Policy Brief
Issue n°01/25 | January 2025

Using data as a production factor: policy ideas for a new EU data strategy

Bertin Martens

Bertin Martens (bertin.martens@bruegel.org) is a Senior Fellow at Bruegel

Executive summary

A MAJOR TASK for the 2024-2029 European Commission will be to reconcile and simplify the European Union's range of data-market laws into a more coherent framework. At the heart of the approach should be the non-rival nature of data, meaning it can be used by many parties for many purposes at the same time. That creates the potential for economic efficiency gains from the re-use and aggregation of data. Exclusive private data control rights and high transaction costs may stand in the way of realising these gains.

WE DEFINE ECONOMIC criteria against which existing data market regulations can be evaluated. These criteria lead to the following recommendations to improve existing EU data regulations:

1. The EU general data protection regulation facilitates re-use of personal data. Machine-readable consent notices and real-time data transfers could reduce high transaction costs that prevent meaningful exercise of informed consent.
2. The European Health Data Space contains an almost-ideal governance regime for health-data re-use and pooling that maximises incentives for data-driven innovation. This regime should be applied to other industrial data-space initiatives.
3. The Data Act facilitates access and re-use of product data but exclusive licensing rights for data holders, monopolistic pricing of third-party data transfers and other anti-competitive measures reduce its impact. Widening its application to services data would make it a truly horizontal data-market regulation.
4. The Digital Markets Act contains several obligations for platforms to grant users access to their own data. Widening access to networked data would facilitate competition between platforms and their users. Mutual instead of unilateral sharing of search-engine data would avoid fragmentation of the search-data pool.
5. The Artificial Intelligence Act imposes unwarranted and costly restrictions on the re-use of copyright-related content data for AI model training data, reducing the innovative impact of AI. The text and data mining exception in the EU Copyright Directive could be broadened to address these new AI technology challenges.

EU DATA MARKET regulations exhibit a tension between exclusive private rights and the realisation of the wider societal value of data. Finding an acceptable balance may involve some redistribution of the efficiency gains between data users and the original data collectors.

Recommended citation
Martens, B. (2024) 'Using data as a production factor: policy ideas for a new EU data strategy', *Policy Brief* 01/2025, Bruegel

1 Introduction

The rapidly growing volume and economic importance of digital data has led European Union policymakers to adopt several data market regulations. Major laws include the general data protection regulation (GDPR, finalised in 2016), the Data Act (2023), the Digital Markets Act (2022), the Artificial Intelligence Act (2024) and the Regulation on the European Health Data Space (agreed in 2024)[1]. All these regulations seek to open up access to data that is locked up in technical silos, to facilitate the emergence of data markets and to stimulate the development of innovative data-driven services.

The sheer number of EU data laws leads to fragmentation, increases compliance costs and may result in inconsistencies

The move to make more efficient use of data and leverage its power as a production factor (Beraja and Yuchtman, 2024), similar to labour and capital, is very welcome (as argued by Martens, 2024a). However, the sheer number of EU data laws leads to regulatory fragmentation, increases compliance costs and may result in inconsistencies between regulations (Martens, 2023a). Rules are often precautionary and not as innovation-focused as they could be. That reduces their impact on data markets and data-driven services in the EU. On artificial intelligence, for example, Draghi (2024) observed that EU restrictions on the storing and processing of data "*create high compliance costs and hinder the creation of large, integrated data sets for training AI models. This fragmentation puts EU companies at a disadvantage relative to the US, which relies on the private sector to build vast data sets, and China, which can leverage its central institutions for data aggregation*".

The view of data as a production factor that drives international competitiveness is gaining traction globally (Diebold, 2023). Bradford (2023) compared the laissez-faire data regime in the United States with China's centralised regime and the EU's somewhere-in-the-middle regime, which has a mixture of private rights and some data-sharing obligations. Regime choices are inspired by domestic political and ideological choices but of course have economic implications. The European Commission is increasingly aware of this. Henna Virkkunen, Commission Vice President for Tech Sovereignty, Security and Democracy, has been given the task of improving EU data market policies and proposing "*a European Data Union Strategy drawing on existing data rules to ensure a simplified and coherent framework to share data seamlessly*"[2].

How should the EU go about designing a data-market regime that maximises the efficient use of data as an economic production factor? Should it continue with the ad-hoc approach, designing specific regulations for specific sectors and issues? Is the variation in data-access rights and conditions across EU data regulations justified by sectoral differences in data-market failures? Or is there scope to harmonise, simplify and generalise some of these, as the task given to Virkkunen suggests?

