GPAI IP Expert

Preliminary Report on Data and AI Model Licensing

November 2022

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 Protecting AI innovation, Intellectual Property (IP). The report reflects the personal opinions of the GPAI Experts and Specialists involved and does not necessarily reflect the views of the Experts' organizations, 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 *Protecting AI innovation, Intellectual Property (IP)* project, with the steering of the project Co-Leads and the guidance of the Project Advisory Group, supported by the GPAI Innovation & Commercialization Working Group. The GPAI Innovation & Commercialization Working Group agreed to declassify this report and make it publicly available.

Co-Leads:
Lee Tiedrich[*], Duke University
Josef Drexl[*], Max Planck Institute for Innovation and Competition

Project Advisory Group:
Teki Akuetteh[*], Africa Digital Rights' Hub
Carolyn Blankenship[†], BigBear.ai
Neeraj Gupta[**], Formulate IP
Matthieu Dhenne[**], IPSILON Dispute & Resolution
Yann Dietrich[**], Atos
Josef Drexl[*], Max Planck Institute for Innovation and Competition
Antoine Dupont[**], CEA
Robert Kroplewski[*], Poland's Minister of Digitalization in Information Society Affairs, Chancellery of Prime Minister
Andreas Liebl[*], appliedAI
Hisao Shiomi[†], Tsukuba University
Rajesh Sreenivasan[†], Technology Media and Telecoms Law Practice, Rajah & Tann Singapore LLP
Lee Tiedrich[*], Duke University
Ulrike Till[†], WIPO
Kat Walsh[†], Creative Commons

Research Assistants:
Marina Bojarski‡, GPAI Junior Investigator - under the direction of Yann Dietrich[**]
Daria Kim‡, under the direction of Josef Drexl[*]
Annmarie Messing‡, under the direction of Lee Tiedrich[*]
Joe O'Brien‡, under the direction of Lee Tiedrich[*]

Yusuke Yogoro‡, under the direction of Hisao Shiomi[†]

The report was written by: **Lee Tiedrich**[*], Duke University; **Josef Drexl**[*], Max Planck Institute for Innovation and Competition and **Daria Kim**‡, under the direction of Josef Drexl[*]; with the significant contribution and involvement of the following individuals in the Project Advisory Group: **Hisao Shiomi**[†], Tsukuba University; **Yusuke Yogoro**‡, under the direction of Hisao Shiomi[†]; **Joe O'Brien**‡, under the direction of Lee Tiedrich and **Annmarie Messing**‡, under the direction of Lee Tiedrich[*]; GPAI would also like to thank the following individuals who provided valuable comments to enrich the report: **Teki Akuetteh**[*], Africa Digital Rights' Hub; **Neeraj Gupta**[**], Formulate IP; **Robert Kroplewski**[*], Poland's Minister of Digitalization in Information Society Affairs, Chancellery of Prime Minister; **Rajesh Sreenivasan**[†], Technology Media and Telecoms Law Practice, Rajah & Tann Singapore LLP; **Ulrike Till**[†], WIPO; **Kat Walsh**[†], Creative Commons. Finally, GPAI would like to acknowledge the efforts of colleagues at the Paris Centre of Expertise at Inria. We are particularly grateful for the support of Kaitlyn Bove, Laetitia Cuignet, Isabelle Herlin, and Catherine Pacherie-Simeral from Inria, and for the dedication of the Working Group Co-Chairs Laurence Liew and Françoise Soulié-Fogelman.

* Expert of GPAI's Innovation & Commercialization Working Group
(n.b. Teki Akuetteh and Josef Drexl are Experts in GPAI's Data Governance Working Group)
** Observer at GPAI's Innovation & Commercialization Working Group
† Invited Specialist
‡ Contracted parties by Inria

Citation

GPAI 2022. 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, November 2022, Global Partnership on AI.

Executive Summary

The premise that access to data is key for data-driven innovation—including for the development of artificial intelligence (AI) systems and applications—is broadly recognized.[1] Yet multiple technical, economic and legal challenges to barrier-free and responsible data sharing persist.[2] It is acknowledged that the standardization of data-sharing agreements may mitigate or help overcome some of such challenges and thereby aid and foster innovation across sectors and jurisdictions. In particular, standard terms and agreements can enhance legal certainty and reduce transaction costs related to contract negotiations and formation.

A variety of approaches to the standardization of data-sharing agreements are conceivable. Some earlier efforts and existing initiatives[3] have attempted to rely on Open Source and/or Creative Commons licenses. The experience has shown that such frameworks and the underlying legal concepts are not readily applicable or transferrable (or otherwise ideal) to situations of data sharing. Efforts to create standardized data-sharing terms and agreements that would better account for the specificities of data and the related technical, economic, and legal aspects are at a nascent stage. Nevertheless, it is clear at this point that a fully-fledged, internationally applicable data-sharing agreement is hardly feasible in the short term. It also is apparent that bespoke agreements that reflect contextual particularities will continue playing a role. This is analogous to situations today where both bespoke agreements and standardized Open Source and Creative Commons agreements all have important roles in fostering transactions.

Based on the preliminary work to date, our findings and recommendations are as follows:

- Given the recognized benefits of facilitating data sharing arrangements, the continuation of work would be encouraged to develop standardized data sharing terms that can help streamline transactions.

- To gain broad acceptance, the standardized terms will likely need to be developed through an inclusive multi-stakeholder process. Various organizations are working on standardized licensing terms, and they are encouraged (and other organizations that decide to pursue this work) to include as many different viewpoints and stakeholders as possible in the process. This should lead to more informed decisions about the content and structure of the standardized terms and broader acceptance and adoption of such terms.

- There likely will continue to be a wide range of different data sharing arrangements and use cases, and a “one size fits all” approach for data licensing may not be optimal, or even feasible. Organizations working on standardized data licensing terms are encouraged to consider developing a menu of different provisions or agreements that provide the community with options. This is similar to the approach used for Open Source and Creative Commons license agreements and already is reflected in some ongoing efforts to develop standardized data licensing terms.

[1] See Section 2.1.

[2] See Section 2.2.

[3] See Section 2.3.

- It is expected that there will continue to be a need for bespoke data licenses, even as standardized terms become more common and accepted. This parallels the experience with Open Source and Creative Commons licenses.

- Section 3.4 of this report highlights some topics to be considered when contemplating standardized terms, including (i) standardizing definitions and developing models for allocating proprietary rights and usage and access rights (including when ethical considerations and/or data sovereignty are relevant considerations), (ii) addressing privacy and confidentiality, including when personally identifiable information (PII) and/or confidential information is being shared and/or privacy enhancing technologies (PETs) may be used, (iii) working toward fostering more data interoperability and better data quality and technical characteristics, (iv) allocating liability and providing for enforcement, (v) addressing software-as-a-service (SaaS) and other business models, and (vi) providing a framework for addressing data governance. It is recognized that developing standardized license terms for all of these topics may not be feasible or easy, particularly in the short term and given the challenges that exist. Therefore, organizations would be encouraged to prioritize work on those terms that seem most feasible and to continue to consider approaches to address the more challenging terms.

- While organizations work to develop standardized data license terms, the global community is encouraged to continue to work on addressing the following issues that make this work more challenging:

  - Technical Matters: As reflected in the report, efforts to develop standards for defining and measuring data quality (including in light of proposed AI regulations), fostering data interoperability, and other technical matters remain relatively nascent. Progress on this front could significantly enhance data sharing and the negotiation of data licenses (including the crafting of standardized terms). Ongoing efforts in this area, both on a sectorial basis as well as more broadly, are encouraged.

  - Legal Uncertainties: As reflected in the report, the evolving legal landscape and need for more cross-border harmonization create further obstacles to data licensing and the crafting of standardized license terms. While this report does not express any views on how the underlying issues should be resolved, it does want to sharpen the focus on how these legal and regulatory issues impact data licensing, so this correlation can be considered as policies continue to evolve. The legal and policy context in which AI innovation takes place is undergoing dynamic developments that should be factored into data-sharing practices. While some of the legal and other policy developments are reflected in the report, a comprehensive account of such developments would go beyond the report's scope given its preliminary character.

  - Business Uncertainties: As also reflected in this report, business uncertainties can impede the negotiation of data sharing arrangements. While this report generally does not express any views on the underlying business issues, it encourages the community to consider whether standardized terms ultimately might be crafted to reflect common business models that may emerge and to provide flexibility for parties to mitigate context-specific business risks and concerns. Among other things, this work could build on efforts to develop terms for AI Software-as-a-Service (SaaS). The work also should take into consideration the need to have options that reflect balanced terms for liability and other risks (e.g., and are not limited to “as is” agreements that disclaim all liability). Crafting of such terms (as well as any other terms) must be undertaken in compliance with competition laws and other applicable laws.

- Data Justice: As data sharing arrangements continue to develop, this report also encourages parties to continue to focus on data justice considerations, which have been highlighted by the GPAI Data Governance Working Group.

The IP Committee of the Innovation and Commercialization (I&C) Working Group of the Global Partnership on Artificial Intelligence (GPAI) remains committed to advancing data licensing work, with the goal of unlocking beneficial data-sharing arrangements, including those that can enhance the development of responsible AI applications. It is anticipated that this report will aid the global community as it undertakes this important work. Given the challenges with this work, it is believed that this Committee is best positioned to help in this effort by focusing on specific use cases. Toward this end, the IP Committee plans to collaborate during the upcoming year with other GPAI Working Groups, such as the Data Institutions Committee within the Data Governance Working Group and the AI and Climate Working Group, as they work on data-sharing projects. More specifically, this Committee can collaborate with these other Working Groups to help identify data licensing terms that can support these broader data-sharing efforts. Through this work, the Committee hopes to make contributions that can inform the broader efforts to develop standardized data licensing terms.

Finally, the IP Committee commends those organizations that are working on developing data licensing terms. The Committee invites those organizations to contact us with further questions about our findings and recommendations and to suggest ways the Committee might be able to assist them in advancing their work in a way that is consistent with our preliminary findings and recommendations.

Report Co-Leads

This report was co-led by Lee Tiedrich and Josef Drexl.

Lee J. Tiedrich is the Distinguished Faculty Fellow in Ethical Technology, at Duke University, with a dual appointment at Duke Law School and Duke Science & Society. Building upon her 30 years of practicing law at a leading global law firm and her electrical engineering studies, Professor Tiedrich focuses on developing practical solutions that help enable society to unlock the benefits of data, AI, and other emerging technologies in a trusted manner that also protects fundamental rights and our national security. She is a member of the Global Partnership on AI (GPAI) Innovation and Commercialization Working Group and co-chair of the GPAI IP Committee and a co-lead of the GPAI AI & Climate Steering Committee. She is co-authoring the first law school case book on Artificial Intelligence law and served on the Biden Campaign Policy Committee. Professor Tiedrich has written and spoken extensively on AI and emerging technologies, including at the Council on Foreign Relations and other prominent venues. She has served as a peer reviewer for *Oxford University Press* and is registered to practice before the United States Patent and Trademark Office. Professor Tiedrich received a B.S.E. in electrical engineering from Duke University (*Phi Beta Kappa and Tau Beta Pi)* and a J.D. from the University of Pennsylvania Law School, where she was an adjunct faculty member before joining the Duke University faculty.

