generators happens in different ways – people may use bespoke applications or sensors to collect new data, or they may also come together at datathons where they generate data collectively. Crowdsourced data is a direct product

[55] Kayra (n.d.), 'Ethical Data Pledge'. https://www.ethicaldatapledge.com/
[56] GPAI (2023). Global Partnership on AI. 'Fairwork AI Ratings 2023: The Workers Behind AI at Sama'. https://gpai.ai/projects/future-of-work/FoW-Fairwork-AI-Ratings-2023.pdf
[57] GPAI (2023). Global Partnership on AI. 'AI for Fair Work, From principles to practices'. https://gpai.ai/projects/future-of-work/FoW2_AI%20Fair%20Work%20.pdf
[58] World Bank (n.d.). 'Crowd-sourced Data'. https://dimewiki.worldbank.org/Crowd-sourced_Data

of many individual contributors who add to, edit and improve the contributions of others. Generally, contributors are required to create an account and agree to terms and conditions of service before they can add to or amend content, although sometimes contributions can be made without signing up to any terms and conditions.

OpenStreetMap, UK

OpenStreetMap is a collaborative database of geospatial data. Launched in 2004, OSM is built by a community of mappers that contribute and maintain data about 'roads, trails, cafés, railway stations, and much more, all over the world'.[59] OSM data is widely used to create digital maps, as well as for navigation, converting GPS coordinates to addresses, and more. Users can contribute information about different services on the map, update street names as well as mapping areas which are not yet covered in the database. Because of the speed at which maps can be updated and edited based on local needs, OSM has played a significant role in disaster response, even spawning another organisation, Humanitarian OpenStreetMap.[60]

Unlike its main competitor, Google Maps, OSM data is available openly: users (including businesses) are free to use it for any purpose as long as they credit OSM and its contributors. This includes the ability to gain profits for private enterprises from the use of OSM.

Co-generator stakeholders and involvement

The following entities are involved in the creation of the ultimate asset – geospatial data – and are therefore co-generators:

- **OpenStreetMap**. OSM created and hosts the infrastructure for people to contribute to the database.

- **The contributors**. In updating information in the OSM database, the contributor is actively generating new data. Contributors to OSM have an account and sign up to the Licence/Contributor Terms, which specify that data belongs to 'the user and the OSM community, and is free and open for everybody, under a creative commons licence'.[61] The process of an individual updating data means that previously inputted data is removed or updated. Contributors will likely be unaware of any changes made to data they have previously provided, but can use online tools like 'Who Did It?'[62] to find out. If there have been incorrect changes, contributors are able to reverse these changes.

- **Users of OSM ('users')**. People use OSM as alternative mapping software to Google Maps. They can access the map via a browser, but there is not currently an OSM app. They can

[59] OpenStreetMap (n.d.). 'About'. https://www.openstreetmap.org/about
[60] Herfort, B. et al. (2021). Scientific Reports. 'The evolution of humanitarian mapping within the OpenStreetMap community'. https://doi.org/10.1038/s41598-021-82404-z
[61] OpenStreetMap Foundation (n.d.). 'Licence/Contributor Terms'. https://wiki.osmfoundation.org/wiki/Licence/Contributor_Terms
[62] WhoDidIt (n.d.). https://simon04.dev.openstreetmap.org/whodidit/

access an app version via a third party supplier. The use of OSM by these users generates data that is collected and used by OSM.

- **Organisations using OSM in their services**. Many organisations use OSM data, including big tech companies like Apple and Amazon, and video games like Pokemon Go. They access OSM data via an application programming interface (API), and it is free as long as they credit OSM and the contributors. It is worth noting that there have been concerns that more and more commercial interests are involved in contributing to OSM as they try to update the data to match their needs, and subsequently there are worries about the impact this has on the dataset.[63]

Legal frameworks

The legal entity behind the OSM project is the OpenStreetMap Foundation, a not-for-profit organisation registered as a company in the UK.

Data rights

OSM collects data from co-generators who use OSM including their IP address, browser and device type, and operates user interaction tracking software.[64] This data is used to improve the OSM dataset, and for research purposes (as anonymised and summarised data), alongside other uses. In this sense, users of OSM (including organisations using OSM services via the API) can be considered as co-generators.

As a legal entity registered in the UK, the primary applicable legal framework for the personal data rights of co-generators who use OSM (hereafter called users) is set out under data protection laws: for UK contributors and users, the UK's data protection regime including the General Data Protection Regulation 2016/679 as implemented in UK law (UK GDPR) and the Data Protection Act 2018 (DPA 2018); for EU contributors and users, the EU data protection regime including the General Data Protection Regulation 2016/679 (GDPR); and for contributors and users in other jurisdictions, the local data protection laws that apply there. These rights typically include rights of access, correction, transfer and erasure. These rights are limited to personal data only.

Intellectual property rights

OSM is open data,[65] licensed under the Open Data Commons Open Database Licence.[66] According to this licence, anyone is free to copy, distribute, transmit and adapt data from OSM, as long as they attribute OSM and its contributors. This also applies to data created by altering or building upon OSM's data. IP rights that subsist in such data may include:

- Copyright in original works, eg, images, drawings, cartography, and in underlying code and algorithms.

[63] Anderson, J. et al. (2019). International Journal of Geo-Information. 'Corporate Editors in the Evolving Landscape of OpenStreetMap'. http://dx.doi.org/10.3390/ijgi8050232
[64] OpenStreetMap Foundation (n.d.). 'Privacy Policy, Data we receive automatically'. https://wiki.osmfoundation.org/wiki/Privacy_Policy#Data_we_receive_automatically
[65] OpenStreetMap (n.d.). 'Copyright'. https://www.openstreetmap.org/copyright
[66] Open Knowledge Foundation (n.d.). 'Open Data Commons Open Database License (ODbL)'. https://opendatacommons.org/licenses/odbl/

- Trademarks and designs – in the logos or names of service providers, shops and public institutions shown on maps.

- Patents in the underlying infrastructure that OSM hosts the platform on (although note that patents for software are rare and software is more likely to be protected by copyright and confidentiality).

Contributions made to OSM are governed by the OpenStreetMap Contributor Terms.[67] As per these terms, the contributor, in contributing to OSM grants the OpenStreetMap Foundation a worldwide, royalty-free, non-exclusive, perpetual, irrevocable licence to any contribution they make, with such contribution allowed to be used for commercial purposes. Additionally, it is stated that the foundation will sub-license the contributions as part of a database and only under the terms of an open database licence or any other free and open licence that is determined by a 2/3 majority vote of active contributors. At the contributor's request, the foundation will attribute the contributor.

These terms are in line with OSM's stated purpose of being an open source, free, collaborative database of geospatial data. This set of co-generators, ie, the contributors, are made aware of the nature of their contributions prior to such contributions being made. Only active contributors can control the foundation's rights to sub-license those rights. Non-active contributors (someone who cannot demonstrate edits on OSM in the last 12 months, does not have a valid email address, and who fails to respond to a request to vote within three weeks) do not have a voting right in determining the licence under which the foundation makes OSM available. Additionally, OSM lists and attributes publishers of data that have allowed OSM to use the data (typically under an open licence).[68]

One additional point to note here is that data from contributors is likely to contain data of individuals, legal persons and communities, or have impacts on them. In this regard, such data contribution involves further co-generation. There is merit in exploring legal frameworks for co-generation in such situations, where these frameworks will not apply to OSM directly but to the individuals or entities that are sharing the data. The Licence/Licence Compatibility OSM guide sets out guidelines for the licences that may apply to data sources utilised for adding objects to OSM,[69] which highlights the legal rights framework that applies in relation to third-party rights in data or other content that is uploaded by contributors. It is notable that the original rights-holder may not be aware of the use of their proprietary content by a contributor, and so may be unaware that this legal rights framework applies to and is accessible by them. This framework for co-generation should be further explored.

Contributions can include data and other content in which IP rights may subsist. This could include:

- **Algorithms**. The Licence/Community Guidelines set out that contributors can make changes algorithmically. The underlying code and algorithms themselves may be protected by copyright if they are original works.

[67] OpenStreetMap Foundation (n.d.). 'Licence/Contributor Term'. https://wiki.osmfoundation.org/wiki/Licence/Contributor_Terms
[68] OpenStreetMap (n.d.). 'Contributors'. https://wiki.openstreetmap.org/wiki/Contributors
[69] OpenStreetMap (n.d.). 'Licence/Licence Compatibility'. https://wiki.osmfoundation.org/wiki/Licence/Licence_Compatibility

- **Produced works** (defined in the Licence/Community Guidelines/Produced Work – Guideline as images, audiovisual, text, sounds). This content could be protected by copyright, and by trademarks and rights in passing off if logos or product or service names are used in this content when provided by a contributor.

- **Databases**. Content that contributors upload may be extracted from existing databases in which the owner may benefit from copyright and/or database rights.

- **Confidential information**. The OSM is open source, so it is unlikely that contributors will upload content if they consider such content to be their own confidential information or trade secrets. There is a risk that contributors could contribute content which a third party would consider to be confidential information, for example, if a contributor uploads code or an algorithm they developed during the course of their employment (and therefore their employer owns or has rights in it) or a contributor uploads an image that shows the inside of a glass building.

- **Patents**. These are unlikely to be relevant in relation to the rights of contributors to OSM.

There are existing rights frameworks in relation to a contributor's ownership of, or ability to use (eg through licensing from third parties), such rights.

Contract law

The Terms of Use[70] note that English law will govern the terms of use of OSM as well as any dispute that may arise between OSM and the contributor or user. Incorporated in the Terms of Use is the Privacy Policy[71] of OpenStreetMap Foundation, which notes that the data collected from users of the service falls under the legitimate interest lawful basis under Article 6.1 (f) of the UK GDPR.

Uncertainties and gaps

Given that OSM is an open source, free, collaborative database, the current legal framework that governs the rights of most contributors is largely adequate. The gap arises in the level of input that non-active contributors (as defined by OSM) have in deciding the licensing regime under which OSM is shared. Additionally, data generated by users of OSM (who are not contributors) is also used to refine OSM's operation. While such users are co-generators, giving such co-generators rights in regard to this data used by OSM is a point of contention. Some legal experts who were interviewed as part of this research put forth the opinion that rights in such cases should be given to users only in cases where the data generation was intentional, and not incidental to the use of a product or service.

The adequacy of legal frameworks in this case study is evidenced by the thriving ecosystem of crowdsourced, open databases and tools created based on the frameworks. They provide the guide rails for co-generation which fairly supports each stakeholder. These resources and technologies have had a large positive impact for social good and innovation, by enabling new advances in technology. OSM itself is a clear example as it is freely used by many organisations who may

[70] OpenStreetMap Foundation (n.d.). 'Terms of Use'. https://wiki.osmfoundation.org/wiki/Terms_of_Use
[71] OpenStreetMap Foundation (n.d.). 'Privacy Policy'. https://wiki.osmfoundation.org/wiki/Privacy_Policy

otherwise not be able to afford to use other online maps, enabling them to build new software, design new services and make better decisions.

Social media platforms, Instagram

Scenario	Case study	Nature of the co-generators	Level of involvement of the co-generators	Type of co-generation	Legal jurisdiction of the data holder and the co-generators
Social media platforms	Instagram	Social media users	Active and passive	Co-generated data, co-generated technology	US, Brazil

Key takeaways:

- Instagram's privacy policy is aligned with the applicable laws to govern the whole range of collected user data.
- However, the legal framework is inadequate as it relies on the notice and consent mechanism in the collection of data and its extensive processing and sharing. The failures of the notice and consent mechanism are well documented, and very often users are unaware that their data from Instagram is being used to train AI models.
- Users are also entitled to protection in respect of 'damages arising from content generated by third parties (that infringes on) copyright or neighbouring rights'.[72] The platform bears civil liability if it fails to take appropriate measures for taking infringing content down. Tracking and reporting such breaches, however, becomes tricky due to the scale of the platform.

Social media platforms have grown in size over the past two decades and are now some of the biggest and most profitable companies in the world.[73] Users interact with these platforms by posting updates, videos or photos. Each individual user has a different experience with these platforms as every user's feed is curated algorithmically. Feeds are informed by data generated by the users themselves (eg, who they follow), and by the platforms about the users (eg, interests derived from activity), both of which ultimately inform the algorithms which underpin the systems and personalised feeds. The feeds are co-generated, not only by the users and the platform, but also by the content that is pulled in from other users. Data gathered from users is also used for psychographic analysis, which allows platforms to display targeted adverts and sponsored content to users.[74]

[72] The Brazilian Civil Framework of the Internet. https://bd.camara.leg.br/bd/bitstream/handle/bdcamara/26819/bazilian_framework_%20internet.pdf
[73] Haqqi, Ty. (2023). Yahoo! Finance. '25 Most Profitable Companies in the World'. https://finance.yahoo.com/news/25-most-profitable-companies-world-192146617.html?
[74] Isaak, J. & Hanna, M. J. (2018). IEEE. 'User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection'. https://ieeexplore.ieee.org/document/8436400

Instagram, Brazil

Instagram is one of the largest social media platforms in the world, with over 2.35 billion-active users. Instagram is owned by Meta, which bought the platform for $1bn (US dollars) in 2012. The app allows users to upload media that can be edited with filters, be organised by hashtags, and be associated with a location via geographical tagging. Instagram currently has a number of different formats for people to share content (both videos and images), including stories (posts which only last for 24 hours, although they can be stored) and reels (TikTok-style short videos).

Co-generator stakeholders and involvement

Instagram involves multiple co-generators through the use of the service. This includes:

- **Instagram**. In addition to providing the service, Instagram curates content to individual users. The Instagram Explore page uses an algorithm to curate content based on a user's interests, behaviour and engagement patterns, which likely takes into account factors such as the accounts followed, the posts liked, the hashtags used, and the content interacted with. In addition, Instagram displays adverts to users based on an analysis of data about them. Instagram stores all of the data and information on the Meta cloud servers.[75]

- **Co-generators who use Instagram ('users')**. Users of Instagram can search and scroll through existing content, comment on others' content and post their own content such as photos and videos. Users can be individuals, businesses or even communities. Users are sometimes paid for their posts on Instagram if they have a certain number of followers, or post product placements for brands.[76]

- **Advertisers whose content is displayed on Instagram**. Posts from advertisers on Instagram can link directly to the advertiser's website. It does not appear that the advertiser needs to have an account on Instagram for their advertisements to appear. Posts from advertisers are typically matched up with users' interests to increase engagement.

Rights

Instagram is available all over the world, and in each jurisdiction users of Instagram will be subject to different rights and legislation based on their location. This scenario will focus on the rights of users in Brazil.