This Policy Brief examines these questions. We start by examining the economic characteristics of data and what distinguishes it from other production factors. Well-defined exclusive private property rights are important to make markets for physical goods, machines, land and labour work efficiently. Physical goods are rival: they can only be used by one party for one purpose at a time. Non-exclusive rights for physical goods would not be a good idea because they would create conflict over their use. Data however is non-rival: it can be used by many parties for many purposes at the same time. Non-rivalry generates externalities or spillovers – value that can potentially be generated by parties other than those involved in the original data collection. A regime that maximises the value of data should enable multiple

1 Respectively, Regulation (EU) 2016/679, Regulation (EU) 2023/2854, Regulation (EU) 2022/1925 and Regulation (EU) 2024/1689. The European Health Data Space was agreed in April 2024 but the final version of the regulation has at time of writing not been published. See European Commission press release of 24 April 2024, 'Commission welcomes European Parliament's adoption of the European Health Data Space and regulation on substances of human origin', https://ec.europa.eu/commission/presscorner/detail/en/ip_24_2250.

2 See Ursula von der Leyen, Mission Letter to Henna Virkkunen, 17 September 2024, https://commission.europa.eu/document/3b537594-9264-4249-a912-5b102b7b49a3_en.

uses and externalities. Exclusive property or control by a single party is not a good solution in that case. Unless the controlling party can benefit from these externalities, it has no incentive to share data. That limits the value that society as a whole can extract from the data.

Private rights cannot be excluded completely however. Data collectors need to earn a return on their investment in data-collection costs. However, the marginal cost of data collection is often close to zero, especially when data is a by-product of ongoing activities and transactions. In that case, re-use by others and for other purposes is not a disincentive for the original data collection. Moreover, data is usually co-produced between at least two parties, for example buyers and sellers. Both parties may have claims to block or limit access to and use of the data by others. The EU GDPR, for example, grants some rights to personal-data collectors and other rights to the natural persons whose behavioural data is collected. On the other hand, the EU Data Act grants some exclusive rights to data holders to charge a price for third-party access that may block data markets. Finding an appropriate balance between private rights and the wider benefits for society is a difficult exercise.

Finding a balance between private rights and the wider benefits for society of date re-use is a difficult exercise

In line with the EU's own better regulation guidelines (European Commission, 2023), we look at data markets from a wider social welfare perspective. How can we achieve more efficient data-driven product and services markets? In section 2, we introduce two economic criteria to assess the potential efficiency gains or externality benefits from non-rival data: economies of scope from the re-use and the aggregation of data. Transaction costs can limit efficiency gains. If private operators are unable to harvest externality benefits, a market failure arises that justifies regulatory intervention in data markets.

Section 3 applies these criteria to the GDPR, the Data Act, the European Health Data Space (EHDS), the Digital Markets Act (DMA) and the AI Act. The latter two are not data regulations per se but contain provisions that affect data markets. We explore what can be done to improve the efficiency of these regulations in terms of exploiting potential data externalities. For the GDPR, which entered into force in 2018, there is considerable empirical evidence on its opportunity costs. Other regulations are not yet in force and therefore cannot generate empirical evidence. We examine these from a more theoretical perspective and find good reasons to suspect continued data-market failures. We recommend policy changes to overcome these regulatory bottlenecks.

Section 4 attempts to generalise from these case studies. When the private value of data for the original data co-generators is lower than the value it can generate for society through re-use and aggregation, tension emerges between exclusive private rights and societal interests. Examples include privacy rights for natural persons and exclusive commercial rights to data, in the form of trade secrets and intellectual property rights. Finding an acceptable trade-off between these conflicting claims to data is a difficult political balancing act that may lead to calls for some redistribution of value from data re-users to the original data collectors. We conclude with a discussion of several redistribution options.

2 The efficient use of data as a production factor

Data non-rivalry generates two potential sources of economic benefits:

- Economies of scope from the re-use of data (Panzar and Willig, 1980; Teece, 1980): once collected, data can be re-used by many parties for many purposes at the same time. For example, the data that Google collects from search queries, data embedded in a bank account or collected by a car, can be re-used for other services and/or by other service providers, to offer complementary and competing services: advertising, payment services, car maintenance services. Re-use by others will not functionally impact the original use but may have an economic impact on the parties that co-generated the original data.
- Economies of scope in data aggregation (Bajari *et al*, 2019; Calzolari *et al*, 2021, Carballa *et al*, 2023): data from many different sources can be pooled and aggregated. Data collected by search engines, navigation apps and medical service providers becomes more valuable when aggregated across more users. Pooled data can reveal patterns and deliver service insights that cannot be extracted from fragmented datasets or individual data. For example, navigation services, social-media newsfeeds or search engine recommendations would not be feasible without data aggregation across users.