Josef Drexl is Director of the Max Planck Institute for Innovation and Competition (Munich), Honorary Professor at the University of Munich and Member of the Bavarian Academy of Science. He was the founding Chair of the Academic Society for Competition Law (ASCOLA) from 2003 to 2013, and he is a Vice-President of the Association Internationale de Droit Economique (AIDE). Josef Drexl is an expert in competition law, intellectual property law and consumer protection law. More recent work focuses on the IP, competition, consumer and data protection law issues of the new digital economy in times of the Internet of Things and Artificial Intelligence. He is a member of the Data Governance Working Group of the Global Partnership for Artificial Intelligence (GPAI).

The following organizations participated in interviews with the IP Committee in connection with this report:

ABEJA
Naohiro Furukawa In-House Counsel at ABEJA

ALEIA: Accélérer vos projets d'Intelligence Artificielle
Antoine Couret President Hub France IA & President @ALEIA: The SaaS collaborative and sovereign AI platform that accelerate and secure AI projects to production

Creative Commons
Kat Walsh General Counsel at Creative Commons

Google
Janel Thamkul Senior Product Counsel at Google
Laura Sheridan Senior Patent Counsel at Google
Ana Ramalho Copyright Counsel at Google
Will Carter Global Policy Lead for Responsible AI, AI Governance and Regulation at Google
Azita Saghafi Corporate Counsel at Google

Linux Foundation
Karen Copenhaver Advisor to Non-Profits and Counsel at The Linux Foundation
Michael Dolan SVP and GM of Projects at The Linux Foundation

METI: Ministry of Economy, Trade and Industry
Professor Hiroki Habuka Research Professor at Kyoto University and former Deputy Director for Global Digital Governance at the Japanese Ministry of Economy, Trade and Industry

Microsoft
Krishna Sood Assistant General Counsel at Microsoft

Noerr
Dr. Jonas Siglmüller Lawyer at Noerr
Dr. David Bomhard Lawyer at Noerr

Contents

1. Introduction

The IP Committee of the I&C Working Group of GPAI has launched this report to examine the feasibility of and progress towards creating standardized agreements for data and AI model licensing. The IP Committee commenced this report in spring 2022, with the goal of facilitating voluntary data sharing, including for purposes of fostering further development of artificial intelligence (AI). The purpose of this preliminary report is to explain the report and share the preliminary findings and recommendations based on work through August 2022.

1.1 Underlying premises

As explained in Section 3.1, a growing consensus has emerged that more tools are needed to facilitate and support the voluntary sharing of data. Standardized or model data licensing terms could be one of the tools that might help achieve this goal. This approach has proved effective in similar contexts. For instance, *Open Source and Creative Commons* agreements have been helpful tools for the voluntarily sharing of software and content, respectively, particularly in situations where more complex or bespoke arrangements are not needed.

Given that "big data" and AI models constitute different subject matter compared to that addressed under the Open Source and Creative Commons licenses, the question is ripe for assessing what new standardized form license agreements should be crafted for data and AI models. As discussed in Section 2.3.2, standardized data license agreements may not replace all bespoke agreements, just as Open Source and Creative Commons licenses have not obviated the need for bespoke license terms in all circumstances. However, standardized license terms could help fill a marketplace need that could streamline some transactions. Even when bespoke data licensing terms are still needed, having standardized data licensing terms could provide a good starting point and facilitate these negotiations.

1.2 Project objective, research questions and report structure

The overall goal of this project is to provide guidance in support of the efforts to develop license templates or standard terms for data-sharing agreements that will facilitate voluntary data sharing, including with a view to developing responsible AI applications. Such templates or terms, even if not adopted as industry standards, could facilitate some common and more straightforward transactions as well as provide a more concrete foundation for drafting bespoke data-sharing provisions.

At this stage, the interim goal of the project, as reflected in the structure of this report, is to map out the following: (i) the methodology used for the project to date (Section 1.3), (ii) the need to facilitate data sharing (Section 2.1) (iii) some of the significant challenges to expanding data sharing (Section 2.2), (iv) certain ongoing efforts to develop standardized data sharing agreements and the need for a variety of approaches (Section 2.3), (v) certain terms and considerations that should be factored into data sharing agreements (including possibly standardized terms) to address some of the challenges discussed in this report (Section 2.4), and (vi) preliminary findings and conclusions and possible next step for the IP Committee to advance data licensing efforts (Section 3).

1.3 Methods and Procedure

To map out issues around the standardization of data-sharing agreements and current developments in this field, the project draws on the following methods.

- *Review* of certain available literature and evidence on data-sharing contractual practices, including studies and reports, industry guidelines and policies, governmental laws and policies, and existing templates and standard contract terms.[4]
- *Semi-structured interviews* based on the questionnaire developed by the Committee members.[5] The main criteria for selecting interviewees were expert knowledge in standardization of contractual terms and hands-on experience in contractual transactions involving data, including for AI-related purposes. In the end, practical realities such as our limited timeframe and availability and willingness to participate also played a role. Given the wide geographical span of the participants both of the interviewers and interviewees, the interviews were conducted in the online virtual meeting format.
- *Drawing upon GPAI Expert knowledge* which is incorporated in the drafting and review of this report and includes both hands-on knowledge and academic perspectives.

The interviewees' knowledge complements the expertise of the Committee participants in various fields of law relevant for data sharing—including contract law, intellectual property, data protection/privacy law, and access-to-data regulatory frameworks—in various jurisdictions.

It is acknowledged that the interviews were not designed as a quantitative empirical assessment of the need for standardization of data-sharing contracts as such. Rather, they were intended to help identify pertinent issues and understand their practical relevance. Findings gained through the interviews are treated as insights that can indicate tendencies but by no means should be generalized.

1.4 Concepts and Terms

1.4.1 Standardization

At the center of the study is the concept of *standardization* of contractual terms. Standardization can be viewed as a scale allowing for variations in degree, ranging from the standardization of the common definition of terminology (e.g. "data", "model", "data uses"), to contractual model provisions (or "standard terms"[6]), to fully-fledged standardized agreements.[7]

1.4.2 Data Sharing

The term "data sharing" refers to a broad and dynamic set of arrangements—there is hardly an all-encompassing taxonomy of data-sharing modalities. Models for data sharing are still emerging, including

[4] The literature and license agreement review are not comprehensive given that this project commenced in spring 2022.

[5] Appendix A contains the Questionnaire on Data and AI Model Licensing used in the interviews.

some led by start-ups and small and medium size enterprises (SMEs). Different dichotomies within data-sharing practices can be observed: standard vs. individually negotiated or bespoke terms, direct vs. intermediary, with or without compensation,[8] individual datasets vs. aggregated or pooled data, and open data vs. controlled modes.[9] Furthermore, data-sharing contracts can be classified by the type of activity performed in relation to data,[10] or by parties to the contract, such as B2G, B2B, C2G, G2B, B2C and C2B.[11]

The research report was not intended to be limited to a particular type of data sharing—the issue of standardization can be relevant to all data-related contractual arrangements. However, the practices examined so far tend to fall primarily within the B2B category.

The terms “data-sharing agreements”, “data license”[12] and “data-licensing agreements” are used in the report synonymously. The same goes for the terms “standardized contract terms”/“model contract terms” and “standardized agreements”/“standard form agreements”.

2. Data and AI Model Licensing

2.1 The Need to Facilitate Voluntary Data Sharing

A growing consensus has emerged that society needs to create tools to facilitate the voluntary sharing of data in a responsible and trusted manner. The OECD emphasized this in its 2021 *Recommendation on Enhancing Access to and Sharing of Data*, which sets forth principles and policy guidance for maximizing the benefits of responsible data access and sharing.[13] Several GPAI and OECD Members have embraced data sharing goals as reflected in their government policies and other initiatives. In addition, other multilateral organizations have called for greater voluntary data sharing, including (i) the United Nations, such as through its efforts to enable data sharing to help achieve the Sustainable Development Goals,[14] (ii) the World Trade Organization, such as in the context of encouraging trade agreements that foster a global data transmission ecosystem,[15] and (iii) the World Health Organization, such as in the context of the sharing and reuse of health-related data.[16] There also are growing efforts in industry and academia to expand data sharing.

[8] The latter are also known as data philanthropy or data donation.

[9] While both “open data” and “controlled” model can be conditional, the distinguishing factor is whether the data holder retains the discretion to refuse a third-party access to and use of data.

[10] For an overview of data-related contractual arrangements and data markets, see Organization for Economic Co-Operation and Development, *Enhancing Access to and Sharing of Data. Reconciling Risks and Benefits for Data Re-use Across Societies* (OECD 2019) 39 ff. See also United Nations Commission on International Trade Law (UNCITRAL), ‘Revised draft legal taxonomy – revised section on data transactions’ (24 May 2021) A/CN.9/1064/Add.2, 2 <https://uncitral.un.org/sites/uncitral.un.org/files/1064_add_2_advance_copy_e.pdf> (distinguishing between two broad categories of data transactions: data provision and data processing).

[11] In these abbreviations, “B” stands for “business”, “C” for “consumer” (or sometimes “customer”), and “G” for government.

[12] The term “data license” is commonly used to refer to data agreements, irrespective of whether exclusive rights, including IP rights, are applicable to data. Where used in this report, the term “data license” does not imply that the data at issue is protected by exclusive rights.

[13] https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0463

[14] https://www.unglobalpulse.org/policy/global-data-access-framework/

[15] https://www.wto.org/english/res_e/booksp_e/tradtechpolicyharddigit0422_e.pdf

[16] https://www.who.int/publications/i/item/9789240044968

Several GPAI Working Groups have launched efforts to help facilitate data sharing. These efforts include, for example, (i) the Data Governance Working Group's project to enable data sharing for social benefit through data trusts[17] and data institutions, and its work on privacy enhancing technologies ("PETs"), (ii) the Responsible AI Working Group's *Climate Change and AI* report,[18] which discusses how data sharing potentially can improve a wide range of existing AI-based responses to climate issues, from urban heat mapping to tracking emissions,[19] and (iii) the AI and Pandemic Response Working Group, which recently released a progress report highlighting how maximizing data sharing can enhance AI-powered responses to COVID-19 and future pandemics.[20]

The Intellectual Property Committee of I&C Working Group seeks to complement these efforts by exploring another essential element of data sharing, namely, how to foster the creation of standardized or model data sharing agreements or contract terms. Standardized contract terms, such as Open Source and Creative Commons licenses, have empowered interested parties to efficiently and effectively share software and other copyrighted materials, respectively. However, as explained in this report, these standardized agreements are not necessarily well suited for the widespread and trusted sharing of data, given some of the specific issues and challenges presented by data. Indeed, in a March 2021 blog post, Creative Commons acknowledged that "there remains significant legal uncertainty about whether copyright applies to AI training, which means it may not always be clear whether a CC license applies".[21] Creative Commons also emphasized that its licenses are not crafted to address the ethical issues presented by data sharing, such as in the context of AI.[22] The challenge thus remains to explore whether there are model license terms that could better facilitate the sharing of data, including to advance the development of responsible AI applications.

