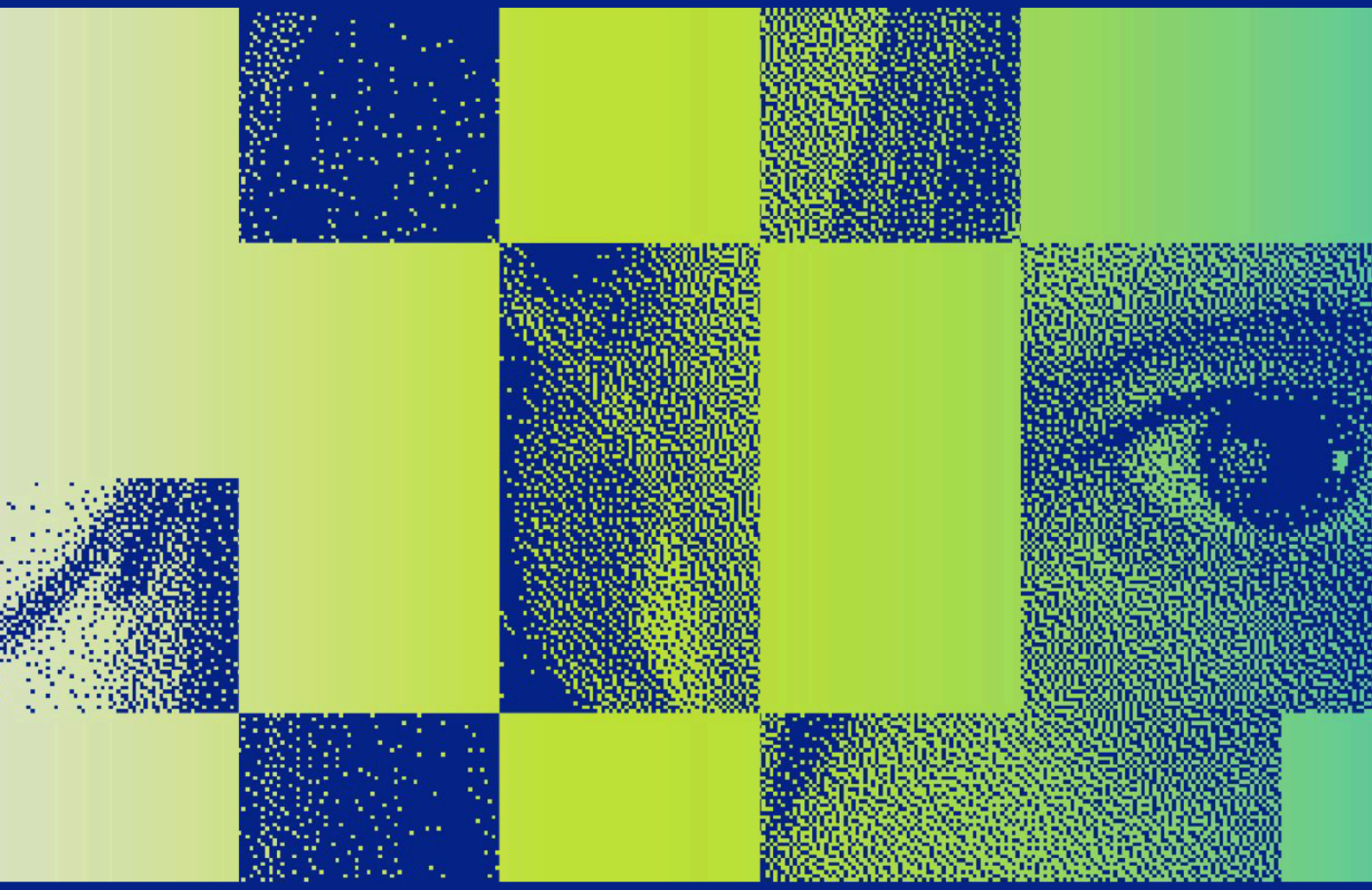


Unlocking data collaboration:
A study on data sharing practices and developing standard data licence terms to promote access and social good

Data-centric AI

ODI Research
ADVANCING TRUST IN DATA

January 2025

Contents

About

This report was researched and produced by the Open Data Institute (ODI) in collaboration with the Duke University Pratt School of Engineering published in January 2025. Its lead authors were Thomas Carey-Wilson, Professor Lee Tiedrich, Dr Gefion Thuermer, Nargiz Kazimova and Professor Elena Simperl. The report was funded through an unrestricted grant by Artificial Intelligence (AI) engineering consortium MLCommons Association. If you want to share feedback by email or would like to get in touch, contact the DCAI project lead Elena Simperl at elena.simperl@theodi.org.

To share feedback in the comments, highlight the relevant piece of text and click the ‘Add a comment’ icon on the right-hand side of the page.

Executive summary

As Artificial Intelligence (AI) continues to drive innovation across sectors and advance social good, the demand for well-governed and appropriate data—encompassing diverse formats such as text, images, audio, video and specialised data—has grown exponentially. To help meet this skyrocketing demand, society needs to find ways to responsibly share data voluntarily and in a legally compliant and ethical manner. Despite this demand and interest, stakeholders face challenges in implementing trusted data sharing arrangements.

The challenges stem from a variety of factors, including the lack of standardised and broadly used data governance techniques and contractual terms or data licence agreements (collectively referred to as 'data licences'). Standard data licences could potentially facilitate greater responsible data sharing in AI—similar to the way that Open Source and Creative Commons licence agreements have better enabled sharing of software and other copyrighted materials, respectively. Some efforts to develop standard data licences have begun. However, we are also increasingly seeing reliance on licence terms (often bespoke) restricting the use of some AI datasets. Strict enforcement of these restrictive terms could limit the availability and diversity of data for AI, potentially causing biases and other harms, and impairing model performance.

To find a path forward, we conducted qualitative research focused on three areas: 1) the factors driving organisations to share data in AI contexts; 2) the challenges these organisations encounter as data holders or data users; and 3) potential improvements to data licensing. This report provides insights on how more standardised data licence terms can be developed to help increase voluntary, legally compliant and ethical data sharing to advance responsible AI development and deployment, including for social good. The report draws on literature research, an electronic survey and oral interviews with data and AI experts to understand both the existing challenges and potential pathways for developing standardised data licence terms that can help overcome such challenges.

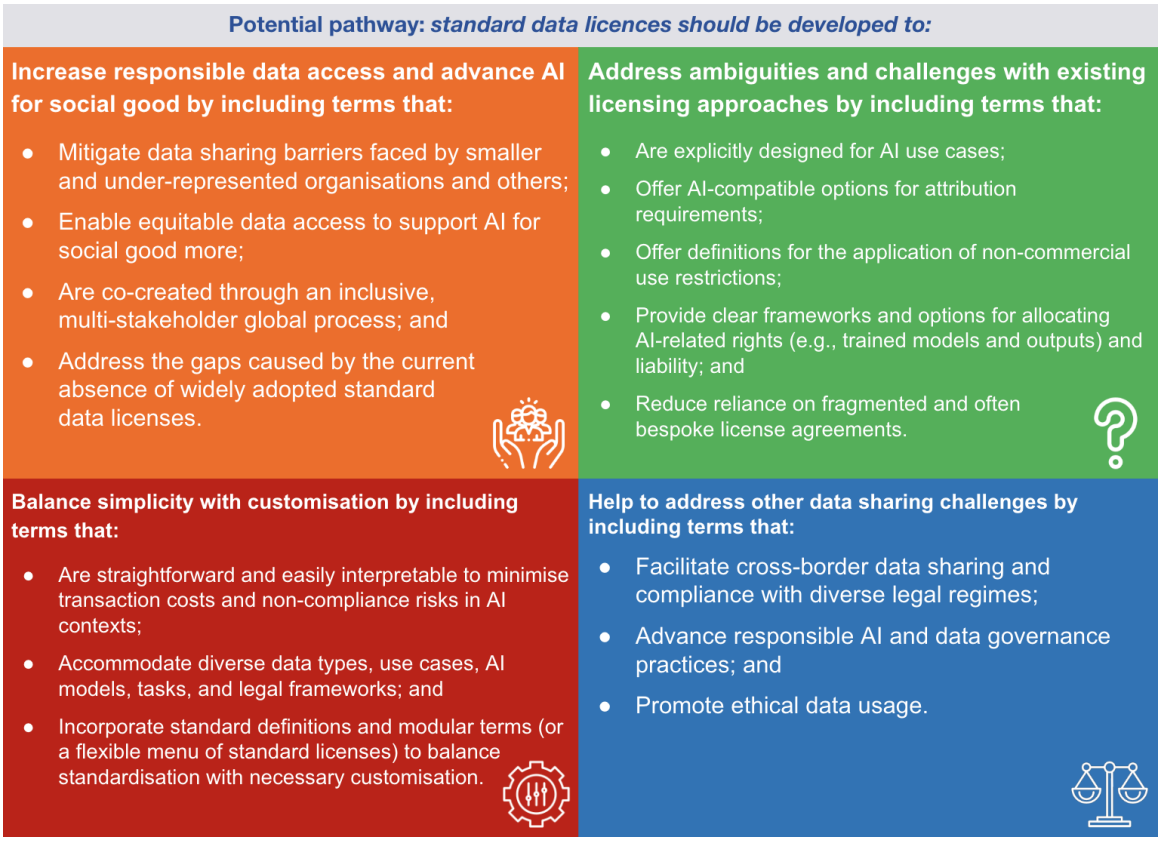
The survey and interviews spanned eight jurisdictions and parts of the data ecosystem, including: 1) data holders seeking to make data available for AI; 2) data users looking to access more data for AI; and 3) data intermediaries aiming to make third-party data available, including for AI. Some organisations that participated in the research play

multiple roles in the ecosystem. Participants included large and small commercial and non-commercial organisations, as well as individual researchers and practitioners.

Our key findings on the challenges and potential paths forward are organised according to four main themes: leveraging new standard data licences to 1) promote greater data access and social good, and 2) address existing ambiguities in current contractual approaches to data sharing, alongside 3) the need for data licensing simplicity and the importance of balancing standardisation with the need for some customisation, and 4) developing a new standard data licensing framework to help address other data sharing challenges, including legal compliance, data and AI governance, and ethical considerations. The development of standard data licence agreements would not preclude parties from entering into bespoke arrangements when such an approach better suits their needs, similar to the way that open-source and Creative Commons licences do not foreclose opportunities for bespoke contracts.

A summary of our recommendations can be found below in **Figure 1.**

Figure 1. Summary of recommendations for developing standard data licences aimed at incentivising broad adoption



1) New standard data licences can promote greater responsible data access and AI for social good.

a) **No standardised data licences have been widely adopted, increasing data sharing challenges, including for under-represented[1] [2] and smaller organisations.** Research confirms that the lack of standard and widely embraced data licence agreements increases barriers to voluntary, legally compliant and ethical data sharing to advance responsible AI. The lack of such licence agreements often increases costs and contractual compliance burdens, which, among other things, makes it harder for smaller or under-represented organisations to reap the benefits of AI, as they do not have the resources to navigate multiple agreements sufficiently.

b) **Data access disparities also impact the advancement of AI for social good.** Some organisations, like research organisations, governments and non-profits, especially in developing regions, also face significant challenges in accessing high-quality datasets needed for responsible AI development. Limited data sharing between different linguistic groups and cultures can also limit the diversity and richness of AI models, particularly due to the reliance on English-language datasets. Data licensing frameworks currently in place do not adequately address these disparities by sufficiently involving these stakeholders in their development. This may hinder the development of improved licence terms acceptable and accessible to these communities, potentially limiting opportunities to leverage AI for social good.

c) **Multi-stakeholder participation can help develop standard data licences that advance equitable data access and AI for social good.** Research indicates that creating an inclusive process is more likely to result in a standard data licensing framework that addresses these concerns and encourages the adoption of data licences resulting from this process. The process should include representatives from different geographic regions and sectors (including commercial, non-commercial and public sectors) and from groups that have faced challenges entering the technology sector. Standard data licensing frameworks should be crafted in a manner that enhances responsible access to data for everyone across geographies, including those who have faced challenges participating in the technology ecosystem and those stakeholders seeking to advance responsible AI for social good.

[1] Here we refer to 'under-represented' organisations as an umbrella term for organisations that typically face barriers to full participation and leadership in international policymaking processes. This includes organisations from the Global South, grassroots and community-based groups, and those representing marginalised populations. These organisations often lack the resources, access and influence of larger, well-established institutions from developed countries, leading to an imbalance in global decision-making forums.
[2] Inclusive Governance Initiative (2021), 'Reflections on building more inclusive governance'.

2) New standard data licences can help address existing ambiguities and challenges with current contractual approaches to data sharing.

a) **The lack of widely embraced standard data licences leads to inconsistency and incompatibility and increases the challenges of using data in compliance with licence terms.** A growing trend in industry is the use of custom terms of use for websites and social media properties that include customised data licences and are tailored to specific AI use cases and models. Research indicates that inconsistent or overly customised terms that do not achieve widespread buy-in can increase legal uncertainties and costs, making it harder for organisations to share data effectively across their ecosystems. Data users also face challenges in using multiple datasets that are collectively subject to many different licence agreements.

b) **Confusion exists about the meaning of non-commercial limitations in data licences, and clarification in new standard licences might encourage more data sharing transactions**. Parties often encounter contractual terms that restrict their use of data to 'non-commercial' purposes, and they have reported difficulty interpreting this term. Interest exists in developing one or more standard data licence agreements that specify or define permitted 'non-commercial' uses in more detail.

c) **New standard data licences could help clarify rights emerging from AI and provide different options for contractually allocating rights and liabilities.** Challenges also arise in allocating usage rights to the licensed data, as well as to the AI models and AI model outputs developed, directly or indirectly, using such data. Several widely-used standard licence agreements do not address these issues, and the law in many cases remains uncertain. New standard data licence agreements can help clarify these rights, to the extent permitted by, and consistent with, applicable laws. New standard data licences can also provide parties with more options for contractually allocating liability.

d) **New standard data licences can help address attribution issues arising under existing agreements.** A key concern with Creative Commons licences in the AI context, particularly CC-BY and CC-BY-SA, is the impracticality of contemplated attribution in AI applications, where data used in AI model training is not directly reproduced in AI model outputs. This challenge extends to other licence agreements, too. New standard data licensing frameworks can potentially help reduce this ambiguity and confusion by developing standard data licence terms tailored to AI's complexities. A possible solution may be incorporating technical solutions like automatic attribution. Standard data licence terms could reference these technical solutions, which in turn may help foster more efficient data sharing.