Data rights

[75] Instagram Engineering (2015). 'Instagration Pt. 2: Scaling our infrastructure to multiple data centers'. https://instagram-engineering.com/instagration-pt-2-scaling-our-infrastructure-to-multiple-data-centers-5745cbad7834

[76] Instagram (n.d.). 'Earn money on Instagram'. https://creators.instagram.com/earn-money?locale=en_GB

The data rights frameworks for Instagram in Brazil are composed of three main items: the Brazilian Civil Framework of the Internet (Marco Civil);[77] the Lei Geral de Proteção de Dados Pessoais (LGPD) (in English: General Personal Data Protection Act);[78] and Instagram's terms and policies.[79]

The application of these frameworks must be analysed to understand their impact on the rights and duties of the various co-generators participating on the platform. A user's right of privacy is safeguarded in both the Marco Civil and the LGPD. The latter recognises additional data rights for individuals, with provisions governing the rights and obligations around access, collection and processing of personal data. However, since the translation of these rights for users of Instagram is determined on the basis of the platform's data and privacy policies, the efficacy of the framework may be questioned.

Instagram's data policy stipulates that by using Instagram, users consent to the collection of a wide range of data.[80] This includes comments as well as content (photos and videos) posted by users, types of content, including adverts that users view or how they interact with them, as well as the time, frequency and duration of a user's activities on Instagram. Instagram also collects information from the user's app or device including what the users are doing on the device (for instance, whether the app is in the foreground, or if the user's mouse is moving) and information the users have shared through device settings, like GPS location.

In addition to the information collected about users from their own device, Instagram also collects data about users from third parties, including cookie data from external links and social plugins about the other apps and games from the device. This external data collection could also extend to data inferred about users based on others' activity to place them within contact lists and suggested groups.

The data collected from users is used to personalise features, content and recommendations, and to improve Meta's products, such as a user's Instagram feed and adverts shown by Instagram itself. User data generated on Instagram is also used by Meta to train generative AI,[81] an issue currently undergoing considerable debate, both in the context of the use of third-party IP rights and the use of personal data. This of course falls within the realm of co-generated technology.

While Instagram's data policy states that Instagram never sells user information to third parties, it does state that the data is shared with the Meta Group Companies, and integrated partners that are selected by users.[82] Additionally, aggregated data on user activities, advert engagement and use of partner products is provided to advertisers and audience network publishers, vendors, and Meta's organisational service providers.

---

[77] The Brazilian Civil Framework Of The Internet. https://bd.camara.leg.br/bd/bitstream/handle/bdcamara/26819/bazilian_framework_%20internet.pdf
[78] Lei Geral de Proteção de Dados Pessoais (LGPD). English translation accessed here: https://iapp.org/resources/article/brazilian-data-protection-law-lgpd-english-translation/
[79] Instagram (2023). 'Terms of use, Instagram'. https://help.instagram.com/581066165581870/?helpref=hc_fnav
[80] Instagram (2022). 'Data policy, Instagram'. https://help.instagram.com/155833707900388
[81] Meta (n.d.). 'How Meta uses information for generative AI models'. https://privacycenter.instagram.com/privacy/genai/
[82] Instagram (2022). 'Data policy, Instagram'. https://help.instagram.com/155833707900388

Intellectual property rights

Instagram does not claim ownership of the content posted by users, but by signing up to use Instagram, users provide Instagram (specifically, Meta) with a non-exclusive, royalty-free, transferable, sublicensable, worldwide licence to host, use, distribute, modify, run, copy, publicly perform or display, translate and create derivative works of the content.[83] This licence ends only when the content has been deleted from Instagram's system. It is also stated that under this licence Instagram has the right to share content posted on Instagram with third parties. However, under Article 8 of the LGPD, doing so would require explicit consent from the users. This policy, specified in the Terms of Use, essentially determines the exercise of IP rights for co-generated data on the platform, allowing Instagram the right to curate users' posts for marketing purposes or towards improving engagement.

Users' IP rights are, however, given some protection not just from Instagram's use of the content, but also by being hosted on the platform. Article 19(2) of the Marco Civil prescribes civil liability on platforms 'for damages arising from content generated by third parties (that infringes on) copyright or neighbouring rights' if they fail to take appropriate measures for taking infringing content down. In such a context, Instagram community guidelines also attempt to account for potential infringement of IP rights of the users by the activities of another.[84] These contain directions against unfairly posting content that may infringe upon third party copyrighted work or trademarks, and a repeat infringer policy which means an account may be disabled or removed from the platform for multiple violations.

Considering the limitations of any individual users in tracking the misuse of their content, and the nature of campaigns and trends on a social media website, these policies may not provide true protection from a co-generation point of view. Accordingly, details on the boundaries of trademark use or copyright licences for collaborative pieces would be better stipulated under bilateral or multilateral contracts between parties to share any platform-generated remuneration, and for big and small business to have general policies around corporate use of social media to protect the branding of assets and trademark use.

Instagram's Terms of Use explicitly state that users are not allowed to post content that would infringe the IP rights of other parties. By posting on Instagram, users represent that they either own or have rights to the content they are sharing.[85] Developing jurisprudence around IP rights over AI-generated content will therefore impact the ability of users to post AI-generated content on Instagram.

Uncertainties and gaps

Given the primacy of consent in the collection of data and its extensive processing and sharing in relation to such platforms, the current legal framework is inadequate from a co-generation perspective. The failures of the notice and consent mechanism are well documented, and consent fatigue means that consent is usually not informed and does not always equal awareness.[86]

[83] Instagram (2023). 'Terms of use, Instagram'. https://help.instagram.com/581066165581870/?helpref=hc_fnav
[84] Instagram (2023). 'Community Guidelines, Instagram'. https://help.instagram.com/477434105621119/?helpref=faq_content
[85] Instagram (2023). 'Terms of use, Instagram'. https://help.instagram.com/581066165581870/?helpref=hc_fnav
[86] World Economic Forum (2020). 'Redesigning Data Privacy: Reimagining Notice & Consent for human technology interaction'. https://www3.weforum.org/docs/WEF_Redesigning_Data_Privacy_Report_2020.pdf;

Additionally, when data and analytics are aggregated and shared across advertisers, vendors and partner organisations, in a manner that affects the same users with algorithmic targeting techniques, risks around profiling and polarising propagandas are intensified and must be accounted for via comprehensive content moderation mechanisms. An additional point of concern is that content posted by a user on Instagram is likely to contain personal information of others in it. This can have implications for the rights of such others in a variety of ways including reputational harm, as well as breach of their privacy rights, with cases of law enforcement agencies having used social media information to track protestors.[87]

While the Marco Civil provides stipulations on the kind and manner of hosting content on the internet, the larger institutional framing on intermediary liability and platform regulation has been under review in Brazil with the proposed Fake News Bill.[88] This version of the law attempts to address rules for risk assessment and duty of care obligations for the platform. It has been criticised for enabling vaguely-defined crisis protocols, and impacting human rights with arbitrary risk assessment and mitigation obligations, as well as broadly criminalising the dissemination of 'untrue facts'.[89]

Internet of Things (IoT), BMW and Europcar

Scenario	Case study	Nature of the co-generators	Level of involvement of the co-generators	Type of co-generation	Legal jurisdiction of the data holder and the co-generators
Internet of Things	BMW connected cars owned by Europcar	Occupants of a care	Passive	Co-generated data	EU
Key takeaways: • The existing legal frameworks around data and IP rights are largely adequate.					

Centre for Communication Governance at National Law University Delhi (n.d.). 'Comments to Niti Aayog on the Draft Discussion Paper on the Data Empowerment and Protection Architecture'. https://ccgdelhi.s3.ap-south-1.amazonaws.com/uploads/ccg-nlu-comments-to-niti-aayog-on-the-draft-discussion-paper-on-the-data-empowerment-and-protection-architecture-238.pdf; Solove, D. J. (2013). Harvard Law Review. 'Privacy Self-Management and the Consent Dilemma'. https://scholarship.law.gwu.edu/cgi/viewcontent.cgi?referer=&httpsredir=1&article=2093&context=faculty_publications

[87] Cooke, K. (2016). Reuters. 'U.S. police used Facebook, Twitter data to track protesters: ACLU'. https://www.reuters.com/article/us-social-media-data-idUSKCN12B2L7

[88] Institui a Lei Brasileira de Liberdade, Responsabilidade e Transparência na Internet (2020). 'Parecer Proferido Em Plenário Ao Projeto De Lei Nº'. https://www.camara.leg.br/proposicoesWeb/prop_mostrarintegra?codteor=2265334&filename=Tramitacao-PL%202630/2020

[89] Alimonti, V. et al. (2023). Electronic Frontier Foundation. 'Settled Human Rights Standards as Building Blocks for Platform Accountability and Regulation: A Contribution to the Brazilian Debate'. https://www.eff.org/deeplinks/2023/07/settled-human-rights-standards-building-blocks-platform-accountability-and

- The primary issue of concern is around the extent of rights (and benefits) for occupants of the vehicle who generate the data, especially given that this data is typically generated merely as a by-product of their use.
- According rights to all occupants for generation of data in all cases would likely have a detrimental impact on research and innovation, both for private gain as well as larger public good.

The Internet of Things (IoT) refers to a 'network of physical devices, vehicles, appliances and other physical objects that are embedded with sensors, software and network connectivity that allows them to collect and share data'.[90] IoTs are commonly known as 'smart [object]', and can range from home appliances to industrial machinery.

IoTs are of particular interest to this research due to the involvement of individuals and multiple entities (through multiple sensors and devices) in the generation of data via technology, and the rights of both the creator of the technology and the individuals over that data. This is particularly important in the EU, where conversations about the Data Act have been heavily focused on access to IoT data and the rights of the co-generators.[91]

BMW cars owned by Europcar, France

This scenario will focus on connected vehicles: cars which use IoT devices to collect data as they are driven and share it via an internet connection. Modern vehicles generate around 25 gigabytes of data every hour.[92] Autonomous cars, or self-driving vehicles, will generate even more – as much as 40 terabytes of data an hour from cameras, radars and other sensors, according to expert forecasts.[93] There is a growing market for data generated from connected vehicles from the manufacturers to insurance companies.[94]

Vehicles can be connected to the internet in two ways, either via an embedded connection (and a chipset and antenna built into the vehicle) or via a tether (connected to a user's phone).[95] They can communicate with each other to share live information about how fast they are moving, direction of travel and more. They can also communicate with the cloud to share diagnostic reports, location

[90] IBM (n.d.). 'What is the internet of things?'. https://www.ibm.com/topics/internet-of-things
[91] Kahl, T. (2023). TaylorWessing. 'Mobility is going digital!'. https://www.taylorwessing.com/en/interface/2023/iot---next-gen/mobility-is-going-digital-what-connected-vehicle-manufactures-need-to-think-about-in-2023
[92] McFarland, M. (2017). CNN. 'Your car's data may soon be more valuable than the car itself'. https://money.cnn.com/2017/02/07/technology/car-data-value/index.html
[93] Naughton, K. (2021). Bloomberg. 'Driverless Cars' Need for Data Is Sparking a New Space Race'. https://www.bloomberg.com/news/articles/2021-09-17/carmakers-look-to-satellites-for-future-of-self-driving-vehicles
[94] Keegan, J. & Ng, A. (2022). The Markup. 'Who Is Collecting Data from Your Car?'. https://themarkup.org/the-breakdown/2022/07/27/who-is-collecting-data-from-your-car
[95] Bull, A. (2022). High Mobility. 'What is a Connected Car?'. https://www.high-mobility.com/blog/what-is-a-connected-car

data and more. Data is generated by occupants of the car through their use of the car. This data can be personal or non-personal (or mixed); it can be individual level or aggregate level data.

Connected vehicles also involve the use of third party products and services, ranging from sensors to the cloud to software. For example, most connected vehicles typically use software for their infotainment system that is provided by a third party, such as Google or Apple. These third parties are also co-generators in such cases.

In this case study, we will examine the applicability of legal frameworks for co-generation situations involving the renting of BMW connected cars from Europcar Mobility Group (Europcar) in France. Bayerische Motoren Werke AG, commonly known as BMW, is a German multinational vehicle manufacturer founded in 1916 with headquarters in Munich, Germany. The BMW Group also owns other car manufacturers like Mini and Rolls Royce. Europcar is a French car rental company founded in 1949 in Paris. It operates a fleet of over 200,000 vehicles at 3,300 locations in 150 countries.

Co-generator stakeholders and involvement

Connected vehicles involve a range of different co-generators. Use of a vehicle, either as a driver or a passenger (together called occupants), generates data that is collected by the manufacturer and rental company, as well as third party providers:

- **BMW is the manufacturer**, it develops and controls the systems in the car which collect the data (called ConnectedDrive), and has access to the data. BMW has collected over a billion kilometres of anonymised real-world driving data from its customers.[96] This data is collected with the drivers' permission, and 80% of BMW owners currently agree to share data from their vehicles.[97] This data is used to develop new driver-assistance features, as well as improving infotainment systems (the BMW infotainment system is called iDrive). BMW also sells this data to third parties as part of their CarData initiative. BMW collects a wide variety of data about or in relation to the use of cars by occupants, including:[98]
    - position of vehicle (including coordinates, orientation and altitude)
    - navigation destination
    - mileage
    - driving style
    - average speed
    - duration of charging and reason for interruption of charging (for electric vehicles).

  This information is stored on-board the car and transmitted directly to BMW. Europcar does not have access to this information.

[96] Atz, T. et al. (2023). Amazon Web Services. 'Scaling Automated Driving data processing and data management with BMW Group on AWS'. https://aws.amazon.com/blogs/industries/scaling-autonomous-driving-data-processing-and-data-management-with-bmw-group-on-aws/

[97] Motor 1 (2022). 'BMW uses customers' driving data to improve its in-car features'. https://uk.motor1.com/news/580364/bmw-collecting-customer-driving-data/

[98] BMW Group (2022). 'BMW CarData Telematics Data Catalogue'. https://www.bmwgroup.com/content/dam/grpw/websites/bmwgroup_com/innovation/Innovation_Mobilitaet/CarData/3PP-CarData--Telematics_Data_Catalogue-en.pdf

- **Europcar owns the vehicles and rents them**. Europcar also has its own sensors installed which collect a wide variety of data about or in relation to the use of the car by the occupants, including:[99]
    - vehicle performance data
    - operational and diagnostic data
    - mileage information
    - acceleration and braking speeds
    - vehicle location.

  In addition, Europcar also collects personal information such as driver's licence and financial information (for example, credit card) when the car is being rented.

- **Occupants of the vehicle**. In order to rent a car from Europcar, both the driver and Europcar must sign a rental agreement, which sets out the terms of the rental including information about what data is collected by Europcar.[100] Data regarding driving habits, vehicle location and the use of the infotainment system is generated as the vehicle is driven. Drivers are typically unaware of the generation of data: a 2020 study by Capgemini found that only 18% of drivers 'know' what data their vehicles transmit.[101] Data is also generated by occupants of the vehicle who are not the driver.

- **Third parties**. In addition to services provided directly by the vehicle manufacturer, most connected vehicles also have services provided by third parties, which result in data generation, with this data being collected by such third parties. BMW has their own infotainment system, iDrive, but also supports Android Auto, Apple CarPlay and Amazon Alexa Car Integration.[102]

Rights

For the purpose of this research, we are focusing on the legal regimes for connected vehicles in France. Connected vehicles are governed by both French and EU regulation.