Teece (1980, 1982) pointed out that the existence of unrealised externalities indicates a failure in markets for complementary inputs required for the production of a service. For example, the holder of car navigation data may not have access to complementary data about hotels and restaurants and is therefore not in a position to offer drivers additional travel services. Collaboration might be tried with a firm that has this additional data but strategic behaviour makes contracting difficult (Schulze *et al*, 2006), especially when there are significant differences between firms in terms of market power. The data collector may fear that the data will be used against their interests. As a result, data-market failures persist and may require regulatory intervention.

In some cases, markets can overcome obstacles to data re-use and aggregation. For example, Google Maps combines road and navigation data with complementary locational data about businesses and services. Advertising revenue gives it an incentive to invest in harvesting the value of data re-use and aggregation. Consumers are incentivised to contribute their data because they get useful and free navigation services in return for accepting ads. In this case, the market realises at least some of the value of navigation data externalities. But there might be more value from further re-use of navigation data that is not realised yet.

Transaction costs often stand in the way of realising the societal value of data

Transaction costs often stand in the way of realising the societal value of data. First, finding partners to share the data with, or to arrange complementary inputs to generate value, may be difficult. Data cannot be exposed in a showroom. The willingness to pay for data may vary between users and service applications. Facilitating exploration of this value may require specific data-market design (Bergemann and Bonatti, 2018). It is hard to determine the value that data contributes to a data-driven service. Negotiated market outcomes often depend on the market power of the partners. Second, data transfers often require intermediary institutions that define data formats and transfer protocols and set the conditions for access and re-use (Martens, 2024b). This can be simple for bilateral data sharing but complex for data aggregation or pooling between many parties.

3 Potential data-market efficiency gains in EU data regulations

3.1 The GDPR

The GDPR is an important 'foundational' data law that regulates markets for personal data collected from natural persons, not from legal entities. It imposes restrictions on the collection of personal data. Firms should ask for the consent of natural persons and should adhere to strict rules on the handling of this data. Personal data cannot be used for purposes other than that for which it was collected. However, the GDPR grants natural persons the right to re-use their personal data for other purposes, or to permit data re-use by other service providers that compete with the original data collector.

That is a pro-competitive and pro-innovation provision: re-use by others increases competition in data-driven services markets in which the original data collector no longer has a data monopoly. There are no explicit provisions in the GDPR on data aggregation. However, data holders collect data from many persons and are therefore *de-facto* data aggregators. Data holders can combine and pool different personal data sources provided it is included in the consent notice.

Unfortunately, use of GDPR rights in practice often runs into high transaction costs. There is ample empirical evidence that GDPR consent notices are too costly and vague for data subjects to be meaningful (for example Barocas and Nissenbaum, 2009; Cate and Mayer-Schonberger, 2013; Utz *et al*, 2019). Data subjects do not read the many consent notices that pop up during daily web surfing because it takes too much time and notices are not intelligible. Moreover, data-subject requests for data access and transfers are often met only with considerable delay or in obscure data formats. The GDPR only requires transfers within three months of a request. That delay greatly diminishes the service market value of the data.

All this results in the so-called privacy paradox (Acquisti *et al*, 2016): natural persons attach importance to privacy but in practice do not use privacy protection tools because the costs of doing so are higher than the expected benefits. Acquisti *et al* (2016) cited many studies that illustrate how privacy costs and benefits vary widely according to the setting and the behaviour of data subjects and collectors. It is very difficult for individuals to know how their privacy choices will affect their welfare. That makes active privacy management very complex.

The GDPR also imposes compliance costs on data service providers. Empirical evidence shows that the GDPR has reduced the supply of digital services in the EU, compared to other regions and to the pre-GDPR period (see Johnson, 2024, for an overview). However, much of that evidence focuses on the supply side. It says little about the impact on consumer welfare or the demand side. Many of the services blocked by the GDPR might have reduced consumer welfare because they use personal data against the interests of the data subject. Others might have increased consumer welfare. How can these two be distinguished? Economists have so far been unable to come up with credible estimates of, or methods to estimate, the economic value of privacy, perhaps because of the wide variation in that value according to circumstances.

Policy recommendations

The GDPR has created the potential for personal-data market efficiency gains through economies of scope in data re-use and aggregation, but policymakers still have some way to go to reduce transaction costs that impede the realisation of these benefits.