3) New standard data licences should embrace simplicity while balancing standardisation with the need for some customisation.

a) **There is a pressing need for simplicity.** Standard licence agreements should include clear easily interpretable terms to reduce transaction-related costs and build trust throughout the ecosystem. This is especially important for any set of standard licence terms which may have to be understood by non-legal practitioners. To help make standard licence agreements more usable, it also may be desirable to develop materials that raise awareness of the licences and explain their purpose and how they can be used.

b) **Standard data licensing frameworks should balance the need for standardisation with the need for some customisation, given the many types of data, use cases, models, tasks, and legal requirements.** While parties expressed interest in having more standardised data licences to facilitate transactions, they also underscored the need to have different data licensing options, particularly given the wide range of types of data, use cases, models, tasks, and possible legal requirements applicable to data and AI, which can vary across jurisdictions. For instance, there are different challenges with AI data licensing across the AI data lifecycle. Data licensed for tasks such as data collection and preprocessing are typically subject to more restrictive licensing terms due to the proprietary or sensitive nature of such data. In contrast, data used for downstream tasks, such as evaluation and deployment, tend to have less restrictive licence terms. Many existing standard licence agreements do not distinguish between different data use cases and thus may not necessarily address stakeholder needs that can vary based on the context in which the data is used. A multistakeholder process could help develop standard licence agreements tailored for using data in different parts of the AI data lifecycle as may be needed to address stakeholder interest or concerns. Some standard licences could permit more narrow uses for data intended to be used upstream, while other standard data licences may allow for broader downstream uses.

c) **Standard modular data licence terms (and/or a menu of standard data licence agreements) could strike a balance between standardisation and some customisation.** Further co-creation (including on a multi-stakeholder basis, as described above and in Section 1c of the findings) is warranted to determine which approach, or combination of approaches, is most suitable to achieve a balanced approach, while also addressing the need for simplicity. It is worth noting that multiple forms of open source and Creative Commons licences exist to help address an array of stakeholder needs when licensing software or other copyrighted materials.

d) **Standard definitions can advance data licensing.** Our research indicates interest in developing standard definitions that could be used across different data licence agreements. Standard definitions could be a key component in striking the appropriate balance between the need for standard data licence terms and some customisation. Standard definitions could also provide more simplicity.

4) **A new standard data licensing framework may help address other data-sharing challenges, including legal compliance, data and AI governance, and ethical considerations.**

a) **Standard data licence agreements could help address other challenges in implementing voluntary, legally compliant data sharing, including by helping advance data and AI governance.** For instance, designing standard data licences to accommodate technical solutions, such as tools for tracking data provenance and lineage, can enhance transparency. Data transparency practices in AI are evolving, with a lot of room for improvement across the industry to document data provenance, lineage, intended use and known limitations. Policymakers have focused on this need too. Sharing data is an opportunity to support responsible data practices in AI, and a new standard data licensing framework should seek to capitalise on this opportunity. For instance, this new framework could potentially reference desired data and AI governance practices. Standard data licensing is not a comprehensive solution to these challenges, but it can potentially support other solutions. As work on standard data licensing frameworks advances, these other considerations should be kept in mind.

b) **Standard data licences may help address cross-border and other compliance challenges.** The diversity of data-related regulations across jurisdictions (eg, data protection) continues to create substantial barriers to data sharing. Standard data licence terms should be prepared with these legal matters in mind. Consideration should be given to whether and how standard contract terms (or a menu of standard modular data licence terms) can help support legally compliant data sharing.

c) **Standard data licences may help advance ethical guidelines for data usage.** Our research confirmed interest in having ethical restrictions on data and AI model usage in at least some scenarios. We found evidence that this could be appropriate for at least some high-risk AI use cases like health and biomedical research. A new standard data licensing framework could reference relevant ethical codes of conduct. Some concern was expressed about incorporating codes of conduct or standards, which could change over time, into standard data licences, as this could increase uncertainty in the contractual terms and costs of compliance. These considerations should be addressed in the multi-stakeholder process for developing standard data licence terms.

These findings provide a framework to guide multi-stakeholder development of standardised data licences that are adaptable to AI applications and are designed with the goal of incentivising broad adoption. Creating a clear set of core terms, with the flexibility to address various AI use cases, models, tasks and data types, should help advance responsible, ethical, legally-compliant and voluntary data sharing and innovation and help build a more equitable, sustainable future for AI, benefitting all stakeholders.

Introduction

Access to data is essential for responsible AI to fulfil its promise.[3] [4] However, as the demand for data escalates, stakeholders increasingly face challenges in accessing well-governed and appropriate data for AI applications. Effective, standardised data licensing can help organisations access more data voluntarily, responsibly, and in a legally compliant manner. This is particularly effective when coupled with robust data governance resources that may also help navigate legal, ethical, and practical concerns.[5]

The current lack of broadly utilised standardised data licences that are mindful of AI use cases and emerging data practices creates obstacles for widespread, effective, and trusted data sharing. Previous studies[6] [7] [8] [9] have highlighted barriers caused by inconsistent data licensing practices and a reliance on agreements not drafted with AI or data licensing in mind.

To address these challenges, we conducted qualitative research focusing on three topics: 1) the factors driving organisations to share data in AI contexts; 2) the challenges these organisations encounter as data holders or users in sharing data for AI; and 3) potential improvements to data licensing to increase voluntary, legally-compliant and ethical data sharing. Through surveys and interviews, we gathered insights from a diverse range of stakeholders, analysed common themes, and derived actionable recommendations.

Our findings indicate a strong need for more standardised data licensing practices that also provide parties with some flexibility to customise standard terms with optional extensions to address different data types, AI models, use cases, tasks and legal requirements. Moreover, by addressing other associated challenges—such as promoting access, inclusivity and legal compliance; addressing ambiguities and challenges in existing licence approaches; simplifying the licensing process; and enhancing data and AI governance practices—we can further reduce barriers to voluntary and responsible data sharing.

---

[3] ScienceDirect, (2023), 'Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation'.
[4] ODI (2023),'Data-centric AI'.
[5] GPAI (2022), 'GPAI IP Expert – Preliminary Report on Data and AI Model Licensing'.
[6] ARXIV (2023), '[2310.16787] The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI'.
[7] GPAI (2022),'GPAI IP Expert - Preliminary Report on Data and AI Model Licensing'.
[8] Mozilla Foundation (2024), 'Training Data for the Price of a Sandwich - Mozilla Foundation'.
[9] CODATA Data Science Journal (2019), 'The Landscape of Rights and Licensing Initiatives for Data Sharing'.

Based on these findings, we recommend having a global and multi-stakeholder process focused on developing modular data licence terms (and/or a menu of standard data licence agreements) that strike an appropriate balance between the need for standardisation and some customisation and promote more simplicity in data licensing. This multistakeholder process should also focus on developing standard definitions to support the data licence terms, which should be designed to incentivise broad adoption. It should also seek to create a new standard data licensing framework that can help address other data sharing challenges, including legal compliance, data and AI governance, and ethical considerations.

Background

The demand for data

AI models' data requirements are growing extremely fast, with a 20% probability that data will become a bottleneck to AI improvements by 2040, according to analysis by MIT FutureTech in March 2024.[10] AI may not fully achieve its potential without access to suitable datasets.

Throughout the report, we refer to 'data' and 'datasets' interchangeably, the only difference is that 'data' refers to raw information in general terms, while 'datasets' are specific, structured collections of data ready for processing. It is important to note that both categories in this context encompass a wide variety of formats, sources and purposes.[11]

This diversity in data types reflects the broad range of AI applications across industries, from healthcare and finance to autonomous vehicles and robotics.[12] Additionally, data is used in AI systems for various tasks other than training occurring at different points in the AI data lifecycle, such as fine-tuning, benchmarking, testing and real-time inference.[13] Whether for training chatbots[14] or diagnosing diseases[15] AI relies on this data to learn, predict and seek to act intelligently.

As the demand for data continues to rise, a complex ecosystem of stakeholders has emerged, engaging in the flow and use of data for AI. Three key stakeholder groups are part of this ecosystem, each playing a distinct role in the AI data lifecycle[16], with organisations sometimes playing multiple roles:

- **Data holders:** They hold and control access to datasets and make them available for use by others. Data holders may include organisations, communities and individuals who collect, curate or generate data. Some data holders steward benchmark datasets, such as MoleculeNet, a collection of molecular datasets used for benchmarking ML models in drug discovery and chemistry.[17]

[10] MIT FutureTech (2024), 'What drives progress in AI? Trends in Data'.
[11] The ODI (2024), 'A data for AI taxonomy'.
[12] ScienceDirect (2024), 'AI revolutionizing industries worldwide: A comprehensive overview of its diverse applications'.
[13] The ODI (2024), 'Understanding data governance in AI: A lifecycle perspective'.
[14] ACM (2021), 'On the Dangers of Stochastic Parrots | Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency'.
[15] Nature Machine Intelligence (2021), 'Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans'.
[16] The ODI (2024), 'Understanding data governance in AI: A Lifecycle Perspective'.
[17] Zhenqin Wu et al (2017), 'MoleculeNet'.

- **Data users:** They access and utilise datasets. Data users may include researchers, developers, businesses, or other parties seeking to extract insights, develop applications, or conduct analyses using the available data.
- **Data aggregators or intermediaries (collectively, "intermediaries"):** Organisations hosting platforms, repositories, marketplaces and/or supporting infrastructures that facilitate the management, distribution, and/or access to datasets. These entities bring together data holders and users, offering locations for sharing data and/or tools to manage usage and /or licensing. While some intermediaries, like Hugging Face[18] or GitHub[19], store and manage data directly, others, like LAION, can provide references to datasets hosted elsewhere, with some metadata .[20] [21]

Data-sharing initiatives

Developing appropriate standard data licence agreements addressing the needs of the AI ecosystem is an important component of incentivising greater voluntary, legally-compliant and ethical data sharing. As the demand for training, fine-tuning, benchmark and other datasets grows, stakeholders across the public, private and third sectors[22] [23] are calling for more data sharing. This can be for driving research and innovation, delivering social good objectives, or providing new revenue streams for commercial organisations. A settled standard licensing framework is a key pillar to help reduce transaction costs and enable greater responsible data flows across parties, sectors and jurisdictions.

Many governments globally are calling for more data sharing to significantly accelerate the responsible and effective development and adoption of AI and/or other technologies.[24] [25] For example, policies such as the European Union's Data Governance Act (DGA)[26], the United States' Federal Data

[18] Hugging Face, n.d., 'The AI community building the future'.
[19] GitHub (2023), 'machine-learning-datasets'.
[20] ArsTechnica (2023), 'Artist finds private medical record photos in popular AI training data set'.
[21] LAION (2022, 'LAION-5B: A new era of open large-scale multi-model datasets'.
[22] An umbrella term for organisations that are neither private (ie, profit-motivated) nor public. This includes charitable, non-profit, civil society, non-governmental or community organisations
[23] National Audit Office, n.d., 'Successful commissioning toolkit: What are third sector organisations and their benefits for commissioners?'.
[24] Gov.uk (2022), 'National AI Strategy - HTML version'.
[25] Department for Science, Innovation & Technology (2023), 'Frontier AI: capabilities and risks - discussion paper'.
[26] The European Data Governance Act.

Strategy[27], and India's Public-Private Partnership Model[28] aim to promote access to certain datasets while maintaining standards of privacy, security and ethics. Several emerging publicly-funded solutions include the Common European Data Spaces, which are intended to enable secure data exchange across sectors as envisioned by the DGA.[29]

Moreover, non-governmental and non-commercial entities, including academia and non-profits, require data but can often face high data acquisition costs, limiting their ability to contribute to AI research and development, and to use AI to advance social good causes.[30] The AI research community has also created large-scale data repositories, such as Papers With Code[31], to promote data sharing and collaboration. Researchers and practitioners work together to create shared benchmark datasets, which are used to systematically assess progress on a given AI task and draw from public data, or data shared by the contributing entities.[32]

Non-governmental organisations (NGOs) use AI to support their missions, like mapping land to monitor deforestation and illegal logging[33] and natural language processing for fundraising.[34] Civil society groups, such as European Digital Rights (EDRi), advocate for stricter data-sharing rules, calling for bans on biometric surveillance and mandatory human rights impact assessments.[35]

As AI companies seek access to diverse and high-quality datasets, several have entered into data licensing transactions with various data holders. For example, Google reportedly secured a (US) $60 million per year deal with Reddit for access to its content[36], while Shutterstock reportedly has ongoing contracts with Meta[37] and Amazon[38] to provide images, videos and music for AI training. Such deals reportedly often involve continuous access to data via APIs or recurring data transfers,[39] providing AI companies with updated datasets to improve model performance.

---

[27] Federal Data Strategy development team (2021), 'Federal Data Strategy 2021 Action Plan'.
[28] Financial Express (2023), 'For making responsible AI models, public datasets will be shared only with trusted firms: Chandrasekhar'.
[29] ELJT (2024), 'The Regulation of Data Spaces under the EU Data Strategy: Towards the 'Act-ification' of the Fifth European Freedom for Data?'.
[30] Nature (2020), 'AI for social good: unlocking the opportunity for positive impact'.
[31] Papers With Code (2024), 'Trending Research'.
[32] Arxiv (2024), 'AI Competitions and Benchmarks: Dataset Development'.
[33] Space Intelligence, n.d., 'Driven by Purpose, Led by Experts. Learn about Space Intelligence's team'.
[34] Nonprofit Tech for Good (2024), 'The Use of AI in Fundraising: Maximizing Donations with Data-Driven Asks'.
[35] European Digital Rights (2021), 'Open letter: Civil society call for the introduction of red lines in the upcoming European Commission proposal on Artificial Intelligence'.
[36] Reuters (2024), 'Exclusive: Reddit in AI content licensing deal with Google'.
[37] Shutterstock (2023), 'Shutterstock Expands Partnership with OpenAI, Signs New Six-Year Agreement to Provide High-Quality Training Data'.
[38] Reuters (2024), 'Inside Big Tech's underground race to buy AI training data'.
[39] ArsTechnica (2024), 'Reddit cashes in on AI gold rush with $203m in LLM training license fees'.