Data rights

In so far as the aforementioned data is personal data (as it is information relating to an identified or identifiable natural person), the applicable legislation is the French Ordinance No. 2018-1125 which is a re-writing of the French Act No. 78-17 of 1978 (on Information Technology, Data Files and Civil Liberties) to incorporate provisions of the GDPR. In addition to this, the European Data Protection

[99] Europcar (n.d.). 'Privacy policy for connected vehicles'.
https://www.europcar.com/files/live/sites/erc/files/connected-cars/privacy-policy.pdf
[100] Europcar Germany (2016). 'Terms and conditions of hire of Europcar Autovermietung GmbH'.
https://oos.glasstec-online.com/medias/AVB-englisch-0617.pdf?context=bWFzdGVyfHJvb3R8MzIyODkyNnxhcHBsaWNhdGlvbi9wZGZ8aDMzL2gyNi85MTk0NjAxNzQyMzY2LnBkZnxiNjMzY2JlYjJmOTUxNjg5OTJhZTczOWQyMmY2MmU5ZWYzNTY2NzYwOWE1NWQ0MjA3ZmRhNDhjOGYyZmM1NWJh
[101] Capgemini (2020). 'Monetizing Vehicle Data: How to fulfil the promise'.
https://www.documentcloud.org/documents/22120767-capgeminiinvent_vehicledatamonetization_pov_sep2020#document/p10/a2130251
[102] BMW (n.d.). 'BMW ConnectedDrive'.
https://www.bmw.co.uk/en/topics/owners/bmw-connecteddrive/overview.html

Board (EDPB) notes in its guidelines on processing personal data in the context of connected vehicles and mobility related applications (guidelines) that the EU ePrivacy Directive (Directive 2002/58/EC as revised by 2009/136/EC) can also apply if data is transferred from the car through a SIM card installed in the car.[103] The ePrivacy Directive is applicable when all of the following conditions are met:

- there is an electronic communications service (ECS)[104]
- this service is offered over an electronic communications network
- the service and network are publicly available
- the service and network are offered in the EU.[105]

BMW notes that the data from the cars is transferred to servers operated by BMW through a private sub-network of a mobile service provider, accessible only to BMW vehicles.[106] Therefore, we do not consider the ePrivacy Directive to be applicable to this case study.

Both BMW and Europcar process personal data relying on the following bases as set out in their respective privacy policies:[107,108]

- legitimate interest as under Article 6(1)(f) of the GDPR (for purposes including product quality assurance, research and development of new products)
- for the performance of contractual obligations as under Article 6(1)(b) of the GDPR (for providing features such as Smart Emergency Call, Concierge Service, RTTI [Real Time Traffic Information], TeleServices)

[103] European Data Protection Board (2021). 'Guidelines: Guidelines 01/2020 on processing personal data in the context of connected vehicles and mobility related application'. https://edpb.europa.eu/system/files/2021-03/edpb_guidelines_202001_connected_vehicles_v2.0_adopted_en.pdf

[104] The term electronic communications service is defined in Article 2(c) of Directive 2002/21/EC (Framework Directive) as 'a service normally provided for remuneration which consists wholly or mainly in the conveyance of signals on electronic communications networks, including telecommunications services and transmission services in networks used for broadcasting, but exclude services providing, or exercising editorial control over, content transmitted using electronic communications networks and services; it does not include information society services, as defined in Article 1 of Directive 98/34/EC, which do not consist wholly or mainly in the conveyance of signals on electronic communications networks'.

[105] European Data Protection Board (2019). 'Opinion of the Board (Art. 64): Opinion 5/2019 on the interplay between the ePrivacy directive and the GDPR, in particular regarding the competence, tasks and powers of data protection authorities'. https://edpb.europa.eu/sites/default/files/files/file1/201905_edpb_opinion_eprivacydir_gdpr_interplay_en_0.pdf

[106] BMW (n.d.). 'Which mobile network does BMW use to transfer my data to third parties/ service providers?'. https://faq.bmw.com.au/s/article/Car-Data-Data-transmission-Mobile-network-9sdaZ?language=en_AU

[107] BMW (n.d.). 'Politique de confidentialité de BMW France'. https://www.bmw.fr/fr/footer/metanavigation/data-privacy.html

[108] Europcar (n.d.). 'Privacy policy for connected vehicles'. https://www.europcar.com/files/live/sites/erc/files/connected-cars/privacy-policy.pdf

- specific consent as under Article 6(1)(a) and Article 7 of the GDPR (for promotional communications and market research).

According to GDPR, occupants of the cars in this scenario have the rights of data access, correction, deletion and portability as set out under the GDPR. These rights have also been called out in the privacy policies of both BMW and Europcar, along with the relevant authority that can be contacted for asserting these rights. The French national data protection authority, Commission nationale de l'informatique et des libertés (CNIL), in its compliance package on Connected Vehicles and Personal Data, discusses these rights,[109] and also notes different scenarios of processing of personal data in relation to connected vehicles and what rights are triggered when.

The EU Data Act[110] imposes an obligation on manufacturers and designers of IoT products to share data with the users of those products and third parties nominated by the users in certain circumstances. The data holder is obliged to make the data available under fair, reasonable and non-discriminatory terms, and in a transparent way. Compensation agreed between a data holder and data recipient must be reasonable, and the basis must be explained. Where the recipient is a micro or small/medium sized enterprise (SME), the data holder can only charge the costs of making the data available, and the Act also addresses fairness of the wider contract terms with such businesses.

The EU Data Act clearly has the potential to have a significant impact on the mobility ecosystem, enabling innovation and competition in aftermarkets and other automotive or electric vehicle-related services, given the amount of data generated by connected vehicles. The Act establishes basic cross-sector horizontal data-sharing requirements. These include data concerning the performance, use or environment obtained, generated or collected by the smart car and related services during common use by the user (including embedded applications) and functions that indicate hardware status or malfunction; data generated during inaction/standby or switch-off mode; and raw data or pre-processed data. The Act does not extend to insights derived from raw data nor to certain products primarily designed to display or play content, or to record and transmit content, for use by an online service (eg, personal computers, smartphones, servers, webcams).

The Act seeks to balance rights of access and use with respecting the IP rights of the data holder. The data holder can impose confidentiality obligations contractually and use technical protection measures and strict terms of access. In addition, the user must not use the data obtained to develop a competing product. Equally, the data holder must not use data generated by use of the product or related service to derive insights that could undermine the user's commercial position in the markets where the user is active.

Additional sector-specific measures are envisaged to build on the EU Data Act. In the mobility and transport sector, there is already a wide variety of data access and sharing rules. Repair and maintenance information for motor vehicles and agricultural machines is subject to specific data

[109] CNIL (2018). 'Connected vehicles and personal data'. https://www.cnil.fr/sites/cnil/files/atoms/files/cnil_pack_vehicules_connectes_gb.pdf

[110] Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act), *Official Journal* L, 2023/2854. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202302854

access/sharing obligations under type approval legislation. Nonetheless, additional rules are envisaged to ensure legislation is fit for the digital age, to enable data-driven mobility services and to promote the development of clean, connected and automated vehicles. Consideration of services that are based on access to car data, such as repair and maintenance, car sharing, mobility as a service and insurance, is underway. It is recognised that it is essential that access to vehicle data, functions and resources does not create new risks for cybersecurity, road safety, IP or data protection.

BMW has access to, and control over, a large and varied amount of data generated by the use of its cars. BMW uses this data to improve its service offerings and develop new features, including autonomous vehicles and AI-based solutions. BMW also uses insights from this data to inform better urban mobility solutions in partnership with city administrations and universities.[111] For example, BMW is working in the Munich Mobile Future Alliance on strategic solutions for the mobility of the future.[112] BMW makes on-board information and maintenance and repair data available to independent operators (as required under Regulation 2018/858) via its CarData initiative, provided customers consent to such sharing.[113] BMW refers to the party giving consent as either customer or driver. In the Europcar scenario, the customer is Europcar while the driver is the individual who rents the car. If consent for sharing data with third parties is given by Europcar, this has to be disclosed and addressed in the agreement between Europcar and the individual renting the car. If consent for sharing data with third parties is obtained by BMW from Europcar, this has to be disclosed by Europcar in the rental agreement with the individual renting the car.

With regard to the use of third-party applications within the car through connecting a smartphone (such as Apple CarPlay or Android Auto), the BMW privacy policy makes it clear that the smartphone does not have any access to vehicle data. The policy further clarifies that for the processing of any other personal data, it is the provider of the application who is the relevant data controller. The processing of personal data in these cases is governed by the privacy policy and terms of use between the occupant of the car and the relevant third party provider of the relevant application. The rights that the occupant has over this data is governed by the GDPR.

Intellectual property rights

Given the nature of data being co-generated, the main types of IP rights involved in this scenario are rights in respect of the database created being either copyright or potentially a *sui generis* database right[114] (or both). The EU Database Directive that grants *sui generis* rights for databases has been incorporated into the French 'Intellectual Property Code' of 1992 by Law 98-536 of July 1, 1998.[115] However, for such database rights to arise, there must be a substantial investment (in terms of quality and/or quantity) in either the obtaining, verifying or presenting of the contents of the database. In this scenario, the data may rather be said to have been passively contributed. In any event, Article 35 of the EU Data Act makes it clear that, when it is in force, databases created in this

[111] BMW Group (n.d.), 'Urban mobility'. https://www.bmwgroup.com/en/innovation/urban-mobility.html
[112] Ibid.
[113] BMW Group (n.d.). 'BMW CarData'. https://bmw-cardata.bmwgroup.com/thirdparty/public/car-data/overview
[114] Directive 96/9/EC Of The European Parliament and of The Council of 11 March 1996 on the legal protection of databases, *Official Journal* L77. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A31996L0009
[115] Code de la propriété intellectuelle replier partie législative (Articles L111-1 à L811-6). https://www.legifrance.gouv.fr/codes/section_lc/LEGITEXT000006069414/LEGISCTA000006161634/2023-03-08/?anchor=LEGIARTI000006278879#LEGIARTI000006278879

scenario will not be eligible for a *sui generis* database right. Article 35 provides that the *sui generis* right provided for in Article 7 of Directive 96/9/EC does not apply to databases containing data obtained from or generated by the use of a product or a related service. The Act states that this is so that the right of users to access and use such data in accordance with Article 4 of the Data Act, or the right to share such data with third parties, is not hindered.

The occupants' data will not result in any ownership of IP rights for occupants, although very large numbers of occupants' data may have contributed to the developments, or at least prompted the developments. IP in technology developments arising from knowledge gained by the vehicle manufacturers, or the original equipment manufacturers (OEM), from the occupants' use of the vehicle will be owned by the manufacturer who developed the relevant technology, together with any co-developers of the technology. Who exercises control over the data and has access to it will depend on the terms of the contract between the third party provider and the manufacturer.

Uncertainties and gaps

Data generated by the use of the car is used by manufacturers like BMW to innovate and create new services ranging from improved voice command assistants to autonomous vehicles, ie, in the co-generation of technology. This has led to debate around the nature of rights (and benefits) occupants should have in relation to the data that is generated by their use of cars. Experts we spoke to noted that there should be sufficient data rights for occupants in relation to this data, but beyond that any rights over the data should only be considered in proportion to the intent of occupants in creating this data with a view to informing co-generated technology. Doing otherwise will definitely have an impact on the ability of manufacturers and researchers to access data and innovate. It is anticipated that the introduction of rights of access and use by the EU Data Act will significantly improve the balance of power and competition.

Conversational generative AI, Whisper and Midjourney

Scenario	Case study	Nature of the co-generators	Level of involvement of the co-generators	Type of co-generation	Legal jurisdiction of the data holder and the co-generators
Conversational generative AI	Whisper and Midjourney	Users and contributors (the creators of the scraped data)	Active and passive	Co-generated technology, AI co-generated works	US, global
Key takeaways: ● Those whose data or content has been used to train generative AI models are subject to a web of different rights (including data and IP rights), but it is currently difficult for individuals to apply them in a practical way.					

- There is currently uncertainty over which co-generators have rights over AI-generated works, however current case law suggests that generative AI models themselves do not hold any rights over outputs.

- There is currently a plethora of lawsuits focused on rights over generative AI, primarily in the US, using a variety of legal mechanisms. Many are waiting on the outcome of these cases to set a precedent going forward.

Conversational generative AI is capable of producing highly realistic and complex content that mimics human creativity, making it a valuable tool for many industries such as gaming, entertainment and product design. Conversational generative AI is of particular interest to this research as each input by a user into these systems (as a prompt) can itself possibly be considered an act of co-generation.[116] There are two instances of co-generation to be explored with such technology: the creation of co-generated technology and the creation of AI-generated works (which are co-generated by the user giving the inputs and the AI system). Conversational generative AI models are trained on huge datasets, which are often scraped from the internet without the consent of those whose data or whose content is included in the scrape. This creates a complex picture to unravel when considering the different co-generators, their rights and how these apply across jurisdictions.

For this scenario, we use two case studies, Whisper and Midjourney, to enable us to look at the differences in co-generation of technology and AI-generated works involving speech data and imagery.

Case study 1: Whisper, US and global

Whisper is an automatic speech recognition (ASR) system, created by OpenAI and 'trained on 680,000 hours of multilingual and multitask supervised data collected from the web'.[117] The model can perform multilingual transcription, speech translation and language detection. Whisper is a free, open source model which everyone can download and run on a computing platform of their choice. Whisper does not have a user interface, and thus there are many applications which have embedded the model for different uses.[118] For example, AI-powered language learning app Speak uses the Whisper API to power a new in-app virtual speaking companion.

65% of the data from the 680,000 hours of audio and corresponding transcripts represents English language audio, and the other 35% contains audio from 98 different languages.[119]

Co-generator stakeholders and involvement

[116] GPAI Working Group on the Responsible Development, Use and Governance of AI (2020). 'A Framework Paper for GPAI's work on Data Governance'. https://gpai.ai/projects/data-governance/gpai-data-governance-work-framework-paper.pdf
[117] OpenAI (n.d.). 'Whisper'. https://openai.com/index/whisper/
[118] Lab Lab AI (n.d.). 'OpenAI Whisper Applications'. https://lablab.ai/apps/tech/openai/whisper
[119] HuggingFace (n.d.). 'Whisper'. https://huggingface.co/openai/whisper-large-v2

For Whisper, the following entities are involved in the creation of the generative AI model and the ultimate asset – an audio output:

- **OpenAI** develops and owns the trained model but does not take ownership of any of the inputs or outputs.[120]

- **Third parties who have integrated Whisper**, like Speak. Users need to interact with Whisper via a third-party service, or-develop and host their own models. These third parties are also therefore co-generators, and hold the data generated from any interactions.

- **The user/prompter**. Given the lack of user interface, users must set up their own version of the model, or use a service which relies on Whisper. Therefore, users are very active in their use of the model as they are actively seeking to translate or transcribe audio. However, given that services may be using Whisper without explicitly saying so, they may not be aware that they are interacting with the generative AI model.

- **Those who have contributed data to the model**. Given that data has likely been scraped from the internet, it is unlikely that many of those voices in the dataset are aware of the use of their voice in this dataset.

- **Data labelling platforms and their workers**. In order to ensure that the outputs of the AI models are less harmful, data collected by the model (such as conversations with prompters) as well as outputs need to be labelled. While this safety labelling can be conducted by the developer of the AI model, it is often outsourced through data labelling platforms (ie, remunerated work, which has been discussed in the first case study).