First, onerous transaction costs for consent notices could be substantially reduced by introducing mandatory standardised and machine-readable consent notices. That could generate a more transparent market for consent services, on top of the market for data, and could enable natural persons to delegate that task to specialised service providers that could

handle it in accordance with users' stated preferences and firms' stated uses of the data. This would reveal privacy preferences for different types of services and consent conditions. A ranking of preferences would be a step towards distinguishing between welfare-augmenting and welfare-reducing personal data services. It would also put pressure on service providers to demonstrate data-sharing benefits for consumers, as a way to move up the ranking.

Second, making personal data available in real-time through application interfaces (APIs) would greatly reduce transfer transaction costs and make transfers to competing service providers more meaningful in an online digital market setting. Some EU data regulations, such as the Data Act and the Digital Markets Act (see sections 3.3 and 3.4), already include these obligations for data collectors. Nothing prevents the GDPR from doing the same.

3.2 The European Health Data Space (EHDS)

In fact, for one of the most sensitive types of personal data – health data – European data regulators have already gone far beyond the GDPR to generate economies of scope in data re-use and aggregation and in reducing transaction costs. The EHDS is the first EU data regulation that distinguishes between market failures in data re-use and in data aggregation. If the GDPR were applied strictly, there would have been no need for the EHDS: health data is probably the most personal data one can think of. The GDPR makes health data accessible and portable, but in a not very practical and operational way. A special health-data regulation was needed to overcome the GDPR's shortcomings.

Tensions between private rights and public benefits arose notably during the COVID-19 pandemic. Individuals wanted to move around freely and refused to give public health authorities access to their health-status data, despite ample evidence that restricting these rights and using health-status data would benefit the public health policy response to the crisis. The EHDS regulation was approved by the European Parliament and Council of the EU because policymakers saw the potential societal benefits of curtailing private rights and leveraging the social value of aggregated data for public benefit.

EHDS provisions regarding 'primary' data transfers reduce transaction costs for one-to-one bilateral data transfers. The EHDS regulation makes personal health data more accessible by defining the health data that should be made freely available for re-use by other health service providers. It establishes intermediary health databases at EU country and EU level that store health data in a common format, and sets out protocols for data transfers.

It also includes provisions for 'secondary' data pooling that go a step further and combine fragmented datasets from multiple parties into a single pool. It requires free access to these health data pools for scientific and policy research. Users only pay the marginal cost of access and processing of the data. This maximises incentives for innovative research. In line with another data law, the Data Governance Act (Regulation (EU) 2022/868), the intermediary aggregator remains neutral and does not monetise value-added from data aggregation.

In some cases, private intermediaries may be in a position to offer incentives for data pooling when they can monetise at least part of the benefits of economies of scope in data aggregation, and re-distribute part of that value to data contributors. For example, online search, navigation and social media platforms have succeeded in doing so. Advertisers pay for the building and running of these platforms; users are incentivised to share their data by being given access to useful free services. In other cases, however, incentivising private data contributors may be difficult because it is difficult to capture and privately monetise economies of scope. Regulatory intervention and mandatory data pooling is required to overcome these market failures. Some cases may also exhibit hybrid characteristics, with partial monetisation and partial dissipation of benefits.

The European Health Data Space should be a template for data regulations in other sectors

The EHDS should be a template for data regulations in other sectors that seek to realise the efficiency gains from economies of scope in data re-use and aggregation. However, some of these gains are constrained by private commercial data rights, including trade secrets and intellectual property rights (Aplin, 2024), that may constitute obstacles to data re-use.

Despite its advantages, the European Commission decided not to apply the EHDS

template in other 'industrial' data-pooling initiatives under the European Strategy for Data (European Commission, 2020). For example, proposals for a Common European Agricultural Data Space (CEADS), designed by farmers' organisations[3], would grant farmers exclusive control rights over farm data, rather than providing shared access rights for data co-producing parties. It essentially confirms prevailing agricultural data-market conditions (Atik and Martens, 2021) and maintains the gap between the private and social value of agricultural data. This is surprising since, under the EU's Common Agricultural Policy (CAP), a massive volume of farm data is already collected and pooled in databases. Rather than complementing these data pools with farm data that currently fall outside CAP reporting requirements, the proposal would keep CEADS and CAP data segmented. Ironically, the CEADS design includes proposals to reduce data transaction costs by prescribing standard data formatting and transmission protocols. But it offers no incentives to effectively use these standard protocols.