Furthermore, many AI companies continue to rely on data scraping to access data. While there is no standard definition, data scraping generally refers to extracting data from third-party websites or social media properties often without obtaining express content from or providing consideration to, or otherwise coordinating with such third-party websites or social media properties.[40] Data scraping can encompass web crawling and/or other techniques, and it has given rise to several pending and threatened legal disputes.[41]

Data licensing 101

Data licensing refers to contractual arrangements that facilitate the sharing of data. Data licences may include a variety of terms, such as for accessing, processing, using, and redistributing datasets.[42] Data sharing can be incentivised and made more efficient by developing standardised data licensing terms that help foster good data governance, transparency and legal compliance, on terms that are widely embraced throughout the ecosystem.[43]

While documentation tools like data cards provide valuable insights into dataset provenance, lineage, intended use, or known limitations[44], data licences can require or encourage use of such tools, in addition to reducing legal and other uncertainties.[45] Appropriate data licensing can build trust between data holders, users and intermediaries by clarifying the terms pursuant to which data may be used and allocating other contractual rights and responsibilities. Having standardised data licence terms should incentivise and facilitate efficient and voluntary data sharing, which is crucial for a sustainable AI data ecosystem.

Data licensing stakeholders, models and examples

Developing standard data licences requires stakeholders within the ecosystem to collaborate to formulate terms that will be widely embraced and accommodate their respective needs in connection with responsible and ethical data sharing. For example, the terms should address data holder needs relating to access, sharing and other use of their data. At the same time, these terms should satisfy the needs of data users and enable parties to easily understand their respective contractual rights and obligations.

---

[40] Arxiv (2024), 'Consent in Crisis: The Rapid Decline of the AI Data Commons'.
[41] Bipartisan Policy Centre (2024), 'Legal Challenges Against Generative AI: Key Takeaways'.
[42] Arxiv (2023), 'An investigation of licensing of datasets for machine learning based on the GQM model'.
[43] Data Science Journal (2019), 'The Landscape of Rights and Licensing Initiatives for Data Sharing'.
[44] Arxiv (2018), 'The Dataset Nutrition Label: A Framework to Drive Higher Quality Data Standards'.
[45] GPAI (2022), 'GPAI IP Expert: Preliminary Report on Data and AI Model Licensing'.

A standardised data licensing framework should also address the role of data aggregators or intermediaries in the AI ecosystem. Presently, data intermediaries rely on a variety of data licensing options to meet different data-sharing needs for their users, ranging from open licences like Creative Commons to more restricted commercial or research-only agreements. For example, platforms like Kaggle offer specific licensing for public and competition datasets[46], while the NeurIPS Datasets and Benchmarks Track require licence information for submitted datasets.[47]

As the AI ecosystem evolves, licence models and attempts to standardise terms associated with them do too. The Global Partnership for Artificial Intelligence (GPAI)[48] discusses efforts to develop standardised AI data licences, including ongoing initiatives such as the Linux Foundation's Common Data License Agreement (CDLA) Permissive 2.0 and Computational Use of Data Agreement (C-UDA) 1.0, among others.[49] Recent developments, such as the Copyright Clearance Center's release of a new collective AI licence in July 2024 (the "CCC License"),[50] [51] highlight ongoing efforts to create AI-specific licences. We conducted a review of licences used by a selection of known data users, holders and intermediaries.

Data licensing challenges

Data licensing in AI presents significant challenges that hinder voluntary, legally-compliant and ethical data sharing. A primary issue is the need for more standardisation in licensing terms, with many datasets from intermediaries like Kaggle either missing licences or containing inaccurate information about the applicable licences.[52] The lack of widely-used standard data licences may exacerbate the legal ambiguities, including around data usage rights, as it could increase the difficulty of identifying the correct licence terms that apply. The situation also often leads to problems such as 'licence laundering', where improper application of licences occurs as datasets are reused downstream.[53]

Additionally, the lack of widely-embraced standard data licences has contributed to a growing trend of including restrictive data licensing terms in website terms of service. Concerned about unwanted data scraping, websites and social media properties increasingly turn to restrictive data licence terms

[46] https://www.kaggle.com/docs/datasets
[47] https://neurips.cc/Conferences/2023/CallForDatasetsBenchmarks
[48] Kaggle, n.d., 'How to Use Kaggle'.
[49] ibid.
[50] https://www.copyright.com/blog/ccc-launches-collective-ai-license/
[51] Since the CCC licence was released a short time ago and after the conclusion of our survey, this report does not address its use.
[52] Arxiv (2023), 'The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI'.
[53] ibid.

(typically within terms of service), as well as robot.txt and other technical measures to prevent these practices. With a 20% probability that data will become a bottleneck to AI improvements by 2040,[54] authors point to this trend as potentially limiting the supply of data for AI.[55] As we outline later in the findings of Section 1c, the development of standard data licence terms through a multi-stakeholder inclusive process could potentially help reverse this trend.

Research further indicates that data holders often seek to tailor restrictions in data licences based on the nature of the data or how their data will be used in the AI data lifecycle. For instance, more restrictive data licence terms often apply to data used for upstream tasks occurring early in the AI data lifecycle, such as data collection and data pre-processing. This is unsurprising, as these tasks more often involve processing proprietary or sensitive data. In contrast, data involved in downstream tasks such as AI model evaluation, may have fewer licence restrictions, as this data may be less sensitive or proprietary than data used upstream.[56] This suggests that it may be appropriate to tailor standard data licence terms based on how the data will be used at different stages in the AI data lifecycle.

Research also points to other confusion arising with current data licensing practices. For instance, some data licences restrict data use to non-commercial purposes, and parties have reported difficulties interpreting this term and distinguishing between commercial and non-commercial contexts.[57] [58] Similarly, many licence agreements used in the AI context include attribution requirements. Parties have faced challenges complying with these requirements because the agreements were not drafted with AI-specific needs and uses in mind. As an example, current licences typically do not address attribution in the AI context where the licensed data is used for AI model training, fine-tuning or evaluation and is not reproduced in AI system outputs.[59] [60]

To address these challenges, there is a need for standardisation and clearer data licence terms that better align with AI system development and deployment, promoting legal compliance, more trust, and equitable data access. These improvements are critical for fostering innovation and ensuring that AI developments serve the broader social good. Table 1 summarises the key challenges outlined in the literature.

---

[54] FutureTech (2024), 'What drives progress in AI? Trends in Data'.
[55] Mozilla (2024), 'Training Data for the Price of a Sandwich'.
[56] Arxiv (2019), 'Towards Standardization of Data Licenses: The Montreal Data License'.
[57] Arxiv (2023), 'The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI'.
[58] Arxiv (2019), 'Towards Standardization of Data Licenses: The Montreal Data License'.
[59] GPAI (2022), 'GPAI IP Expert: Preliminary Report on Data and AI Model Licensing'.
[60] EUIPO, n.d., 'Study on the impact of artificial intelligence on the infringement and enforcement of copyright and designs'.

Table 1. Summary of key challenges

Challenge	Description
Lack of standardisation in licensing terms	Many datasets lack licences, contain inaccurate information about the applicable licences, or have bespoke licence terms. The lack of widely-used standard data licences may increase compliance challenges, legal ambiguities and issues like 'licence laundering', where incorrect licences are applied to downstream data reuse.[61] [62]
Restrictive licensing in website terms of service	Websites and social media platforms increasingly impose restrictive data licensing terms and implement technical measures to prevent data scraping. These practices significantly limit the availability of data for AI development.[63]
Tailored restrictions based on data nature and use	Data holders often seek restrictions that vary based on the data type or its intended use. More stringent restrictions are often wanted for data used upstream, such as in data collection and preprocessing, as compared to data used for downstream tasks existing data licences typically do not provide desired options for tailoring the permitted uses.[64]
Confusion over non-commercial use restrictions	Non-commercial use restrictions in licences often lead to confusion among parties, as it can be difficult to distinguish between commercial and non-commercial applications, which hinders data sharing and increases compliance risks.[65] [66]
Challenges with attribution requirements	Attribution requirements in existing licence agreements often fail to account for AI-specific contexts. Challenges often arise when licensed data is used for training or evaluation, where direct reproduction in AI models and/or AI outputs is absent.[67] [68]
Need for standardisation and clearer licence terms	There is a critical need for standardised and clearer data licence terms that align with AI system development and deployment. Such standardisation could help promote more legal compliance and trust and enable more equitable data access, fostering innovation and serving the broader social good.[69] [70]

[61] ibid.
[62] GPAI (2022), 'GPAI IP Expert: Preliminary Report on Data and AI Model Licensing'.
[63] Mozilla (2024), 'Training Data for the Price of a Sandwich'.
[64] Arxiv (2019), 'Towards Standardization of Data Licenses: The Montreal Data License'.
[65] Arxiv (2023), 'The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI'.
[66] Arxiv (2019), 'Towards Standardization of Data Licenses: The Montreal Data License'.
[67] GPAI (2022), 'GPAI IP Expert: Preliminary Report on Data and AI Model Licensing'.
[68] EUIPO, n.d., 'Study on the impact of artificial intelligence on the infringement and enforcement of copyright and designs'.
[69] Mozilla (2024), 'Training Data for the Price of a Sandwich'.
[70] Arxiv (2019), 'Towards Standardization of Data Licenses: The Montreal Data License'.

Research findings and recommendations

The challenges summarised in **Table 2** informed the design of our research. We combined insights from a survey with **51** participants with in-depth interviews with **15** participants from all relevant stakeholder groups across **five** countries. A detailed account of our research methodology is available in **Annex A**. We report and explain below the four main themes that emerged across both survey and interview responses:

1) New standard data licences can increase responsible data access and AI for social good
2) New standard data licences can help address existing ambiguities and challenges in current contractual approaches to data sharing
3) New standard data licences should embrace simplicity while balancing standardisation with the need for some customisation
4) New standard data licences may help address other data-sharing challenges, including legal compliance, data and AI governance, and ethical considerations

1) New standard data licences can promote greater responsible data access and AI for social good

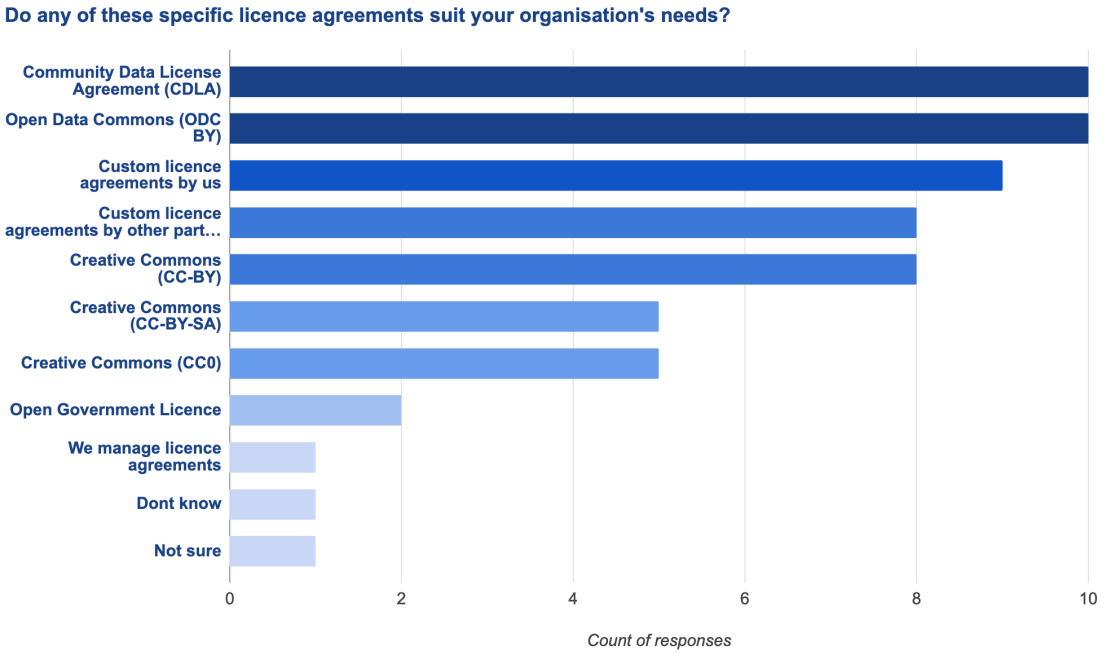
a) No standardised data licences have been widely adopted, increasing data-sharing challenges, including for under-represented and smaller organisations

Our research highlights that the absence of widely adopted, standardised data licences exacerbates data-sharing challenges. This is particularly the case for smaller and under-represented organisations, who may lack sufficient resources to negotiate, navigate and/or manage many different licence agreements. Firstly, several interviewees highlighted that the trend of using non-standardised licence agreements is creating fragmentation, which increases workload and decreases

clarity. As one practitioner noted: ‘We don't have time to read all of [the licences]. We need to provide basic guidance.’

Our results for the question, ‘Do any of these specific license agreements suit your organization’s needs?’ (Q11), support this, as we find a significant range of licences in use by all respondents. According to our results, the most popular licences in use for AI system development are the Community Data Licence Agreement (CDLA) with **10** (**15%** of total responses), Open Data Commons (ODC BY) with **9** (**14%** of total responses). The survey insights are shown in **Figure 2**.

Figure 2. All respondents’ preference in licences, n = 22



It is also notable that a combined **17** (**25%**) of the survey responses were either ‘Custom agreements by us’ or ‘Custom agreements by other parties’. This suggests that a significant number of the organisations we surveyed find bespoke licence agreements suitable for their organisation, adding to the complexity of licences in use. This is because bespoke licence terms can vary significantly from agreement to agreement: As one interviewee explained, ‘Everybody is trying to make up their own licence... You have as many licences as you have lawyers.’ Moreover, although some survey respondents reported that some Creative Commons licences suit their needs, as explained in **Section 2a** of this report, the survey also indicates that some organizations find them challenging. A description of the various Creative Commons licence agreements can be found on the Creative Commons website.[71]

[71] Creative Commons, n.d., ‘About CC Licenses’.