2.2 Challenges to Voluntary Data Sharing

Despite the growing consensus around the need to facilitate voluntary data sharing, several challenges persist. While a detailed literature review on this subject would go beyond the report's scope, this section summarizes some key challenges drawing on certain available studies and insights from the interviews and GPAI Experts. Understanding these challenges is important when considering how to draft data licensing agreements as it identifies some of the key topics that need to be addressed and can help inform approaches for contractually addressing such topics.

By way of background, as a general principle, the parties' willingness to complete a business transaction depends on whether the parties can agree upon terms which they both believe will have expected benefits that are likely to outweigh potential costs and risks. In the case of data transactions, the cost-benefit analysis can appear highly uncertain, which in turn can pose barriers to forming data sharing agreements. These uncertainties center on the following four main considerations and challenges: (i) assessing what is economically viable (below at 2.2.1), (ii) determining what is legally permissible in order to comply with

[17] https://gpai.ai/projects/data-governance/data-trusts-in-climate-interim-report.pdf
[18] Produced in collaboration with Climate Change AI and the Centre for AI and Climate.
[19] https://www.gpai.ai/projects/climate-change-and-ai.pdf
[20] https://gpai.ai/projects/ai-and-pandemic-response/gpai-ai-pandemic-response-wg-report-november-2021.pdf
[21] https://creativecommons.org/2021/03/04/should-cc-licensed-content-be-used-to-train-ai-it-depends/
[22] Ibid.

applicable laws and reduce liability risks (below at 2.2.2), (iii) developing data sharing technical solutions that can be implemented efficiently and effectively to facilitate data sharing agreements (below at 2.2.3), (iv) addressing data justice considerations (below 2.2.4), and (v) developing model contract terms that address the foregoing challenges and other topics in a way that streamlines negotiations (below at 2.2.5).

2.2.1 Uncertainty regarding economic costs and benefits

The value of data as an innovation input often is realized through data processing and analysis, including data mining, business intelligence and machine learning (ML).[23] Since data processing and analysis can take some time, neither the data holder nor the data user may have adequate information at the point of negotiations about the prospective outcomes and the value that data might have in terms of commercially relevant knowledge. This in turn can lead to impasses in data contract negotiations.

On the data holder's side, data-sharing might be associated with a loss of control over data, which in turn might be perceived as the loss of competitive advantage and the inability to earn returns on investment—especially in the creation and curation of high-quality data—due to positive spillovers.[24] Hence, the data holder might have low motivation towards data sharing, or be inclined to charge a high price, especially for curated data. At the same time, the data recipient, who also needs to invest in resources and capacity to carry out the data analysis, might perceive the price charged by the data holder for making data available as unreasonably high. This inability to agree upon terms reflecting the valuation of the data sharing arrangement may stymie the negotiations.

Unequal bargaining power may further reduce the likelihood of reaching an agreement on mutually beneficial contract terms. According to a survey conducted by the European Commission in 2018-2019 among SMEs,[25] the majority of respondents who experienced difficulties in acquiring data from other companies indicated "unfair or unreasonable practices regarding access to data [such as] unreasonably high licensing fees […] for granting access to data" as "the most pressing issue".[26]

Yet it is important to put the Commission's finding into perspective: only one-third of the total number of the survey participants indicated that they had tried to acquire data from another company, out of which 39% reported that they had difficulties in doing so.[27] The following reasons were submitted by companies for not engaging in obtaining and using data from other companies: (i) not using data in their business models or producing the necessary data in-house (52%); (ii) the seeming non-availability of data on the market (21%); and (iii) a lack of the necessary expertise (8%).[28]

[23] Organization for Economic Co-Operation and Development, *Data-Driven Innovation: Big data for growth and well-being* (OECD 2015) 131, 143.

[24] ibid 191-192 (specifying that high upfront costs are incurred in datafication, data collection, data cleaning, data curation); OECD (n 10) 95 ff; Zillner, S. et al., 'A Roadmap to Drive Adoption of Data Ecosystems' in Curry, E. et al. (eds) *The Elements of Big Data Value* (Springer 2021), doi: 10.1007/978-3-030-68176-0_3.

[25] European Commission (n **Error! Bookmark not defined.**) 4-5.

[26] ibid.

[27] ibid. In total, 979 SMEs replied to the survey.

[28] ibid.

According to some commentators, the lack of necessary skills and competences "for delivering big data value within applications and solutions" is "a key challenge" for leveraging the potential value of data,[29] especially for SMEs.[30] Furthermore, the lack of transparency concerning which data is available and under which conditions has been viewed as one of the major factors constraining the developing of data markets and the uptake of data-driven business models.[31]

The Committee recognizes that the survey was conducted a few years ago, and it is possible that the respondents' experiences have changed since that time. Nevertheless, it suggests that only a small share of data appears to be utilized. Our interviewees also observed that data currently is not fully utilized. Indeed, one interviewee commented that data is by and large "sitting in silos."

2.2.2 Uncertainty arising from complex and evolving legal requirements and liability risks

a) The complexity of navigating the legal landscape

As with any business transaction, contractual data sharing requires the assessment of the applicable legal requirements. Such assessment can help the parties ensure compliance with applicable laws and reduce their liability risks. Liability risks can include, for example, breach of contract claims and third-party claims, such as in the B2B context, claims arising from a violation of data subjects' rights with respect to personal data. Liability risks also can include government enforcement actions. If this legal assessment is too difficult or does not shed light on how to share data in a way that is both commercially reasonable and in compliance with applicable laws, the parties will likely refrain from sharing data in order to protect themselves from liability and/or to protect themselves from reputational harm often arising from legal claims and enforcement actions.

In the case of data transactions, the legal landscape can be highly complex and uncertain, and this can deter parties from engaging in voluntary data-sharing transactions. Factors contributing to such complexity and uncertainty include the following.

- <u>First, multiple legal regimes are potentially applicable</u>: Due to its ubiquitous nature, data cuts across different areas of law, including data protection and privacy law, contract law, intellectual property, cybersecurity, tortious liability and the emerging access-to-data regulatory frameworks.

- <u>Second, the existence and scope of legal rights in data are often uncertain</u>. Given that many parties are typically involved in a data value chain, there is a risk that multiple upstream rights—unknown at the point of the transaction by the data user and even the data holder—might possibly be infringed or otherwise violated.

[29] Zillner et al. (n 24) (further pointing out that "[m]any European organizations lack the skills to manage or deploy data-driven solutions with global competition for talent under way"). On this issue, see also OECD (n 10) 90 ff.
[30] Martina Barbero et al., *Study on Emerging Issues of Data Ownership, Interoperability, (Re-)usability and Access to Data, and Liability* (Publications Office of the European Union 2017) 56.
[31] OECD (n 10) 96.

- Third, in the case of cross-border data transactions, jurisdictional differences make the assessment of legal risks even more complicated and dubious, due to the variations in the legal criteria of eligibility for and standards of protection applicable to data.

- Fourth, there is considerable uncertainty regarding potential liability claims, including by the downstream data users against upstream data providers. Statutory laws do not (yet) directly address those novel AI-related issues, and case-law is still missing in practically all jurisdictions.

- Fifth there is considerable uncertainty regarding the availability of appropriate enforcement mechanisms, including to provide support for applicable liability regimes (see section 2.4.8).

Not surprisingly, the complexity of the legal framework applicable to data transactions and uncertainty as to “who can do what with data on which conditions”[32] and evaluating and managing the potential liabilities are often posited as a major challenge, especially for individuals and SMEs.[33] Some materials seek to provide options for addressing the complex legal framework. Nevertheless, the complexity of the applicable legal frameworks continues to pose challenges to reaching agreements that advance the realization of the full value of data as an innovation resource in a commercially reasonable manner. Several interviewees underscored this point.

b) Uncertainty regarding the existence and scope of legal rights in data

Parties may be less inclined to enter into data sharing transactions if they cannot readily identify, define, and protect their respective rights to the data. The question of how rights in data are defined and allocated is highly jurisdiction-specific,[34] and the answer is still evolving, which can make it even more challenging to agree upon commercially reasonable contractual data sharing terms. The scope of this report allows us only to sketch the key persisting challenges, in line with the distinction between personal and non-personal data.

- **Risks related to personal data/personally identifiable information (PII)**

According to the OECD, risks related to the disclosure of confidential information, including personal data protection, are often indicated by individuals and organizations as the main reasons not to share data.[35] The risk-avoiding strategy can be explained, first, by substantial penalties for the breach of personal data protection/PII. Second, the applicability of data protection/PII law is often uncertain as the definition of what constitutes “personal data” hinges on whether a natural person might be identified,[36] which is

[32] Commission Staff Working Document. Impact Assessment Report accompanying the Proposal for a Regulation of the European Parliament and of the Council on harmonized rules on fair access to and use of data (Data Act) SWD(2022) 34 final (23 February 2022) 15.

[33] OECD (n 10) 17.

[34] See e.g. ALI-ELI Principles (n **Error! Bookmark not defined.**) Principle 29(2) (pointing out that the extent to which t hird-party rights might “limit data activities, as well as the effect of such limitations, is determined by the applicable law”).

[35] OECD (n 10) 17 (further also pointing out that individuals are “increasingly wary of the re-use of their personal data”).

[36] See e.g. the definition of “personal data” under Article 4(1) GDPR; Recital 26 GDPR states that the “principles of data protection should […] not apply to anonymous information, namely information which does not relate to an

susceptible to state-of-the-art data anonymization and re-identification techniques. On the one hand, Privacy Enhancing Technologies (PETs) may provide advanced technical solutions for contractual assurances and facilitate transactions involving anonymized data. On the other hand, re-identification techniques continue to develop as well,[37] limiting the effect of such assurances and exposing the parties to liability and other risks. In view of such uncertainty and potentially severe penalties, it is not surprising that the anticipated risks related to data protection are in many instances likely to outweigh possible gains from a data transaction.[38]

As much as streamlined legal certainty regarding personal data might be desirable, the global harmonization of protection standards does not appear feasible, at least for the time being. In the US alone, data protection laws have been enacted in some states, with differing terms of the standard of protection. While the US has some sector-specific federal privacy laws, it has yet to enact broadly applicable federal privacy legislation. The U.S. Federal Trade Commission, however, has issued an Advanced Notice of Proposed Rulemaking seeking comments about the possible adoption of federal privacy regulations. Additionally, the White House Office of Science and Technology Policy released a Blueprint for an AI Bill of Rights that addresses privacy and other topics.[39] In sum, the US approach to data privacy remains rather piecemeal and evolving.