Case study 2: Midjourney, US and global

Midjourney is a generative AI model which converts text prompts into images, which is similar to other image generators such as DALL-E and Stable Diffusion. Midjourney is a closed source model which relies on two machine learning technologies, large language and diffusion models. The model is developed by Midjourney Inc., and is currently in beta form, with over two million users.[121] Midjourney describes itself as an 'independent research lab exploring new mediums of thought and expanding the imaginative powers of the human species', which reflects in a more 'artistic' style of output in comparison to other image generators whose outputs tend to be more lifelike.[122] The CEO, David Holz, says that, 'Midjourney is designed to unlock the creativity of ordinary people by giving them tools to make beautiful pictures just by describing them'.[123]

---

[120] OpenAI (2023). 'Introducing ChatGPT and Whisper APIs'. https://openai.com/blog/introducing-chatgpt-and-whisper-apis

[121] Salkowitz, R. (2022). Forbes. 'Midjourney Founder David Holz On The Impact Of AI On Art, Imagination And The Creative Economy'. https://www.forbes.com/sites/robsalkowitz/2022/09/16/midjourney-founder-david-holz-on-the-impact-of-ai-on-art-imagination-and-the-creative-economy/?sh=b14a0aa2d2b8

[122] Midjourney (n.d.). 'Community Showcase'. https://www.midjourney.com/showcase/recent/

[123] Salkowitz, R. (2022). Forbes. 'Midjourney Founder David Holz On The Impact Of AI On Art, Imagination And The Creative Economy'.

The organisation is self-funded and the model is financed via a subscription model, starting from $10 (US dollar) a month. This is different to many of the other models which have at least some level of free access. Users can also access the Midjourney model via a bot which operates on Discord.

Co-generator stakeholders and involvement

For Midjourney, the following entities are involved in the creation of the generative AI model and the ultimate asset – an image:

- **Midjourney Inc.** who own the trained AI-based model. By default, every prompt and generation is public, and is stored and shared in a public gallery.[124]

- **The prompter**. Users of Midjourney sign up to terms of service when they join the platform,[125] as well as three guiding rules (including 'don't be a jerk'). The prompter is aware and active in their use of Midjourney, as they are ultimately using the platform to create a new image.

- **Those who have contributed data or images to the model (with or without consent).** Midjourney's model has been trained on millions of images, with text descriptions. These images have been scraped from the internet, and it is likely that these contributors are largely unaware and passive in their co-generation. However, services like 'Have I Been Trained?' can help users to identify which systems have used their images and artworks for training algorithms.

- **Data labellers.** As discussed in the Karya case study, AI-based systems rely heavily on labelled data. For Midjourney, labelling inputs is particularly important to ensure the system does not generate extremely graphic or illegal outputs.

Rights

Unlike the other scenarios that we analysed for this report, the legal frameworks for AI are still being developed and there are currently a number of uncertainties and gaps. Therefore, unlike the previous scenarios, we have not outlined these uncertainties and gaps; they are documented throughout the discussion that follows. Further, it is important to note that the analysis in this section also looks at rights frameworks in regards to co-generated technology and AI co-generated works in addition to co-generated data. With the Whisper and Midjourney models trained on vast amounts of data (which includes co-generated data), the models are co-generated technology. The outputs of these models in and of themselves are considered AI co-generated works.

Data rights

https://www.forbes.com/sites/robsalkowitz/2022/09/16/midjourney-founder-david-holz-on-the-impact-of-ai-on-art-imagination-and-the-creative-economy/?sh=b14a0aa2d2b8

[124] Dallery Gallery (2022). 'Everything you wanted to know about MidJourney'. https://dallery.gallery/midjourney-guide-ai-art-explained/

[125] Midjourney (2023). 'Terms of Service'. https://docs.midjourney.com/docs/terms-of-service

OpenAI states in its Terms of Use[126] that it assigns to the user all its rights, title and interest in and rights to outputs, subject to compliance with the Terms of Use. It also states that inputs (prompts) into the system are owned by the user. However, it must be noted that, as detailed in its Privacy Policy,[127] OpenAI uses information provided by the user to improve its services and conduct research. This includes:

- personal information that is included in the inputs or file uploads
- types of content inputted, viewed or engaged with
- the features used and the actions taken.

In this sense, while the user is the owner of its input(s) and, as between OpenAI and the user, has all rights in the output(s), both are still used to further train OpenAI's models. Furthermore, third parties may have rights in the generated content as explained below.

Midjourney does not explicitly explain where the data which the model is trained on comes from. However, in an interview with Forbes, CEO David Holz stated that: 'It's just a big scrape of the Internet. We use the open data sets that are published and train across those. And I'd say that's something that 100% of people do. We weren't picky'.[128]

The personal data rights of the user in both Whisper and Midjourney are governed by applicable laws,[129] including the GDPR, or equivalent legislation.

The framework on the rights of parties that contribute data or assets to the model is not only tricky but has led to several litigations. Whisper is trained through several thousand data points. In an OpenAI paper,[130] there is no mention of the sources of Whisper's training datasets apart from a reference to its data collection pipeline, which is sourced primarily from English-centric parts of the internet. A likely assumption is that a lot of the data was scraped from the internet.[131] What this means is that the training data is likely to have been obtained without consent. This is likely to have implications for data rights as well as IP rights of individuals, and users' rights to use AI co-generated outputs freely.

There are also significant considerations for data rights in text-based generative AI models such as ChatGPT. While users may be made aware that content may be posted publicly online, they are unlikely to have the knowledge or understanding that they will be used to train a model such as

[126] OpenAI (2023). 'Terms of use'. https://openai.com/policies/terms-of-use
[127] OpenAI (2023). 'Privacy policy'. https://openai.com/policies/privacy-policy
[128] Salkowitz, R. (2022). Forbes. 'Midjourney Founder David Holz On The Impact Of AI On Art, Imagination And The Creative Economy'. https://www.forbes.com/sites/robsalkowitz/2022/09/16/midjourney-founder-david-holz-on-the-impact-of-ai-on-art-imagination-and-the-creative-economy/?sh=b14a0aa2d2b8
[129] OpenAI (2023). 'Privacy policy'. https://openai.com/policies/privacy-policy
[130] OpenAI (2022). 'Robust Speech Recognition via Large-Scale Weak Supervision'. https://cdn.openai.com/papers/whisper.pdf
[131] Mahelona, K. et al. (2023). Papa reo. 'OpenAI's Whisper is another case study in Colonisation'. https://blog.papareo.nz/whisper-is-another-case-study-in-colonisation/

ChatGPT. Data protection authorities across Europe have also been examining the risks for individuals' privacy and have provided statements as to the transparency requirements and other protections required of platform providers. The European Data Protection Board (EDPB) has established a ChatGPT taskforce to examine the issues and provide guidance to ensure the protection of individuals' privacy rights.

The use of text and data to train generative AI models is not the only instance of privacy concerns. The concept of Privacy as Contextual Integrity relates to instances where adequate protection for privacy is tied to norms of specific contexts;[132] demanding that information gathering and dissemination be appropriate to that context and obey the governing norms of distribution within it. As discussed in the subsequent sections, there is value in identifying the specific dimensions of data privacy that need acknowledgement and regulating in the context of generative AI models.

Case study: Whisper and the Māori language[133]

These issues are exemplified in an article written by Te-Hiku Media, a non-profit Māori organisation in Aotearoa New Zealand, which explores how Whisper is an example of data colonisation. The Whisper model works for Māori, however it is not immediately clear where the data for the language comes from. Needless to say, Whisper doesn't explicitly state where this data was collected from. The article authors ask: 'who gave them access, and who gave them the right to create a derived work from that data and then open source the derivation?'.

The Māori language has faced years of oppression and was previously forbidden to be spoken in New Zealand. As such, the Māori community faces a difficult path; recognising the benefits of these technologies in supporting the survival of the language, it denounces the lack of Māori involvement in the development of them. The lack of involvement of Indigenous people can lead to further harms to the survival of the language if translations are incorrect or if words are mispronounced in generated audio. Furthermore, the interest in the Māori language by big tech companies more broadly is seen as an opportunity to create economic opportunities, further extracting value from the country and the Indigenous community.

The message from the community is clear: *'if anyone is to profit from te reo Māori it should be Māori and Māori alone, especially considering the fact that non-Māori once sought to make the Māori language extinct'.*

IP rights over inputs and outputs

The overarching legal frameworks over input data used to train the AI-based systems are the applicable copyright laws. Additionally, at the internal policy level, Midjourney's Terms of Service include a DMCA Takedown policy, which sets out the procedure for a takedown request if anyone believes that material located on or linked to by Midjourney violates their copyright or trademark.[134]

[132] Nissenbaum, H. (2004). Washington Law Review. 'Privacy as Contextual Integrity'. https://digitalcommons.law.uw.edu/wlr/vol79/iss1/10

[133] Mahelona, K. et al. (2023). Papa reo. 'OpenAI's Whisper is another case study in Colonisation'. https://blog.papareo.nz/whisper-is-another-case-study-in-colonisation/

[134] Midjourney (2023). 'Terms of Service'. https://docs.midjourney.com/docs/terms-of-service

Midjourney has also been the subject of a lawsuit[135] alleging that Midjourney, along with other generative-AI providers such as Stable Diffusion, has infringed the rights of 'millions of artists' by training their AI tools on five billion images scraped from the internet 'without the consent of the original artists'. Stable Diffusion has also been sued by Getty Images in the UK and the US for copyright infringement.[136]

With more and more instances of artists having their work fed into AI-based models such as Stable Diffusion to create very similar looking artwork,[137] the question of IP rights over images is of utmost importance, and the same issues arise in relation to literary and musical works, as well as other copyright-protected materials.

In the EU, the 'Copyright in the Digital Single Market (CDSM) Directive' has a specific call out for text and data mining (TDM) activities, noting in Article 4(1) that limitations will be provided on copyrights for the purpose of TDM.[138] Article 4(3) does note that rights holders have the right to expressly reserve their rights, in which case such exceptions will not apply. In the US, the question is whether such data scraping activities qualify as fair use.[139] Existing jurisprudence seems to suggest that the doctrine of fair use allows for a significant range of TDM.[140]

In the UK, Section 29A of the Copyright Designs and Patents Act 1988 exempts TDM from copyright infringement if there is lawful access to the work and the sole purpose of copying is non-commercial research. It is anticipated that this would not exempt TDM for the training of AI for commercial purposes, although this is likely to be considered soon by the UK High Court in the Getty v. Stability AI litigation. The exception has been in the spotlight over the past two years as the UK government, following a consultation process and with the training of AI-based models in mind, initially proposed its expansion to enable TDM for both non-commercial and commercial research purposes. However, at the start of 2023 it was announced that this would not proceed having taken account of the fact that 'IP is the lifeblood of many creative industries'.[141] Sir Patrick Vallance's report on 'Pro-Innovation Regulation of Technologies Review' published in March 2023 advocated for a pro-TDM landscape and recommended a clear policy position on the relationship between IP and

[135] Vincent, J. (2023). The Verge. 'AI art tools Stable Diffusion and Midjourney targeted with copyright lawsuit'. https://www.theverge.com/2023/1/16/23557098/generative-ai-art-copyright-legal-lawsuit-stable-diffusion-midjourney-deviantart

[136] Vincent, J. (2023). The Verge. 'Getty Images sues AI art generator Stable Diffusion in the US for copyright infringement'. https://www.theverge.com/2023/2/6/23587393/ai-art-copyright-lawsuit-getty-images-stable-diffusion

[137] Baio, A. (2022). Waxy. 'Invasive Diffusion: How one unwilling illustrator found herself turned into an AI model'. https://waxy.org/2022/11/invasive-diffusion-how-one-unwilling-illustrator-found-herself-turned-into-an-ai-model/

[138] Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC, *Official Journal* L 130/92. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32019L0790

[139] US Copyright Office (2023). 'U.S. Copyright Office Fair Use Index'. https://www.copyright.gov/fair-use/index.html

[140] Quintais, J. P. (2023). Copyright Blog. 'Generative AI, Copyright and the AI Act'. https://copyrightblog.kluweriplaw.com/2023/05/09/generative-ai-copyright-and-the-ai-act/

[141] Department for Digital, Culture, Media & Sport & Lopez, J. MP (2021). 'Minister Lopez speech to the Advertising Standards Authority Parliamentary Breakfast'. https://www.gov.uk/government/speeches/minister-lopez-speech-to-the-advertising-standards-authority-parliamentary-breakfast

AI,[142] and publication of a code of practice to 'support AI developers to access copyrighted work as an input to their models' with the focus on 'clarifying a simple process' around that.

Japan has developed and revised AI-related regulations with the goal of maximising AI's positive impact on society, rather than suppressing it out of potentially overestimated risks. Article 30-4 of Japan's Copyright Law permits the use of copyrighted material such as text and images to train AI models, including for commercial use, but does contain a provision stating that such material cannot be used if it would 'unreasonably prejudice the interests of the copyright owner'.

On the output front, there are a number of questions that are being debated around rights over the outputs, including:[143]

- Is an output from a generative AI model protected by copyright or IP licences?
- Does such an output infringe a copyrighted work of a third party, especially if those works are 'ingested' during the training stage of the AI model?

Here too, under Midjourney's terms of use, the rights of a user over the output(s) created using the service is dependent on their subscription model with Midjourney and the rights of third parties in the generated content.[144] For free users, Midjourney grants a licence to the outputs under the 'Creative Commons Noncommercial 4.0 Attribution International License'. Paid users, however, own all outputs they create using Midjourney services. Employees or owners of a company with more than $1m (US dollars) a year in gross revenue using Midjourney on behalf of the company must purchase a 'Pro' membership to own the outputs. Regardless of the type of subscription, when any user signs up to use Midjourney, they provide Midjourney and its successors with a perpetual, worldwide, non-exclusive, sublicensable, no-charge, royalty-free, irrevocable copyright licence to reproduce, prepare derivative works of, publicly display, publicly perform, sublicense, and distribute text and image prompts they input into Midjourney, or outputs produced by Midjourney from their prompts. As a result, by default, images created on Midjourney are publicly viewable and remixable – with the exception of Pro users who choose certain settings. This limits the users' exclusive rights of use and reproduction in relation to the generated images. Any content added to, or used to alter, the generated content outside of the tool will, however, be owned by the author and remain protected from unauthorised reproduction by copyright, provided it involves sufficient original creative content from the author.