Multiple interviewees directly involved in AI development express a desire for standardised licence terms to enable more efficient data sharing. For instance, one industry expert emphasised: *'Having more standardized terms is good... I think having standard terms to a large extent is helpful... To me, those are the really valuable things.'*

Our interviews confirmed that the complexity of the current landscape can be especially problematic for smaller organisations, which lack the in-house legal expertise to navigate it. For instance, one interviewee explained: 'We see a lot of [startups] and they tend to come to us sooner about questions on the legal end of the spectrum largely because they don't have the resources in-house for that.'

As shown in **Figure 3,** our survey results support this point. When respondents were asked to rate the statement, 'The terms and conditions of licence agreements are often difficult to understand' (Q12), a disproportionate number of those who agreed were from organisations with 0–50 employees (the values in each cell represent the proportion of the total responses).

Figure 3. Data users' rating of the statement 'The terms and conditions of licence agreements are often difficult to understand' against company size, n = 23

	Disagree	Neutral	Agree	N/A
0-50	4.3%	4.3%	21.7%	0.0%
1001+	8.7%	4.3%	13.0%	4.3%
201-1000	8.7%	0.0%	13.0%	0.0%
51-200	8.7%	4.3%	4.3%	0.0%

One interviewee from the international development sector pointed out that similar challenges exist for organisations from underrepresented countries and regions, which can lead to unintentional contract breaches: *'We have projects [in the Global South] whose teams do not know much about licensing. So they need to learn what it means, and what the different types of licences mean, because they can easily breach licences without even knowing.'*

In addition to the risk of unintentional contract breaches, some interviewees explained how effectively selecting a licence or exercising legal rights to use the data in a desired way is a challenge given the range of licences available. This is especially the case on the platforms hosted by data intermediaries, where the abundance of different data licence options can become daunting for some users. As an interviewee explained: *'People are just publishing data... under the Apache licence, MIT, Creative Commons, it's data everywhere. If you look at Hugging Face, it's all over the map... putting a licence agreement on data that gives you the legal rights to do what you want is a challenge.'*

Furthermore, through the qualitative survey responses to the question, 'What, if any, measures does your organization currently implement to ensure that your organization's data is used ethically and in compliance with agreed terms?' (Q13), we saw numerous responses outlining that the lack of standard licences for certain, particularly sensitive, use cases increases the challenge of ethical and legally compliant data use.

For instance, one respondent deploying textual analysis tools to patient health information (PHI) mentioned that their organisation is 'proceeding cautiously' because there are no standard licences to guide the ethical sharing of sensitive data for this use case. Some interviewees mentioned that this lack of standardisation requires them to extensively consult legal experts and rely on custom licence terms, highlighting the challenges which some organisations face when standard licences are unavailable for their specific use cases.

One interviewee from the tech industry also mentioned that existing licences designed to prevent misuse or harm in open-source models are not fully standardised: *'With the deployment of open source models, we are seeing an evolution of standardised licences, that are actually not standardised because they have bespoke restrictions to mitigate against different kinds of misuse and harm and abuse… it renders those licences not truly standardised.'*

Other interviewees explained this further, confirming that organisations typically license data from different third parties, and that the applicable licence agreements often include various non-standardised terms seeking to prevent data misuse or other undesired uses. The data obtained from these various sources is often aggregated for use in AI development. This presents data users with the challenge of having to comply with a patchwork of different, often complex, data licences that can become unmanageable. As one senior practitioner from the tech industry highlighted: *'I don't think the ethical restriction of some should be in any licences… If I start to have data licensing with 10, 20, 40 different licences, which by the way, some of them are not that adapted... then it becomes unmanageable.'*

This patchwork of data licences increases the risk of non-compliance with applicable contractual terms. This in turn, may deter stakeholders, including data providers and users, from sharing data. Unsurprisingly, cost is a factor impacting organisations' ability to comply with the terms of many different licence agreements. This was consistently highlighted for smaller-scale organisations that may lack the resources to comprehensively map the landscape of data licences. For example, one interviewee from an organisation that stewards a licence mentioned: *'Some organizations who are more on the startup SME scale express that they simply lack the resources to understand in detail the system of licences out there.'*

Furthermore, underrepresented countries can also experience increased costs related to the complexity and patchwork of licences. One interviewee from the

international development sector said that organizations they work with in these geographies do not know much about data licensing in the first place, increasing the risk of contract breaches: *'We have projects whose teams do not know much about licensing. So they need to learn what it means, and what the different types of licences mean, because they can easily breach licences without even knowing... we try to make sure that the funds they receive through us also include funding for capacity strengthening, not only for data collection and analysis but also data sharing and licensing."*

As confirmed by various interviewees, standardised data licence terms could help alleviate these challenges by clarifying rights and obligations, reducing customisation needs, and facilitating broader AI data usage. Some interviewees highlighted that, for general purposes, some elements of standard licences like CC-BY-SA are 'crystal clear' insofar as they 'help users understand what they can and cannot do', adding that this eliminates ambiguity regarding rights and obligations. However, as we show later throughout Section 2, this does not imply that these licences are necessarily suitable for data in the context of AI use cases, or that these terms are sufficiently clear when used broadly in the AI context.

b) Data access disparities also impact the advancement of AI for social good

Some large commercial AI developers have entered into bespoke commercial arrangements to gain access to data. This is evident in data-sharing transactions such as those between companies like Google and Reddit.[72] as well as transactions between Open AI and the Associated Press[73], or Meta and Shutterstock.[74]

However, based on information available to us, many entities do not appear to be able to gain lawful access to this data. This could adversely impact their ability to develop competitive AI systems or leverage AI systems for social good. While most of these bespoke deal terms have not been shared publicly, one interviewee highlighted this as a potential issue: *'Data is a key bottleneck... we have seen in the last months and years an increasing number of exclusive licence agreements coming from some large stakeholders… making sure that those deals are not exclusive [is crucial] because again, if those are made exclusively, it prevents other stakeholders from accessing this valuable data."* [75]

[72] Reuters (2024), 'Exclusive: Reddit in AI content licensing deal with Google'.
[73] AP (2023), 'ChatGPT-maker OpenAI signs deal with AI to license news stories'.
[74] Variety (2024), 'AI Content Licensing Deals with Publishers: Complete Updated Index'.
[75] Some interviewees characterised these licensing arrangements as exclusive. Because the terms of these transactions are not publicly available, we do not know whether or to what extent they are exclusive.

Our research also uncovered that access to data is crucial for advancing AI research and addressing societal impacts in underrepresented regions like those in Africa.[76] Through our interviews, we heard how social good uses of AI technologies, such as scientific research, would benefit from fair access to datasets via licence terms to this effect. One interviewee proposed: 'Making sure that there's strong terms about fair access to certain data could be enforced... to ensure that deals are not exclusive.' As mentioned, most of these bespoke transaction terms have not been disclosed. However, some have noted that this lack of transparency may prevent stakeholders from understanding the full scope of data-sharing agreements, potentially leading to unfair advantages for certain parties.[77] This observation is consistent with concerns outlined by several interviewees.

Some interviewees explained how cost can contribute to data access disparities, as access to certain datasets can be prohibitively expensive, including those used for social good (eg, scientific or humanitarian) purposes. High-resolution meteorological or satellite data, for instance, is often too expensive for low-resource research institutions in the Global South. An interviewee explained: *'There are many challenges. For sure the availability of high-quality data in the fields that I work or I kind of support... especially on the ground. There's a lot of satellite data... some are open access, some of them are not. The ones that are not open... are very expensive, and our grantees [from the Global South] usually don't have the resources to access them. They do have access through conventions and partnerships that we have with space agencies or other agencies... or through Northern grantees that form consortiums with the Southern grantees.'*

The interviews show that entities from the Global South can provide useful data for AI, but that equitable frameworks are needed to govern arrangements between the Global North and South. As an interviewee from the international development sector mentioned: *'From data that people in Africa thought to be maybe useless… it's now useful because of this dearth of data…Northern organizations can transfer capacity [to harness data] through equitable consortia with Southern partners.'*

The interviews also highlight additional negative impacts of failing to equitably unlock data from the Global South, such as limiting access to data in different languages and reflecting different cultural perspectives. Such data could be used to localise AI models for under-served regions, helping to address digital divides. Interview participants highlighted the need for diverse data sources to enrich AI models: *'We need different kinds of cultural aspects… and we do train our models [with] open public data that would be accessible without effort… So, what we need is a large corpus of data, and that's where we believe that actually, we need data from diverse sources.'*

---

[76] UNDP (2024), 'Africa Development Insights: Theme: Artificial Intelligence for Development'.
[77] GPAI (2023), 'Fostering Contractual Pathways for Responsible AI Data and Model Sharing for Generative AI and Other AI Applications'.

Some interviewees further described an overreliance on English-language textual datasets, which limits the diversity and richness of AI models, affecting their suitability across different cultures and languages: *'We know that 65% of the data available in the world is in English. So there's also a question of making sure that in the process of training our model, we can reflect the diversity of the world.'*

c) Multi-stakeholder participation can help develop standard data licences that advance equitable data access and AI for social good

Our interviews indicate that well-structured standard data licence agreements, developed with the buy-in of diverse stakeholders, will likely increase equitable data sharing and help advance AI for social good. As one interviewee explained: 'We have organisations who want to work on something together and we help put together an intellectual property framework around it… [Licences] help level the playing field, so smaller entities can participate without fear of being exploited by larger ones.' This suggests that a multi-stakeholder approach to developing standard data licences can play an important role in addressing bargaining power imbalances and potentially increase opportunities for smaller participants to engage in collaborations on terms more acceptable to them.

The multistakeholder process should include organisations of multiple scales and sectors from all over the world to help address disparities and drive data driven innovation on mutually agreeable terms.[78] Many interviewees stressed the importance of creating equitable and inclusive partnerships between organisations from the Global North and South when working on AI and data-related projects: 'A lot of our effort is to make sure that when a partnership includes Northern and Southern organisations, it's an equitable partnership... the partnership is well-designed so that the weight is always towards capacity building for the Southern organisation.' Thus, it is important that any global efforts to develop new standard data licences are designed with the goal of creating mutually acceptable terms that might help to equitably mitigate the risks of one-way data extractivism.[79]

This highlights the need for engaging multiple stakeholders (including proactively engaging smaller and underrepresented entities) to raise awareness and help ensure that standard data licences reflect diverse perspectives and are accessible and understandable for a broad community. As mentioned in one interview by an organisation hosting a data licence: *'It needs to be multi-stakeholder... [for example] we had 250 lawyers reviewing and*

[78] GPAI (2022), 'GPAI IP Expert: Preliminary Report on Data and AI Model Licensing'.
[79] Springer Nature (2020), 'The Nooscope manifested: AI as instrument of knowledge extractivism'.

commenting on our agreement.' This notion was supported by other interviews, as a multistakeholder approach can ensure that licences are clear and flexible enough to cater to various 'cultural and economic perspectives'.

2) New standard data licences can help address existing ambiguities and challenges with current contractual approaches to data sharing

a) The lack of widely embraced standard data licences leads to inconsistency and incompatibility, and increases the challenges of using data in compliance with a patchwork of licence terms

Next, our research suggests that reliance on the existing patchwork of standard licence agreements, many of which were drafted for non-AI purposes, can create ambiguity and confusion in AI development and deployment. This is for several reasons. Firstly, even if licences are tailored to AI data types, models, use cases, tasks and legal requirements, the increasing use of organisation-specific bespoke licence agreements creates complexity. As mentioned by one interviewee from the tech sector: *'If you're taking data under many different licenses or structures... what kind of chaos do you get at the end when you try to figure out all your obligations?'* In other words, as the range of available licences expands, so does the complexity associated with understanding and complying with them.

Regarding bespoke licences used in pharmaceutical research, one interviewee mentioned: *"I've got a pool of licences here that are different than the pool of licences they have over here. So that's part of the challenge... I want these to be standard licences."* This suggests that inconsistent or overly customised terms that do not achieve widespread buy-in can create legal challenges and/or uncertainties, making it harder for organisations to share data effectively across the ecosystem.

This issue is exacerbated as AI solutions often require data from multiple sources, each potentially governed by different licences. For example, practitioners have raised concerns about how well CDLA-Permissive-2.0, a permissive open data licence, interoperates with other widely used licences

like CC-BY-4.0, which is commonly used in AI research.[80] As one interviewee explained: *'We are never using one dataset... you combine the licence you have on the train model [with the] licence on the dataset. You need to make sure that all of this is working together.'*

Our research also showed that smaller organisations like startups struggle with the complexities of navigating between terms from multiple licences because 'they don't have the resources in-house for that'. One interviewee using data for AI flagged the high transaction costs: *'You have this issue of proliferation of licences and compatibility of all these licences... the high transaction cost here, it's absolutely the most critical.'* This, and the other factors mentioned here, work together to create risks due to complex and incompatible licences, making it increasingly difficult to use data in compliance with licence terms throughout the AI ecosystem.

There is evidence that these challenges may also deter data holders from sharing data. Without applying simple standard licence terms to the data throughout the AI data lifecycle, organisations may face legal and operational risks,[81] which can diminish their willingness to share data: *'The major aggregated fine-tuning datasets tend to have one licence left on top and that licence...is a mishmash of things, some of which are non-commercially licensed, some of which are commercial, every licence under the sun... I think the main people who are hesitant to share data are private companies because they know... Everybody's going to train on anything that they release, regardless of what terms or licence are attached to it.'* Given this, the development of standard AI-specific licences may help to address some of these concerns.

b) Confusion exists about the meaning of non-commercial limitations in existing licences, and clarification in new standard licences might encourage more data-sharing transactions

Our research indicates challenges in understanding and complying with licence terms that permit using data for non-commercial purposes only. In our survey, when data users were asked to respond to the statement, 'Non-commercial use is clearly defined' (Q12), a significant **56.5%** disagreed, while only **17.3%** agreed, **13%** were neutral and **13.2%** preferred not to answer. Some of our interviewees

[80] Open Knowledge Forums (2021), 'License compatibility as imperative?'.
[81] Global Partnership for Sustainable Development Data (2021), 'Effective and Ethical Data Sharing at Scale'.

went further to describe how this ambiguity can present a barrier to organisations sharing or acquiring certain datasets: *'Non-commercial is subjective... It is both subjective in what people consider commercial and in the risk profile of companies around that. There are some datasets that we've used that companies don't want to touch because of this ambiguity.'*

Survey responses to the question, 'Which, if any, part of non-commercial licences do you find challenging to engage with for data shared with you?' (Q15), support this. One respondent noted that the classification of data used in academic or non-profit AI development is 'unclear'. Many echoed this sentiment, including one interviewee from the non-profit sector who said: *'There is a little bit of frustrating ambiguity with what "non-commercial use" means... If you were a non-profit and you developed a huge AI model that happened to be competitive... non-commercial is subjective in these cases.'"* This suggests that when academic or non-profit data usage is later commercialised, some questions may emerge about whether, and to what extent, such data usage continues to fall within non-commercial use contractual limitations.