Under the EU approach, personal data transfers to non-EU countries can be allowed only if such countries ensure protection equivalent to the fundamental rights and freedoms of data subjects in the EU.[40] To enable such transfers, the GDPR provides for the instruments of the adequacy decisions and standard contractual clauses (SCCs).[41] The incorporation of SCCs into a contract between an exporting controller and an importing controller or processor suggests that the transfer of data to a non-EU country is deemed to be in accordance with EU data protection law. Such instruments might be viewed as an indirect harmonization of the personal data protection standards. Yet, the recent CJEU decision in *Schrems II*[42] that suspended the EU-US Privacy Shield[43] highlights that such mechanisms are not full proof. . It demonstrates that, even where certain instruments of legal "interoperability" have been established, their workability and solidity cannot be taken for granted. At present, the full impact of the CJEU decision in *Schrems II* is somewhat unknown...".[44] In response to *Schrems II,* the US and the EU jointly announced

identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable".

[37] Alexander Bernier, Hanshi Liu and Bartha Maria Knoppers, 'Computational Tools for Genomic Data De-identification: Facilitating data protection law compliance' (2021) Nat Commun. 29;12(1):6949. doi: 10.1038/s41467-021-27219-2; Luc Rocher, Julien M Hendrickx and Yves-Alexandre de Montjoye, 'Estimating the Success of Re-identifications in Incomplete Datasets Using Generative Models' (2019) Nat Commun. 23;10(1):3069. doi: 10.1038/s41467-019-10933-3.

[38] OECD (n 10) 17 (with further references).

[39] Blueprint for an AI Bill of Rights | The White House

[40] Recital 104 GDPR.

[41] Standard contractual clauses may be laid down by the European Commission or adopted by a supervisory authority in accordance with the conditions set out under the GDPR.

[42] Case C-311/18, *Data Protection Commissioner v Facebook Ireland and Schrems* ('*Schrems II*') ECLI:EU:C:2020:559.

[43] The EU-US Privacy Shield refers to an agreement between the EU and the US that used to allow for the transfer of personal data between the two countries.

[44] *World Development Report 2021. Data for Better Lives* (World Bank Group 2021) 251.

plans in March 2022 to establish a new Trans-Atlantic Privacy Framework to address these uncertainties.[45] In October 2022, President Biden issued an Executive Order to implement the EU-US Privacy Shield.[46]

- **Risks related to non-personal data**

The legal status of non-personal data may vary substantially among jurisdictions,[47] also raising barriers to reaching data sharing agreements. Even though literature and commentators often use the term "proprietary data" and "data ownership" in relation to non-personal data, it is often unclear whether there is any legal basis for exclusive rights in such data, and whether such terms will denote any more than factual control over data, if no trade secrets or other intellectual property protection applies. For data transactions, the existence of exclusive rights in data would dictate a "license" approach, as opposed to a "sales" approach: the former implies that anything that is not explicitly allowed is prohibited; the latter, in contrast, means that what is not explicitly forbidden is allowed.[48] The European Commission explicitly rejected the concept of "data ownership" in device-generated data and thus has taken a distinct turn towards the access-based framework for data.[49]

In some cases, the datasets at issue may qualify for IP protection.[50] According to some interviewees, where EU copyright law applies, parties often rely on the text and data mining exception[51] for carrying out data processing activities. Interviewees also shared that when US copyright law applies, parties often seek to rely on the fair use exception to carry out the data processing activities, even though the scope of fair use exceptions remains somewhat uncertain. The interviewees also identified challenges presented by the differing approach in the US and the EU to copyright exceptions.

It was also mentioned that uncertainty persists as to whether and to what extent database *sui generis*[52] protection in the EU might apply to aggregated datasets. Such legal uncertainty is acknowledged by the

[45] FACT SHEET: United States and European Commission Announce Trans-Atlantic Data Privacy Framework - The White House, https://www.whitehouse.gov/briefing-room/statements-releases/2022/03/25/fact-sheet-united-states-and-european-commission-announce-trans-atlantic-data-privacy-framework/.

[46] FACT SHEET: President Biden Signs Executive Order to Implement the European Union-U.S. Data Privacy Framework | The White House

[47] ALI-ELI Principles 194 (noting that the Principles "take no position as to whether data constitutes "property" that can be 'owned'"). On uncertainties regarding "data ownership", see OECD (n 10) 98 ff.

[48] On this issue, see also ALI-ELI Principles 10 (recommending the policy choice based on the "sales approach" under which the default position is that the supplied or shared data "may be used by the recipient for any lawful purpose that does not infringe the rights of third parties").

[49] See the Commission's Proposal for a Data Act Recital 6.

[50] OECD (n 10) 99.

[51] 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, [2019]OJ L130/92. While this report does not aim to assess the effectiveness of the text-and-data-mining exception/limitation in the context of AI, it should be noted that the availability of such exception/limitation – for *other than scientific purposes* – can be excluded upon the express reservation by the right holders (Article 4(3) of the Digital Single Market Directive). See also J Drexl, RM Hilty et al., 'Artificial Intelligence and Intellectual Property Law. Position Statement of the Max Planck Institute for Innovation and Competition of 9 April 2021 on the Current Debate', Max Planck Institute for Innovation & Competition Research Paper No. 21-10, p. 8, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3822924 (concluding that "the current system of [copyright] exceptions and limitations alone cannot solve the unbalance problem in the AI context").

[52] Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases, [1996] OJ L 77/20.

proposal for the EU Data Act[53] that limits the availability of database *sui generis* protection, albeit only in the situations where such protection would prevent the exercise of the IoT data access and use right under Chapter II of the Data Act. Such limitation (posed as a clarification) appears to be largely irrelevant for situations where data is shared for the purposes of developing AI systems.[54] In the US, there is no *sui generis* database protection. This is another area where the lack of cross-border harmonization may be contributing to the data sharing challenges.

As for trade secrets protection, the main challenge subsists in sharing data in a manner that does not jeopardize such protection. For instance, trade secret protection may be compromised if the data recipient does not take sufficient measures to protect the secrecy of the data. However, current law may not necessarily provide clear guidance on the types of measures that are needed to satisfy this standard. Furthermore, there might be jurisdictional specificities as to the scope of protection: for instance, the EU Trade Secrets Directive introduced extra-contractual liability for unauthorized access, acquisition and use of trade secrets. These legal uncertainties can deter parties from engaging in data sharing transactions involving confidential information.

c) Data quality and liability issues

The current lack of commonly understood terms and measures to define and assess data quality was highlighted in several interviews as another inhibitor of data sharing arrangements. In some transactions, the provision of data with the agreed technical characteristics and quality may fall within the contractual obligations of the data provider. This may be desirable to help foster greater data interoperability, in addition to promoting better data quality. However, negotiating these contractual terms can be very challenging, particularly when there are not widely accepted data standards and measures that can be easily referenced in the agreement (see sections 2.4.4 and 2.4.5). As noted below, another option is for data providers to agree to make data available on an "as-is" basis, without any guarantees or assurances, including with respect to data quality. While this approach may reduce the need to agree upon technical data contractual terms, it also could increase regulatory compliance challenges and impose more burdens (such as data hygiene and standardization burdens) and legal risks on the data recipient (see sections 2.4.4 and 2.4.5).

Deciding how to allocate legal risks can pose further barriers to negotiating data sharing transactions. In this regard, the unique characteristics of data—especially its dynamic nature and propensity to change over time—increases uncertainty regarding how concepts of liability might apply to data and data-based applications.[55]

Parties engaging in data transactions also face risks of potential third-party claims or extra-contractual liability. For example, a third party (such as a data subject) might bring a claim alleging that neither party to the data sharing contract has sufficient rights to use the data in accordance with such contract. The

[53] Below.

[54] For an analysis, see Drexl et al. para 333 ff.

[55] Barbero et al. (n 30) 17, 47; Timan, T., van Oirsouw, C., Hoekstra, M., 'The Role of Data Regulation in Shaping AI: An overview of challenges and recommendations for SMEs' in: Curry, E. et al. (eds), *The Elements of Big Data Value* (Springer 2021), https://link.springer.com/chapter/10.1007/978-3-030-68176-0_15.

parties to the data sharing agreement may want to allocate the risk and liability associated with such third-party claims contractually. For instance, if the data provider contractually agrees to certain data quality terms (such as obtaining all necessary rights and consents), such party may assume relatively more risks for these types of claims. In contrast, if the data provider makes the data available on an "as-is" basis without any quality commitments, it may have minimal (or no) contractual obligations with respect to such claims.

The risk of third-party claims may increase as the data supply chain becomes more complex. For instance, a data provider may aggregate data from multiple upstream sources and enter into an agreement to license the aggregated data to a data recipient. This raises the question of whether and to what extent the data provider and the data recipient could be responsible for third-party (e.g., extra-contractual) claims brought by the upstream data providers. It also poses the question of whether and to what extent the data provider should bear contractual responsibility to the data recipient for claims stemming from the quality of an upstream providers' data. In sum, navigating these issues can be challenging, which in turn, may impede efforts to reach agreement on data sharing terms.

Most (if not all) interviewees identified allocating liability as an impediment to forming data sharing agreements. It is worth noting, however, that the European Commission studied this topic in a 2017 online consultation. According to the 2017 online consultation by the European Commission, fear of liability and data security were indicated as the reasons for not using data from other companies by 5% and 4% of the respondents, respectively.[56] Another study reports that uncertainty regarding 'liability costs in case of damage caused by the data shared was reported by 15% of the surveyed companies as an obstacle to data sharing'.[57]

It is unknown whether the survey respondents were fully aware of the liability risks. In any case, one would assume that the appreciation of liability risks might be changing with the increasing regulation of the digital economy, including in the field of AI, especially given that high-quality data is key for developing reliable AI systems. Several interviewees pointed out uncertainty about the potential impact of emerging AI regulations[58] on data transactions. Yet some studies suggest that the negative effect of safety regulations and liability regimes on innovation cannot be presumed and the relationship between liability and innovation is more complex than 'the view that 'liability chills innovation'.[59]

Notably, safety requirements – including the requirements for data governance under Article 10 of the draft EU AI Act – are treated under the draft EU AI Liability Directive as "due care standards". The proof of noncompliance with them triggers the application of the presumption of 'the causal link between the fault of the defendant and the output produced by the AI system or the failure of the AI system to produce an output'. [60] This underscores that , further research on data-sharing practices should take into account

[56] European Commission 5.