Policy recommendations

- The EHDS regulation addresses data-market failures that are very similar across sectors and may not require specific sectoral regulations. EU policymakers could therefore use the EHDS as an almost-ideal template for many industrial data-pooling initiatives that seek to leverage the benefits of data as a production factor.
- EHDS data requirements, formats and protocols should be adapted to specific settings in other sectors.

3.3 The Data Act

The Data Act will apply from September 2025 – there is at time of writing no empirical evidence yet on its impact. Rather than filling the regulatory gap left by the GDPR for non-personal data, it created a new category, 'product' data, ie data generated by the use of tangible devices that can communicate data wirelessly. This is a fuzzy category since all data requires a tangible carrier for interaction with users, whether held by users or located remotely. The Data Act facilitates economies of scope in data re-use by making it mandatory for manufacturers of devices to: (a) inform users about the raw data generated during use of a product, (b) make this data accessible to the user, free of charge and in real-time, and (c) allow the user to transfer the data to a third-party service provider.

The Data Act also introduces a number of obstacles to data re-use (Martens, 2023b)[4]. First, the data holder can charge third-party data recipients a monopolistic marked-up price, though this is somewhat attenuated by fair, reasonable and non-discriminatory (FRAND) conditions, a controversial concept taken from standard essential patent pricing. The interpretation of FRAND in data pricing remains to be defined. The third-party service provider may (partially) recuperate this price from the product user. In that case, the user pays twice for the same data: once at the time of purchase of the device or related service, and a second time to transfer the data from the device or related service.

The right of manufacturers to charge a license price for third-party access to product data introduces a quasi-ownership right for the manufacturer (Kerber, 2024). If that right were applied to third-party transfers of personal data, it would be a violation of the GDPR, which requires free transfers of personal data, including to third parties. Data-pricing provisions illustrate how the EU wavers between exclusive ownership rights for one party and a fair distribution of rights between data co-generators[5].

3 See European Commission news of 2 May 2024, 'Blueprint proposal for the Common European Agricultural Data Space', https://digital-strategy.ec.europa.eu/en/library/blueprint-proposal-common-european-agricultural-data-space.

4 For a more detailed discussion of the Data Act, see Sattler and Zech (2024).

5 The 1996 EU Database Directive (96/9/EC) first introduced exclusive ownership rights on databases.

Second, the Data Act restricts competition in the re-use of product data. It forbids the re-use of data to design new products that compete with the product manufacturer that initially collected or generated the product data. Data should not be transferred to the platform services of companies designated as 'gatekeepers' (meaning very large, hard-to-avoid platforms) under the EU Digital Markets Act (section 3.4). This prevents a user from transferring data from, for example, smart home appliances to a Google Android or Apple iOS smartphone, or to a Windows computer. It prevents welfare-enhancing network effects in data re-use and aggregation in digital ecosystems.

Policy recommendations

- The Data Act is the only EU data regulation that allows monopolistic pricing of third-party data transfers and puts anti-competitive restriction on these transfers. That should be abolished. It distorts data markets in favour of product manufacturers and re-introduces the concept of (quasi) ownership rights.
- More fundamentally, the EU should take a clear position against exclusive data ownership rights that have no place in a digital economy, with non-rival data that is co-generated between two or more parties, each with claims to the data. Data should be treated as a co-generated commons to facilitate the efficient use of data as a production factor and to benefit from economies of scope in re-use and aggregation of data.
- The fuzzy category of 'product data' is bound to create confusion. All data resides on a tangible physical device, whether held by the user at the place of activity or located remotely on a server. To avoid further regulatory fragmentation in data markets, the data transfer provisions of the Data Act should be extended to all product and services data. Services data also resides on tangible devices. That extension would also fill a gap in EU data regulation for services markets. It would make the Data Act a truly horizontal data regulation.

3.4 The Digital Markets Act (DMA)

Online platforms are a prime example of private market-based efficiency gains from economies of scale and scope in data aggregation. They bring together many types of market users, including buyers and sellers, advertisers, payment services and logistics companies, in a multi-sided online market. Platforms pool market-interaction data from all these users and extract insights from that pool that could not be extracted from partial or fragmented market datasets. That gives them a privileged and comprehensive market overview that can be used to guide and match users.

Users are attracted to platforms by these data-driven network effects (Prüfer and Schottmuller, 2021), which yield additional benefits compared to bilateral transactions in ordinary markets. Users get a wider market overview at lower information cost. There is more transparency and competition in the market. At the same time, these network effects may result in monopolistic winner-takes-all platforms that dominate markets. Monopolistic behaviour may be exploitative and reduce user welfare. Finding a good balance between these countervailing forces is difficult (Cabral *et al*, 2021).