Some Creative Commons licences have non-commercial use restrictions but no clear definition of what 'commercial' or 'non-commercial' usage entails. This can lead to confusion.[82] One survey respondent confirmed this, as their response to the aforementioned question (Q15) listed exclusively Creative Commons licences (CC-BY-NC, CC-BY-ND-NC, CC-BY-SA-NC) as examples of agreements where defining non-commercial use is challenging.

To address these challenges, we identified interest in developing data licence agreements that provide clearer definitions of 'non-commercial' use: *'What kind of categories do you think kind of fit in that non-commercial category? Because we're really keen to kind of break that down.'* Several interviewees stated that specifying licensed non-commercial uses in more detail could reduce ambiguities and better facilitate data sharing.

c) New standard data licences could potentially help clarify rights emerging from AI and provide different options for contractually allocating liabilities

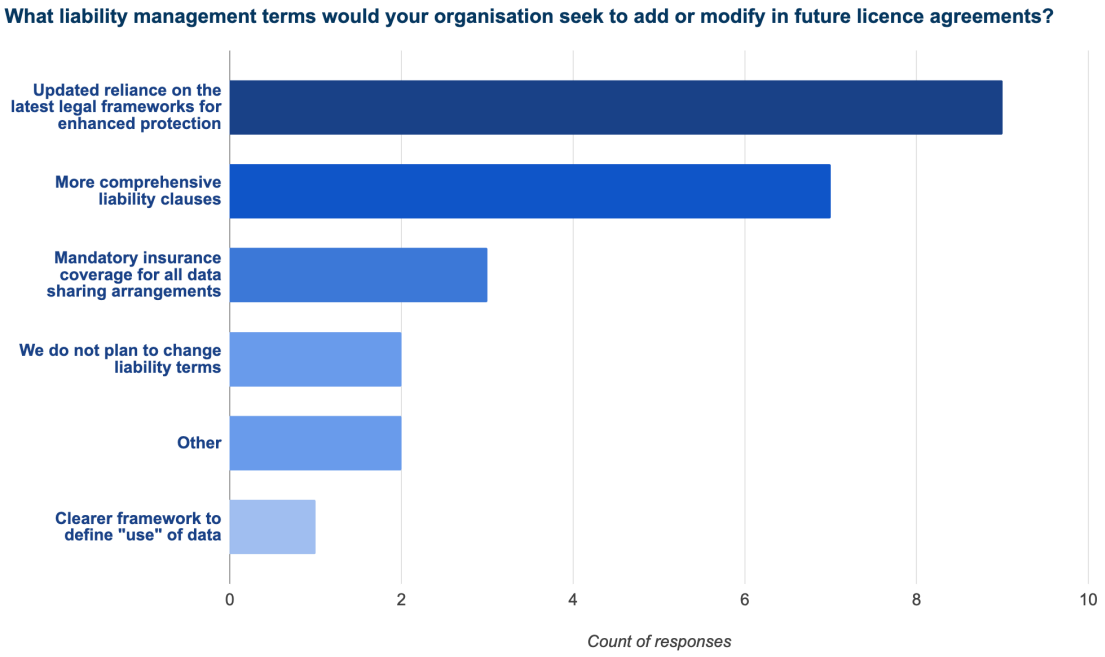
Contractually allocating liabilities is another critical aspect of data licensing in AI. New standard data licences can provide parties with different contractual options for managing liabilities, consistent with applicable laws. One interviewee from a non-profit organisation elaborated on their own efforts to develop a licence that gave companies some options to contractually reduce any liabilities associated with the data: *'We came up with [a licence], the idea was to put together a set of agreements specific to data... It was meant to give companies a clear line [to*

[82] Springer Nature (2022), 'For Open Data, think twice before applying Non-Commercial conditions'.

reduce] any liabilities or warranties they might be passing on." However, as described in **Section 1c**, clarity over licence terms can also help to mitigate inequities in bargaining power between high- and low-resource entities. This implies that clear and varied options for contractually allocating liabilities can help to both reduce liabilities and reduce uncertainty for smaller entities entering into such agreements.

Our survey also highlighted interest in having standard contractual options to help manage contractual liabilities. Survey responses to the question, 'What liability management terms would your organisation seek to add or modify in future licence agreements?' (Q35), demonstrate that data holders want *'more comprehensive liability clauses'* and *'updated reliance on the latest legal frameworks for enhanced protection'.* By taking these aspects into consideration, new standard licence agreements could help provide contractual options for allocating liabilities in a sometimes uncertain legal landscape. These insights are shown in **Figure 5.**

Figure 5. Data holders' preferences for future liability management terms, n = 16



Challenges also arise in allocating usage rights not only to the licensed data used to train or fine-tune AI models, but also to the trained AI models and the AI model outputs. One interviewee described this gap: *'We started off with just using Creative Commons, that didn't work very well for a number of reasons [...] the number one question was "what are the continuing rights or obligations as you pass through each machine learning phase and if you're many different licences to the data?".'* This is unsurprising since, as noted earlier in **Section 2a** and **2b** and later in **2d**, Creative Commons licences were not drafted with today's AI systems in mind.

To help address this situation, organisations are developing agreements that include terms that better reflect how AI systems work. Some ongoing efforts are mentioned in the earlier section 'Data licensing 101' under subsection 'Data licensing stakeholders, models and examples' above. An interviewee from the tech sector acquiring data to train AI models described their approach: *'We obtain [data] via contractual agreement. Those are more custom bespoke arrangements for our model training.'* A set of standard definitions, and a menu of contract terms, could provide parties with options for contractually allocating rights to licensed data, trained AI models and AI model outputs. The standard definitions could provide a common framework for developing a series of alternative standard contract clauses that parties can choose from to allocate rights in a way that aligns with their mutual intent.

d) New standard data licences can help address attribution issues arising under existing agreements

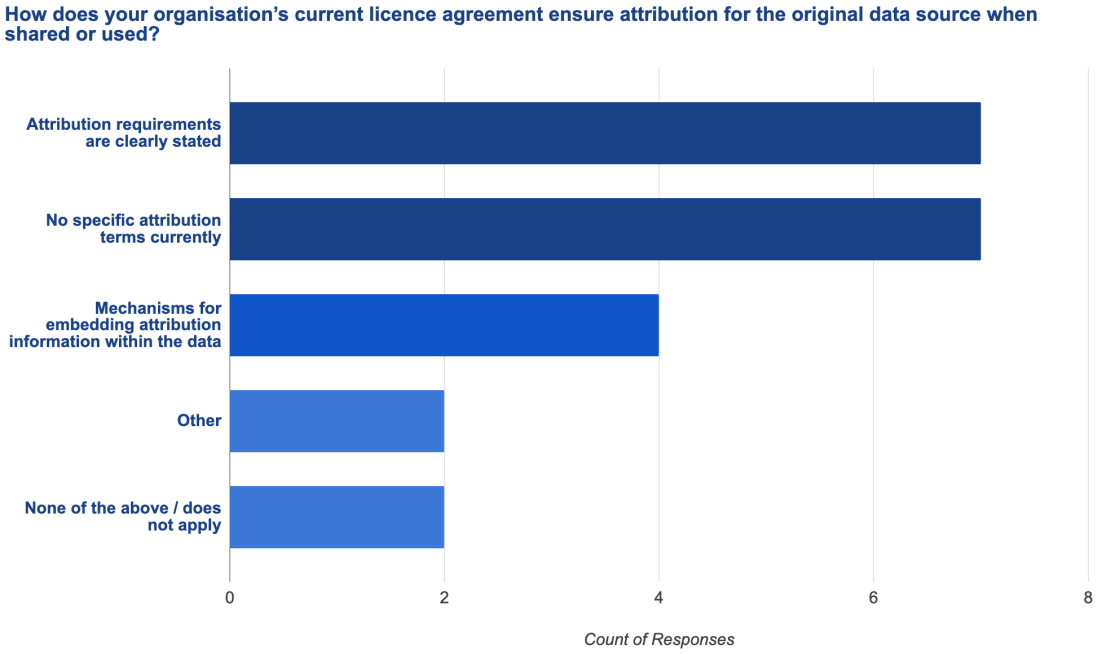
Some licences currently used for data sharing include attribution requirements, whereby users must provide credit to the original creators when using or distributing their work. However, interviewees mentioned that attribution in AI applications can pose challenges to data sharing for several reasons, especially with licences like Creative Commons (CC-BY) and ODC BY.[83]

Firstly, some challenges in applying attribution requirements stem from using licensed data for AI model training and not having the licensed data fully reproduced in AI model outputs. One technical expert from industry explained: *'The attribution requirement really doesn't make sense in the context of training... you're not reproducing the original work.'* We heard through numerous other interviews with other technical experts that some existing licences used for data are unsuitable for AI because often data used in training is not directly reproduced in outputs, making the contractual attribution requirements impractical to implement in these cases.

Also, several steps exist both before and after model training whereby the data can undergo various transformations and processing[84] making granular attribution throughout more difficult. Our survey responses to the question, *'How does your organization's current licence agreement ensure attribution for the original data source when shared or used?'* (Q30) were divided, with an equal number of respondents answering,*"attribution requirements are clearly stated' and 'no specific attribution terms currently'.* **Figure 6** indicates that even attribution terms related to data provenance (ie, typically defined during the early stages of the AI data lifecycle[85]) can be missing, or challenging to apply, a significant portion of the time.

[83] Open Mod (2018), 'Choosing an Open Data License - ODC BY vs. CC BY'.
[84] The ODI (2024), 'Understanding data governance in AI: A lifecycle perspective'.
[85] ibid.

Figure 6. Data Holders, Attribution Requirements, n = 18



However, some standard licences do not contain any attribution requirements. For instance, the Creative Commons CC0 1.0 Universal (CC0 1.0) public domain dedication licence[86] allows creators to waive all their copyright and related rights in their works to the fullest extent allowed by law. This means that works under CC0 can be freely used for any purpose, including AI training, without any attribution requirements. The qualitative survey responses confirmed that simpler licences like CC0[87] and the Open Government License (OGL)[88] are more usable for AI due to their *'minimal attribution requirements'*.

Nevertheless, this issue with the practicability of attribution in AI use cases is still important for two reasons. Firstly, calls from content creators to have proper credit assigned for their material used as training data for AI systems are ongoing, especially where this data may be subject to copyright.[89] [90] In other words, many content creators may not agree to have the CC0 licence govern their works, as its terms may not reflect how they want their works to be used. Secondly, licences that do include attribution clauses are still widely used. Many commonly used open licences, such as Creative Commons Attribution (CC BY) and Open Data Commons Attribution (ODC-BY), require attribution when the licensed material is used or shared.[91] [92]

---

[86] 'Deed – CC0 1.0 Universal – Creative Commons'.
[87] 'CC0 – Creative Commons'.
[88] National Archives n.d. 'Open Government Licence for public sector information'.
[89] Forbes (2024), 'Adobe Launches New AI Credit And Creator Attribution Tool'.
[90] Forbes (2022), 'Who Ultimately Owns Content Generated By ChatGPT And Other AI Platforms?'.
[91] CreativeCommons n.d., 'About CC Licenses - Creative Commons'.
[92] ODC-By n.d., 'Open Data Commons Attribution License v1.0'.

Simplified and AI-specific standard licence agreements, which tailor attribution requirements for AI applications (when such terms are desired), can help reduce the discussed complexities in data sharing for AI model development, deployment and ongoing inference. Licences often used for AI data (**Figure 1**) frequently include contractual terms that are challenging to implement in the AI context, such as attribution clauses that do not align with how data is used in AI systems.[93]

These findings further underscore an interest in developing standard data licence agreements specifically designed for AI applications, accommodating the unique ways in which data is sourced, used and processed at different stages of the AI data lifecycle. As one interviewee states: *'I think that credit is the thing that almost everybody wants with their data if they are okay with it being used... Many models will use data to answer many different questions and many want to know this attribution information. And then of course you get into other areas where there's privacy concerns in medical use cases.'*

Moreover, a possible alternative to strategies like simple minimisation of attribution terms (e.g., via the adoption of licences like the aforementioned CC0 and OGL) is the incorporation of technical solutions enabling automatic attribution. In **Figure 6**, we see that a notable number of respondents selected 'Mechanisms for embedding attribution information within the data' are currently in use. This general point was also brought up in the interviews: *'For me, we need more technical solutions... We are building metadata so we can ensure you're complying with all the licence obligations.'* Therefore, new standard data licences can potentially reduce ambiguity and foster more efficient data sharing by clarifying these points in an AI context. Also, licences for certain use cases may potentially benefit from being designed to interoperate with existing technical solutions in use for logging metadata to ensure proper attribution can be made.

More research is needed into machine learning (ML) architectures and methods that allow for such attribution to happen once the data has been used in training. For example, some types of xAI techniques such as SHAP (SHapley Additive exPlanations) can attribute model outputs to specific input features, potentially tracing model outputs back to training data.[94] These are additional considerations to be addressed in the global and multi-stakeholder process discussed in **Section 1c**.

[93] ARXIV (2023), 'The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI'.
[94] ScienceDirect (2023), 'Explainable, trustworthy, and ethical machine learning for healthcare: A survey'.