[57] Catarina Arnaut et al., *Study on Data Sharing Between Companies in Europe* (Publications Office of the European Union 2018) 78, https://op.europa.eu/en/publication-detail/-/publication/8b8776ff-4834-11e8-be1d-01aa75ed71a1/language-en.

[58] Such as the AI Act in the EU and the Algorithmic Accountability Act in the US.

[59] Andrea Bertolini, *Artificial Intelligence and Civil Liability* (European Parliament 2020) https://data.europa.eu/doi/10.2861/220466, 35 ff (with further references).

[60] . (Article 4 of the draft EU AI Liability Directive)

the emergence of AI-specific liability frameworks[61] and their interaction with AI-specific safety regulations.[62]

To summarize, data-sharing contracts should be drafted with the awareness of the applicable legal framework, the existence of third-party legal rights with *erga omnes* effect that might be affected by the contract execution and how such rights are protected under the existing laws (a failure to conduct due diligence regarding the existence of third-party rights and their scope could expose the contracting parties to the risk of legal disputes).

2.2.3 Barriers relating to technical usability

Many interviewees identified the lack of widely embraced technical protocols relating to data sharing as hindering efforts to negotiate data sharing agreements. The technical feasibility of implementing access to data, carrying out data transmission, data portability and (re-)use hinges upon the availability of technical infrastructure, commonly accepted and accessible data formats, protocols for data collection and processing, and application programming interfaces (APIs).[63] Their presence allows *inter alia* for flexible, coherent and scalable processing of datasets from different sources and makes the implementation of data sharing arrangements more practical and efficient.

The importance of such technical enabling factors for promoting data sharing was highlighted in almost every interview. While the need to develop tools that can facilitate the usability of data, such as industry-led standards for data sharing, is widely acknowledged, it remains to be seen how the pathways to more concrete solutions will emerge. Potentially, data intermediaries can play a role in the wide adoption of technical standards, given that they usually provide data in certain data formats that enable syntactic and semantic interoperability.[64]

2.2.4 Addressing Data Justice

As highlighted by the work of the GPAI Data Governance Working Group, data justice needs to be considered in the context of data sharing arrangements.**[65]** While data justice is outside the scope of this report, the Committee recognizes the importance of addressing these important concerns.

[61] See e.g. Proposal of the European Commission of 28 September 2022 for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive), COM(2022) 496 final and Proposal of the European Commission of 28 September 2022 for a Directive of the European Parliament and of the Council on liability for defective products, COM(2022) 495 final (adapting the rules on strict liability for defective products to the digital age and AI).

[62] Notably, safety requirements – including the requirements for data governance under Article 10 of the draft AI Act – are treated under the draft AI Liability Directive as "due care standards". The proof of noncompliance with them triggers the application of the presumption of 'the causal link between the fault of the defendant and the output produced by the AI system or the failure of the AI system to produce an output' (Article 4 of the draft AI Liability Directive).

[63] OECD (n 10) 91 ff; Zillner et al. (n 24); COM/2014/0442 final para 3.1 (pointing out that the availability of "good quality, reliable and interoperable datasets and enabling infrastructure' is a characteristic of a thriving data-driven economy").

[64] OECD (n 10) 94.

[65] Advancing research and practice on data justice - GPAI

2.2.5 *Uncertainty about contracting practices*

Overall, while there is a consensus that guidance regarding model contracts or provisions could support and facilitate contractual data-sharing practices,[66] it should also be acknowledged that the standardization of contractual terms might mitigate only some of the challenges outlined in this section. Section 2.4 outlines some types of model contractual terms that could help streamline the negotiation of data-sharing agreements. As also noted by many interviewees, the increasing national regulation of the data economy and the emphasis on "data sovereignty" make the task of designing model contracts, standard terms and definitions that are internationally usable challenging—"quick-fix" or "one-size-fits all" solutions can hardly be envisaged. Against this background, it needs to be further explored to what extent and in what way standardization might be feasible and mutually beneficial from a balance-of-interests perspective.

2.3 Preliminary Project Findings on the Pathways to Develop Standardized Terms

The following summarizes some potential pathways that may help create standardized or model agreements or contract terms that can be used to facilitate data sharing. More specifically, Section2.3.1 highlights some ongoing efforts to develop such terms, and Section2.3.2 describes how creating a variety of terms (as opposed to a "one-size-fits-all" approach) could help advance a broad range of different types of data sharing arrangements.

2.3.1 *Greater Coordination and Collaboration Are Needed to Advance Relatively Nascent Efforts*

While several organizations are working on data licensing templates, these efforts remain relatively nascent.[67] Indeed, to our knowledge, no data licensing forms have achieved widespread acceptance or use comparable to Open Source or Creative Commons licenses. The following are some examples of data licensing forms or similar guidance that have been published, and this list is not intended to be exhaustive.

- **The Linux Foundation** has developed data licensing templates through a stakeholder-based process.[68] These include the Community Data Licensing Agreement Permissive 2.0 (CDLA-Permissive 2.0),[69] which permits licensees to broadly use, analyze, modify, and share data, and the Computational Use of Data Agreement (C-UDA 1.0),[70] which permits licensors to share data for computational use purposes such as AI/ML and text and data mining (TDM).[71] Notably, the Linux Foundation designed the CDLA 2.0 to be more streamlined and brief, based on community feedback that version 1.0 was too complex, particularly for non-lawyers to effectively use.[72]

[66] Barbero et al. (n 30) 153.

[67] This list may not be exhaustive, as many templates exist for various jurisdictions and materials (e.g. data, software, and content). For example, the Open Knowledge Foundation maintains a list of >100 open licenses.

[68] https://www.linuxfoundation.org/press-release/enabling-easier-collaboration-on-open-data-for-ai-and-ml-with-cdla-permissive-2-0/

[69] https://cdla.dev/permissive-2-0/

[70] https://cdla.dev/computational-use-of-data-agreement-v1-0/

[71] https://github.com/microsoft/Computational-Use-of-Data-Agreement#:~:text=The%20C%2DUDA%20is%20a,and%20text%20and%20data%20mining.

[72] https://www.linuxfoundation.org/press-release/enabling-easier-collaboration-on-open-data-for-ai-and-ml-with-cdla-permissive-2-0/

- **Microsoft** has also developed several licensing templates. Its Data Use Agreement for Open AI Model Development (DUA-OAI) is designed for parties sharing data for the limited purpose of training an AI model, especially in situations where that data cannot be made public due to privacy and/or business concerns.[73] Under the Microsoft DUA-OAI license, the model trained on the shared data is required to be licensed on an open-source basis. Microsoft's Data Use Agreement for Data Commons (DUA-DC) is designed to support multiple parties sharing data sets in a common, API-enabled database.[74]
- **Responsible AI Licenses** **(RAIL)** has developed license templates to address ethical risks of AI by posing restrictions on the sharing and distribution of AI software (including clauses regarding potentially harmful applications in surveillance, computer-generated media, health care, and criminal justice).[75]
- **Open Data Commons** hosts several licenses for data and databases, including an attribution license, an attribution share-alike license, and a public domain dedication.[76] Its host organization, the Open Knowledge Foundation, maintains the Open Definition, a set of principles intended to define "openness" in relation to data and content.[77]
- **Creative Commons** licenses often are used to license data, although they are not tailored for this purpose .[78]
- **METI and Japan Patent Office ("JPO") efforts** METI and the JPO have published some template agreements and guidance for data and AI contracts, that offer various options for approaching these arrangements. https://www.meti.go.jp/press/2019/04/20190404001/20190404001-2.pdf and https://www.jpo.go.jp/support/general/open-innovation-portal/index.html (which also includes AI SaaS terms).][79]
- **UK License**. The United Kingdom has published the Open Government License (nationalarchives.gov.uk) for the sharing of public data.
- **Singapore Information Media Development** Authority has published the Trusted Data Sharing Framework (Link 2 – Framework PDF)

There was general agreement among interviewees on the need for a coordinated and inclusive process to develop standardized or model data sharing forms or contract terms. As noted above, one of the challenges is to develop forms and terms that stakeholders will embrace and want to use. By including a broad range of stakeholders in the process of developing standardized or model agreements or terms, there is a greater likelihood that the resulting product will be widely embraced and used. The Committee knows that organizations working on standardized or model forms to date have strived to be inclusive in their efforts and the Committee encourages these efforts. Creative Commons also underscored the importance of this goal in its March 2021 blog post, which states as follows:

> … to promote the use of CC-licensed content to train AI, we need a community-led, coordinated and inclusive approach to consider not only the copyright system in which CC licenses operate, but also issues of accountability, responsibility, sustainability, cultural rights, human rights, personality rights, privacy rights, data protection, and ethics. As one actor in a vibrant community of open

[73] https://news.microsoft.com/wp-content/uploads/prod/sites/560/2019/07/DUA-OAI-README.pdf
[74] https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE4Rjfv
[75] https://www.licenses.ai/ai-licenses
[76] https://opendatacommons.org/licenses/
[77] https://opendefinition.org/od/2.1/en/
[78] https://creativecommons.org/2021/03/04/should-cc-licensed-content-be-used-to-train-ai-it-depends/
[79] https://www.meti.go.jp/english/press/2019/0404_001.html

> advocates defending the interests of the millions of people who use CC licenses, we want to engage in rich conversations on AI's multiple facets to promote better sharing in the public interest.[80]

2.3.2 Need for a Variety of Standardized or Model Agreements or Contractual Terms

Our research and interviews have highlighted that there is broad range of potential data licensing use cases, and that the desired data licensing terms may vary based on the use case as well as on parties' underlying objectives and the nature of the data sharing arrangement. For example, if the parties are exchanging confidential information, the agreements should include confidentiality and cybersecurity clauses. If this information includes personally identifiable information, then privacy regulations need to be addressed as well. And if the parties are using federated learning or another form of PETs, the contract needs to address these applications.

For all these reasons, the Committee foresees a need to work toward developing a range of alternative standardized or model data licensing terms that parties can select from based on their needs for a particular data licensing arrangement. This approach is analogous to what has evolved with Open Source and Creative Commons licenses, where a variety of forms exist, and parties can select the one that is appropriate for their specific use case.[81, 82] At least one interviewee expressed the view that there may be too many Open Source and Creative Commons licenses, and that there should be more consolidation for data licensing. The question of the appropriate number of standardized data licensing terms merits further inquiry and evaluation.