The DMA imposes a number of obligations on very large gatekeeper platforms and their core services, to avoid these exploitative behaviours. This includes several data-sharing obligations to reduce information asymmetry between gatekeepers and users of their services, and to enable users to better position themselves in the market:

- Gatekeepers are required to give business users and end users (consumers) free real-time access to their data collected by the platform.

- Vertically integrated e-commerce platforms that sell products in competition with independent sellers should not use data that is not available to their competitors, to prevent distortions of the competitive level playing field.
- Gatekeeper search engines should share query, click and view data with other search engines that request access. However, data sharing is not free and can be subject to FRAND pricing.

These data-sharing obligations are a first step towards greater data sharing by platforms, beyond the narrow unpaid search and paid advertising data channels that they usually offer to users. The obligations could be extended to encompass a wider set of platform interaction data, beyond first-party 'own' direct interaction data. For example, consumers usually browse e-commerce platforms and look at several products before deciding on a purchase. Browsing data across products and sellers may provide very useful information for sellers to better understand their competitors. Platforms have this data but often don't share it with sellers. Including networked interactions would give a much better market overview to buyers and sellers on platforms, putting them on par with the quality of the market overview that the platform has. Making this market data sharable between competing platforms would increase competition in otherwise monopolistic platform markets.

Policy recommendations

- Giving business users access to 'their' data implies access to first-party click-and-view data only. That still leaves the platform operator in a privileged position with more fine-grained market insights. Extending access to second- and third-party network interaction data would enable business users to identify their nearest competing products and sellers and to modify their commercial strategies (Petropoulos *et al*, 2023). This could be done using privacy-preserving data access techniques.
- The obligation not to use certain valuable market information is a welfare-reducing lost opportunity to have more efficient markets. It would be better to share that information equally with all relevant market players, rather than not allowing any party to use it.
- Unilateral search-engine data sharing, with data going from the gatekeeper to others, risks fragmenting the search-engine data pool that is important for the efficiency of search. Mutual data sharing between search engines, irrespective of market shares or size, would be a more efficient solution (Martens, 2023a). However, the competitive landscape for search engines may be about to change rapidly under pressure from new AI-driven search tools. This should be taken into account when enforcing this obligation.

3.5 The AI Act

The AI Act's data re-use and aggregation provisions affect the availability of AI training data and thus the quality of AI models

The AI Act is not a data regulation. But it contains provisions on the re-use and aggregation of AI model training data that is subject to private rights under the EU Copyright Directive (CDSM, Directive (EU) 2019/790) and the GDPR. These provisions affect the availability of AI training data and thus the quality of AI models.

The performance of AI models is subject to so-called "*scaling laws*" (Kaplan *et al*, 2020): the quality and reliability of model responses to queries varies with the volume of data inputs used for training the model, the number of internal model parameters and the computational capacity available to train the model. With exponential growth in the size of AI models, model developers are running out of training data. So far, publicly available Common Crawl webpages data[6] has constituted the bulk of training data, sometimes complemented with other sources. However, a considerable part of that public content is subject to copyright.

Longpre *et al* (2024) showed how copyright holders are increasingly claiming their right

6 See https://commoncrawl.org/.

to opt-out of free use of these materials under Article 4(3) EU CDSM, thereby reducing the amount of text available for training by 20-25 percent or more. A September 2024 court judgement from Hamburg suggested, however, that commercial AI developers can circumvent the opt-out if the AI training dataset was compiled by a non-commercial entity and made available to the public on non-commercial terms under Article 3(1) CDSM (Pukas and Nordemann, 2024).

To increase the volume of training data, AI model developers are now shifting their attention to much larger volumes of publicly-posted social-media messages (this is especially important for smaller language communities with a relatively limited volume of webpages and other sources of written text). There is debate in the EU whether these private posts fall under the 'legitimate use' clause in the GDPR (Article 6(1)) and can therefore be used for AI model training without the explicit consent of the persons who posted the messages.

Policy recommendations

- Reducing copyright protection for AI model training purposes can enable positive spill-overs from copyright-protected training data to the wider economy, in which AI models are used as a general-purpose technology. The Hamburg court ruling takes a step in this direction and the Code of Practice for the implementation of the AI Act should take this into account[7].
- The same reasoning applies to the use of social-media posts for AI model training. Since these posts are already in the public domain, re-use for AI model training will not reduce the welfare of the data subjects as long as model developers take care to avoid any direct publication of the posts in AI outputs.