These questions are also the subject matter of lawsuits against generative AI providers.[145] In the recent judgement of Thaler v. Perlmutter,[146] Judge Beryl Howell briefly opined on the question of

[142] HM Treasury (2023). 'Pro-innovation Regulation of Technologies Review: Digital Technologies'. https://www.gov.uk/government/publications/pro-innovation-regulation-of-technologies-review-digital-technologies

[143] Quintais, J. P. (2023). Copyright Blog. 'Generative AI, Copyright and the AI Act'. https://copyrightblog.kluweriplaw.com/2023/05/09/generative-ai-copyright-and-the-ai-act/

[144] Midjourney (2023). 'Terms of Service'. https://docs.midjourney.com/docs/terms-of-service

[145] Vincent, J. (2023). The Verge. 'Getty Images sues AI art generator Stable Diffusion in the US for copyright infringement'. https://www.theverge.com/2023/2/6/23587393/ai-art-copyright-lawsuit-getty-images-stable-diffusion

[146] Stephen Thaler, vs. Shira Perlmutter (2022). https://ecf.dcd.uscourts.gov/cgi-bin/show_public_doc?2022cv1564-24

whether the user of an AI programme could own the copyright of the output. In his decision, it was indicated that 'the increased attenuation of human creativity from the actual generation of the final work will prompt challenging questions regarding how much human input is necessary to qualify the user of an AI system as an "author" of a generated work', suggesting that with sufficiently detailed prompts, authorship may be established, resulting in the protection of copyright granted to a generative AI user on the outputs produced creatively. Furthermore, the US Copyright Office has issued a policy statement requiring human authorship for eligibility for copyright protection.[147] According to the US Copyright Office, US copyright law only protects material that is the result of human creativity. However, where an individual user materially modifies AI-generated content as a result of the user's own human creativity, the resulting work may be protected by copyright. In China, AI-generated images have been granted copyright protections in a recent court ruling, in recognition of the 'intellectual achievement' involved in generating the image.[148]

There is significant legal uncertainty around the copyright regimes for AI-generated content,[149,150] resulting in scholarship about what aspects of input and output data need protections within the copyright framing, and how the fair use doctrine within the US should be interpreted in this context. Recognising the societal benefits of opening access to AI datasets[151] on the input level and the practical viability of licensing all underlying copyrighted work within massive AI datasets,[152] studies argue that the fair use exception to copyright should be expanded in scope to cover 'fair learning' as well.[153] As per this argument, copyright law should permit generative AI models to copy works not just for non-expressive purposes within the fair use doctrine as provided for TDM use-cases, and for transformative uses, but to take a 'pluralist view of fair use' and acknowledge the value of permitting copyrighted works for meaningful social purposes such as education and learning. In the context of AI applications, 'fair learning' would permit the use of input datasets which include copyrighted works, as long as the purpose of the AI's use is 'to access, learn, and use the unprotectable parts of the work' and not to incorporate the copyright elements of a work in the output.[154] This would strike a balance between protecting the interests of authors while not hindering innovation.

[147] United States Copyright Office (2023). 'Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence'. https://copyright.gov/ai/ai_policy_guidance.pdf
[148] Baker McKenzie (2023). 'China: A landmark court ruling on copyright protection for AI-generated works'. https://insightplus.bakermckenzie.com/bm/data-technology/china-a-landmark-court-ruling-on-copyright-protection-for-ai-generated-works#:~:text=In%20brief,by%20Artificial%20Intelligence%20(AI)'.
[149] Guadamuz, A. (2023). 'A Scanner Darkly: Copyright Liability and Exceptions in Artificial Intelligence Inputs and Outputs'. https://ssrn.com/abstract=4371204
[150] Vincent. J. (2022). The Verge. 'The scary truth about AI copyright is nobody knows what will happen next'. https://www.theverge.com/23444685/generative-ai-copyright-infringement-legal-fair-use-training-data; Margoni, T., & Kretschmer, M. (2022). GRUR International. 'A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology'. https://doi.org/10.1093/grurint/ikac054
[151] National Science and Technology Council Committee on Technology (2016). Executive Office of the President. 'Preparing for the future of Artificial Intelligence'. https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf
[152] Lemley, M. & Casey, B. (2021). Texas Law Review. 'Fair Learning'. https://texaslawreview.org/fair-learning/
[153] Ibid.
[154] Ibid.

Key findings

The case studies covered in this report focus on different types of co-generation in different legal jurisdictions. Across the six scenarios explored, co-generators are subject to an overlapping web of different rights and protections. This research makes clear that recognising adequate rights around co-generation needs to leverage different types of legal rights. Reliance cannot be placed on a single type of right, as different rights provide different types of protection or recognition. Here we summarise some of the key findings drawn from the analysis of the case studies.

Co-generators are subject to a complex web of rights, but gaps remain

For those involved in the co-generation of data, both legal and data rights which apply are generally quite clear and often adequate. For the co-generators covered in the scenarios based in the EU, there are relatively strong personal data rights thanks to the efforts of the European Commission in regulating the data ecosystem with legislation like the GDPR. In Brazil, the right of privacy is safeguarded in both the Marco Civil and the LGPD, with the LGPD recognising further rights around personal data. In India, the new Digital Personal Data Protection Act recognises rights around personal data. However, personal data is only one piece of the data rights puzzle. Large swathes of data are collected and processed as non-personal data, and regulations like the Data Act and Data Governance Act are nascent attempts in the EU at governing non-personal data sharing. IP regimes offer protection over literary, graphical and audio works but usually only apply to the creators of the databases rather than the other co-generators.

For co-generated technology and AI co-generated works, there is significantly less clarity around the rights of different co-generators. In some cases, copyright is seen as a potential solution to providing rights to co-generators, and is currently the subject of numerous lawsuits with creators of generative AI models. However, applicability is still being debated and nuances are being fleshed out. Many expect these cases to continue to proliferate, and for more cases to come to courts around the world in the coming years.[155] In 2023, the US copyright office launched a consultation to support its thinking on how copyright applies to generative AI, demonstrating the complexity and fluidity of the situation.[156]

In addition to data and IP rights, there are other legal frameworks that also apply in certain cases. For instance, labour rights frameworks are of critical importance in the space of data labelling for AI training and content moderation, where working conditions are generally poor. Also, there continue to be gaps in existing frameworks. One area in particular is collective and community rights, which rarely apply to co-generation scenarios. Data and IP rights are usually framed through an individualistic approach, and generally do not address rights of communities; notable exceptions include instruments such as traditional knowledge and indigenous knowledge protection. For digital rights, some regulations seek to reassert individual control for data subjects over the terms of their datafication, while others aim to maximise financial gain. But these proposals share a common conceptual flaw: according to Salome Viljoen, they miss the point of data production in a digital economy, which is to put people into population-based relations with one another.[157] This relational

[155] Sheng, E. (2023). CNBC. 'In generative AI legal Wild West, the courtroom battles are just getting started'. https://www.cnbc.com/2023/04/03/in-generative-ai-legal-wild-west-lawsuits-are-just-getting-started.html
[156] US Copyright Office (2023). 'Artificial Intelligence Study'. https://www.copyright.gov/policy/artificial-intelligence/
[157] Viljoen, S. (2020). Yale Law Journal. 'A Relational Theory of Data Governance'. http://dx.doi.org/10.2139/ssrn.3727562

aspect of data production drives much of the social value as well as the social harm of data production and use in a digital economy.[158] Current regulation provides no recourse for communities, which is especially concerning given the ability of data insights to operate at a group level. This is particularly relevant in the context of aggregated datasets, which are typically considered as non-personal data. While there are efforts to address this gap,[159] there remains a long way to go.

Where rights do exist, the practical application of them is not always straightforward

The relative power of large technology companies is well documented.[160] This power influences the relationship between the creators of technology and other co-generators. Where rights exist, exercising or applying these rights is often a complicated task that relies on co-generators having the necessary literacy and capacity to exercise them. In the example of Instagram, the translation of rights for users of Instagram is determined on the basis of the platform's data and privacy policies. To use Instagram, users must consent to these policies. However, the failures of the notice and consent mechanism are well documented, and consent fatigue means that consent is usually not informed and does not always equal awareness.[161] Therefore many users are likely to sign up to Instagram without being able to make an informed decision based on a full understanding of their rights and of the extent of the co-generation of data and technology which using the platform entails. This is particularly problematic when co-generators are not aware that their actions are generating data.

Similarly, the ability of co-generators to apply copyright protections is difficult, and is not always set up in the interest of the co-generators. For example, where existing IP laws recognise rights over works, the use of standard contractual documents such as terms and conditions can lead to the transfer of IP rights from creators to bigger entities, leaving co-generators without protection.[162] These dynamics also manifest in much of the discussion around legal protections for co-generators with AI. For example, the discourse around IP prioritises the holders of copyright and IP. In some cases, this will be the creators themselves, but oftentimes big companies such as record labels, film studios or publishers are the actual holders of the IP rights.[163] There are wider concerns that the current push to use IP as a tool to tackle these complications around AI may further entrench the established power dynamics in the data ecosystem.

---

[158] Ibid.

[159] Lubin, A. (2023). Temple Law Review. 'Collective Data Rights and their Possible Abuse.' https://www.templelawreview.org/essay/collective-data-rights-and-their-possible-abuse/; Singh, P. J. & Vipra, J. (2019). Development. 'Economic Rights Over Data: A Framework for Community Data Ownership'. https://link.springer.com/article/10.1057/s41301-019-00212-5; Ausloos, J. et al. (2022). Ada Lovelace Institute. 'The case for collective action against the harms of data-driven technologies'. https://www.adalovelaceinstitute.org/blog/collective-action-harms/

[160] Eavis, P. & Lohr, S. (2020). The New York Times. 'Big Tech's Domination of Business Reaches New Heights'. https://www.nytimes.com/2020/08/19/technology/big-tech-business-domination.html

[161] World Economic Forum (2020). 'Redesigning Data Privacy: Reimagining Notice & Consent for human technology interaction'. https://www3.weforum.org/docs/WEF_Redesigning_Data_Privacy_Report_2020.pdf

[162] Neuffer, B. & Kresz, M. (2020). American Bar Association. 'Don't Give Away Your Intellectual Property'. https://www.americanbar.org/groups/construction_industry/publications/under_construction/2020/winter2020/dont-give-away-your-intellectual-property/

[163] Knox, R. (2021). Wired. 'Big Music Needs to Be Broken Up to Save the Industry'. https://www.wired.com/story/opinion-big-music-needs-to-be-broken-up-to-save-the-industry/

The act of scraping data and content from the internet to train AI models is the crux of some of these issues. It means that co-generators are unaware of their involvement, and do not have a chance to consent to the use of the data or content for training AI models. This data may not be covered by any legal basis, and as publicly available information which is posted online, there is ambiguity over its use. This impacts not only those whose data is being used for training purposes, but also those creating technology. The complexities around which rights may or may not apply has led to a plethora of legal battles focused on rights over AI co-generated content.

The lack of clarity of access to large-scale datasets for training AI models limits the potential of AI development. Those developing these models require access to huge amounts of data, but the complexity around various rights results in a chilling effect on innovation, where creators may choose not to pursue certain technology due to these complexities. More broadly, the current developments in AI, and the discourse around access to data, has led to a broader chilling effect around the sharing of data.[164] Data holders are more resistant to sharing or opening up data because of concerns about how it might be used beyond their control. This resistance limits the amount of data available to organisations to use for innovation, and to tackle key societal challenges, which ultimately negatively impacts society at large.

Where rights don't currently exist, deciding which legal rights should apply, if any, in different jurisdictions and co-generation situations can be challenging

Access to data is critical in tackling many of the challenges facing society, but the same data ecosystem can simultaneously cause harm for individuals, communities and society more broadly. This remains true for co-generation scenarios where there is a significant trade off between access to data for innovation in technology and science, and the need for rights and protections for co-generators. For example, in the discussions around access to IoT data, there is fierce debate about whether individuals should have control over data generated by them, or not.[165] The data generated through the use of IoT devices is crucial for the development of new and improved technologies, and some argue that creating new rights for co-generators will limit this innovation. Debates around scraping of the internet to train AI models follow a similar pattern: without data there is no AI.

Discussions about how to solve these issues are intrinsically connected with the moral and philosophical imperatives of the technologies, legislators and policymakers involved. For instance, one perspective is to view data as labour,[166] which means it should primarily benefit the individuals and communities generating the data, as opposed to the existing paradigm where data is treated as an exhaust for tech companies to build and innovate with for private profit.[167] Another perspective is that, in some circumstances, co-generators are making no effort to intentionally generate data and thus should not be entitled to specific rights over that data. The example of driving a car as explored in one of the scenarios illustrates this: the driver typically has no intention of generating data to further innovation; the real effort in collecting and analysing data is done by manufacturers and third

[164] Verhulst, S. (2024). 'Are we entering a "Data Winter"?' https://sverhulst.medium.com/are-we-entering-a-data-winter-f654eb8e8663

[165] Janeček, V. (2018). 'Ownership of personal data in the Internet of Things'. Computer Law & Security Review, Volume 34, Issue 5. https://doi.org/10.1016/j.clsr.2018.04.007.

[166] Arrieta Ibarra, I. et al. (2017). American Economic Association Papers & Proceedings. 'Should We Treat Data as Labor? Moving Beyond "Free"'. https://ssrn.com/abstract=3093683

[167] Crawford, K. (2023). Green European Journal. 'Mining for Data: The Extractive Economy Behind AI'. https://www.greeneuropeanjournal.eu/mining-for-data-the-extractive-economy-behind-ai/

parties. Furthermore, the free flow of this data is crucial for innovation and creates benefits for society more broadly. For example, data collected by car manufacturers about how users drive their cars, including their speed, how they brake and more, can support the development of new, greener and safer cars.

Different political economic leanings have significant implications for the nature of the digital ecosystem, and for how benefits are distributed.[168] At a national level, decisions about legislation stem from the fundamental values of each jurisdiction, alignment in the global community and appetite for risk. These underlying values shape approaches to governance, and we have already seen this in the different approaches to data governance taken by the US compared with the EU.[169] Other countries, like Japan and China, have taken different approaches again, based on their respective values. As new legislation comes into force, the landscape of rights will become more complex across moral and geographical boundaries.

Where co-generators *should* have rights, deciding what they should look like, and for whom, remains difficult. Claiming rights as co-generators is complex for two reasons. First, as discussed by Wendehorst and Cohen,[170] there is a variance in how rights should apply across factors like level of involvement in co-generation, awareness of co-generation, harm and financial gain. This was also highlighted by interviewees during this research. Second, the value of contributions made by different co-generators becomes more complex as the number of co-generators involved increases. Generative AI makes both of these issues significantly more complicated still, specifically with regards to co-generated technology. With generative AI, the models have been trained with datasets which are often:

- **Not disclosed.** Developers of generative AI models do not always disclose where the data has been collected from, making it difficult to know with certainty that data or content has been included. This has knock-on effects for attributing contributions within outputs.

- **From multiple sources**. Most training datasets are built using data from multiple sources, which means they are likely to consist of data with varying and incompatible licences. Similarly, the publication and use of data for AI doesn't naturally respect national boundaries (which is the level at which most legal frameworks are applied and enforced), so there is huge jurisdictional fragmentation.

- **Enormous in volume**. The huge scale of data needed to train some AI models creates difficulties for developers in verifying ownership and licences across millions of pieces of data or checking for inclusion of personal data without a legal basis.

[168] Ibid
[169] Komaitis, K. & Sherman, J. (2021). Brookings. 'US and EU tech strategy aren't as aligned as you think'. https://www.brookings.edu/articles/us-and-eu-tech-strategy-arent-as-aligned-as-you-think/
[170] Wendehorst, C. & Cohen, N. (2023). American Law Institute and European Law Institute. 'ALI-ELI Principles for a Data Economy - Data Transactions and Data Rights'. https://www.europeanlawinstitute.eu/fileadmin/user_upload/p_eli/Publications/ALI-ELI_Principles_for_a_Data_Economy.pdf

Areas for further research

The findings of this research highlight the way existing legal frameworks operate, their gaps and the complexities that arise out of their operation. Here we set out three broad areas that require further exploration and research.