Several interviewees also stated that the need for bespoke data license agreements will continue, even if standardized or model data licensing agreements or contract terms are developed. Again, this is analogous to the software and copyright licensing context where standardized forms exist, and in some cases, parties still opt for bespoke licensing arrangements. In other words, both types of agreements serve important needs. Many interviewees expressed the view that standardized, or model data licensing terms could serve as a useful starting point for negotiating bespoke data licensing agreements.

Against this backdrop, the Committee continues to evaluate whether the goal should be to develop (i) a series of standardized data licensing form agreements, (ii) a set of model data licensing terms that can be used in bespoke or more complex agreements (such as data commons and/or for data intermediaries), or (iii) both such standardized forms and model terms. The Committee also continues to evaluate which types of standardized terms are most feasible, particularly given the uncertainties outlined in Section 2.2. All interviewees agreed that data licensing and sharing would be much easier if there were some standardized approaches that the community can draw from.

2.4 Description of Data and Related Licensing Terms

This section identifies some of the terms that possibly could be addressed in model data licensing provisions or in a form agreement, and it is not necessarily intended to be exhaustive. The Committee offers this information to help advance ongoing efforts to develop standardized data licensing terms

[80] https://creativecommons.org/2021/03/04/should-cc-licensed-content-be-used-to-train-ai-it-depends/
[81] https://opensource.org/licenses
[82] https://creativecommons.org/about/cclicenses/

and/or perhaps to inspire new efforts. also hope that this information will help educate the community about data licensing, which can also help facilitate data sharing transactions.

2.4.1 Standardizing the Nomenclature and Allocation of Proprietary Rights

Right now, there are not widely accepted definitions of the key terms and rights that need to be addressed in data licensing agreements. One objective should be to standardize some terms and develop a menu of model provisions that can be used to allocate rights depending upon the goals of the data sharing arrangement. The following describes some examples of possible topics or concepts that potentially could be addressed with standardized terms or definitions (but for clarity, the text below is not intended to serve as model terms or definitions, since drafting model terms or definitions is beyond the scope of this report).

Original Input Data. It could be helpful to have a definition associated with the data in the form provided by the data provider to the data recipient pursuant to the contract. This definition could provide a useful framework for addressing each parties' respective rights to the data, as addressed more fully in Section 2.4.2.

Processed Data or Results. It could be helpful to have one or more definitions associated with data developed by the data recipient using the original input data, as described above. Depending upon the circumstances, there could be one broad definition or various definitions to differentiate among the following: (i) the cleansed data created by the data recipient that has undergone data hygiene and is based on the original input data, (ii) any data compilation, database, insights, or metadata developed by or on behalf of the data recipient (in whole or in part) using or otherwise based on the original input data, and (iv) outputs of any AI model trained on any of the foregoing or the original input data.

Untrained Model. It could be helpful to have a definition associated with an AI model that has not been trained on data that is the subject of the data sharing arrangement. This definition could provide a useful framework for defining each parties' rights to such model.

Trained Model. It could be helpful to have a definition associated with an AI model that has been trained on data that is the subject of the data sharing arrangement, including Original Input Data and/or Processed Data or Results.

In addition to developing a potential menu of standard definitions, it might be desirable to develop a menu of terms for allocating rights to each of the items discussed above. The interviewees generally agreed that a menu of options would be helpful, as the way rights are allocated may vary among transactions based on the underlying objectives. They also agreed that developing standardized provisions could advance data licensing, including by providing important alternatives to relying on current frameworks allocating rights based on "derivative works". There was general consensus that allocating rights to "derivative works" in the data context may not be applicable or the best approach, particularly since 'derivative works' is a copyright concept, and the data may not be copyrightable.

2.4.2 Usage and Access Rights

As with most license agreements, standardized and model data license agreements will likely need to enable parties to specify the scope of both permitted and prohibited uses of data and/or AI models.

Depending upon the context, this principle potentially applies not only to the licensed data (e.g. .the original input data, as described above), but also to (i) the cleansed data that has undergone data hygiene, (ii) any data compilation, database, insights, or metadata developed (in whole or in part) using such data, (iii) untrained models, (iv) any trained models trained on the data, and (v) outputs of any trained models(and (i), (ii) and (v) could be forms of Processed Data or Results, as described above).

Several of the license agreements described in Section 2.3.1 include usage restrictions. For instance, both Microsoft's DUA-OAI agreement[83] and Linux's C-UDA license[84] contain these types of terms. The Microsoft DUA-OAI agreement permits sharing of data only to train AI models. The C-UDA limits data use strictly to computational activities (e.g., machine learning or TDM).

In addition to addressing usage rights, the parties also might want to include terms for specifying how to access the data, such as through a designated API, and reserve the right to revoke access if certain events occur. It merits further examination how model terms addressing usage and access rights can be crafted that provide some degree of standardization but also give parties the flexibility to tailor them for particular use cases.

As highlighted above, some parties also may also wish to limit access to and use of data and AI models due to ethical concerns. There are efforts to address this challenge in licensing agreements, such as the Responsible AI Licenses (RAIL) that include model terms for limiting use of AI models (such as prohibitions on the use of the shared product to predict the likelihood that a person will commit a crime, diagnose a medical condition without human oversight, or impersonate a person or entity, among other potentially harmful uses).[85, 86] As also noted above, Creative Commons has identified ethical considerations as an important factor in data licensing as well. How to approach this in the context of standardized or model agreements or contract terms merits further consideration.

Finally, "data sovereignty" is the term often used in the context of national and international data flows, yet there is hardly unanimity regarding its meaning and the delineation from related concepts such as "network sovereignty" and "internet sovereignty".[87] Data sovereignty may refer to "the self-determination of individuals and organizations with regard to the use of their data"[88] and, in this sense, be contrasted with the concept of data privacy.[89] Alternatively, data sovereignty might refer to "the right of a nation to collect and manage" data[90] or to the national legislation regarding the geolocation of data.[91] There are also more context-specific understandings of data sovereignty, such as "indigenous data

[83] https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE4RlP7

[84] https://cdla.dev/computational-use-of-data-agreement-v1-0/

[85] https://dl.acm.org/doi/abs/10.1145/3531146.3533143

[86] https://www.licenses.ai/enduser-license

[87] AK Woods, 'Litigating Data Sovereignty' (2018) 128(2) The Yale Law Journal 328, 360; P Hummel, M Braun and M Tretter, 'Data Sovereignty: A review' (2021) Big Data & Society, https://doi.org/10.1177/2053951720982012.

[88] M Jarke, B Otto and S Ram, 'Data Sovereignty and Data Space Ecosystems' (2019) 61 Bus Inf Syst Eng 549, 550.

[89] ibid (2019) 61 Bus Inf Syst Eng 549, 550 (explaining that personal data protection laws such as GDPR 'sees the citizen in a rather passive role to be protected against powers they cannot confront on an equal footing'; in contrast, "data sovereignty aims at enabling 'data richness' by clearly negotiated and strictly monitored data usage agreements").

[90] Hummel, Braun and Tretter (n 87) (with further references).

[91] ibid (with further references).

sovereignty".[92] Given a wide variety of interpretations, whenever parties refer to "data sovereignty" in the context of a legal framework or contractual obligations, they should specify in what sense the term is used. These considerations should be kept in mind in the context of preparing model data sharing terms.

2.4.3 Privacy and Confidentiality

Parties holding data often face tensions between the desire to share data, on the one hand, and the need to protect the privacy or confidentiality of that data on the other. As noted in Section 2.3.2 preserving confidentiality can be challenging with data sharing. To help resolve the tension and address this challenge, many data sharing arrangements require privacy, confidentiality and/or cybersecurity provisions, including (i) when sharing proprietary data that is intended to be subject to trade secret protection, and/or (ii) sharing personal data or other data that is subject to regulation. As discussed above, the sharing of personal data is complicated further in the cross-border context due to the current need for more regulatory harmonization.

Navigating the regulatory, confidentiality and cybersecurity concerns pertaining to data often is challenging, including in the data-licensing context. While, as discussed above, there is no easy solution, it is worth considering whether and to what extent standardized or model terms can be crafted that could help streamline negotiations, including with respect to these concerns.

Several existing license templates address privacy, cybersecurity, and confidentiality, at least to some degree. Microsoft's DUA-OAI agreement[93] and DUA-DC agreement[94] both have relevant provisions, including: (i) terms prohibiting users from attempting to identify individuals through de-identified or anonymized data; (ii) attachments where parties can specify applicable privacy laws or frameworks, such as GDPR or HIPAA; and/or (iii) attachments where parties can negotiate data security requirements. Linux's CDLA-Sharing-1.0 license[95] also includes confidentiality and privacy terms, though Linux has since developed an updated CDLA license which omits these terms for the sake of simplicity.[96]

To protect confidentiality and address regulatory concerns, parties increasingly are turning to privacy enhancing technologies (PETs). *PETs* encompasses a range of technologies such as differential privacy (a mathematical definition for privacy in statistical and machine learning analysis),[97] federated learning (a method for training AI models in a decentralized manner that does not require the transfer of data)[98] and synthetic data (the replacement of real-world data with computer-generated data, such as synthetic electronic health records).[99] Consideration should be given as to whether standardized or model terms can be drafted to assist parties desiring to use PETs. Since PETs can differ based on their design and deployment, it may be desirable to have terms tailored for specific PETs.

[92] Defined as "the ability for Indigenous peoples to control their data", including DNA/genomics and community health data. See National Library of medicine, 'Data Glossary', https://nnlm.gov/guides/data-glossary/data-sovereignty.

[93] https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE4RlP7

[94] https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE4Rjfs

[95] https://cdla.dev/sharing-1-0/

[96] https://www.linuxfoundation.org/press-release/enabling-easier-collaboration-on-open-data-for-ai-and-ml-with-cdla-permissive-2-0/

[97] https://privacytools.seas.harvard.edu/files/privacytools/files/pedagogical-document-dp_new.pdf

[98] https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9084352

[99] https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-020-00977-1

2.4.4 Data Interoperability and Quality

The OECD has noted that efficient and effective data sharing depends in part on data interoperability, which may be supported through the development and use of common licensing arrangements .[100] Our interviewees shared this view. Consequently, consideration should be given to how data quality, interoperability, and other technical matters might be addressed in standardized or model contract terms. These terms could serve as templates for requiring (i) baseline data quality requirements, (ii) the use of specified standard techniques for data hygiene and formatting, (iii) formats for tracking data provenance and lineage (and fostering traceability) as well as data usage restrictions, and (vi) the use of common APIs for accessing data. As explained in Sections 2.2.3 and 2.4.5, the Committee recognizes that crafting these terms may not be easy, particularly given the lack of widely used and embraced standards and practices in this area. Nevertheless, the Committee believes this merits further consideration.