4 Conclusion: finding a balance between private rights and social benefits

This tension between private rights and public benefits is not unique to health data during the COVID-19 period; it pervades all EU data-market regulations discussed in this Policy Brief. The COVID-19 example illustrates in a very intuitive way why the overall objective of any data-market regulation should be to maximise the social value of data, leveraging economies of scope in data re-use and aggregation, while minimising transaction costs. The EU data regulations discussed above take significant steps in this direction because they open up access to data and facilitate re-use by others. However, there is considerable scope to improve these regulations, as suggested in the policy recommendations set out above.

EU data regulations open up access to data and facilitate its re-use; however, there is considerable scope to improve these regulations

The recommended policy changes revolve around the balance between exclusive rights of data holders and the granting of access, re-use and aggregation rights to data co-generators and intermediaries. This is where political data-regime choices (section 1) must be made. Policymakers decide on the trade-offs between individual and social welfare. Bradford (2023) argued that the US tends to leave it all to the market and individual bargaining power; China is inclined towards the collective side and the EU is somewhere in between. However, the two sides are not necessarily opposed. Pursuing social welfare does not necessarily imply weakening private rights to data. Technologies exist that can combine the two objectives, at least to some extent. For example, privacy-preserving technologies may still enable personal or trade secrets data to be made available for socially useful purposes.

7 The General-Purpose AI Code of Practice is being drawn up by the European AI Office, which was established by the AI Act. For an overview, see https://digital-strategy.ec.europa.eu/en/policies/ai-code-practice.

Finding an acceptable trade-off between these conflicting claims to data is a political balancing act that cannot be based on economics only. That trade-off may vary with the evolution of digital technologies and the extent to which they enable data-driven externalities. When new technologies emerge that increase the social value of data because they generate new insights, applications and innovations that were previously not feasible, the pressure mounts to "*turn straw into gold*" (Beraja and Yuchtman, 2024) and tone down the exercise of exclusive private rights to enable more extraction of social value from data. Technological progress can also ease the tension when it produces new technologies that make it easier to protect private rights while harvesting the public benefits. For example, federated learning techniques in AI and machine learning, and various types of privacy sandboxes in personal data protection, leave private data in its own protected setting, while still enabling model training. The balance between private rights and social benefits is thus a constant political tightrope to be walked, driven by technological progress.

Finding a politically acceptable balance may involve some degree of redistribution of the efficiency gains from data markets between data users and the original data collectors. There are three options. The first leaves the gains in the hands of data users – as in the GDPR, the EHDS and the DMA. That maximises incentives for the innovative re-use of data. The second allows data collectors to set a monopolistic price for access to the data, as in the Data Act, the preliminary design of the CEADS and copyright provisions in the AI Act. That reduces competition and innovation incentives, and the efficiency of data as a production factor. An intermediate regime would somewhat soften monopolistic pricing with FRAND conditions, at the cost of substantial administrative intervention and market uncertainty to achieve this. Only the first option would be fully in line with the EU's oft-stated aim of maximising innovation.

References

Acquisti, A., C. Taylor and L. Wagman (2016) 'The Economics of Privacy', *Journal of Economic Literature* 54(2): 442–92

Aplin, T. (2024) 'The Data Act and trade secrets: an experiment in compulsory licensing', in A. Sattler and H. Zech (eds) *The Data Act: First Assessments*, Institut für Recht und Digitalisierung, University of Trier, available at https://doi.org/10.25353/ubtr-04b0-0969-2b7a

Atik, C. and B. Martens (2021) 'Competition Problems and Governance of Non-personal Agricultural Machine Data: Comparing Voluntary Initiatives in the US and EU', *Journal of Intellectual Property, Information Technology and Electronic Commerce Law* 12(3): 370-396

Bajari, P., V. Chernozhukov, A. Hortaçsu and J. Suzuki (2019) 'The impact of big data on firm performance: An empirical investigation', *AEA Papers and Proceedings* 109: 33–37

Barocas, S. and H. Nissenbaum (2009) 'On Notice: The Trouble with Notice and Consent', *Proceedings of the Engaging Data Forum*, October, available at https://ssrn.com/abstract=2567409

Beraja, M. and N. Yuchtman (2024) 'Turning Straw into Gold: Novel productive factors and innovation under contested property rights', mimeo, National Bureau of Economic Research, available at https://conference.nber.org/conf_papers/f209446/f209446.pdf

Bergemann, D. and A. Bonatti (2018) 'Markets for information: an introduction', *CEPR Discussion Paper* DP13148, Centre for Economic Policy Research

Bradford, A. (2023) *Digital Empires: the global battle to regulate technology*, Oxford University Press