3) New standard data licences should embrace simplicity while balancing standardisation with the need for some customisation

a) There is a pressing need for simplicity

In addition to being mindful of the nuances of distinct AI models, use cases, tasks and legal requirements, there is a need for simplicity in standard data licence terms, as a wide range of stakeholders depend on understanding these terms. As one interviewee noted: *'We have projects whose teams do not know much about licensing. So they need to learn what it means and what different types of licences mean. They can easily breach licences without even knowing.'*

Another interviewee echoed this concern: *'I think that health and bio are so far behind in this data licensing question that a lot of users don't fully understand the limitation of the licence... It's difficult because the first reaction to these licences is to assume that they're fine, but you don't always know what you can or can't do with the data. In practice, people often ignore what it means, or they don't even bother to check until it's too late and it becomes an issue.'* This suggests that the complexity of licence terms can leave users unaware of the limitations of their permitted data use, including in the health and biomedical research fields.

A lack of clarity in licence terms not only increases the risk of accidental non-compliance but also drives up transaction costs. One interviewee from the tech sector elaborated: *'The problem is the transaction cost; also, if people don't understand your licence, don't understand what you're trying to achieve...you get a problem.'* Several interviewees mentioned that organisations need to spend additional time and resources ensuring proper interpretation and adherence to licences if they are too complex.

As shown in **Figure 7**, data users rated the ease of understanding and implementing various licence agreements, highlighting potential challenges associated with complex terms. Here, we observe once again that the CDLA, ODC BY and Creative Commons licences, as well as custom licence agreements, pose particular difficulties for data holders in terms of comprehension and implementation. Although simpler licences like CC0 have no attribution requirements and are sometimes considered more usable for AI, survey responses reveal that understanding and applying CC0's full legal implications can still be challenging for some. This aligns with the broader issues noted earlier in Section 2a—2d regarding the use of Creative Commons and custom licence agreements for AI purposes.

Figure 7. Data user's rating of the statement *'The terms and conditions of licence agreements are often difficult to understand'* against licences used, n = 23

	Agree	Neutral	Disagree	N/A
Community Data License Agreement (CDLA)	3.9%	0.0%	11.8%	0.0%
Creative Commons (CC-BY)	2.0%	2.0%	7.8%	0.0%
Creative Commons (CC-BY-SA)	3.9%	2.0%	2.0%	0.0%
Creative Commons (CC0)	0.0%	2.0%	5.9%	0.0%
Custom license agreements by other parties	3.9%	0.0%	7.8%	0.0%
Custom license agreements by us	3.9%	0.0%	7.8%	0.0%
Not sure	0.0%	2.0%	0.0%	0.0%
Open Data Commons (ODC BY)	5.9%	0.0%	7.8%	0.0%
Open Data Commons Licenses (ODCL)	2.0%	0.0%	0.0%	0.0%
Open Database license (ODbl)	5.9%	0.0%	3.9%	0.0%
Open Government Licence	2.0%	0.0%	0.0%	0.0%
Open Government Licence (OGL)	0.0%	0.0%	2.0%	0.0%
We manage license agreements	0.0%	0.0%	2.0%	0.0%

It should be noted that some interviewees highlighted that CDLA Permissive version 1.0 was considered overly *'complex'* and *'legalistic'*, which led to the development of the simpler CDLA Permissive version 2.0 to better meet users' needs for clarity.

Overall, simplifying licence agreements reduces transaction costs and builds trust, especially when non-legal practitioners like technical experts need to review and implement these terms. Through the interviews, we found that initiatives like the Linux Foundation's CDLA Permissive 2.0 and Microsoft's DUA-OAI may provide useful examples of efforts to standardise AI data licences with clear, easily interpretable terms.[95] For instance, the DUA-OAI is part of a suite of Microsoft contracts aimed at reducing ambiguity and fostering responsible data sharing.[96] Feedback from the interviews indicates that such agreements are valued for their clarity and practicality, especially in environments where researchers and engineers must navigate complex legal requirements to advance AI development.

To help make standard licence agreements more usable, it also may be desirable to develop materials that raise awareness of the licences and explain their purpose and how they can be used. The Open Source Initiative and Creative Commons, for example, publish information about licences on their websites.[97]

[95] Variety (2024), 'How Licensing Models Can Be Used for AI Training Data'.
[96] Query Pod n.d.,'Data Use Agreement for Open AI Model Development (DUA-OAI) Goal Overview Contemplated use case'.
[97] 'Licenses – Open Source Initiative' and 'About CC Licenses - Creative Commons'.

b) New standard data licensing frameworks should balance the need for standardisation with the need for some customisation, given the many types of data, use cases, models, tasks and legal requirements

As discussed earlier in Section 2d, the need to handle data under various, use case-specific licences was a recurring theme in the interviews. For instance, one interviewee noted: *'We're seeing the need for more vertical-specific open datasets as organisations realise they don't have enough data for refining models. These use cases go beyond basic applications and may require specialised licences.'*

Moreover, many existing licences assume a homogeneous definition of 'use', which overlooks the complexities of AI development and deployment.[98] Data is used differently across the AI data lifecycle, including at the training, fine-tuning and inference stages.[99] Data licensed for tasks such as data collection and preprocessing are typically subject to more restrictive licensing terms due to the proprietary or sensitive nature of such data. In contrast, data used for downstream tasks, such as evaluation and deployment, tend to have less restrictive licence terms.

Many existing standard licence agreements do not distinguish between different data use cases and thus may not necessarily address stakeholder needs that can vary based on the context in which the data is used. A multi-stakeholder process could help develop standard licence agreements tailored for using data in different parts of the AI data lifecycle, as may be needed, to address stakeholder interest or concerns. Some standard licences could permit more narrow uses for data intended to be used upstream, while other standard data licences may allow for broader downstream uses.[100]

Interviews also indicate that the challenges of AI data licensing often vary depending on the stage of the data pipeline. As one interviewee mentioned: 'Different phases of the AI lifecycle—training, fine-tuning and inference—need distinct licensing terms. It's important to make sure licences cover the right portions of datasets, like training data versus test data.' The need for licensing terms for different tasks is also present in the support literature. Benjamin et al. (2019) note that upstream tasks like data collection often have stricter licences due to data sensitivity, while downstream tasks, such as evaluation and deployment, face fewer restrictions as the data is less sensitive or processed.[101]

---

[98] Arxiv (2019), 'Towards Standardization of Data Licenses: The Montreal Data License'.
[99] The ODI (2024), 'Understanding data governance in AI: A lifecycle perspective | The ODI'.
[100] ibid.
[101] Arxiv (2019), 'Towards Standardization of Data Licenses: The Montreal Data License'.

In response to these complexities, the demand for flexible licence agreements that are to some extent customisable is driven by the rapid evolution of AI technologies. While some interviewees expressed interest in having more standardised data licences to facilitate data transactions, they also underscored the need for some customisation: *'Having more standardized terms is good...you might even have a menu of standard options… Some folks have articulated, "Hey, I'm okay with accessing or getting the training portion of the dataset, but not the test portion".'*

c) Modular standard data licence terms (and/or a menu of standard data licence terms) can potentially strike a balance between standardisation and some customisation

A balance between standardisation and some customisation is important given the wide range of data types, use cases, AI models, tasks and legal requirements applicable to data and AI, which can vary across jurisdictions. As one interviewee stated: *'AI models are evolving quickly, and one licensing model doesn't fit all purposes anymore. We need modular licences that strike a balance between standard terms and task-specific provisions, accounting for different data types, tasks and industries.'*

A modular approach to licensing—where standard terms can be augmented with task-specific or sector-specific provisions—could help to ensure that licences meet the varying needs of different sectors and applications. It should be noted that various open-source[102] and Creative Commons[103] licenses do address diverse stakeholder needs for licensing software and copyrighted materials. However, as discussed in **Section 2a**, we find that they may not be well suited for AI data use.

During the interviews, multiple stakeholders involved in developing AI applications emphasised that existing licensing models often fall short of supporting the diverse requirements that AI applications demand. One interviewee said: *'We are offering customisation to our customers, making sure we're not in a place where one size fits all.'* Further co-creation through multi-stakeholder collaboration is warranted here to explore the best approach for balancing standardisation with the need for some customisation.

[102] JISA (2018), 'Organizing for openness: six models for developer involvement in hybrid OSS projects | Journal of Internet Services and Applications'.
[103] CreativeCommons n.d., 'Software: Free and Open-Source Code – Creative Commons'.

d) Standard definitions can help advance data licensing

Our research suggests that developing standard definitions for use across standard data licence agreements could help increase simplicity and reduce barriers to sharing data. As highlighted in GPAI (2022), the standardisation of licensing terms can enhance legal certainty and reduce transaction costs, facilitating more effective data-sharing practices across sectors and jurisdictions.[104]

Many interviewees indicated that using consistent terminology means that organisations can more easily navigate the complexities of choosing between, and/or complying with, different agreements (or agreement extensions for different use cases, tasks and legal requirements as previously described): *'Having standard terms...is helpful...it's a commonly understood shorthand that has no ambiguity.'*

This need for standard definitions was also echoed by a senior practitioner from the tech sector: *'I know that a lot of companies are working on specific language for licences... but at the moment, you have as many languages as you have lawyers.'* One interviewee from a commercial organisation emphasised that 'having consistent terminology' may ultimately lower transaction costs by allowing organisations to work with greater confidence across varied datasets.

4) A new standard data-licensing framework may help address other data-sharing challenges, including legal compliance, data and AI governance, and ethical considerations

a) Standard data licence agreements could help address other challenges in implementing voluntary, legally-compliant data sharing

Our research suggests that standard data licences can also help address other challenges for implementing voluntary and legally-compliant data sharing. Interviewees emphasised that 'machine-readable' metadata could streamline compliance with licence terms at scale: *'I would love to be in a world where limitations are machine-readable... they could be embedded in*

[104] GPAI (2022), 'GPAI IP Expert — Preliminary Report on Data and AI Model Licensing'.

the metadata and then when you get the datasets, [you know] exactly what you can use it for and what not to use it for.' However, the lack of consistency in current metadata practices makes it difficult to maintain consistent information, especially across platforms.[105]

As noted earlier, the work by Longpre et al. (2023) highlights the potential room for improvement in metadata practices reflecting data provenance, lineage, use, limitations and other categories supporting the responsible use of data.[106] These efforts are important to consider in the data licensing context because they can enable greater oversight and accountability in how data is used, for instance, throughout AI development and deployment pipelines.[107] Additionally, policymakers are increasingly imposing data governance requirements in the AI context.[108] For example, the EU AI Act introduces obligations related to data governance and transparency for certain AI systems.[109] [110] As standard contract terms are drafted, these legal requirements should be considered, alongside available data governance techniques.

Tools for tracking data provenance and lineage, can help to support a resolution to this inconsistency by automatically updating metadata across the development of AI applications.[111] [112] Our interviewees expressed interest in use of these tools to enhance tracking of dataset characteristics: *'It would be nicer... if that [metadata] was built into the tools in the systems that people were using... to track [the data].'* One example raised during the interviews was Cloudera Machine Learning, which uses Apache Atlas to automatically collect and visualise data lineage for ML workflows—from training data to model deployments.[113]

Furthermore, the range of data governance techniques continues to expand. For instance, industry and open-source communities are developing data governance frameworks to help document the provenance, lineage and usage of AI datasets. Examples include standardised formats like Google's Data Cards[114] and Microsoft's Datasheets for Datasets[115], which offer structured information on dataset creation, biases and intended use. Meanwhile,

---

[105] ibid.
[106] Arxiv (2023), 'The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI'.
[107] Opendatasoft (2024), 'What is metadata and why is it as important as the data itself?'.
[108] Ethics and Information Technology (2023), 'The landscape of data and AI documentation approaches in the European policy context'.
[109] EU Artificial Intelligence Act (2024), 'High-level summary of the AI Act | EU Artificial Intelligence Act'.
[110] EU Artificial Intelligence Act (2024), 'Article 53: Obligations for Providers of General-Purpose AI Models'.
[111] IBM (2024), 'What is Data Provenance?'.
[112] IBM (2024), 'What Is Data Lineage?'.
[113] Cloudera, n.d., 'Model governance using Apache Atlas'.
[114] ACM Digital Library (2022), 'Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI'.
[115] Arxiv (2021), 'Datasheets for Datasets'.

open-source tools such as the Data Nutrition Project[116] and ABOUT ML[117] provide quick insights into datasets, while academic contributions like the University of Oxford's capAI[118] can help to assess datasets for bias.

Designing standard data licences to interoperate with these technical solutions and frameworks was a consistent desire, particularly among the technical experts we interviewed. For instance: *'What we [need and] are trying to build at the moment is a technical solution where we can tag the data... You log all the metadata for your data licensing.'* Yet other interviewees outlined that the widespread implementation of standardised metadata tags remains a significant technical and practical challenge.

A new standard data licence framework presents an opportunity to promote responsible data sharing and governance, potentially by incorporating or referencing external frameworks like those described. One participant from the tech sector noted common transparency issues related to the use of data from intermediaries: *'The issues we see are that anyone will take a dataset, upload it to Hugging Face, and copy a licence they found on GitHub. But the GitHub licence might be for the code or the model, not the data, because no-one thought about this... Transparency is important to point these issues out.'* Therefore, establishing a new standard licensing agreement that references metadata documentation frameworks could help address transparency issues, promote responsible data sharing, and advance AI governance practices. These are additional considerations to be addressed in the global and multi-stakeholder process discussed in **Section 1c**.

b) Standard data licences may help address cross-border and other compliance challenges

The diversity of regulations (e.g., in domains like AI, data protection and cybersecurity) across jurisdictions continues to create substantial barriers to data sharing, and standard licence terms should aim to help reduce these barriers.[119] [120] One interviewee from the tech sector showed hesitation over whether they would be able to comply given the heavy toll on the required legal expertise: *'The fragmentation in general, whether regulatory or territorial, is a strong problem for a company… when you have to be an expert in 10, 15, 20 regulatory frameworks, it's obviously a difficulty.'*

---

[116] Data Nutrition, n.d., 'The Data Nutrition Project'.
[117] Partnership on AI, n.d., 'ABOUT ML'.
[118] IBM Research Trusted AI, n.d., 'AI Fairness 360'.
[119] LSE (2024), 'New data protection and privacy laws have changed the regulatory landscape for researchers in the Global North'. s
[120] GPAI (2022), 'GPAI IP Expert - Preliminary Report on Data and AI Model Licensing'.