It also is worth noting that there appears to be no standard approach to allocating responsibility for data quality, interoperability and other similar obligations in data agreements. In some cases, the data providers may bear some or all of these responsibilities, yet in others, the data recipient might assume these roles. These various arrangements should be taken into consideration when undertaking the preparation of standardized or model data licensing terms. It also underscores the importance of having a menu of model terms so parties can choose the ones that align with the underlying objectives of their contemplated data-sharing arrangement.

2.4.5 Disclaimers, Liability and Enforcement of License Terms

As discussed above, there is a general perception that uncertainty concerning how national liability frameworks apply to data transactions is perceived as a barrier to data sharing, especially in cross-border settings.[101] While acknowledging that drafting liability clauses in data contracts is a broad and complex topic of its own, this section reviews interim findings regarding how liability issues are currently addressed in data-sharing practices.

a) Approaches adopted in practice according to certain literature

Studies on how liability issues are addressed in data-sharing contracts are rare. Data-sharing approaches were examined through a survey among companies interested in acquiring access to data held by others, already using data acquired from others, and active in both gaining data from and sharing it with others.[102] With certain statistically rather insignificant variations in the percentages within those subgroups, the survey results show that some companies examine liability assurance regarding the shared datasets on a case-by-case basis; others often accept data as provided, even with potential errors; while many companies do not consider as relevant negotiations with individual data providers about additional liability assurance.[103] The survey also found that companies interested or active in sharing data with third parties ‘contractually limit what people can do with their data and do not accept liability if they use it for

[100] https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0463.

[101] Above at 3.2.

[102] Barbero et al. (n 30).

[103] ibid 392-394.

a different purpose)' and, more generally, 'try to exclude liability as far as possible in their contracts or terms and conditions'.[104] This is in line with our findings: as mentioned by some interviewees, disclaiming or limiting warranties and/or liability in data-sharing agreements, including for data quality, is a common practice.

The questions of whether the examined sample was representative enough to identify statistically meaningful tendencies, and how the dynamics might have developed more recently, go beyond the scope of this report. However, one would assume that the apportionment of liability in individual negotiations would likely depend on the individual circumstances of a data transaction, including the parties' relative bargaining power. In this regard, it needs to be further examined whether and/or to what extent the maximum disclaimer of warranties and/or liability by the data provider is an appropriate solution, or whether standard agreements or model provisions on liability may establish different ways of allocating responsibilities and/or liability risks. It also merits further consideration whether there should be standardized terms for different approaches to allocating liability, particularly given the wide range of potential data-sharing arrangements.

b) *Approaches to liability in standard terms and guidance documents*

As far as apportioning liability is concerned, the 'easiest' approach from the data providers' perspective would be to disclaim warranties and liability altogether, as illustrated by the Montreal Data License, which states: "Unless otherwise agreed in writing by the parties, the data is licensed as is and as available. Licensor excludes all representations, warranties, obligations, and liabilities, whether express or implied, to the maximum extent permitted by law."[105]

The European Commission's Guidance on the preparation and/or negotiation of data usage agreements does not go further than suggesting to specify liability provisions under certain circumstances. In particular, it recommends including "rules on liability provisions for supply of erroneous data, disruptions in the data transmission, low quality interpretative work, if shared with datasets, or for destruction/loss or alteration of data (if it is unlawful or accidental) that may potentially cause damages".[106]

A more substantive approach can be found in the ALI-ELI Principles, which outline three possible approaches to strike 'the difficult balance between third-party protection and the protection of data recipients':

(i) a protected third party can enforce the same rights against a downstream recipient as against upstream parties in the data value chain;

(ii) the data supplier has a due diligence duty to choose the recipient who will comply with the same restrictions that the suppler has to abide by, and has to undertake safeguarding measures vis-à-vis protected parties (hence, the data supplier can be liable only for the breach of such due diligence duties and safeguarding measures); or

[104] ibid 399.

[105] Misha Benjamin et al., 'Towards Standardization of Data Licenses: The Montreal Data License' (2019) 15, https://ui.adsabs.harvard.edu/abs/2019arXiv190312262B/abstract.

[106] SWD(2018) 125 final 7.

(iii) strict vicarious liability is imposed on the data supplier for wrongful data activities that may occur downstream.[107]

Furthermore, the ALI-ELI Principles formulate default contract rules on liability that are adjusted to a type of data contract. For instance, the default rule in contracts for mere authorization to access holds that the recipient must indemnify the authorizing party—"whose role is quite passive"[108]—for any liability vis-à-vis third parties that may follow "from the authorizing party's authorization to access the data unless such liability could not reasonably be foreseen by the recipient".[109] Alternatively, where "the authorizing party were to qualify as a normal 'supplier', it would be subject both to any duties it owes vis-à-vis third parties under Principle 32 and to potential liability where these duties are breached".[110]

Furthermore, Principle 32 states that "[n]othing in this Principle precludes strict vicarious liability of a controller for data activities by a processor under the applicable law".

c) *Insights from interviews*

From the perspective of the interviewees and several GPAI Experts, allocating liability as well as responsibility for data quality often are difficult issues that can impede transactions. Some of the complexities are explained in Section 2.2. In addition, to the extent that *data quality* constitutes part of the data provider's contractual obligations, the definition of what "quality" means is of paramount importance for the interpretation of the scope of the respective obligations. As pointed out in one interview, there is hardly a standard definition of data quality. As a multifaceted concept, it can include reliability, relevance, representativeness, completeness, lack of harmful bias, being free from third-party rights such as IP rights, etc. More specifically, what quality data means can be defined only relative to the purpose for which the shared datasets are intended to be used. This peculiarity may explain why standard data quality clauses might not go into much detail. While it might be useful to specify the above-mentioned meta-characteristics of data quality (completeness, freedom from bias or IP rights etc.), in individual cases they would need to be supplemented with more concrete descriptions of datasets in a standardized format. How to address and balance these issues and tensions merits further attention in connection with data licensing.

Furthermore, concerns were expressed regarding the potential impact of *AI regulations* on contractual data-sharing practices; several interviewees referred to the upcoming EU AI Act[111] as a highly pertinent example. Motivated by safety concerns and dedicated to protection of fundamental rights,[112] the Act undertakes a precautionary approach based the differentiated risks associated with AI systems and envisages compliance obligations related to the development of "high-risk" AI systems,[113] based on strict liability rules. in view of the paramount importance of data quality for developing robust ML models and

[107] ALI-ELI Principles 208-9.

[108] ibid 86.

[109] ibid Principle 10, para 2(e).

[110] Principle 32 defines the duties of a data supplier in the case of the onward supply of data.

[111] Proposal for a Regulation of the European Parliament and of the Council laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts, COM/2021/206 final (21.4.2021) (hereinafter EU AI Act).

[112] Recital 43 EU AI ACT.

[113] That is, systems that 'pose significant risks to the health and safety or fundamental rights of persons'. COM/2021/206 final (n 111) 3.

ML-based applications, the Act requires *inter alia*[114] that "high-risk" AI systems be trained, tested and validated on data sets[115] that meet the quality criteria—such as being "free of errors" and "complete"[116]—and are subject to "appropriate data governance and management practices".[117]

This raises the question of how the compliance and enforcement of the data quality requirements would interact with "how as-is-where-is" liability provisions, which seem to be a favored option.[118] As pointed out by the interviewees, this is far from being clear-cut and it remains to be seen which viable strategies would be adopted by businesses as a response to such compliance obligations. In theory, the downstream data user might raise a contributory liability claim against the upstream data supplier. The success of such claim would depend on the applicable law and individual circumstances of a case.

It needs to be further examined whether an alternative—more balanced compared to the one disclaiming the data provider's liability altogether—approach to the standardization of contractual provisions on liability might be viable.

2.4.6 Data Governance

GPAI and the OECD have recognized that some parties may want to include data governance terms in data sharing agreements, to ensure that data is shared responsibly, transparently, and in an accountable manner.[119, 120] This may be especially critical in multi-party data-sharing agreements, in which responsibilities for data stewardship and sharing may be complex and require careful management.[121]

Existing templates demonstrate the desirability for addressing data governance terms, particularly in multi-party agreements. For example, Microsoft's DUA-DC multi-party data sharing form addresses data governance by prompting parties to specify in the agreement: (i) the formation and operation of a Governance Committee to oversee the data sharing activities; (ii) the Data Governance Committee's composition, decision making processes, and requirements for approval or removal of data contributors

[114] 'Other requirements that are 'strictly necessary to mitigate the risks to fundamental rights and safety posed by AI and that are not covered by other existing legal frameworks' include high-quality data, documentation and traceability, transparency, human oversight, accuracy and robustness. COM/2021/206 final (n 111) 7.

[115] For regulatory definitions, see Article 3(29), (30) and (31) EU AI Act. At the same time, the definition of training, validation and testing data sets 'shall take into account, to the extent required by the intended purpose, the characteristics or elements that are particular to the specific geographical, behavioural or functional setting within which the high-risk AI system is intended to be used' (Article 10(4) EU AI Act).

[116] Article 10(3) EU AI Act.

[117] In particular, under Article 10(2) EU AI Act such data governance and management practices concern (a) the relevant design choices; (b) data collection;(c) relevant data preparation processing operations, such as annotation, labelling, cleaning, enrichment and aggregation; (d) the formulation of relevant assumptions, notably with respect to the information that the data are supposed to measure and represent; (e) a prior assessment of the availability, quantity and suitability of the data sets that are needed; (f) examination in view of possible biases; (g) the identification of any possible data gaps or shortcomings, and how those gaps and shortcomings can be addressed.

[118] Above (n 104) and the accompanying text.

[119] https://www.oecd-ilibrary.org/agriculture-and-food/issues-around-data-governance-in-the-digital-transformation-of-agriculture_53ecf2ab-en.

[120] https://gpai.ai/projects/data-governance/data-trusts-in-climate-interim-report.pdf

[121] https:/AI /dl.acm.org/doi/pdf/10.1145/3531146.3534637

or operators, among other responsibilities and (iii) the effects of dissolution of the data sharing arrangement.[122]

The Linux Foundation takes an alternative approach by explicitly placing governance aspects outside of its license templates to avoid imposing on other governance tools that it identifies as potentially more effective and/or adaptive for governing the use of data and AI.[123] Regardless of approach, designers of licensing templates should consider the interaction between licensing and other data governance tools.

2.4.7 SaaS AI

Data and AI increasingly are being made available on a software-as-a-service (SaaS) basis. Given the prevalence of SaaS services, it might be helpful to advance efforts on developing standardized terms for data and SaaS AI.

2.4.8 Enforcement

One data sharing challenge that flows through to licensing is the question of enforcement. Assuming that parties can agree upon contracts or standard terms to address the issues discussed above, the question may remain as to how the contractual terms should be enforced. For example, if a party breaches contract terms by using data in an impermissible manner, the data provider will want to have mechanisms to protect and enforce its rights and to be entitled to remedies, which potentially might include injunctive relief and/or monetary damages. Without these mechanisms and rights, data holders might be reluctant to participate in data sharing arrangements.