Cabral, L., J. Haucap, G. Parker, G. Petropoulos, T. Valletti and M. Van Alstyne (2021) *The EU Digital Markets Act: A Report from a Panel of Economic Experts*, Publications Office of the EU, available at https://dx.doi.org/10.2760/139337

Calzolari, G., A. Cheysson and R. Rovatti (2023) 'Machine data: market and analytics', mimeo, available at https://ssrn.com/abstract=4335116

Carballa-Smichowski, B., N. Duch-Brown, S. Höcük, P. Kumar, B. Martens, J. Mulder and P. Prüfer (2022) 'Economies of scope in data aggregation: evidence from health data', *TILEC Discussion Paper* 2022-020, available at https://dx.doi.org/10.2139/ssrn.4338447

Cate, F. and V. Mayer-Schönberger (2013) 'Notice and consent in a world of Big Data', *International Data Privacy Law* 3(2): 67–73

Diebold G. (2023) *Comparing Data Policy Priorities Around the World*, Center for Data Innovation, September

Draghi, M. (2024) *The future of European competitiveness – A competitiveness strategy for Europe*, European Commission

Johnson, G. (2024) 'Economic Research on Privacy Regulation: Lessons from the GDPR and Beyond', in A. Goldfarb and C. Tucker (eds) *The Economics of Privacy*, University of Chicago Press

Kaplan, J., S. McCandlish, T. Henighan, T.B. Brown, B. Chess, R. Child … D. Amodei (2020) 'Scaling Laws for Neural Language Models', mimeo, available at https://arxiv.org/abs/2001.08361

Kerber, W. (2024) 'The EU Data Act: Will New User Access and Sharing Rights on IoT Data Help Competition and Innovation?' *Journal of Antitrust Enforcement* 12(2): 234-240, available at https://doi.org/10.1093/jaenfo/jnae011

Longpre, S., R. Mahari, A. Lee, C. Lund, H. Oderinwale, W. Brannon … S. Pentland (2024) 'Consent in Crisis: The Rapid Decline of the AI Data Commons', mimeo, available at https://arxiv.org/abs/2407.14933

Martens, B. (2023a) 'Are new EU data market regulations coherent and efficient?' *Working Paper* 21/2023, Bruegel, available at https://www.bruegel.org/system/files/2023-12/WP%202023%2021%20181223%20final.pdf

Martens, B. (2023b) 'Pro- and anti-competitive provisions in the proposed European Union Data Act', *Working Paper* 01/2023, Bruegel, available at https://www.bruegel.org/sites/default/files/2023-01/WP%252001.pdf

Martens, B. (2024a) 'Memo to the commissioner responsible for digital affairs', in M. Demertzis, A. Sapir and J. Zettelmeyer (eds) *Unite, defend, grow: Memos to the European Union leadership 2024-2029*, Bruegel, available at https://www.bruegel.org/memo/memo-commissioner-responsible-digital-affairs

Martens, B. (2024b) 'An institutional economics approach to data spaces as data market intermediaries', mimeo, available at https://dx.doi.org/10.2139/ssrn.4731126

Panzar, J.C. and R.D. Willig (1981) 'Economies of scope', *American Economic Review* 71(2): 268–272

Petropoulos, G., B. Martens, G. Parker and M. Van Alstyne (2023) 'Platform competition and information sharing', *Working Paper* No. 10663, CESifo, available at https://www.cesifo.org/DocDL/cesifo1_wp10663.pdf

Prüfer, J. and C. Schottmuller (2021) 'Competing with big data', *Journal of Industrial Economics* 69(4): 967-1008, available at https://doi.org/10.1111/joie.12259

Pukas, J. and J.B. Nordemann (2024) 'German Regional Court of Hamburg paves the way to treat the reproduction of works as AI training data under the EU text and data mining exceptions', *Kluwer Copyright Blog*, 25 October

Sattler, A. and H. Zech (eds) (2024) *The Data Act: first assessments*, Institut für Recht und Digitalisierung, University of Trier, available at https://doi.org/10.25353/ubtr-04b0-0969-2b7a

Teece D.J. (1980) 'Economies of scope and the scope of the enterprise', *Journal of Economic Behavior and Organization* 1(3): 223-247

Teece, D.J. (1982) 'Towards an economic theory of the multiproduct firm', *Journal of Economic Behavior and Organization* 3(1): 39–63

Utz, C., M. Degeling, S. Fahl, F. Schaub and T. Holz (2019) '(Un)informed Consent: Studying GDPR Consent Notices in the Field', *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security: 973-990*, available at https://doi.org/10.1145/3319535.3354212