Additionally, our survey gathered responses to the statement, 'Ensuring compliance with privacy laws is challenging.' Respondents operating on a global scale reported fewer difficulties, contributing **41%** of the total 'disagree' responses, while regional respondents overwhelmingly agreed, making up **86%** of the 'agree' responses (**Figure 8**).

Figure 8. Data users' responses to the statement *"Ensuring compliance with privacy laws is challenging,"* by licences used, n = 23

	Agree	Neutral	Disagree	N/A
Africa (e.g. Uganda, Egypt, South Africa)	6.7%	0.0%	3.3%	0.0%
Central America and Caribbean (e.g. Costa Rica, Panama, Dominican Republic)	3.3%	0.0%	0.0%	0.0%
East Asia (e.g. China, Japan, India)	3.3%	0.0%	3.3%	0.0%
Europe (e.g. UK, Germany, France)	10.0%	0.0%	6.7%	0.0%
Global (all regions)	6.7%	6.7%	16.7%	3.3%
Middle East (e.g. UAE, Qatar, Saudia Arabia)	3.3%	0.0%	0.0%	0.0%
North America (e.g. United States, Canada, Mexico)	6.7%	0.0%	10.0%	0.0%
Oceania (e.g. Australia, New Zealand, Papua New Guinea)	6.7%	0.0%	0.0%	0.0%
South America (e.g. Brazil, Argentina, Peru)	3.3%	0.0%	0.0%	0.0%

Furthermore, consideration should be given to whether and how standard contractual terms, or a menu of standard modular data licence terms, can help support legally compliant data sharing. For instance, organisations operating in the EU under GDPR often need to specify their legal basis for processing personal data, which may involve terms set out under 'contracts', including standardised contract clauses that facilitate compliance.[121] Data security is an important consideration too.[122] Therefore, standard data licence terms should be prepared with these legal matters in mind.

[121] EDPB n.d., 'Data protection guide for small businesses – Process personal data lawfully'.
[122] European Commission n.d., 'What data can we process and under which conditions?'.

c) Standard data licences may help advance ethical guidelines for data usage

Ethical guidelines can help to support more accountable and sustainable data-sharing practices that benefit all stakeholders.[123] While ethical guidance was important for our interviewees, the majority expressed a preference for keeping it distinct from the licence terms for simplicity: *'One of the biggest ones that comes up is what should we do around the ethical privacy side of things? This comes up all the time... What our ecosystem has said in the overwhelming feedback has been: keep that separate from the data licensing agreement… That's a layer on top of the licence agreement.'*

Several interviewees highlighted the need to resolve issues related to AI models trained on sensitive datasets, such as those used in health and biomedical research. One interviewee explained: *'We are starting to become more user-oriented… in terms of training and fine-tuning. We're just starting to see folks making use-case requests [for health and biomedical research], and we are confronted with the question of how to share the models that come from that data, especially when the data is… not fully in the open. So, do the models undergo the same kind of restrictions as the original datasets?'* A set of ethical guidelines may therefore help clarify desired practices for such high-risk AI applications involving vulnerable populations or sensitive domains such as healthcare.

Our survey respondents hinted at existing strategies to answer the question, *'What measures does your organisation implement to ensure ethical data use?'* (Q28), highlighting ethical data sourcing and use as critical priorities for data holders. Respondents emphasised the need for clearly defined terms of use, with practices such as posting *'terms of use'* on websites and limiting data access only to reputable entities. One organisation noted that they *'assess each client's track record'* before entering into data-sharing agreements, not offering data services to all users.

Internal policies were also frequently mentioned during the interviews and survey responses as key resources for ensuring ethical practices, with one survey respondent requiring regular legal review of all policies for adherence to ethical standards (Q28). At the same time, some organisations flagged a lack of clear guidance for data users on what ethical practices look like. Interviewees revealed the practical challenges of enforcing ethical data-sharing measures, noting that embedding them in licences can complicate compliance. One interviewee discussed how differing ethical perspectives across regions and legal frameworks add complexity: *'The distribution of what is considered ethical is very broad... California and Texas have wildly different views, and if you extrapolate that globally, it becomes even harder.'*

---

[123] Global Partnership for Sustainable Development Goals (2024), 'Effective and Ethical Data Sharing at Scale'.

One possible solution to this challenge is the development of an ethical code of conduct that exists separately from the licence and would be flexible and adaptable across contexts: *'What our ecosystem has said in the overwhelming feedback has been to keep [ethical considerations] separate from the data licensing agreement... [It's] a challenge because ethics and privacy vary from jurisdiction to jurisdiction... The suggestion is to make that part of a collaborative framework or a code of conduct that everyone can adhere to.'* The ethical code potentially could be referenced in the standard licence terms. However, some concern was expressed about incorporating codes of conduct or standards, which could change over time, into standard data licences, as this could increase uncertainty in the contractual terms and costs of compliance. These considerations should be addressed in the global and multi-stakeholder process discussed in **Section 1c** for developing standard data licence terms.

As with previous findings, a multi-stakeholder effort to develop this code of conduct is key. Engaging stakeholders from multiple jurisdictions will help to ensure that a set of common ethical principles is as applicable as possible across diverse contexts. Examples of such codes include the Data Science Association's Code of Professional Conduct and IEEE's Ethically Aligned Design principles. The Data Science Association's code emphasises confidentiality, data security and responsible data use, stressing the protection of confidential information from creation to disposal,[124] while IEEE's framework focuses on human rights, wellbeing, data agency and transparency, offering guidance for ethical AI development and data use.[125] Each of these efforts engaged global experts in their development.[126] [127]

[124] Data Science Association n.d., 'Code of Conduct'.
[125] IEEE (2019), 'ETHICALLY ALIGNED DESIGN'.
[126] University of Cambridge (2019), 'IEEE Ethically Aligned Design published'.
[127] International Journal of Data Science (2015) 'The professionalisation of data science'.

Concluding remarks

In conclusion, this report underscores the critical role that standardised data licences could play in advancing responsible data-sharing practices to support AI development. The increasing demand for data to fuel AI, coupled with the existing barriers stemming from a patchwork of licensing frameworks, emphasises the need for accessible and clear licensing standards. Standard data licences can foster transparency, reduce legal ambiguities and potentially lower the cost and complexity of data sharing, which is particularly beneficial for smaller or under-resourced organisations and entities within underrepresented regions.

Moreover, these licences, developed through multi-stakeholder collaboration, could support equitable access to data by addressing the unique challenges of various sectors, such as non-commercial, research, and social good-oriented uses of AI. A balanced approach that combines standardisation with some flexibility for some customisation can help ensure that data sharing aligns with diverse legal, ethical and governance considerations across global jurisdictions. As AI continues to grow as an impactful force, implementing widely-accepted data licensing frameworks will be essential for building a fair, inclusive and sustainable data ecosystem that benefits all stakeholders.

Annex A: Methodology

The primary research questions guiding this study is:

What are the motivations, challenges and best practices for data sharing within the AI ecosystem?

How can data licensing practices be optimised to support these activities?

We aim to uncover the underlying factors that drive organisations to participate in data sharing, identify the obstacles they face, and explore how data licensing frameworks can be structured to enhance data accessibility and collaboration across various sectors. Based on our insights, we provide actionable recommendations to inform the development of effective data licensing strategies, thereby fostering a more open and efficient data-sharing environment within the AI community.

Research design

This research employed a mixed-methods approach, integrating qualitative and quantitative techniques to thoroughly explore data-sharing motivations, practices and challenges in the AI domain. The design was structured around two key components: the survey and interviews.

Survey

Survey design and deployment

The survey was designed to capture a wide array of perspectives on data-sharing practices, drawing insights from the background literature and initial expert consultations. Developed using Microsoft Forms for its user-friendly interface and robust data collection capabilities, the survey included both closed and open-ended questions to support the triangulation approach. Key topics addressed in the survey included motivations for data sharing, perceived benefits and challenges, types of data shared, and the impact of data licensing on sharing practices.

The survey was pre-tested with a small group of stakeholders to identify potential ambiguities and technical issues. Based on their feedback, the survey was refined for clarity and effectiveness. The final survey was disseminated through multiple channels, including direct emails to expert practitioners and professional networks such as ODI, Duke and MLCommons, as well as public postings on social media. This strategy ensured broad reach and diverse participation, ultimately yielding 51 responses, including 23 data users, 17 data holders and 11 data intermediaries.

Survey data analysis

The collected survey data was subjected to both quantitative and qualitative analyses. Quantitative data was analysed using descriptive statistics to identify key trends, patterns and correlations. Qualitative responses underwent inductive thematic analysis, extracting meaningful insights that highlighted challenges such as the definition and operationalisation of non-commercial use, the complexities of cross-border data sharing, the need for ethical data use policies, and the growing demand for modern, adaptable licence agreements.

Interviews

Interview design and collection

Following the survey analysis, semi-structured interviews were conducted to delve deeper into the themes identified. The interview questions were developed by analysing the preliminary survey responses, focusing on notable trends and areas of disagreement among respondents. Key themes explored in the interviews included the high costs associated with using licences, the challenges of cross-border data sharing, and the demand for transparency in data sourcing.

In-depth interviews were conducted with 15 stakeholders from seven different sectors, ensuring representation from various geographical regions, including the global South. Interviewees were selected based on their organisational affiliations, relevance to the AI data ecosystem, and their roles as data holders, users, platforms, or intermediaries. This selection process aimed to ensure a comprehensive understanding of the diverse perspectives within the AI data-sharing ecosystem.

Interview data analysis

The interview transcripts were analysed using thematic analysis, which involved coding the data to identify recurring themes and patterns. These were then organised into broader categories, providing nuanced insights into the technical and behavioural factors underpinning data-sharing practices. This analysis offered a deeper understanding of the motivations and challenges associated with data sharing, complementing the findings from the survey.

Ethical considerations

All research activities adhered to strict ethical guidelines. Participants provided informed consent, and their data was anonymised to protect confidentiality. Ethics approval was secured from Duke University (**protocol: 2024-0348**) for its portion of the research, with equivalent procedures followed by the ODI as required.

Data sharing and security

Data collected during the project was shared securely between the ODI and Duke teams, with a commitment to full deletion of sensitive data, including personally identifiable information (PII), after the project term.

Annex B: Survey content

Section	#	Question	Help text	Question type	Response options
	Intro				This questionnaire is part of a research project conducted by Duke University's Pratt School of Engineering and the Open Data Institute. The project aims to understand the current and desired licence agreements facilitating data sharing for AI. Specifically, it seeks to understand how licensing and other agreements are, or could be, structured to create incentives for voluntary data sharing to advance social good, such as assisting organisations desiring to develop and implement AI applications that benefit society, or to achieve the United Nations' Sustainable Development Goals. This survey also seeks to understand concerns that entities may have about sharing their data. More specifically, the survey inquires about the types of contractual provisions that might deter parties from sharing data. The survey seeks perspectives from different organisations in the data ecosystem, including data contributors, data users, and data hosting platforms; and within those, from many different stakeholders, including the research community, industry (including creative industry), non-profits, academia, and the government sector. The goal is to analyze the effectiveness of current agreements for sharing data to advance social good, to identify potential improvements to these agreements, and ultimately develop and make available common templates for data licences for AI. In this survey, we use terms relating to contractual arrangements supporting the sharing of data (such as 'data licences', 'licence agreements', or 'data-sharing agreements'), to refer to any formal agreements or permissions that govern the access, use and distribution of data between parties. We use these terms interchangeably, regardless of whether the data in question is subject to intellectual property and/or other proprietary rights. We are seeking responses from decision-makers responsible for the collection, use

					and management of data.
1 - Organisation Profile	Intro		The following questions concern general information about your organisation and their use of data.		
1 - Organisation Profile	1	Please select the role or function(s) within your organisation that best matches the way that you personally work with data.	Please select all that apply.	Multi choice	Data management (managing or overseeing data) Policymaking (creating, influencing or implementing policies related to data) Research (conducting research that involves the use of data) Technical development (technical aspects of data processing, analysis or infrastructure) Legal (legal services for the organization's use of data) None of the above Other: Please specify: _____
1 - Organisation Profile	2	What is your organisation's role in the data ecosystem?	Please select all that apply.	Multi choice	Data User: Your organisation uses a significant amount of data provided by other organisations. Data Contributor: Your organisation makes a significant amount of data available to other organizations. Hosting Platform: Your organisation provides a platform for enabling other organizations to share data. None of the above Other: Please specify: _____

1 - Organisation Profile	3	What types of data does your organisation share, have shared with it, or host on behalf of others for the purpose of sharing?	Please select all that apply.	Multi choice	Personal data (e.g. names, contact details or other information that identify a person) Sales data (e.g. transactions, customer interactions) Sensor data (e.g. IoT devices, environmental monitors) Health data (e.g. patient records, clinical trials) Financial data (e.g. banking records, investment portfolios) Operational data (e.g. logistics, supply chain information) Public data (e.g. government datasets, public records) Web data (e.g. scraped datasets) None of the above Other: Please specify: _____
1 - Organisation Profile	4	How, if at all, can the data you contribute, use, or host, be used to advance social good?	'Social good' refers to actions, policies or initiatives that benefit the largest of people in the most significant possible way; this can encompass a wide range of issues such as healthcare, education, community development, environmental sustainability and economic empowerment.	Long text	
2 - Data Users	Intro		The following questions concern your organisation's ecosystem role as a data user.		