The landscape of legal rights and how to best apply them in a practical setting
Across the six scenarios explored in this report, co-generators are subject to an overlapping web of different rights and protections. This research has shown that recognising rights around co-generation needs to leverage different legal frameworks; there is no single framework, as different rights frameworks provide different types of protection or recognition. In practice, these rights take the form of terms and conditions, contracts and licensing agreements, which feature in each of the scenarios covered in this research.

Further study is needed to explore how to make these mechanisms more effective at an implementation stage in protecting the rights of co-generators, while also allowing for data to be made available for training AI models. Enabling people to consent to using a system, or for it to use data about them, should be a crucial part of enabling access to data for innovation while protecting the rights of the co-generators. Exploring how to make these mechanisms accessible, informative and scalable with guides on the implications of its use is important. In some cases, particularly with generative AI where consent currently doesn't exist, for example with training datasets scraped from the internet, new approaches will be required for people to opt out. Further research and experimentation around standard contractual terms is needed to understand how reliability and equity around rights can be balanced against innovation interests for co-generation at scale.[171]

Closing the legal gaps and clarifying the overlaps
While legal frameworks cover different aspects of co-generation, there are gaps in these frameworks. When it comes to co-generated data, the legal frameworks are quite clear as far as personal data is concerned, with issues arising more around the practical application. The large gaps that remain, however, are around legal frameworks for non-personal data and collective rights. A large portion of the value of data is the ability to derive population-level insights from it, and the relational nature of data drives this. There are attempts to close the gaps on how these collective rights in the context of non-personal data could look, but there is a need for further study to arrive at models that find a good balance between protecting the interests of communities and unlocking data for broader societal value.

For co-generated technologies and AI-generated works, there is even more of a need for clarity, for example on whether IP protected works can be used to train AI models; what AI models can be used for; what protection can be accorded to AI-generated works; when such protection can be accorded; and who is the correct rightsholder. Different jurisdictions are also adopting different approaches to answering these questions, and this is likely to result in a degree of confusion. Blurred jurisdictional boundaries are not novel in the regulation of the digital economy, and this

[171] GPAI (2023). Global Partnership on AI. 'Fostering Contractual Pathways for Responsible AI Data and Model Sharing for Generative AI and Other AI Applications'.
https://gpai.ai/projects/responsible-ai/IC_Intellectual%20Property%20project.pdf

continues to be the case when attempting to determine the landscape of digital rights for co-generated technologies and AI-generated works. There have been developments, through case law, policy and legislation, with new positions being reached even over the course of this study. There is a need to explore the implications of these new developments and for research that provides clarity on the different overlapping legal frameworks.

Non legal mechanisms

As demonstrated in this report, the complexity of the landscape of different rights is vast. For many co-generators this complexity is a hindrance. For those developing new technologies, uncertainty around rights creates friction during the innovation process, and can limit the breadth and quality of data which could be used for further research and development. For those using these technologies, there are limited opportunities for exercising control over data that they have co-generated. Below are some mechanisms outside of national legal frameworks for further exploration which could go some way to mitigating these issues:

- There will continue to be a need for bespoke data licences, as well as those built around standardised terms agreed upon by a broad community of co-generators, which resonates with the experience of Open Source and Creative Commons licences.[172] There are new licences being developed, like the Responsible AI License (RAIL), to empower developers to restrict the use of their AI technology to prevent irresponsible and harmful applications.[173] The Kaitiakitanga License has been developed by Te Hiku Media,[174] which defines the rules for collaborations based on the Māori principle of kaitiakitanga, or guardianship. The licence only grants data access to organisations that agree to respect Māori values, maintain consent, and share any benefits derived from the use of the licence with the Māori people.[175]

- Deploying alternative forms of data governance, such as data trusts,[176] or data cooperatives,[177] which already work to steward data on behalf of different communities, such as gig workers, are an option. The Bennett Institute has proposed the creation of a national data trust to steward data on behalf of a nation, negotiate access with technology companies, and return a Digital Commons Dividend to its members.[178]

[172] GPAI (2022). Global Partnership on AI. 'Protecting AI innovation, Intellectual Property (IP): GPAI IP Expert: (I) Guidelines for Scraping or Collecting Publicly Accessible Data and (II) the Preliminary Report on Data and AI Model Licensing, Report'. https://gpai.ai/projects/innovation-and-commercialization/intellectual-property-expert-preliminary-report-on-data-and-AI-model-licensing.pdf

[173] Responsible AI License (n.d.). https://www.licenses.ai/

[174] Kaitiakitanga-License (n.d.). https://github.com/TeHikuMedia/Kaitiakitanga-License?ref=blog.papareo.nz

[175] Hao, K. (2022). MIT Technology Review. 'A new vision of artificial intelligence for the people'. https://www.technologyreview.com/2022/04/22/1050394/artificial-intelligence-for-the-people/

[176] GPAI (2021). Global Partnership on AI. 'Enabling data sharing for social benefit through data trusts'. https://gpai.ai/projects/data-governance/data-trusts/enabling-data-sharing-for-social-benefit-data-trusts-interim-report.pdf

[177] Kapoor, A. & Vaitla, B. (2022). Brookings. 'Data co-ops: How cooperative structures can support women's empowerment'. https://www.brookings.edu/articles/data-co-ops-how-cooperative-structures-can-support-womens-empowerment/

[178] Chan, A. & Bradley, H. (2023). Bennett Institute. 'Reclaiming the digital commons'. https://www.bennettinstitute.cam.ac.uk/blog/reclaiming-the-digital-commons/

- Increasing the involvement of communities in how AI is designed, developed and deployed. This may mean involving people in public engagements and deliberations at policy level to decide how certain technologies such as AI should be used, or at a data level to decide when co-generated data should be collected and shared. It could also involve people in the design of these systems.[179] In a scenario where consent is often not serving its purpose, involving people in consultations in the design of these systems, around which data is collected and used, will create a better chance of creating a meaningful outcome.

[179] Groves, L. et al. (2023). FAccT 'Going public: the role of public participation approaches in commercial AI labs'. https://doi.org/10.1145/3593013.3594071

Conclusion

The concept of co-generation in the context of the digital economy is not new. From open source software to gig work, the digital realm has always featured co-generation of data and technologies. Legal regimes too have accounted for co-generation of different types, even before new scenarios created by digital innovation. Our research has shown that legal frameworks for co-generation of data in most contexts (outside of AI and social media) are largely clear and adequate, barring a few gaps around collective rights and non-personal data. Even with social media, the complications arise not so much with the legal frameworks themselves but with how they are operationalised and the challenges of enforcing them at immense scale.

AI models have introduced new complications for these existing regimes and have given rise to questions that require a refined understanding of long-standing concepts such as the notion of free use, originality, and artistic expression and effort. AI technologies have already contributed to social good and have demonstrated immense potential for addressing sustainable development goals. However, there is a need for clarity on how existing legal rights operate in the context of these new co-generated technologies and AI-generated works. There is also a need for new mechanisms that speak to the complexities and specificities of these new technologies. These mechanisms are not restricted to the legal realm and there is potential and reason to explore non-legal mechanisms to provide clarity and robustness around the data of individuals and communities. Doing so is critical, not just in terms of equity and fairness for those whose data is used to train AI models, but also in ensuring the availability of good quality, comprehensive data to build AI models that can help tackle some of humanity's major challenges.

Methodology

Different examples of co-generation scenarios occur all over the world, whether in the generation of new data points, training datasets, algorithms or AI-generated works. This report is concerned with the different legal mechanisms which apply to co-generators in different jurisdictions. There is of course an enormous number of co-generation scenarios, with an expansive number of different legal regimes applying to them, depending on the jurisdiction. This research takes a case study approach, focusing on six different examples of co-generation to explore the legal ramifications of co-generation in practice. These case studies cover different types of co-generation, legal landscapes across jurisdictions and co-generation activities. By conducting a deep dive into these case studies, this research draws key takeaways for the wider landscape of co-generation, in particular considering the new challenges posed by generative AI. Here we detail the different steps in the methodology of this project.

Selecting legal frameworks to focus on

We initiated this research to understand the various rights and legal protections of different co-generators and how AI – particularly generative AI – complicates these situations. We examined literature from legal, technical, and data perspectives to deepen our understanding of the field and map out the diverse legal mechanisms connected to co-generation. The literature review covered both scientific and grey literature. We complemented this literature review with regular working sessions with the GPAI project team, who provided an expert steer and suggestions for the team. Following this initial process, we generated the following, non-exhaustive longlist of rights and legal protections for co-generators:

- data rights, including rights of privacy, data access, portability and more
- intellectual property rights, including copyrights, trademarks and trade secrets
- contracts (such as licences and terms and conditions)
- other legal frameworks such as:
    - fundamental rights not under data rights frameworks (such as right to privacy)
    - collective and community rights, including Indigenous data sovereignty rights, moral and ethical rights.

Selecting co-generation scenarios to focus on

Building on the literature review, and in collaboration with the GPAI project working group, we researched a variety of different co-generation scenarios to compile a longlist. We then narrowed this down to six, and through these six we sought to explore the breadth of different types of co-generation. The rights frameworks that apply in each scenario depend on the following criteria:

- **The nature of the co-generators**. This criterion looks at who the co-generators are; whether the co-generators are private businesses, governments, members of the public, communities, or even AI models themselves.

- **The level of involvement of the co-generators**. This criterion assesses whether the generation of data was a result of the active or passive involvement of the co-generator. As noted in the 'Comment to Principle 18 of the ALI-ELI Principles',[180] the share/role the co-generator had in co-generation has a bearing on the co-generator's claim or justification for a right over co-generated data.

- **The type of co-generation**. This criterion refers to the three types of co-generation considered in this report: co-generated data, co-generated technology and AI-generated works.

- **The legal jurisdiction of the data holder and the co-generators**. This criterion is important to understand some of the nuances of different legal jurisdictions around the world, and how they apply to co-generators.

The scenarios selected for this research are as follows:

Scenario	Case study	The nature of the co-generators	The level of involvement of the co-generator	The type of co-generation	Legal jurisdiction of the data holder and the co-generators
Remunerated work	Karya	Data labellers	Very active	Co-generated data, co-generated technology	India, global
Crowdsourced data collection	OpenStreet Map	Contributors to OpenStreetMap	Very active	Co-generated data	UK, global
Social media platforms	Instagram	Social media users	Active and passive	Co-generated data, co-generated technology	US, Brazil
Internet of Things	BMW connected cars owned by Europcar	Occupants of a car	Passive	Co-generated data	EU
Conversational generative AI	Whisper	Users and contributors (the creators of the scraped data)	Active and passive	Co-generated technology, AI-generated works	US, global

[180] Wendehorst, C. & Cohen, N. (2023). American Law Institute and European Law Institute. 'ALI-ELI Principles for a Data Economy - Data Transactions and Data Rights'. https://www.europeanlawinstitute.eu/fileadmin/user_upload/p_eli/Publications/ALI-ELI_Principles_for_a_Data_Economy.pdf

Analysis of the scenarios by applicable legal frameworks impacting co-generation related rights

For each of the six scenarios, we began with some background contextual research, specifically looking at the companies involved, the co-generators' jurisdiction, and the different co-generating activities. Then, building on the desk research, we conducted additional research and analysis on each scenario and covered company documents, national legislation, and academic and grey literature. This analysis involved examining:

- the different co-generators involved in each scenario, their degree of involvement and their awareness of involvement in co-generation
- the applicable data rights in each scenario, as well as how the rights are reflected through contracts, licences, and terms and conditions
- the IP rights that apply in each scenario, as well as looking at how the rights are reflected through contracts, licences, and terms and conditions
- any other frameworks (including collective rights, fundamental rights and more) that apply in each scenario, and whether these rights are best in practice.

Identifying complications created by AI and reflections on the scenarios

In August and September 2023, the team conducted interviews to explore and confirm our understanding of the co-generation scenarios above, as well as the complications for co-generators created by AI. The team conducted 13 interviews with experts from around the world and from a variety of disciplines. We sought to balance legal expertise – those with an understanding of that specific area of law or rights in a given jurisdiction – with expertise from industry – those with knowledge of the realities of co-generation and impacts of generative AI, alongside experts from civil society. The full list of participants can be found in Appendix 2.

We conducted 45-minute to one-hour semi-structured interviews. The specific questions in each interview were adapted based on the interviewee's expertise, and are included in Appendix 1. They generally covered the following topics:

- the current legal landscape for co-generators
- the current legal landscape around AI for co-generators
- looking forward to potential solutions to support co-generators.

The interviews were recorded and uploaded to Dovetail (user research software enabling collaborative coding). The interviews were thematically coded and insights were drawn to contribute to the analysis of the case studies, the key findings and the areas for further study.

In January 2024 we hosted two workshops with a group of experts to test the findings of this research, and to gain additional feedback for the report.

Limitations of this research

The world of co-generation is expansive, and instances of co-generation occur in lots of different spaces. This research has two main limitations directly related to this:

First, given the timescale of this project, we were only able to cover six scenarios in significant depth, meaning that many other jurisdictions, sectors and situations of co-generation still need to be covered. While we chose a set of jurisdictions and scenarios that could generally account for the broad variety of contexts in this space, it would be worthwhile exploring other scenarios and legal regimes in future studies.

Second, the legal frameworks and sectors explored in this research each command incredible nuance and are composed of profound discourse and expertise. We have summarised the nuances and debates as best as possible in the interest of brevity, but each of these is worth exploring further in future studies.

Process of report drafting and review This interim report was first drafted in July, August and September 2023, with further updates in January, February, March and April 2024.

Appendix 1: Interview guide

Broad research questions

- Do current legal frameworks adequately cover rights of co-generators?
- Do current regimes take into account the specificities of AI?
- How does AI impact co-generated data, and the rights of co-generators?

Background

- Could you tell us a bit more about the role and the organisation that you work for?
- How does your work interact with data and AI: practically, legal, research etc?

Co-generation

- Should co-generators have rights over co-generated data? In what scenarios should they/should they not have rights?
- Do you think there are currently legal frameworks that provide rights for those involved in co-generation?
- If yes:
    - Do you think they are effective?
    - What is needed to make them effective?
    - Are there specific approaches which have worked well?
    - Do current frameworks recognise collectives/communities as having rights as co-generators?
- If no:
    - What can rights for co-generators look like?
    - In addition to data governance laws and IP laws, do you think there are any other legal frameworks that are relevant to co-generation?
    - What can rights for collectives/communities as co-generators look like? How would you define collectives/communities?

Current legal frameworks for AI

- What are your reflections on the current legal discourse around AI?

- Do current movements of legal frameworks for generative AI reflect an accurate understanding of the development and implementation of generative AI?

- Are existing legal frameworks that deal with co-generation adequate to deal with situations involving the use of AI?

Looking forward

- Are there non-legal approaches or solutions which could support co-generators?