There are many factors to take into consideration in connection with enforcement and remedies. For instance, cross-border data sharing raises questions about where the rights should be enforced and the governing law. Additionally, practical considerations need to be addressed. For example, initiating a judicial action to seek remedies can be both time consuming and expensive. Some data holders may lack the resources to enforce their rights. Others who do enforce their rights may find the remedies inadequate because the data already has been used in an impermissible manner that financial awards may not necessarily address. The bottom line is that the need to have appropriate enforcement mechanisms for data licensing, particularly in a multi-party context, merits further attention.

3. Conclusion and Looking Forward

The preliminary work highlights that there is significant interest in developing standardized data licensing terms to facilitate data sharing, but that this work is challenging for various reasons. Nevertheless, it is thought that this work is beneficial and can potentially advance many important goals, so it is encouraged

[122] https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE4Rjfs
[123] https://cdla.dev/context/.

that it continues. To help advance these efforts, the following summarizes some of the challenges and potential paths forward for addressing them:

- To gain broad acceptance, the standardized terms will likely need to be developed through an inclusive multi-stakeholder process. Various organizations are working on standardized licensing terms, and they are encouraged (and other organizations that decide to pursue this work) to include as many different viewpoints and stakeholders as possible in the process. This should lead to more informed decisions about the content and structure of the standardized terms and broader acceptance and adoption of such terms.
- There likely will continue to be a wide range of different data-sharing arrangements and use cases, and a "one-size-fits-all" approach for data licensing may not be optimal, or even feasible. It would be encouraged that organizations working on standardized data licensing terms to consider developing a menu of different provisions or agreements that provide the community with options. This is similar to the approach used for Open Source and Creative Commons license agreements and already is reflected in some ongoing efforts to develop standardized data licensing terms.
- It is expected that the need for bespoke data licenses will continue, even as standardized terms become more common and accepted. This parallels the experience with Open Source and Creative Commons licenses.
- Section 2.4 of this report highlights some topics should be considered when contemplating standardized terms, including (i) standardizing definitions and developing models for allocating proprietary rights and usage and access rights (including when ethical considerations and/or data sovereignty are relevant considerations), (ii) addressing privacy and confidentiality, including when PII and/or other confidential information is being shared and/or PETs may be used, (iii) working toward fostering more data interoperability and better data quality and technical characteristics, (iv) allocating liability and providing for enforcement, (v) addressing SaaS and other business models, and (vi) providing a framework for addressing data governance. It is recognized that developing standardized license terms for all of these topics may not be feasible or easy, particularly in the short term and given the challenges that exist. Therefore, it would be encouraged that organizations prioritize work on those terms that seem most feasible, and continue to consider approaches for addressing the more challenging terms.
- While organizations work to develop standardized data license terms, it is encouraged that the global community continues to work on addressing the following issues that make this work more challenging:
    - Technical Matters: As reflected in the report, efforts to develop standards for defining and measuring data quality (including in light of proposed AI regulations), fostering data interoperability, and other technical matters remain relatively nascent. Progress on this front could significantly enhance data sharing and the negotiation of data licenses (including the crafting of standardized terms). Ongoing efforts in this area both on a sectorial basis as well as more broadly are encouraged.

    - Legal Uncertainties: As also reflected in the report, the evolving legal landscape and need for more cross-border harmonization creates further obstacles for data licensing and the crafting of standardized license terms. While this report does not express any views on how the underlying issues should be resolved, the Committee does want to sharpen the focus on how these legal and regulatory issues impact data licensing, so this correlation can be considered as policies continue to evolve. The legal and policy context in which AI innovation takes place is undergoing dynamic developments that should be factored into data-sharing practices. While some of the legal and other policy developments are reflected in the report, a comprehensive account of such developments would go beyond the report's scope given its preliminary character.

- Business Uncertainties: As also reflected in this report, business uncertainties can impede the negotiation of data sharing arrangements. While the Committee generally does not express any views on the underlying business issues, the Committee encourages the community to consider whether standardized terms ultimately might be crafted to reflect common business models that may emerge and to provide flexibility for parties to mitigate context-specific business risks and concerns. Among other things, this work could build on efforts to develop terms for AI Software-as-a-Service (SaaS). The work also should take into consideration the need to have options that reflect balanced terms for liability and other risks (e.g., and are not limited to “as is” agreements that disclaim all liability). Crafting of such terms (as well as any other terms) must be undertaken in compliance with competition laws and other applicable laws.

- Data Justice: As data sharing arrangements continue to develop, the Committee also encourage parties to continue to focus on data justice considerations, which have been highlighted by the GPAI Data Governance Working Group.

The IP Committee remains committed to advancing data licensing work, with the goal of unlocking beneficial data-sharing arrangements, including those that can enhance the development of responsible AI tools and applications. The Committee hopes that this report will aid the global community as it undertakes this important work. Given the challenges with this work, the Committee believes that this Committee is best positioned to help in this effort by focusing on specific use cases. Toward this end, the IP Committee plans to collaborate during the upcoming year with other GPAI Working Groups, such as the Data Institutions Committee within the Data Governance Working Group and the AI and Climate Working Group, as they work on data sharing projects. More specifically, this Committee can collaborate with these other Working Groups to help identify data-licensing terms that can help support the broader data-sharing efforts. Through this work, the Committee hope to make contributions that can inform the broader efforts to develop standardized data-licensing terms.

Finally, the IP Committee commends those organizations that are working on developing data licensing terms. The Committee invites those organizations to contact us with further questions about our findings and recommendations and to suggest ways the Committee might be able to assist them in advancing their work in a way that is consistent with our preliminary findings and recommendations.

4. Appendix A: Interview Questionnaire

Last updated: March 17, 2022

Questionnaire on Data and AI Model Licensing (GPAI IP Committee)

The Global Partnership on Artificial Intelligence - GPAI is a multi-stakeholder initiative that includes 25 countries and aims to bridge the gap between theory and practice on AI by supporting cutting-edge research and applied activities on AI-related priorities. The GPAI IP Committee (which is part of the Innovation & Commercialization Working Group) has launched a project that examines the need for and progress toward creating standardized agreements that will facilitate the voluntary sharing of data and algorithms, with the goal of fostering further development of AI. Open Source and Creative Commons agreements have been helpful tools for freely sharing software and content. Given that AI involves data and raises other considerations, the question is ripe for assessing what new form agreements should be crafted for data and AI models and determining the best pathways forward for developing them.

To assist us with our work in evaluating these issues, the GPAI IP Committee is inviting organizations to participate in a 90- minute informational interview focused on the questions below. We would be delighted for your organization to participate in such an interview. If you are interested, please contact Kaitlyn Bove (kaitlyn.bove@inria.fr) to schedule a time. Interviewees also are welcome to submit written information in response to the questions below. Interviewees can use Power Point or another similar tool during their interview, if they so choose. We are interested in gathering insights and perspectives, and it is fine to participate in an interview if you do not have information on some of the questions below.

1) Please specify whether you act more, and to what extent, as a licensor or a licensee of data and AI models or whether you act as a provider/draft of standard licensing contracts. Describe also in general terms the character and functions of the data as well as the later use of the AI models.
2) Data is increasingly being licensed or otherwise shared in order to train AI algorithms and for other purposes. How do you currently share data and models? How do you expect your practices to evolve over time?
   a) For software and content, open licensing models exist, such as open source licenses and Creative Commons. Work is underway to develop forms for sharing data. What types of licenses/agreements currently are being used to license or share data, including for the purpose of training AI models? Please share the experiences of your organization, to the extent applicable.
   b) What types of licenses/agreements are currently being used to license trained models? Please share the experiences of your organization, to the extent applicable.
   c) Are these licenses/agreements sufficient for these purposes from a legal and business perspective, and have they been adopted by a sufficient number of users?
   d) What is the current state of play in the standardization of the licensing agreements for sharing data for training AI models? Do the standardization attempts take place at the level of a sector/ individual technology/specific types of data or use-cases? Describe the efforts of companies/organizations working on these terms and your thoughts about their efforts.
3) What are some of the key issues that should be addressed in standardized agreements for i) sharing data to train AI models, ii) sharing data more broadly, and iii) sharing AI models?
   a) Are there standardized definitions of “data,” “model” commonly used in the agreements? Have you experienced problems due to the ambiguous contractual definitions of “data”, “model”, or licensed usage rights? How should “data” and “model” be defined? Is there a need to identify sub-categories of data and models (e.g., “untrained model” and “trained model”)?

   b) Do we need specific forms of agreement to address specific technologies, such as federated learning and/or other privacy enhancing technologies? If so, please describe.

4) Are training datasets usually defined as confidential information in data sharing agreements? Is there a clear distinction between data sharing licenses and confidentiality agreements, or do confidentiality clauses normally form part of data sharing licenses?
   a) In your view, can the requirements for the protection of confidential commercial information (including trade secrets) be effectively addressed under the standardized data sharing agreements, or does such protection require individually tailored agreements?
   b) What types of contract terms should be considered to protect confidentiality of the licensed information and data? How do these terms differ from those in agreements for sharing data that is not confidential?
   c) What are other critical intellectual property and proprietary rights issues, including from the perspective of addressing rights to i) underlying training data, ii) trained models, and iii) outputs from the trained models? What are the limitations of existing standardized agreements in this context, including reliance on the concept of "derivative works?"
   d) What are the key considerations for developing a form that enables parties to specify the permitted uses of the data and models (and/or prohibited uses)? How feasible is it to have a standard form that can be adaptable for different contexts and/or usage scenarios? If usage rights are clearly defined in the agreement, to what extent do you see a need to address ownership issues?
   e) What are the critical privacy considerations and how might they be addressed?
   f) Should ethical considerations be addressed, and if so, which ones and how?
   g) Is there a need to adapt standardized agreements to different jurisdictions in light of diverging approaches to IP and data protection or can a harmonized form be created?
   h) Please describe other considerations, including with respect to liabilities and remedies, that you think are applicable.
5) Do you think there should be a single form agreement or multiple forms with different terms (similar to Open Source and Creative Commons licenses)? Please explain.
6) Do you have suggestions for other organizations who should be interviewed in connection with this project?
7) What do you see as the best pathway for creating standardized agreements for sharing data and training algorithms? Please identify organizations that are leading in this effort as well as projects that might be well suited to pilot new form agreements.
8) What advice do you have on developing new standard form agreements?
9) Should the agreements be adaptable to enable parties to specify a common data format or standard (or how the data should be labelled) and/or an API for the data sharing arrangement?
10) Should the agreements be adaptable to enable the parties to specify data quality requirements?
11) How critical is it to develop these new standardized agreements?
12) Is there anything else that the GPAI I&C Working Group should consider for its work?