			Please note that there will be a separate set of questions about data intermediaries.		
2 - Data Users	1	For which purposes does your organisation use data that is shared with it?	Please select all that apply.	Multi choice	Improving products/services Decision-making Customer insights Research and development Compliance and reporting Efficiency and optimization Risk management None of the above Other: Please specify: _____
2 - Data Users	2	Approximately how many different sets of licence agreement terms does your organization use to license data?	Please select the option that best describes the data you use.	Single choice	None 1-5 6-10 More than 10 Unsure / prefer not to answer
2 - Data Users	3	Which part of your development pipeline do you use shared data in?	Please select all that apply.	Multi choice	Planning (incl. problem definition, stakeholder engagement, etc) Data collection/creation Data exploration/analysis Data pre-processing (incl. cleaning, transformation, etc) Training of AI/ML models Evaluation (incl. validation, test, etc) Deployment (incl. post-deployment tasks like ongoing monitoring, etc) None of the above Other: Please specify: _____

2 - Data Users	4	Does your organisation obtain data from any of the following sources	Please select all that apply.	Multi choice	Common Crawl LAION Eleuther.ai ML Commons HuggingFace Kaggle None of the above Other: Please specify: _____
2 - Data Users	5	How impactful are licence agreements for the development, supply or marketing of your goods and services?	Please select the option that fits best.	Single choice	Very impactful A little impactful Not impactful None of the above Other: Please specify: _____
2 - Data Users	6	Do any of these specific licence agreements suit your organisation's needs?	Please select all that apply.	Multi choice	Creative Commons (CC-BY-SA) Creative Commons (CC-BY) Creative Commons (CC0) Open Data Commons (ODC BY) Open Database licence (ODbl) Community Data License Agreement (CDLA) Custom licence agreements by us Custom licence agreements by other parties None of the above Other standard licence. Please specify: _____

2 - Data Users	7	To which degree does your organisation agree with the following statements regarding challenges your organisation may have faced with existing licence agreements.	Please rate each statement on a scale from 1 (strongly disagree) to 5 (strongly agree)	Please rate each statement on a scale from 1 (strongly disagree) to 5 (strongly agree)	The terms and conditions of licence agreements are often difficult to understand. Current licence agreements do not accommodate our organisation's specific needs or use cases well. Ensuring compliance with and enforcing licence agreements is challenging. Ensuring compliance with, and enforcing, privacy laws is challenging. Ensuring compliance with and enforcing other legal requirements is challenging. The costs associated with licence agreements are a significant barrier to accessing or using the data we need. Existing licence agreements do not provide adequate protection for our intellectual property or proprietary information. The terms and conditions of licence agreements do not impact our use of data. The differences between the data licences we could use are difficult to understand. 'Non-commercial' use is clearly defined. Licences are useful to ensure that the data is suitable for the organisation’s intended purposes.
2 - Data Users	8	In relation to data privacy, security and liability, which measures or considerations are prioritized by your organisation?	Please select all that apply.	Multi choice	Unauthorised access Data breaches Compliance with privacy laws Data misuse Liability for data errors Reputational damage Transparency in data collection and use Ethical sourcing of data Environmental impact of data storage and processing Equity and fairness in data usage Integration of privacy-by-design None of the above Other: Please specify: _____
2 - Data Users	9	Does your organisation ever encounter situations where the data you intend to use is shared under	Please select the option that fits best.	Single choice	Yes No I don't know

		non-commercial license agreements?			
2 - Data Users	10	Which of the following scenarios would your organisation consider to be ‘non-commercial’ ?	Please select all that apply.	Multi choice	Using data for internal testing to ensure product safety within a commercial organisation. Using data for internal testing to ensure product safety within a non-commercial organisation (incl. government). Using data for research purposes only within a commercial organisation. Using data for research purposes only within a non-commercial organisation (incl. government). Employing a testing benchmark internally within an organisation without direct profit generation within a for-profit organisation. Employing a testing benchmark internally within an organisation without direct profit generation within a non-profit organisation. Strictly non-profit activities. Activities not directly generating revenue. Educational or research purposes regardless of the institution's nature. None of the above Other: Please specify: _____
2 - Data Users	11	Which, if any, part of non-commercial licences do you find challenging to engage with for data shared with you?	Please elaborate on any relevant aspects.	Long text	
2 - Data Users	12	How, if at all, do you typically address rights and/or ownership for AI models developed using shared data?	Please select the option that fits best.	Single choice	We retain ownership of the AI models. Ownership is shared between us and others. Others retain ownership. AI models are in the public domain. None of the above Other: Please specify: _____

2 - Data Users	13	How, if at all, do you typically address rights and/or ownership for the outputs generated by the AI models developed with shared data?	Please select the option that fits best.	Single choice	We retain ownership of the AI models. Ownership is shared between us and others. Others retain ownership. AI outputs are in the public domain. None of the above Other: Please specify: _____
3 - Data Contributors	Intro		The following questions concern your organisation's ecosystem role as a data contributor. Please note that there will be a separate set of questions about the perspectives of data users and intermediaries.		
3 - Data Contributors	1	What licences does your organisation currently use to share data?	Please select all that apply.	Multi choice	Creative Commons (CC-BY-SA) Creative Commons (CC-BY) Creative Commons (CC0) Open Data Commons (ODC BY) Open Database licence (ODbl) Community Data License Agreement (CDLA) Custom licence agreements by us Custom licence agreements by other parties None of the above Other standard licence. Please specify: _____

3 - Data Contributors	2	To what degree does your organisation agree with the following statements regarding licence agreements for data that you share?	Please rate each statement on a scale from 1 (strongly disagree) to 5 (strongly agree)	Likert cluster	Overall, I am currently satisfied with the licence agreements that I use The licence agreements offer the level of protection we need for data. The licence agreements increase transaction costs associated with data sharing. The licence agreements clearly and fairly allocate intellectual property rights. The licence agreements properly allocate liabilities and risks. The licence agreements are sufficiently enforceable. The licence agreements sufficiently address privacy and legal compliance. The licence agreements are flexible and adaptable to changing business needs. The licence agreements are easy to understand and implement. The licence agreements clearly define attribution requirements.
3 - Data Contributors	3	What, if any, specific challenges arise for your organisation through sharing data across international borders?	Please provide a brief overview.	Short Text	
3 - Data Contributors	4	How, if at all, do you license combined datasets?	Please select all that apply.	Multi choice	We do not combine any datasets. We do not have any policies governing these practices. We use standard form licence agreements. We use a customised licence agreement. We use various different licence agreements. Other: Please specify: _____
3 - Data Contributors	5	How does your organisation typically address rights and/or ownership for AI models developed using your organisation's	Please select the option that fits best.	Single choice	Our data is not used for AI model development Our organisation retains ownership of the AI models. Ownership is shared between our organisation and the developers of the AI models. The developers of the AI models retain ownership. AI models are in the public domain. Other: Please specify: _____

		data?			
3 - Data Contributors	6	How does your organisation typically address rights and/or ownership for the outputs generated by the AI models developed using your organisation's data?	Please select the option that fits best.	Single choice	Our data is not used for AI model development Our organisation retains the right to the outputs for any use. Rights are shared, allowing both our organization and others to use the outputs. Users of the AI model retain rights to the outputs. AI model outputs are in the public domain. Other: Please specify: _____
3 - Data Contributors	7	How concerned is your organization about the following scenarios:	Please rate each statement on a scale from 1 (not at all concerned) to 5 (very concerned).	Likert cluster	Privacy and security measures for the data shared by your organisation? Data scraping of your organisation's data by others without permission? Using scraped data within your organisation?
3 - Data Contributors	8	What, if any, measures does your organisation currently implement to ensure that your organisation’s data is used ethically and in compliance with agreed terms?	Please elaborate on any relevant aspects.	Long text	

3 - Data Contributors	9	Which cybersecurity measures are explicitly covered in your organisation's current licence agreements?	Please select all that apply.	Multi choice	We do not specify cybersecurity measures in our licence agreements. Compliance with specified cybersecurity standards. Requirement for security assessments or audits. Defined cybersecurity obligations and responsibilities. Provisions for ongoing monitoring of security practices. None of the above Other: Please specify: _____
3 - Data Contributors	10	How does your organisation’s current licence agreement ensure attribution for the original data source when shared or used?	Please select all that apply.	Multi choice	Attribution requirements are clearly stated. Mechanisms for embedding attribution information within the data. No specific attribution terms currently. None of the above Other: Please specify: _____
3 - Data Contributors	11	How does your organisation's current licence agreement address liability issues related to data misuse or unintended consequences?	Please select all that apply.	Multi choice	Inclusion of specific liability clauses Insurance coverage requirements for data-related risks Dependency on existing legal frameworks for liability coverage Unsure or not applicable None of the above Other: Please specify: _____
3 - Data Contributors	12	In your practice, do you differentiate between commercial and non-commercial uses?	Please select the option that fits best	Single choice	Yes Sometimes No Don't know

3 - Data Contributors	13	Which of the following scenarios would your organisation consider to be ‘non-commercial’ ?	Please select all that apply.	Multi choice	Using data for internal testing to ensure product safety within a for-profit organisation. Using data for internal testing to ensure product safety within a non-profit organisation. Employing a testing benchmark internally within an organization without direct profit generation within a for-profit organisation. Employing a testing benchmark internally within an organization without direct profit generation within a non-profit organisation. Strictly non-profit activities. Activities not directly generating revenue. Educational or research purposes regardless of the institution's nature. None of the above Other: Please specify: _____
3 - Data Contributors	14	Does your organisation have an interest in new, alternative license agreements for data sharing?	Please select the option that fits best.	Single choice	Yes Maybe No Don't know
3 - Data Contributors	15	What liability management terms does your organisation seek to add or modify in future licence agreements?	Please select all that apply.	Multi choice	We do not plan to change liability terms. More comprehensive liability clauses. Mandatory insurance coverage for all data-sharing arrangements. Updated reliance on the latest legal frameworks for enhanced protection. Other: Please specify: _____
3 - Data Contributors	16	What cybersecurity measures does your organisation intend to include in future licence agreements?	Please select all that apply.	Multi choice	We do not plan to change cybersecurity terms. Stricter compliance with advanced cybersecurity standards. Mandatory and more frequent security assessments or audits. Enhanced cybersecurity obligations and detailed responsibilities. Stronger provisions for continuous monitoring and reporting. Other: Please specify: _____

3 - Data Contributors	17	What attribution terms does your organisation plan to incorporate in future?	Please select the option that fits best.	Single choice	We do not plan to change attribution terms. More detailed and stringent attribution requirements Technical solutions for automatic attribution within the data Other: Please specify: _____
3 - Data Contributors	18	What terms would your organisation like to incorporate in future to address the allocation of intellectual property and/or other proprietary rights?	Please select all that apply.	Multi choice	Planning (incl. problem definition, stakeholder engagement, etc) Data collection/creation Data exploration/analysis Data pre-processing (incl. cleaning, transformation, etc) Training of AI/ML models Evaluation (incl. validation, test, etc) Deployment (incl. post-deployment tasks like ongoing monitoring, etc) None of the above Other: Please specify: _____
3 - Data Contributors	19	What other features would you like to see in alternative licence agreements?	Please elaborate on any relevant aspects; your considerations might include terms concerning Privacy, Intellectual Property, Liability, Data usage governance, Duration, Transparency, Attribution and Cybersecurity	Long text	

4 - Data Intermediaries	Intro		The following questions concern your organisation's ecosystem role as a data-hosting platform. Please note that data contributor and user perspectives were already covered in earlier sections of this survey.		
4 - Data Intermediaries	1	What services does your platform provide for data sharing?	Please select all that apply.	Multi choice	Data storage and management Data sharing and distribution Support for AI development Analytical tools provision Community engagement and support Data security and compliance services Data integration and interoperability Data cleaning and preparation services Metadata management and documentation Audit trails and change management None of the above Other: Please specify: _____
4 - Data Intermediaries	2	How important are standard licence agreements for your organisation?	Please select the option that fits best.		Very important Somewhat important Neutral Not very important Not important at all

4 - Data Intermediaries	3	What are the primary challenges your users experience related to licence agreements for data sharing that your organisation facilitates?	Please select all that apply.	Multi choice	Balancing data acceptability with protection Crafting clear and understandable terms Ensuring compliance with diverse legal requirements Encouraging participation from data owners None of the above Other: Please specify: _____
4 - Data Intermediaries	4	What are the primary challenges that you experience as a data intermediary, which are related to licence agreements in data sharing that your organisation facilitates?	Please select all that apply.	Multi choice	Balancing data acceptability with protection Crafting clear and understandable terms Ensuring compliance with diverse legal requirements Encouraging participation from data owners None of the above Other: Please specify: _____
4 - Data Intermediaries	5	How does your organisation implement and enforce licence agreements to ensure responsible data sharing and use?	Please select all that apply.	Multi choice	Automated licensing to identify breaches Third-party reporting of alleged breaches Monitoring for breaches and enforcement decisions We do not have a process / policies for implementation Staff training and audits None of the above Other: Please specify: _____
4 - Data Intermediaries	6	What measures does your organisation include in licence agreements to	Please select all that apply.	Multi choice	We do not accept liability for third-party data We do not accept liability for actions of downstream data users Other: Please specify: _____

		address concerns of liability?			
1 - Organisation Profile			Nearly there! We just need some details about the organisation you responded for, to wrap things up.		
1 - Organisation Profile	5	Please select the type of organisation you responded for.	Please select the option that fits best.	Single choice	For-profit Non-profit Academic institution Government Supranational (e.g. EU, UN) None of the above Other: Please specify: _____

1 - Organisation Profile	6	Please select the sector your organisation operates in.	Please select all options that apply.	Multi choice	Healthcare Environment Agriculture Transportation Financial services Legal services Manufacturing Consumer products Food and beverages Creative industries Education Energy Information technology Telecommunications Construction Hospitality Public administration Utilities None of the above Other: Please specify: _____
1 - Organisation Profile	7	How many offices does your organization have?	Please select the option that fits best.	Single choice	0-3 4-10 11+
1 - Organisation Profile	8	How many employees does your organization have?	Please select the option that fits best.	Single choice	0-50 50-200 200-1000 1000+

1 - Organisation Profile	9	Where does your organisation provide goods and/or services geographically?	Please select all that apply.	Multi choice	North America (eg, United States, Canada, Mexico) Central America and Caribbean (eg, Costa Rica, Panama, Dominican Republic) South America (eg, Brazil, Argentina, Peru) Europe (eg, UK, Germany, France) Asia (eg, China, Japan, India) Middle East (eg, UAE, Qatar, Saudi Arabia) Africa (eg, Uganda, Egypt, South Africa) Oceania (eg, Australia, New Zealand, Papua New Guinea) Global (all regions) Other: Please specify: _____
5 - Conclusion	1	Would you be happy to join an interview with us to provide further insight into the use of data licenses?	Please select the option that fits best.	Single choice	Yes No
5 - Conclusion	2	Please provide your email address - this will only be used to contact you about an interview. Alternatively, you can send an email to thomas.careywilson@theodi.org	Please provide a brief overview.	Short text	