- How would you define and understand co-generators?

Appendix 2: Research participants

Expert Interviews
Awi Mona, National Tsing Hua University,
Bertrand Monthubert, National Council of Geographic Information
Carolina Rossini, Datasphere Initiative
George Oates, Flickr Foundation
Jacob Rogers, Wikimedia Foundation
Jeni Tennison, Connected by Data
Linnet Taylor, Tilburg University
Maja Bogataj, Intellectual Property Institute
Rafael Zanatta, Data Privacy Brasil Research Foundation
Rahul Matthan, Trilegal
Safiya Husain, Karya
Teki Akutteh Falconer, Nsiah Akutteh & Co.; Africa Digital Rights Hub
Vikneswaran Kumaran, Lawyer, Singapore

Workshop 1
Awi Mona, National Tsing Hua University
Carolina Rossini, Datasphere Initiative
Eliza McCullough, Partnership on AI
Jai Vipra, Center for Technology and Society
Lucas Costa Dos Anjos, Sciences Po
Michael O'Sullivan, University of Auckland
Micaela Mantegna, Fellow at Datasphere Initiative
Nidhi Kulkarni, Karya
Safiya Husain, Karya

Workshop 2 with GPAI experts
Abdul Majeed
Adrian Weller, Alan Turing Institute
Alison Gillwald, Research ICT Africa
Anurag Agrawal, Council of Scientific and Industrial Research
Avik Sarkar, Indian School of Business
Bertrand Monthubert, Conseil National de l'Information Géolocalisée
Camille Seguin, CEIMIA
Christiane Wendehorst, European Law Institute
Dafna Feinholz, UNESCO
Inma Martinez, Independent Expert in industrial and societal digital transformation
Ivan Bratko, University of Ljubljana
Jaco DuToit, UNESCO
Kyoko Yoshinaga, Keio University
Maja Bogataj, Intellectual Property Institute
Maya Sherman, American India Foundation
Monica Lopez, HOLISTIC AI
Nava Shaked, Holon Institute of Technology.

Naohiro Furukawa, Hitachi Ltd
Paola Ricaurte Quijano, Tecnológico de Monterrey
Paula Garneron, Chief of Staff's Office (Argentina)
Ren Bin Lee Dixon, Center for AI and Digital Policy
Shameek Kundu, Truera
Stefan Janusz, CEIMIA
Sundar Sundareswaran, World Economic Forum
Toshiya Jitsuzumi, Chuo University
Yeong Zee Kin, Infocomm Media Development Authority of Singapore
Zümrüt Müftüoğlu, Yildiz Technical University

Appendix 3: About the organisations involved

Aapti Institute is a non-profit public research organisation that works at the intersection of technology and society, to build solutions that enhance societal impact, justice and equity, building policy-relevant and actionable insights on the digital economy. It was founded in 2019 in Bangalore, India. Through its two labs – the Data Economy Lab and the Digital Public Lab – Aapti conducts research, develops solutions, and embeds them in policy for a more responsible data economy.

The **Open Data Institute (ODI)** is a dynamic non-profit company founded in 2012 by Sir Tim Berners-Lee and Sir Nigel Shadbolt. Its vision is to create a world where data works for everyone. The ODI builds an open, trustworthy data ecosystem, collaborating with businesses, governments and civil society to generate social and economic value. It equips leaders with data skills, shapes public policy, and influences global data-enabled initiatives through training, expert consultancy, and applied research.

Pinsent Masons LLP is a multinational law firm specialising in the technology, science, industry, energy, infrastructure, financial services and real estate sectors.

The **Global Partnership on Artificial Intelligence** (GPAI) is a multi-stakeholder initiative that aims to bridge the gap between theory and practice on AI by supporting cutting-edge research and applied activities on AI-related priorities. Built around a shared commitment to the OECD Recommendation on Artificial Intelligence, GPAI brings together engaged minds and expertise from science, industry, civil society, governments, international organisations and academia to foster international cooperation.

Bibliography

1. Ada Lovelace Institute (2021). 'Participatory data stewardship'. https://www.adalovelaceinstitute.org/report/participatory-data-stewardship/
2. Alimonti, V. et al. (2023). Electronic Frontier Foundation. 'Settled Human Rights Standards as Building Blocks for Platform Accountability and Regulation: A Contribution to the Brazilian Debate'. https://www.eff.org/deeplinks/2023/07/settled-human-rights-standards-building-blocks-platform-accountability-and
3. Anderson, J. et al. (2019). International Journal of Geo-Information. 'Corporate Editors in the Evolving Landscape of OpenStreetMap'. http://dx.doi.org/10.3390/ijgi8050232
4. Arrieta Ibarra, I. et al. (2017). American Economic Association Papers & Proceedings. 'Should We Treat Data as Labor? Moving Beyond "Free"'. https://ssrn.com/abstract=3093683
5. Atz, T. et al. (2023). Amazon Web Services. 'Scaling Automated Driving data processing and data management with BMW Group on AWS'. https://aws.amazon.com/blogs/industries/scaling-autonomous-driving-data-processing-and-data-management-with-bmw-group-on-aws/
6. Ausloos, J. et al. (2022). Ada Lovelace Institute. 'The case for collective action against the harms of data-driven technologies'. https://www.adalovelaceinstitute.org/blog/collective-action-harms/
7. Baio, A. (2022). Waxy. 'Invasive Diffusion: How one unwilling illustrator found herself turned into an AI model'. https://waxy.org/2022/11/invasive-diffusion-how-one-unwilling-illustrator-found-herself-turned-into-an-ai-model/
8. Baker McKenzie (2023). 'China: A landmark court ruling on copyright protection for AI-generated works'. https://insightplus.bakermckenzie.com/bm/data-technology/china-a-landmark-court-ruling-on-copyright-protection-for-ai-generated-works#:~:text=In%20brief,by%20Artificial%20Intelligence%20(AI)'.
9. Bearne, S. (2023). BBC. 'New AI systems collide with copyright law'. https://www.bbc.co.uk/news/business-66231268
10. BMW (n.d.). 'BMW ConnectedDrive'. https://www.bmw.co.uk/en/topics/owners/bmw-connecteddrive/overview.html
11. BMW (n.d.). 'Politique de confidentialité de BMW France'. https://www.bmw.fr/fr/footer/metanavigation/data-privacy.html
12. BMW (n.d.). 'Which mobile network does BMW use to transfer my data to third parties/ service providers?'. https://faq.bmw.com.au/s/article/Car-Data-Data-transmission-Mobile-network-9sdaZ?language=en_AU
13. BMW Group (2022). 'BMW CarData Telematics Data Catalogue'. https://www.bmwgroup.com/content/dam/grpw/websites/bmwgroup_com/innovation/Innovation_Mobilitaet/CarData/3PP-CarData--Telematics_Data_Catalogue-en.pdf
14. BMW Group (n.d.), 'Urban mobility'. https://www.bmwgroup.com/en/innovation/urban-mobility.html
15. Brownlee, M. (1993). Columbia Law Review. 'Safeguarding Style: What Protection Is Afforded to Visual Artists by the Copyright and Trademark Laws?'. https://www.jstor.org/stable/1122961

16. Bull, A. (2022). High Mobility. 'What is a Connected Car?'. https://www.high-mobility.com/blog/what-is-a-connected-car
17. Capgemini (2020). 'Monetizing Vehicle Data: How to fulfil the promise'. https://www.documentcloud.org/documents/22120767-capgeminiinvent_vehicledatamonetization_pov_sep2020#document/p10/a2130251
18. Centre for Communication Governance at National Law University Delhi (n.d.). 'Comments to Niti Aayog on the Draft Discussion Paper on the Data Empowerment and Protection Architecture'. https://ccgdelhi.s3.ap-south-1.amazonaws.com/uploads/ccg-nlu-comments-to-niti-aayog-on-the-draft-discussion-paper-on-the-data-empowerment-and-protection-architecture-238.pdf
19. Chakravorti, B. (2020). Harvard Business Review. 'Why It's So Hard for Users to Control Their Data'. https://hbr.org/2020/01/why-companies-make-it-so-hard-for-users-to-control-their-data
20. Chan, A. & Bradley, H. (2023). Bennett Institute. 'Reclaiming the digital commons'. https://www.bennettinstitute.cam.ac.uk/blog/reclaiming-the-digital-commons/
21. Chui, M. et al. (2018). McKinsey. 'What AI can and can't do (yet) for your business'. https://www.mckinsey.com/capabilities/quantumblack/our-insights/what-ai-can-and-cant-do-yet-for-your-business
22. Clutton-Brock, P. et al. (2021). Global Partnership on AI Report, Climate Change AI, & the Centre for AI & Climate. 'Climate Change & AI: Recommendations for Government'. https://www.gpai.ai/projects/climate-change-and-ai.pdf
23. CNIL (2018). 'Connected vehicles and personal data'. https://www.cnil.fr/sites/cnil/files/atoms/files/cnil_pack_vehicules_connectes_gb.pdf
24. Code de la propriété intellectuelle replier partie législative (Articles L111-1 à L811-6). https://www.legifrance.gouv.fr/codes/section_lc/LEGITEXT000006069414/LEGISCTA000006161634/2023-03-08/?anchor=LEGIARTI000006278879#LEGIARTI000006278879
25. Cooke, K. (2016). Reuters. 'U.S. police used Facebook, Twitter data to track protesters: ACLU'. https://www.reuters.com/article/us-social-media-data-idUSKCN12B2L7
26. Copyright Act, 1957. https://www.indiacode.nic.in/bitstream/123456789/1367/1/A1957-14.pdf
27. Crawford, K. (2023). Green European Journal. 'Mining for Data: The Extractive Economy Behind AI'. https://www.greeneuropeanjournal.eu/mining-for-data-the-extractive-economy-behind-ai/
28. Dallery Gallery (2022). 'Everything you wanted to know about MidJourney'. https://dallery.gallery/midjourney-guide-ai-art-explained/
29. Department for Digital, Culture, Media & Sport & Lopez, J. MP (2021). 'Minister Lopez speech to the Advertising Standards Authority Parliamentary Breakfast'. https://www.gov.uk/government/speeches/minister-lopez-speech-to-the-advertising-standards-authority-parliamentary-breakfast
30. Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC, *Official Journal* L 130/92. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32019L0790
31. Directive 96/9/EC Of The European Parliament and of The Council of 11 March 1996 on the legal protection of databases, *Official Journal* L77, p. 20. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A31996L0009
32. Dzieza, J. (2023). The Verge. 'AI Is a Lot of Work'. https://www.theverge.com/features/23764584/ai-artificial-intelligence-data-notation-labor-scale-surge-remotasks-openai-chatbots

33. Eavis, P. & Lohr, S. (2020). The New York Times. 'Big Tech's Domination of Business Reaches New Heights'. https://www.nytimes.com/2020/08/19/technology/big-tech-business-domination.html
34. Europcar (n.d.). 'Privacy policy for connected vehicles'. https://www.europcar.com/files/live/sites/erc/files/connected-cars/privacy-policy.pdf
35. Europcar Germany (2016). 'Terms and conditions of hire of Europcar Autovermietung GmbH'. https://oos.glasstec-online.com/medias/AVB-englisch-0617.pdf?context=bWFzdGVyfHJvb3R8MzIyODkyNnxhcHBsaWNhdGlvbi9wZGZ8aDMzL2gyNi85MTk0NjAxNzQyMzY2LnBkZnxiNjMzY2JlYjJmOTUxNjg5OTJhZTczOWQyMmY2MmU5ZWYzNTY2NzYwOWE1NWQ0MjA3ZmRhNDhjOGYyZmM1NWJh
36. European Data Protection Board (2019). 'Opinion of the Board (Art. 64): Opinion 5/2019 on the interplay between the ePrivacy directive and the GDPR, in particular regarding the competence, tasks and powers of data protection authorities'. https://edpb.europa.eu/sites/default/files/files/file1/201905_edpb_opinion_eprivacydir_gdpr_interplay_en_0.pdf
37. European Data Protection Board (2021). 'Guidelines: Guidelines 01/2020 on processing personal data in the context of connected vehicles and mobility related application'. https://edpb.europa.eu/system/files/2021-03/edpb_guidelines_202001_connected_vehicles_v2.0_adopted_en.pdf
38. European Union Agency for Fundamental Rights (n.d.). 'What are fundamental rights?'. http://fra.europa.eu/en/about-fundamental-rights
39. Gomer, R., & Simperl, E. (2020). Cambridge University Press. 'Trusts, co-ops, and crowd workers: Could we include crowd data workers as stakeholders in data trust design?'. doi:10.1017/dap.2020.21
40. GPAI (2020). Global Partnership on AI. 'A Framework Paper for GPAI's work on Data Governance'. https://gpai.ai/projects/data-governance/gpai-data-governance-work-framework-paper.pdf
41. GPAI (2021). Global Partnership on AI. 'Enabling data sharing for social benefit through data trusts'. https://gpai.ai/projects/data-governance/data-trusts/enabling-data-sharing-for-social-benefit-data-trusts-interim-report.pdf
42. GPAI (2022). Global Partnership on AI. ' Protecting AI innovation, Intellectual Property (IP): GPAI IP Primer, Report'. https://gpai.ai/projects/innovation-and-commercialization/gpai-intellectual-property-primer-2022.pdf
43. GPAI (2022). Global Partnership on AI. 'Protecting AI innovation, Intellectual Property (IP): GPAI IP Expert: (I) Guidelines for Scraping or Collecting Publicly Accessible Data and (II) the Preliminary Report on Data and AI Model Licensing, Report'. https://gpai.ai/projects/innovation-and-commercialization/intellectual-property-expert-preliminary-report-on-data-and-AI-model-licensing.pdf
44. GPAI (2023). 'Fostering Contractual Pathways for Responsible AI Data and Model Sharing for Generative AI and Other AI Applications'. https://gpai.ai/projects/responsible-ai/IC_Intellectual%20Property%20project.pdf
45. GPAI (2023). Global Partnership on AI. 'AI for Fair Work, From principles to practices'. https://gpai.ai/projects/future-of-work/FoW2_AI%20Fair%20Work%20.pdf
46. GPAI (2023). Global Partnership on AI. 'Fairwork AI Ratings 2023: The Workers Behind AI at Sama'. https://gpai.ai/projects/future-of-work/FoW-Fairwork-AI-Ratings-2023.pdf

47. GPAI Working Group on the Responsible Development, Use and Governance of AI (2020). 'A Framework Paper for GPAI's work on Data Governance'. https://gpai.ai/projects/data-governance/gpai-data-governance-work-framework-paper.pdf
48. Groves, L. et al. (2023). FAccT 'Going public: the role of public participation approaches in commercial AI labs'. https://doi.org/10.1145/3593013.3594071
49. Guadamuz, A. (2023). 'A Scanner Darkly: Copyright Liability and Exceptions in Artificial Intelligence Inputs and Outputs'. https://ssrn.com/abstract=4371204
50. Hao, K. (2022). MIT Technology Review. 'A new vision of artificial intelligence for the people'. https://www.technologyreview.com/2022/04/22/1050394/artificial-intelligence-for-the-people/
51. Haqqi, Ty. (2023). Yahoo! Finance. '25 Most Profitable Companies in the World'. https://finance.yahoo.com/news/25-most-profitable-companies-world-192146617.html?
52. Herfort, B. et al. (2021). Scientific Reports. 'The evolution of humanitarian mapping within the OpenStreetMap community'. https://doi.org/10.1038/s41598-021-82404-z
53. His Holiness Kesavananda Bharati v. State Of Kerala (1973). Supreme Court of India. Judgement. https://judgments.ecourts.gov.in/KBJ/
54. HM Treasury (2023). 'Pro-innovation Regulation of Technologies Review: Digital Technologies'. https://www.gov.uk/government/publications/pro-innovation-regulation-of-technologies-review-digital-technologies
55. Hu, K. (2023). Reuters. 'ChatGPT sets record for fastest-growing user base'. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/
56. HuggingFace (n.d.). 'Whisper'. https://huggingface.co/openai/whisper-large-v2
57. IBM (n.d.). 'What is the internet of things?'. https://www.ibm.com/topics/internet-of-things
58. Instagram (2022). 'Data policy, Instagram'. https://help.instagram.com/155833707900388
59. Instagram (2023). 'Community Guidelines, Instagram'. https://help.instagram.com/477434105621119/?helpref=faq_content
60. Instagram (2023). 'Terms of use, Instagram'. https://help.instagram.com/581066165581870/?helpref=hc_fnav
61. Instagram (n.d.). 'Earn money on Instagram'. https://creators.instagram.com/earn-money?locale=en_GB
62. Instagram Engineering (2015). 'Instagration Pt. 2: Scaling our infrastructure to multiple data centers'. https://instagram-engineering.com/instagration-pt-2-scaling-our-infrastructure-to-multiple-data-centers-5745cbad7834
63. Institui a Lei Brasileira de Liberdade, Responsabilidade e Transparência na Internet (2020). 'Parecer Proferido Em Plenário Ao Projeto De Lei Nº'. https://www.camara.leg.br/proposicoesWeb/prop_mostrarintegra?codteor=2265334&filename=Tramitacao-PL%202630/2020
64. International Labour Office (2018). 'Digital labour platforms and the future of work: Towards decent work in the online world'. https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/---publ/documents/publication/wcms_645337.pdf
65. Ipsos (2022). NHS AI Lab Public Dialogue on Data Stewardship. https://www.ipsos.com/sites/default/files/ct/news/documents/2022-11/22-033229-01%20NHS%20AI%20Lab%20Data%20Stewardship%20Dialogue%20-%20Report_0.pdf

66. Isaak, J. & Hanna, M. J. (2018). IEEE. ‘User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection’. https://ieeexplore.ieee.org/document/8436400
67. Janeček, V. (2018). ‘Ownership of personal data in the Internet of Things’. Computer Law & Security Review, Volume 34, Issue 5. https://doi.org/10.1016/j.clsr.2018.04.007.
68. K&S Partners and J. Sagar Associates (2021). Thomson Reuters. ‘Intellectual property right assignments Q&A: India’. https://www.jsalaw.com/wp-content/uploads/2021/07/Intellectual-property-right-assignments-QAndA-India.pdf
69. Kahl, T. (2023). TaylorWessing. ‘Mobility is going digital!’. https://www.taylorwessing.com/en/interface/2023/iot---next-gen/mobility-is-going-digital-what-connected-vehicle-manufactures-need-to-think-about-in-2023
70. Kaitiakitanga-License (n.d.). https://github.com/TeHikuMedia/Kaitiakitanga-License?ref=blog.papareo.nz
71. Kapoor, A. & Vaitla, B. (2022). Brookings. ‘Data co-ops: How cooperative structures can support women’s empowerment’. https://www.brookings.edu/articles/data-co-ops-how-cooperative-structures-can-support-womens-empowerment/
72. Karya (2022). ‘The Future is Data (Cooperatives)’. https://karya.in/resources/blog/the-future-is-data/
73. Kayra (n.d.), ‘Ethical Data Pledge’. https://www.ethicaldatapledge.com/
74. Keegan, J. & Ng, A. (2022). The Markup. ‘Who Is Collecting Data from Your Car?’. https://themarkup.org/the-breakdown/2022/07/27/who-is-collecting-data-from-your-car
75. Knox, R. (2021). Wired. ‘Big Music Needs to Be Broken Up to Save the Industry’. https://www.wired.com/story/opinion-big-music-needs-to-be-broken-up-to-save-the-industry/
76. Komaitis, K. & Sherman, J. (2021). Brookings. ‘US and EU tech strategy aren’t as aligned as you think’. https://www.brookings.edu/articles/us-and-eu-tech-strategy-arent-as-aligned-as-you-think/
77. Lab Lab AI (n.d.). ‘OpenAI Whisper Applications’. https://lablab.ai/apps/tech/openai/whisper
78. Leffer, L. (2023). Scientific American. ‘Your Personal Information Is Probably Being Used to Train Generative AI Models’. https://www.scientificamerican.com/article/your-personal-information-is-probably-being-used-to-train-generative-ai-models/
79. Lei Geral de Proteção de Dados Pessoais (LGPD). English translation accessed here: https://iapp.org/resources/article/brazilian-data-protection-law-lgpd-english-translation/
80. Lemley, M. & Casey, B. (2021). Texas Law Review. ‘Fair Learning’. https://texaslawreview.org/fair-learning/
81. Lubin, A. (2023). Temple Law Review. ‘Collective Data Rights and their Possible Abuse.’ https://www.templelawreview.org/essay/collective-data-rights-and-their-possible-abuse/
82. Mahelona, K. et al. (2023). Papa reo. ‘OpenAI's Whisper is another case study in Colonisation’. https://blog.papareo.nz/whisper-is-another-case-study-in-colonisation/
83. Margoni, T., & Kretschmer, M. (2022). GRUR International. ‘A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology’. https://doi.org/10.1093/grurint/ikac054
84. McFarland, M. (2017). CNN. ‘Your car's data may soon be more valuable than the car itself’. https://money.cnn.com/2017/02/07/technology/car-data-value/index.html
85. Meta (n.d.). ‘How Meta uses information for generative AI models’. https://privacycenter.instagram.com/privacy/genai/

86. Midjourney (2023). ‘Terms of Service’. https://docs.midjourney.com/docs/terms-of-service
87. Midjourney (n.d.). ‘Community Showcase’. https://www.midjourney.com/showcase/recent/
88. Mizushima, K. & Ikawa, Y. (2011). IEEE. ‘A structure of co-creation in an open source software ecosystem: A case study of the eclipse community’. https://ieeexplore.ieee.org/document/6017787
89. Motor 1 (2022). ‘BMW uses customers' driving data to improve its in-car features’. https://uk.motor1.com/news/580364/bmw-collecting-customer-driving-data/
90. National Science and Technology Council Committee on Technology (2016). Executive Office of the President. ‘Preparing for the future of Artificial Intelligence’. https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf
91. Naughton, K. (2021). Bloomberg. ‘Driverless Cars’ Need for Data Is Sparking a New Space Race’. https://www.bloomberg.com/news/articles/2021-09-17/carmakers-look-to-satellites-for-future-of-self-driving-vehicles
92. Neuffer, B. & Kresz, M. (2020). American Bar Association. ‘Don’t Give Away Your Intellectual Property’. https://www.americanbar.org/groups/construction_industry/publications/under_construction/2020/winter2020/dont-give-away-your-intellectual-property/
93. Nissenbaum, H. (2004). Washington Law Review. ‘Privacy as Contextual Integrity’. https://digitalcommons.law.uw.edu/wlr/vol79/iss1/10
94. ODI (2023). ‘Responsible data stewardship’. https://www.theodi.org/article/defining-responsible-data-stewardship/
95. ODI (2023). ‘What do we mean by “without data, there is no AI”?’. https://theodi.org/news-and-events/blog/what-do-we-mean-by-without-data-there-is-no-ai/
96. Open Knowledge Foundation (n.d.). ‘Open Data Commons Open Database License (ODbL)’. https://opendatacommons.org/licenses/odbl/
97. OpenAI (2022). ‘Robust Speech Recognition via Large-Scale Weak Supervision’. https://cdn.openai.com/papers/whisper.pdf
98. OpenAI (2023). ‘Introducing ChatGPT and Whisper APIs’. https://openai.com/blog/introducing-chatgpt-and-whisper-apis
99. OpenAI (2023). ‘Privacy policy’. https://openai.com/policies/privacy-policy
100. OpenAI (2023). ‘Terms of use’. https://openai.com/policies/terms-of-use
101. OpenStreetMap (n.d.). ‘Contributors’. https://wiki.openstreetmap.org/wiki/Contributors
102. OpenStreetMap (n.d.). ‘Copyright’. https://www.openstreetmap.org/copyright
103. OpenStreetMap (n.d.). ‘Licence/Licence Compatibility’. https://wiki.osmfoundation.org/wiki/Licence/Licence_Compatibility
104. OpenStreetMap Foundation (n.d.). ‘Licence/Contributor Terms’. https://wiki.osmfoundation.org/wiki/Licence/Contributor_Terms
105. OpenStreetMap Foundation (n.d.). ‘Privacy Policy’. https://wiki.osmfoundation.org/wiki/Privacy_Policy
106. OpenStreetMap Foundation (n.d.). ‘Terms of Use’. https://wiki.osmfoundation.org/wiki/Terms_of_Use
107. Perrigo, B. (2022). Time. ‘Inside Facebook's African Sweatshop’. https://time.com/6147458/facebook-africa-content-moderation-employee-treatment/
108. Quintais, J. P. (2023). Copyright Blog. ‘Generative AI, Copyright and the AI Act’. https://copyrightblog.kluweriplaw.com/2023/05/09/generative-ai-copyright-and-the-ai-act/

109. Reed, R. (2024). Harvard Law Today. ‘ChatNYT: Harvard Law expert in technology and the law says the New York Times lawsuit against ChatGPT parent OpenAI is the first big test for AI in the copyright space’. https://hls.harvard.edu/today/does-chatgpt-violate-new-york-times-copyrights/
110. Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act), *Official Journal* L, 2023/2854. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202302854
111. Reuters (2024). ‘OpenAI's ChatGPT breaches privacy rules, says Italian watchdog’. https://www.reuters.com/technology/cybersecurity/italy-regulator-notifies-openai-privacy-breaches-chatgpt-2024-01-29/
112. Richard Kadrey, Sarah Silverman & Christopher Golden v. Meta Platforms, Inc. (2023). United States District Court Northern District of California. Complaint. https://llmlitigation.com/pdf/03417/kadrey-meta-complaint.pdf
113. Routley, N. (2023). World Economic Forum. ‘What is generative AI? An AI explains’. https://www.weforum.org/agenda/2023/02/generative-ai-explain-algorithms-work/
114. Salkowitz, R. (2022). Forbes. ‘Midjourney Founder David Holz On The Impact Of AI On Art, Imagination And The Creative Economy’. https://www.forbes.com/sites/robsalkowitz/2022/09/16/midjourney-founder-david-holz-on-the-impact-of-ai-on-art-imagination-and-the-creative-economy/?sh=b14a0aa2d2b8
115. Sarah Silverman, Christopher Golden & Richard Kadrey v. OpenAI, Inc. & Others (2023). United States District Court Northern District of California. Complaint. https://llmlitigation.com/pdf/03416/silverman-openai-complaint.pdf
116. Sheng, E. (2023). CNBC. ‘In generative AI legal Wild West, the courtroom battles are just getting started’. https://www.cnbc.com/2023/04/03/in-generative-ai-legal-wild-west-lawsuits-are-just-getting-started.html
117. Singh, P. J. & Vipra, J. (2019). Development. ‘Economic Rights Over Data: A Framework for Community Data Ownership’. https://link.springer.com/article/10.1057/s41301-019-00212-5
118. Solove, D. J. (2013). Harvard Law Review. ‘Privacy Self-Management and the Consent Dilemma’. https://scholarship.law.gwu.edu/cgi/viewcontent.cgi?referer=&httpsredir=1&article=2093&context=faculty_publications
119. Stephen Thaler, vs. Shira Perlmutter (2022). https://ecf.dcd.uscourts.gov/cgi-bin/show_public_doc?2022cv1564-24
120. Thapelo, S. T. et al. (2021). Data Science Journal. ‘SASSCAL WebSAPI: A Web Scraping Application Programming Interface to Support Access to SASSCAL’s Weather Data’. https://datascience.codata.org/articles/10.5334/dsj-2021-024.
121. The AI Patent Blog (2023), ‘Legal protection from generative AI in Japan’. https://www.theaipatentblog.com/legal-protection-from-generative-ai-in-japan
122. The Brazilian Civil Framework of the Internet. https://bd.camara.leg.br/bd/bitstream/handle/bdcamara/26819/bazilian_framework_%20internet.pdf
123. The Economist (2024). ‘Generative AI is a marvel. Is it also built on theft?’. https://www.economist.com/business/2024/04/14/generative-ai-is-a-marvel-is-it-also-built-on-theft

124. United States Copyright Office (2023). ‘Artificial Intelligence Study’. https://www.copyright.gov/policy/artificial-intelligence/
125. United States Copyright Office (2023). ‘Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence’. https://copyright.gov/ai/ai_policy_guidance.pdf
126. United States Copyright Office (2023). ‘U.S. Copyright Office Fair Use Index’. https://www.copyright.gov/fair-use/index.html
127. Universal Declaration of Human Rights. https://www.un.org/en/about-us/universal-declaration-of-human-rights
128. van Doorn, N. & Badger, A. (2020). Antipode: A Radical Journal of Geography. ‘Platform Capitalism’s Hidden Abode: Producing Data Assets in the Gig Economy’. https://doi.org/10.1111/anti.12641
129. Verhulst, S. (2024). ‘Are we entering a “Data Winter”?’. https://sverhulst.medium.com/are-we-entering-a-data-winter-f654eb8e8663
130. Viljoen, S. (2020). Yale Law Journal. ‘A Relational Theory of Data Governance’. http://dx.doi.org/10.2139/ssrn.3727562
131. Vincent. J. (2022). The Verge. ‘The scary truth about AI copyright is nobody knows what will happen next’. https://www.theverge.com/23444685/generative-ai-copyright-infringement-legal-fair-use-training-data
132. Wendehorst, C. & Cohen, N. (2023). American Law Institute and European Law Institute. ‘ALI-ELI Principles for a Data Economy - Data Transactions and Data Rights’. https://www.europeanlawinstitute.eu/fileadmin/user_upload/p_eli/Publications/ALI-ELI_Principles_for_a_Data_Economy.pdf
133. WIPO (n.d.). ‘Copyright’. https://www.wipo.int/copyright/en/
134. WIPO (n.d.). ‘What is Intellectual Property?’. https://www.wipo.int/about-ip/en/
135. World Bank (n.d.). ‘Crowd-sourced Data’. https://dimewiki.worldbank.org/Crowd-sourced_Data
136. World Economic Forum (2020). ‘Redesigning Data Privacy: Reimagining Notice & Consent for human technology interaction’. https://www3.weforum.org/docs/WEF_Redesigning_Data_Privacy_Report_2020.pdf